Chapter 06 / Capítulo 06

New literacies in the age of AI: Ethics, teaching, and writing
ISBN: 978-9915-9854-3-5
DOI: 10.62486/978-9915-9854-3-5.ch06

©2025 The authors. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Generative feedback: causal effects of LLM's on writing quality and evaluative equity in higher education

Retroalimentación generativa: efectos causales de los LLM en la calidad de la escritura y la equidad evaluativa en educación superior

Carluys Suescum Coelho[1], Car-Emyr Suescum Coelho[2], Carlysmar Suescum Coelho[1], Carelys Suescum Coelho[1]

[1]Centro de Estudios Gerenciales Avanzados (CEGA). Caracas, Venezuela.

[2] Universidad Metropolitana, Universidad Central de Venezuela. Caracas, Venezuela.

Autor de correspondencia: Carluys Suescum Coelho

ABSTRACT

This chapter examines, within the framework of SDG 4, how large-scale generative *feedback* using large language models (LLMs) impacts the quality of scholarly writing and evaluative equity in higher education. The aim is to estimate improvements in coherence, argumentation, use of evidence, and clarity, and to determine whether gaps between subgroups (L1/L2, first generation, and baseline performance) are reduced. An experimental or quasi-experimental design (*stepped-wedge* cluster trial) with blinded assessment based on analytical rubrics and secondary metrics such as time to *feedback*, revision iterations, editing distance, self-efficacy, and teaching load is recommended. A synthesis of recent evidence suggests consistent gains when LLM *feedback* is effectively integrated into explicit rubrics and revision microtasks; furthermore, it increases engagement and can decrease inter-rater variability. However, risks of dependency, stylistic homogenization, and algorithmic bias are noted. Therefore, this chapter proposes a framework for didactic governance with traceability (registration of *prompts*, versions, and logs), open science principles (pre-registration, repositories), and a dual human-AI evaluation scheme with bias audits and human-in-the-loop thresholds. Among the limitations acknowledged are the heterogeneity of contexts, sensitivity to model versions, and external validity. The conclusion is that, under ethical safeguards and with teacher training, generative *feedback* can raise writing standards and broaden inclusion; further research is suggested on differential effects, disciplinary transferability, and institutional sustainability of scaling up.

Keywords: Evaluative Equity; Academic Writing; Large-Scale Language Models (LLMS); SDG 4 (Quality Education); Generative *Feedback*; Algorithmic Biases.

RESUMEN

Este capítulo examina, en el marco del ODS 4, cómo la retroalimentación generativa a escala mediante modelos de lenguaje grandes (LLM) incide en la calidad de la escritura académica y en la equidad evaluativa en educación superior. El objetivo es estimar mejoras en coherencia, argumentación, uso de evidencia y claridad, y determinar si se reducen brechas entre subgrupos (L1/L2, primera generación y rendimiento basal). Se recomienda un diseño experimental o cuasiexperimental (ensayo por clúster tipo *stepped-wedge*) con evaluación ciega basada en rúbricas analíticas, y métricas secundarias como tiempo a *feedback*, iteraciones de revisión, distancia de edición, autoeficacia y carga docente. La síntesis de evidencia reciente sugiere

ganancias consistentes cuando el *feedback* del LLM se integra efectivamente a rúbricas explícitas y microtareas de revisión; además, aumenta el involucramiento y puede disminuir la variabilidad interevaluador. Sin embargo, se advierten riesgos de dependencia, homogeneización estilística y sesgos algorítmicos. Por ello, el capítulo plantea un marco de gobernanza didáctica con trazabilidad (registro de *prompts*, versiones y logs), principios de ciencia abierta (pre-registro, repositorios) y un esquema de evaluación dual humano-IA con auditorías de sesgo y umbrales human-in-the-loop. Entre las limitaciones, se reconocen la heterogeneidad de contextos, la sensibilidad a versiones de modelos y la validez externa. Se concluye que, bajo salvaguardas éticas y formación docente, la retroalimentación generativa puede elevar estándares de escritura y ampliar la inclusión; se sugieren además, líneas futuras sobre efectos diferenciales, transferibilidad disciplinar y sostenibilidad institucional del escalamiento.

Palabras clave: Equidad Evaluativa; Escritura Académica; Modelos de Lenguaje a Gran Escala (LLM); ODS 4 (Educación de Calidad); Retroalimentación Generativa; Sesgos Algorítmicos.

INTRODUCCIÓN

El fenómeno de la masificación de la educación superior y la diversidad creciente de estudiantes han resaltado un problema estructural clásico, como lo es la provisión de una retroalimentación académica oportuna, consistente y altamente personalizada (Wambsganss et al., 2022; Zhuang et al., 2025). Precisamente es ante esta imposibilidad práctica de atender individualmente a una cohorte amplia de estudiantes, que la irrupción de la inteligencia artificial generativa (IA Gen) basada en modelos de lenguaje a gran escala (LLM) ofrece un avance significativo.

Estos sistemas prometen brindar retroalimentación generativa a escala, es decir, permiten generar comentarios automatizados ricos en ejemplos y explicaciones concretas sobre los textos de los estudiantes (Benner, 2024; Thomas et al., 2025). Tal retroalimentación inmediata y oportuna puede mejorar sustancialmente la calidad de la escritura estudiantil, al señalar con claridad criterios de coherencia, evidencia y claridad argumental. Aunado a ello, la uniformidad de los criterios empleados por la IA favorece la equidad evaluativa, evitando exclusiones o desigualdades, pues todos los alumnos reciben el mismo nivel de atención independiente de su origen, género o condición social, reduciendo las brechas tradicionales (González et al., 2024; Jovic et al., 2025).Es aquí cuando la retroalimentación generativa representa un cambio de fase, pues combina escalabilidad y trazabilidad del proceso evaluativo, sin sacrificar bajo ningún concepto la especificidad pedagógica

La relevancia de este enfoque recae en dos grandes aportes potenciales, el primero de ellos estima que puede mejorar la calidad escritural de los estudiantes, en este sentido investigaciones recientes han encontrado efectos positivos en tareas de escritura al incorporar *feedback* automatizado detallado. Wambsganss et al., (2022) observaron que los estudiantes que recibieron una retroalimentación (*feedback*) automatizada de argumentación junto con un empujón (*nudge*) de comparación social escribieron textos mucho más convincentes y mejor estructurados.

De manera similar, Glüsing et al., (2025) hallaron en un experimento controlado aleatorio (ECA) que el *feedback* individualizado generado por GPT-4 mejoró la calidad de las revisiones de abstractos académicos y aumentó el tiempo de revisión de los alumnos. En contratas, en entornos de aprendizaje de idiomas, Zhuang et al., (2025) demostraron que un sistema adaptativo de IA elevó significativamente las puntuaciones de escritura de cientos de alumnos de secundaria, hallazgos que están en consonancia y armonía con los hallazgos de Mekheimer

(2025), quien encontró que los estudiantes de postgrado de inglés como lengua extranjera obtuvieron puntajes post-test significativamente más altos y mayor mejora en contenido, organización y cohesión tras usar retroalimentación asistida por IA.

En segundo lugar, se enfatiza el aporte a la equidad evaluativa, pues al aplicar criterios constantes, la IA puede nivelar las condiciones de los estudiantes con menor dominio lingüístico o de primera generación, teniendo acceso al mismo *feedback* constructivo que sus pares que cuentan con mayor experiencia, reduciendo de este modo la dispersión en resultados. Esto genera múltiples efectos como una mayor autoconfianza y compromiso que los estudiantes reportan tras haber usado herramientas de IA, los cuales sugieren y recalcan que estas tecnologías pueden servir como andamiajes uniformes.

El presente capítulo tiene como objetivo analizar en profundidad cómo la retroalimentación generativa influye en la calidad de la escritura académica y en la equidad evaluativa, todo ello enmarcado en el ODS 4 (Educación de Calidad). Para ello se pretende evaluar cómo este tipo de *feedback* puede no solo estimar, sino mejorar la calidad de la escritura y seguidamente examinar su impacto en la equidad, particularmente en la reducción de brechas asociadas al idioma nativo, desempeño inicial o condición de primera generación.

Para ello se plantean preguntas de investigación centradas en el impacto en la calidad del producto escrito, la oportunidad y utilidad del *feedback* generado y la heterogeneidad de efectos entre subgrupos. Se definen operativamente los conceptos clave: retroalimentación generativa del *feedback* personalizado creado por IA, calidad escritural con la capacidad de argumentar con coherencia, evidencia y claridad, y equidad evaluativa, entendida como justicia distributiva del *feedback* en términos de acceso, oportunidad y utilidad uniforme (Jovic et al., 2025). También se introducen nociones de andamiaje (soporte educativo progresivo) y trazabilidad (registro técnico del proceso de generación de *feedback*) para enmarcar la discusión metodológica.

Para ello se ofrece un *roadmap* del capítulo, desarrollando el marco teórico y evidencia empírica sobre *feedback* automatizado; luego se propone un diseño metodológico con enfoques experimentales recomendados; se presenta una síntesis de resultados; se discuten las implicaciones pedagógicas e institucionales; se abordan consideraciones éticas e integridad académica; para finalmente enunciar las limitaciones y agenda futura.

DESARROLLO

Marco conceptual

Desde una perspectiva teórica, la literatura destaca la importancia del aprendizaje autorregulado y la escritura iterativa para mejorar el rendimiento escritural. El modelo de aprendizaje autorregulado (Zimmerman, 2000) implica que los estudiantes planifican, monitorean y evalúan su propio proceso de escritura. En entornos de retroalimentación, los autores señalan que las tareas de revisión basadas en *feedback* son altamente autorreguladas, pues el alumno debe emplear estrategias cognitivas y metacognitivas para incorporar las sugerencias (Alcudia Ólan & Morales Vázquez, 2025). Tian & Liu (2022) evidencian en su investigación que al recibir *feedback*, ya sea automático, de pares o docente, los escritores despliegan diversas estrategias de regulación (planificación o autorrevisión) para mejorar sus textos. Esto sugiere que un *feedback* bien diseñado, inmediato y procesable estimula el ciclo autorregulado de revisión; mientras que por el contrario, un *feedback* insuficiente dificulta en gran medida la autorrevisión.

En concordancia con Hattie & Timperley (2007), se reconoce que la retroalimentación es más efectiva cuando es inmediata, correcta y elaborada. Por ejemplo, el *feedback* que no solo

indica errores sino que explica su corrección y brinda ejemplos concretos produce mayores ganancias de aprendizaje. En este sentido, la retroalimentación generativa de los LLM puede actuar como un andamiaje cognitivo en tanto que los modelos pueden suministrar ejemplos y explicaciones específicas, lo que ayuda a los estudiantes a comparar su escrito con criterios establecidos en cuanto a la argumentación, evidencias y claridad, pero sobre todo a reescribir de manera más efectiva (Quesada & Salinas, 2021).

Además, al ser un proceso automatizado, permite la retroalimentación procesable y consistente, es decir, todos los alumnos reciben orientaciones basadas en una rúbrica analítica explícita que define indicadores observables. Esta consistencia metodológica del *feedback* automatizado apunta también a la equidad, pues al aplicar los mismos estándares uniformemente se busca minimizar el sesgo humano.

El concepto de equidad evaluativa se entiende aquí como justicia distributiva del *feedback*, siguiendo una visión de accesibilidad y utilidad igualitaria. Esto implica que todos los estudiantes, independientemente de su perfil o condiciones socioculturales tengan igual acceso a comentarios relevantes y oportunos. En términos prácticos, las métricas de equidad incluyen la reducción de brechas, tanto absolutas, como relativas en los resultados escriturales entre estos grupos. En la literatura sobre evaluación se ha planteado que la equidad no siempre lleva implícita o equivaldrá a igualdad de resultados, sino a una justa aplicación de criterios (Romero-González, 2024).

Jovic et al., (2025) definen la equidad como la justicia y consistencia del *feedback* a través de distintos grupos de estudiantes, sean nativos o no nativos). Partiendo de los preceptos concebidos por estos autores, un sistema de *feedback* automatizado debe registrar indicadores como el tiempo de respuesta a *feedback* o la variabilidad en las mejoras de subgrupos, lo cual es crucial para identificar si efectivamente se reducen las disparidades presentes.

Los LLM también funcionan como tecnología educativa con potencial y límites propios, entre sus ventajas se encuentra la capacidad de escalar el *feedback* para adaptarlo dinámicamente al texto escrito por el estudiante. Sin embargo, estudios recientes advierten que estos sistemas requieren supervisión humana, por ejemplo, la evidencia indica que el *prompting* conversacional (guiar al LLM con preguntas específicas) puede mejorar la calidad de la retroalimentación, pero los LLM presentan fallos frecuentes como alucinaciones (inventar contenido inexistente) o respuestas genéricas cuando se enfrenta a argumentos mucho más complejos (Sandoval et al., 2024).

Por ello, ante este panorama, es necesario integrar rúbricas claras, tareas de revisión guiada y validación por expertos humanos, de modo que la IA sirva para complementar y no termine por reemplazar la labor docente (Valenzuela & Pérez, 2025). Del mismo modo, la trazabilidad del proceso a través del registro de los *prompts*, versiones de modelos y logs de uso, es clave para auditar y corregir sesgos algorítmicos, así como para ajustar los umbrales o niveles de intervención humana. En este marco, la retroalimentación generativa se configura como un recurso pedagógico innovador, pero que, en aras de garantizar la calidad educativa debe estar inserto en estructuras de supervisión, para asegurar relevancia y justicia en la evaluación (Xia et al., 2024).

La investigación empírica sobre retroalimentación generativa en educación superior está en expansión reciente, y es que, hasta la fecha, varios estudios han mostrado efectos positivos de la retroalimentación automatizada en la calidad escritural y en la eficiencia del proceso de

retroalimentación. En su estudio con 124 estudiantes, Wambsganss et al. (2022) hallaron que aquellos que recibieron *feedback* automatizado de argumentación combinado con un incentivo de comparación social mejoraron significativamente la calidad de sus textos argumentativos comparado con grupos control (Afrin., et al., 2021). Estudios de diseño cuasiexperimental, como el desarrollado por Glüsing et al., (2025), han confirmado que los alumnos que recibieron *feedback* de GPT-4 alcanzaron revisiones de abstracts de investigación con mayor calidad y mostraron mayor compromiso conductual en la tarea (más tiempo de revisión, más distancia de edición) que un grupo sin *feedback* automatizado.

En contexto de la enseñanza de idiomas el estudio desarrollado por Zhuang et al., (2025) reporta mejoras estadísticamente significativas en las puntuaciones de escritura de estudiantes de secundaria tras usar un sistema de AI multimedial para retroalimentación adaptativa. Del mismo modo, Mekheimer (2025) encontró que estudiantes de inglés como lengua extranjera (EFL) con *feedback* asistido por la IA Grammarly, superaron al grupo de control en pruebas de redacción, aunado a ello, revisaron sus textos más frecuentemente, además de reportar un mayor grado de satisfacción y confianza con el proceso.

Este cuerpo de evidencia tal como sugieren Gombert et al., (2024) demuestra que el *feedback* generativo, cuando se diseña apropiadamente, puede mejorar significativamente el ciclo de revisión y elevar la calidad de los escritos, asimismo, la disponibilidad inmediata de *feedback* reduce el tiempo de espera para los estudiantes, lo cual es crucial en cursos masivos con grandes matrículas. Cabe destacar que en general, los alumnos han valorado positivamente la utilidad del *feedback* generado por IA, especialmente cuando se percibe como específico y relevante para su proceso de aprendizaje (Thomas et al., 2025).

En contraposición es menester indicar que la investigación sobre efectos heterogéneos es todavía incipiente pues pocos estudios han analizado sistemáticamente cómo estas intervenciones benefician de modo diferenciado a subgrupos. No obstante, Jovic et al., (2025) observan que las respuestas que brinda la AI tienden a ser más consistentes entre estudiantes de diferentes orígenes, aunque carecen de profundidad en el análisis de argumentos complejos.

De acuerdo con estas evidencias reflejadas por estas investigaciones, las brechas por cerrar incluyen la validación de la equidad de los algoritmos de *feedback*, pues existe un gran temor por la replicabilidad de sesgos de datos, así como de la trazabilidad explícita para auditar posibles favoritismos. Precisamente ante este panorama el estado del arte apunta a que la retroalimentación generativa puede mejorar aprendizajes, reduciendo las disparidades existentes de manera escalable, sin embargo, aún se requiere más investigación causal y analítica por subgrupos. Por ello, la investigación aportará valor al enfocarse explícitamente en el enfoque causal y equitativo, así como en dimensiones institucionales de escalado y gobernanza del *feedback* automatizado, aspectos que, hasta ahora, han sido poco explorados.

Del diseño metodológico recomendado para la retroalimentación generativa

Para evaluar rigurosamente los efectos de la retroalimentación generativa, se recomienda un diseño empírico experimental o cuasiexperimental en entornos reales de educación superior. En concreto, un ensayo por clúster o diseño escalonado (*stepped-wedge*) equilibra la validez interna y factibilidad logística (Chen et al., 2021). Por ejemplo, se podrían seleccionar varios cursos universitarios, o asignaturas que tengan similares asignaciones escritas (entrenadas por cohorte o clases), y aleatorizarlos por grupos o secuencias para recibir la intervención.

La condición de control correspondería al método habitual de retroalimentación (profesoral

o correcciones manuales), mientras que los tratamientos incluirían al menos dos variantes, a saber: 1) LLM + rúbrica estructurada, y 2) LLM + rúbrica + micro tareas de revisión guiada (andamiaje explícito). La población objetivo serían estudiantes de licenciatura en cursos de escritura o materias con componentes escritos sustanciales, preferentemente se considerarían muestras grandes y diversas para garantizar de este modo un mayor poder estadístico y un mejor análisis por subgrupos.

Los resultados primarios arrojados serían métricas de calidad escritural obtenidas mediante evaluación ciega con una rúbrica analítica, siendo que dicha rúbrica incluiría criterios clave como argumentación (coherencia de la tesis y su correspondiente estructura lógica), coherencia textual, uso de evidencia y claridad expresiva, la cual debe ser consistente con estándares pedagógicos (Glüsing et al., 2025). Estas rúbricas deben desarrollarse a partir de literatura previa sobre buenas prácticas de escritura y, en la medida de lo posible, adaptarse por expertos lingüistas o profesionales docentes antes del estudio.

Además, se recolectarán indicadores de resultado secundarios como: tiempo desde la entrega de texto hasta recepción de *feedback* (para medir oportunidad), número de iteraciones de revisión que realiza el estudiante, distancia de edición (medida cuantitativa de cuánto cambia el texto tras *feedback*), autoeficacia escritural (encuestas de los alumnos sobre su confianza), y percepción de carga docente (cuestionarios a profesores sobre esfuerzo de supervisión); estos *outcomes* cubrirán tanto dimensiones de efectividad como las de eficiencia.

Para evaluar la equidad, se realizaría análisis de heterogeneidad de efectos; siendo que los datos sociodemográficos que se incluirían serían: lengua materna, condición de primera generación, decil o cuartil de rendimiento base, género, entre otros que se estimen pertinentes. Se estimarán efectos promedio (ATE) y efectos condicionales (CATE) con modelos estadísticos que permitan interacciones, como las regresiones con variables ficticias (dummy) de subgrupo y sus interacciones con tratamiento (Holgersson et al., 2015).

En particular, se examinará si el tratamiento con *feedback* generativo reduce significativamente la brecha de puntajes entre grupos respecto al control, como sería el caso de cambios en diferencia absoluta y relativa de medias entre estudiantes de un nivel (L1) y estudiantes de un nivel superior (L2). El análisis estadístico utilizará modelos multinivel (por ejemplo, con alumno en nivel 1 y sección / clase en nivel 2) para controlar la dependencia jerárquica de los datos (Impey et al., 2025). Se aplicarán métodos de correcciones estadísticas por comparaciones (Bonferroni o FDR) al probar múltiples *outcomes*, y se calcularán tamaños del efecto e intervalos de confianza para interpretar magnitudes prácticas (Ren & Ren , 2025).

Como buena práctica de ciencia abierta, se recomienda pre registrar el protocolo de investigación, como por ejemplo siguiendo CONSORT-AI para intervenciones con IA (Chen et al., 2025), y disponer un repositorio público con materiales y datos anonimizados como *prompts*, versión del modelo LLM, hiperparámetros usados, rúbricas completas, scripts de análisis y logs de interacción (secuencias de *prompts* y respuestas). Esta transparencia facilita la reproducibilidad y futura auditoría de sesgos.

Sobre la presentación de resultados y síntesis de evidencia

En la sección de resultados se presentaría la síntesis tanto descriptiva como inferencial de los hallazgos. Primero, es crucial que el investigador describa la muestra: número de alumnos, distribución por características clave (idioma nativo, género, entre otras), así como nivel basal o punto de referencia que se considerará para la determinación de la calidad escritural (puntajes

iniciales). En estudios previos se observa variabilidad inicial que justifica controles por línea base (Wambsganss et al., 2022).

Seguidamente se mostrarían los efectos principales del tratamiento LLM versus el control en calidad escritural y tiempos. Por ejemplo, se esperarían aumentos en las puntuaciones ciegas de calidad estadísticamente significativos y de magnitud práctica relevante; el tamaño del efecto comparado (empleando la d de Cohen) se interpretaría en términos pedagógicos considerando cuánto avanza el puntaje medio de escritura. (Sullivan & Feinn, 2012). Asimismo, se evaluaría el cambio en tiempo del *feedback*, idealmente, los estudiantes en el grupo LLM recibirían correcciones casi instantáneas (tiempo cercano a cero), en contraste con días o semanas en el grupo control.

Respecto a efectos diferenciales, se reportaría cómo el *feedback* influye en cada subgrupo. En este caso se examinaría si la brecha absoluta entre alumnos del nivel (L1) y (L2) en puntuación de escritura se reduce tras recibir *feedback* generativo (indicando mayor equidad). También se analizarían diferencias por cuartiles iniciales, revelando si algunos grupos obtienen un beneficio mayor o si, por el contrario, la intervención mantiene uniformidad en mejoras (situación deseable para equidad).

En estudios análogos (Jovic et al., 2025) a veces se observan mayores ganancias relativas en grupos tradicionalmente rezagados, pues la IA puede elevar la base más baja. Sin embargo, también se considerará la posibilidad de que se presenten sesgos inversos cuando los LLM privilegian patrones lingüísticos de hablantes nativos, en el caso de que sea aplicado, por ejemplo, en el aprendizaje de una segunda lengua. De modo complementario, se evaluaría la robustez de estos resultados a partir de un análisis de sensibilidad y controles para contaminación entre grupos para evitar que el grupo control obtenga *feedback* extra.

Posteriormente, se presentarían indicadores complementarios, como, por ejemplo, cuando en el experimento de Glüsing et al., (2025), el *feedback* del LLM aumentó el tiempo de edición (indicador de involucramiento) y esto mediaba en una mayor calidad revisada. Se reportarían asimismo las valoraciones de estudiantes sobre la utilidad del *feedback* (mediante escalas de Likert) para entender la percepción de usabilidad. En todas las tablas se incluirían intervalos de confianza del 95% para los efectos estimados y criterios estadísticos (p-valor, tamaño del efecto). De este modo la sección resumiría los hallazgos clave, a saber, los cambios en puntajes ciegos de calidad y tiempos de respuesta, junto con la reducción o aumento de brechas entre subgrupos, apoyándose en la interpretación de magnitudes prácticas.

Los hallazgos hipotéticos descritos tendrían varias implicaciones pedagógicas y evidentemente institucionales. En primer lugar, un efecto positivo en la calidad escritural indicaría que el *feedback* generativo actúa como un andamiaje cognitivo efectivo. Y esto, precisamente puede interpretarse a través de mecanismos de aprendizaje autorregulado al recibir explicaciones y ejemplos, de este modo, los estudiantes asimilan criterios explícitos (metacognición) y mejoran su escritura en iteraciones sucesivas.

Los resultados de Wambsganss et al., (2024) muestran que el componente motivacional (*nudge* social) combinado con *feedback* explícito desencadenó procesos psicológicos favorables (autoeficacia, motivación). De igual manera, el uso de ejemplos específicos en el *feedback* (contenido obtenido vía LLM) supone un andamiaje directo, pues el alumno ve modelos de estructura de argumento o citas apropiadas, lo que facilita las reescrituras. En la práctica, esto sugiere que las intervenciones pedagógicas con IA deberían incluir *prompts* que generen

aclaraciones y ejemplos concretos, no solo limitarse a la realización de observaciones generales.

En cuanto a la equidad, los resultados indicarían hasta qué punto la tecnología niveló el terreno. Siendo que un hallazgo de reducción de brechas entre los estudiantes de L1 y L2, o de mayores mejoras en primeros deciles, respaldaría la idea de que la IA puede promover justicia distributiva. No obstante, se debería interpretar con cautela: la validez interna del experimento (aleatorización, control) fortalece la inferencia causal de estos efectos promedio, pero la generalización (externa) dependerá en gran medida del contexto institucional. Por ejemplo, si el estudio se realizó en universidades que cuentan con una alta infraestructura tecnológica, su traslado a contextos de universidades que cuentan con menos recursos requeriría un conjunto de adaptaciones. Igualmente, la consistencia en la inferencia causal es reforzada por el diseño multinivel y análisis rigurosos aplicados (Impey et al., 2025), pero la variabilidad de los LLM, ya sean con la aparición de nuevas versiones o con el cambio de *prompts* puede limitar la replicabilidad exacta.

En todo caso, los resultados tendrían implicaciones sumamente claras, pues si el *feedback* generativo muestra transparencia y robustez, se recomendaría su despliegue institucional. Esto implica, primeramente, de la implementación de copilotos didácticos basados en IA que asistan a estudiantes y docentes, como chatbots de escritura, asistentes de revisión o agentes IA. Además, las instituciones para potenciar estos resultados deberán invertir en formación docente, enseñando a los profesores a interpretar y complementar el *feedback* automatizado, y a usar dashboards de trazabilidad que les permitan observar qué *feedback* se dio.

Y, en segundo lugar, desde el punto de vista de la gobernanza, se requeriría de la definición de protocolos de mejora continua que se encarguen de la revisión periódica de los *prompts* y de los ajustes del sistema según datos efectivamente reales de su rendimiento. En definitiva, se trata de un enfoque sistémico de toma de decisiones que al basarse en datos sería clave, en el sentido que se apalancaría en el uso de la analítica de los logs de *feedback* para refinar las políticas de enseñanza y así poder asegurar la excelencia escritural con el uso de una IA profundamente responsable.

Consideraciones éticas, integridad y trazabilidad

El uso de IA en la educación conlleva grandes desafíos éticos y de integridad académica que deben abordarse de manera explícita. En cuanto a ética y privacidad, se requiere obtener el consentimiento informado de los estudiantes para poder procesar sus escritos con IA, garantizando, ante todo, la anonimización de los datos. Aunado a ello, se debe aplicar el principio de minimización, es decir, se debe recolectar solo la información estrictamente necesaria para evaluar la efectividad del sistema y eliminar identificadores personales. Del mismo modo, se debe garantizar que los estudiantes que opten y tomen la decisión de no participar, reciban un apoyo equivalente con retroalimentación manual, y sin tomar ningún tipo de represalia o sanciones.

En el plano de integridad académica, es crucial subrayar el rol activo que juega el estudiante en la generación de su escritura, en efecto, es cierto que la IA brindará *feedback*, pero el trabajo intelectual final es del alumno y su producción intelectual. Se promueve, por tanto, la autoría humana, por ende, los sistemas de IA no deben figurar como autores, en ningún producto derivado, y ello parte del exigir a los estudiantes una declaración del uso de las herramientas de IA en sus procesos para salvaguardar la transparencia. Además, es aconsejable la integración de módulos de capacitación sobre el uso responsable de la IA, enfatizando que éstas son asistentes, más no, sustitutos de la reflexión crítico - analítica. Mekheimer (2025) halló que los alumnos

enfatizan la necesidad de evitar la sobre dependencia en IA y por esta circunstancia que en la documentación del estudio se requerirá reforzar la alfabetización digital y ética, priorizando habilidades de pensamiento crítico aun cuando media el uso de la IA.

Respecto a la trazabilidad técnica, se deben registrar metadatos detallados de cada ciclo de retroalimentación, conservando los *prompts* utilizados con los LLM, la versión del modelo empleado, los parámetros de generación, y logs de las interacciones sin incluir información personal. De esta forma se facilitará la auditoría de sesgos, pudiendo re-ejecutar escenarios si se sospecha de respuestas injustas del modelo.

Esto parte del establecimiento de un umbral de revisión en donde se permita que la IA al detectar ambigüedades o incoherencias en un texto, active automáticamente la intervención humana ("human-in-the-loop"), de modo que un profesor revise o complemente el *feedback*. Además, en este punto sería deseable la aplicación de auditorías periódicas de sesgos al sistema como sugieren guías de AI responsable en la educación para identificar patrones discriminatorios (UNESCO, 2024) pues la transparencia y trazabilidad son las que garantizan que el uso de IA no comprometa la equidad ni la calidad de la enseñanza, alineándose con estándares éticos internacionales.

Como todo estudio, la investigación propuesta tiene limitaciones, primeramente, los datos y contextos específicos como los cursos, asignaturas, instituciones y cohortes, pueden restringir la generalización. Los resultados dependerán de la versión del modelo LLM disponible durante la fase del experimento, siendo que las futuras actualizaciones del modelo podrían alterar la efectividad del *feedback* (Saini et al., 2024). Aunado a ello, la medición de calidad escritural estaría sujeta a la confiabilidad de las rúbricas analíticas, y es que, aunque se blindará la calificación, siempre habrá margen de subjetividad. Y, en términos de equidad, las categorías analizadas para los niveles de los estudiantes L1 y L2, podrían estar correlacionadas con factores no observados como el socioeconómico, el cual complicaría la interpretación de las brechas.

Es importante recalcar que el capítulo plantea una agenda futura centrada en la profundización de estas exploraciones, por ello se recomienda realizar seguimientos longitudinales para evaluar si las mejoras escriturales perduran en el tiempo o se acumulan con retroalimentación continua. También sería valioso extender el estudio a otras disciplinas donde la escritura tiene funcionalidades muy distintas. La integración de esta retroalimentación con analítica del aprendizaje más amplia puede enriquecer la comprensión de los mecanismos de cambio, y ante ello, se sugiere evaluar el costo, el beneficio y efectividad institucional, pues aunque la tecnología promete escalabilidad, su implementación requiere a su vez, de inversión en infraestructura y capacitación docente, situación de gran trascendencia para que se pueda materializar una adopción sostenible en el marco del ODS 4.

CONCLUSIONES

La evidencia recopilada muestra que la retroalimentación generativa alimentada por LLM tiene un potencial considerable para elevar la calidad escritural en la educación superior, al ofrecer comentarios detallados y oportunos que apalancan el aprendizaje autorregulado del alumno. Asimismo, este enfoque puede contribuir a una mayor equidad evaluativa al aplicar criterios constantes, pues de este modo se reducen las disparidades de base entre grupos y se amplía el acceso a *feedback* significativo. Sin embargo, es importante destacar que estos efectos positivos dependen de un diseño cuidadoso que requiere de la supervisión humana y de una arquitectura transparente para evitar sesgos automáticos, pues en la literatura actual faltan todavía investigaciones robustas sobre los efectos diferenciales precisos por subgrupos y

sobre el despliegue institucional a gran escala.

Como recomendaciones, se exhorta a docentes y programas de estudios a adoptar gradualmente el copilotaje pedagógico con IA; esto implica definir criterios claros de adopción como la exigencia de rúbricas explícitas y el entrenamiento a profesores para interpretar el *feedback* automatizado. A nivel institucional, se debe planificar un escalado con auditoría continua con comités de ética de IA educativa, vinculando la estrategia de IA a los objetivos de equidad establecidos en el ODS 4. Principalmente, se enfatiza la necesidad latente de continuar generando evidencia científica, evidencia que parte de la creación de redes de investigación universitaria que comparta datos de forma protegida sobre sus lecciones aprendidas.

Finalmente, y de cara a la hoja de ruta para los años que restan para alcanzar el *deadline* de la agenda 2030, este capítulo vislumbra un futuro en el cual la excelencia escritural y la equidad se realzan gracias a una IA responsable. Para lograrlo, se deberán trazar rutas de mejora iterativas que monitoricen los nuevos desarrollos de LLM, integre *feedback* automatizado en planes de estudio, y forme a toda la comunidad académica en su uso crítico. Sólo así la educación de calidad prevista en el ODS 4 podrá alcanzar su meta de aprendizajes relevantes y efectivos para todos los estudiantes, impulsada por una retroalimentación escalable y equitativa.

REFERENCIAS

Afrin, T., Kashefi, O., Olshefski, C., Litman, D., Hwa, R., & Godley, A. (2021). Effective interfaces for student-driven revision sessions for argumentative writing. In Proceedings of the 2021 CHI conference on human factors in computing systems. https://doi.org/10.1145/3411764.3445683

Alcudia Olán, Z. V., & Morales Vázquez, E. (2025). Aprendizaje Autorregulado y Aplicaciones Digitales. Estudios Y Perspectivas Revista Científica Y Académica. 5(3), 3815–3830. https://doi.org/10.61384/r.c.a.v5i3.1434

Benner, D. (2024). Introducing Generative *Feedback* to Chatbots in Digital Higher Education. Proceedings of WI2024, 99. https://aisel.aisnet.org/wi2024/99

Chen, D-G., Pan, W., & Kainz, K. (2021). *Stepped-wedge* Cluster Randomized Controlled Trial for Intervention Research: Design and Analysis. Journal of the Society for Social Work and Research. 12 (2). https://doi.org/10.1086/714135

Chen, D., Arnold, K., Sukhdeo, R., Farag J., & Raman, S. (2025). Concordance with CONSORT-AI guidelines in reporting of randomised controlled trials investigating artificial intelligence in oncology: a systematic review. BMJ Oncology. 2025;4:e000733. https://doi.org/10.1136/bmjonc-2025-000733

Glüsing, R., Fleckenstein, J., Schmidt, F. T. C., & Möller, J. (2025). LLM *feedback* for academic writing: Effects on students' performance and engagement. https://doi.org/10.2139/ssrn.5445319

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the automated assessment of student essay content to highly informative *feedback*: A case study. International Journal of Artificial Intelligence in Education, 34(3). https://doi.org/10.1007/s40593-023-00387-6

González, L., O'Neil-Gonzalez, K., Eberhardt-Alstot, M., McGarry., M., & Van Tyne, G. (15 de agosto de 2024). Leveraging Generative AI for Inclusive Excellence in Higher Education. EDUCASE REVIEW. https://n9.cl/85lrnr

Hattie, J., & Timperley, H. (2007). El poder de la retroalimentación. Revista de Investigación Educativa , 77 (1), 81-112. https://doi.org/10.3102/003465430298487

Holgersson, T., Nordström, L., & Öner, Ö. (2015). On regression modelling with dummy variables versus separate regressions per group: comment on Holgersson et al. Journal of Applied Statistics, 43(8), 1564–1565. https://doi.org/10.1080/02664763.2015.1092711

Impey, C., Wenger, M., Garuda, N., Golchin, S., & Stamer, S. (2025). Using Large Language Models for Automated Grading of Student Writing about Science. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-024-00453-7

Jovic, M., Papakonstantinidis, S., & Kirkpatrick, R. (2025). From red ink to algorithms: investigating the use of large language models in academic writing *feedback*. Language Testing in Asia. 15 (59). https://doi.org/10.1186/s40468-025-00389-2

Mekheimer, M. (2025). Generative AI-assisted *feedback* and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Education 4, 170. https://doi.org/10.1007/s44217-025-00602-7

Quezada, S., & Salinas, C. (2021). Modelo de retroalimentación para el aprendizaje: Una propuesta basada en la revisión de literatura. Revista mexicana de investigación educativa, 26(88).225-251. https://ojs.rmie.mx/index.php/rmie/article/view/219

Ren, F., & Ren, C. (2025). Optimizing physical education strategies through circular intuitionistic Fuzzy Bonferroni based school policy formulation. Scientific reports, 15(1), 21616. https://doi.org/10.1038/s41598-025-03363-3

Romero-González, L. (2024). La evaluación inclusiva para una evaluación justa. Know and share psychology, 5(3), 16-22. https://doi.org/10.25115/kasp.v5i3.10137

Saini, A. K., Cope, L., Kalantzis, M., & Zapata, R. (2024). The future of *feedback*: Integrating peer and generative AI reviews to support student work. https://doi.org/10.35542/osf.io/x3dct

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. Journal of graduate medical education, 4(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Thomas, S. L., Penn, A., Hauser, M., Severyn, A., Radev, D., & Liu, R. (2025). LLM-Generated *Feedback* Supports Learning if Learners Choose to Use It. Arvix. https://doi.org/10.48550/arXiv.2506.17006

Tian, L., & Liu, Q. (2022). Self-regulated writing strategy use when revising upon automated, peer, and teacher *feedback* in an online English as a foreign language writing course. Frontiers in Psychology, 13, 873170. https://doi.org/10.3389/fpsyg.2022.873170

UNESCO (2024). Guía para el uso de IA generativa en educación e investigación. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000389227_spa

Valenzuela, R., & Pérez, A. . (2025). Inteligencia artificial en educación superior: ¿un reemplazo para los profesores o una herramienta de apoyo?. Revista Iberoamericana De Investigación En Educación, (9). https://doi.org/10.58663/riied.vi9.221

Wambsganss, S., Günther, A., Ketelhut, D. J., & Scheiter, K. (2022). Enhancing argumentative writing with automated *feedback* and social comparison nudging. Computers & Education, 191, 104644. https://doi.org/10.1016/j.compedu.2022.104644

Xia, Q., Weng, X., Fan, O., Lin, T., & Chiu, T.K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, 21, 1-22. https://doi.org/10.1186/s41239-024-00468-z

Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI *feedback* on picture-cued writing tasks. Computers & Education: Artificial Intelligence, 9, Art. 100450. https://doi.org/10.1016/j.caeai.2025.100450

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. En M. Boekaerts, P.R. Pintrich y M. Zeidner (Eds.), Handbook of self-regulation.13-40. https://doi.org/10.1016/B978-012109890-2/50031-7

CONFLICTO DE INTERESES

No se presentan conflictos de interés de ninguna índole.

FINANCIACIÓN

El trabajo no ha recibido subvención de organismos de financiación en los sectores públicos, comerciales o sin fines de lucro.

USO DE INTELIGENCIA ARTIFICIAL

Se empleó IA para la traducción del resumen y corrección de estilo.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Carluys Suescum Coelho.
Curación de datos: Car-Emyr Suescum Coelho, Carluys Suescum Coelho.
Análisis formal: Carelys Suescum Coelho, Carlysmar Suescum Coelho.
Investigación: Carelys Suescum Coelho, Car-Emyr Suescum Coelho, Carluys Suescum Coelho, Carlysmar Suescum Coelho.
Metodología: Car-Emyr Suescum Coelho, Carluys Suescum Coelho.
Gestión del proyecto: Carluys Suescum Coelho.
Supervisión: Carluys Suescum Coelho.
Validación: Carluys Suescum Coelho, Car-Emyr Suescum Coelho.
Redacción - borrador original: Carluys Suescum Coelho, Car-Emyr Suescum Coelho.
Redacción - corrección y edición: Carluys Suescum Coelho, Carlysmar Suescum Coelho.