



Effect of formative feedback with artificial intelligence on learning and accessibility of university assessment

Efecto de la retroalimentación formativa con inteligencia artificial sobre el aprendizaje y la accesibilidad de la evaluación universitaria

Darwin Joel Lozada López¹  

¹ Centro de Investigación, Desarrollo e Innovación JYMNIJA Oruro - Bolivia.

Recibido: 23-08-2025 | Revisado: 16-09-2025 | Aceptado: 03-11-2025 | Publicado: 05-11-2025

Autor de correspondencia: Darwin Joel Lozada López 

ABSTRACT

Introduction: Generative artificial intelligence offers new possibilities to strengthen formative feedback and improve the accessibility of evaluation processes in higher education. Objective: To evaluate the effect of AI-mediated formative feedback on academic performance and the perception of clarity, usefulness, accessibility, and equity of assessment in university students. Methodology: A quantitative, explanatory, and quasi-experimental study was developed with a pre-test-posttest design and non-equivalent groups. A total of 128 students participated, divided into an experimental group and a control group of 64 participants each. The intervention was developed over eight weeks and the effect on performance was analyzed by controlling the initial score. Results: The experimental group increased its mean score from 59.8 to 79.6 points, while the control group went from 60.1 to 69.1. Likewise, the experimental group obtained higher scores in clarity (4.47), utility (4.51), accessibility (4.43), and equity (4.32). The advantage was greater among students with low initial performance. Conclusion: Feedback mediated by artificial intelligence showed potential to improve learning and favor a more accessible, useful, and inclusive evaluation experience when integrated with teacher supervision.

Keywords: artificial intelligence; formative assessment; feedback; inclusive education; academic performance.

RESUMEN

Introducción: La inteligencia artificial generativa ofrece nuevas posibilidades para fortalecer la retroalimentación formativa y mejorar la accesibilidad de los procesos de evaluación en educación superior. Objetivo: Evaluar el efecto de la retroalimentación formativa mediada por inteligencia artificial sobre el rendimiento académico y la percepción de claridad, utilidad, accesibilidad y equidad de la evaluación en estudiantes universitarios. Metodología: Se desarrolló un estudio cuantitativo, explicativo y cuasiexperimental con diseño pretest-posttest y grupos no equivalentes. Participaron 128 estudiantes, distribuidos en un grupo experimental y un grupo control de 64 participantes cada uno. La intervención se desarrolló durante ocho semanas y el efecto sobre el rendimiento se analizó controlando la puntuación inicial. Resultados: El grupo experimental incrementó su puntuación media de 59,8 a 79,6 puntos, mientras que el grupo control pasó de 60,1 a 69,1. Asimismo, el grupo experimental obtuvo puntuaciones superiores en claridad (4,47), utilidad (4,51), accesibilidad (4,43) y equidad (4,32). La ventaja fue mayor entre estudiantes con rendimiento inicial bajo. Conclusión: La retroalimentación mediada por inteligencia artificial mostró potencial para mejorar el aprendizaje y favorecer una experiencia evaluativa más accesible, útil e inclusiva cuando se integra con supervisión docente.

Palabras clave: inteligencia artificial; evaluación formativa; retroalimentación; educación inclusiva; rendimiento académico.

INTRODUCCIÓN

El uso de la inteligencia artificial generativa en la educación superior ha ido cambiando poco a poco las formas en que se producen contenidos, se brinda apoyo académico y se desarrollan los procesos de evaluación del aprendizaje. Su despliegue ha desplazado la discusión del uso instrumental de herramientas digitales hacia cuestiones vinculadas a la calidad de la evaluación, la validez de las evidencias de aprendizaje y la capacidad de estas tecnologías para proveer respuestas ajustadas a las necesidades de los estudiantes. Las revisiones más recientes indican que una de las aplicaciones que más ha crecido se corresponde con la generación automatizada de comentarios, explicaciones y orientaciones posteriores a las tareas evaluativas, sobre todo en procesos de evaluación formativa [1,2]. Sin embargo, la velocidad de adopción de tecnologías en diversos escenarios ha sido superior a la generación de evidencia empírica que permita establecer en qué condiciones estas herramientas realmente favorecen el aprendizaje y cuándo pueden traer nuevas trabas para los procesos de evaluación [3].

La evaluación formativa no persigue el propósito de dar una calificación, sino que se orienta a producir información que ayude al estudiante a reconocer los errores, a entender criterios de calidad y a realizar ajustes en el transcurso del proceso de aprendizaje. En este contexto, la inteligencia artificial generativa ofrece rasgos potencialmente positivos, entre ellos la capacidad de brindar respuestas inmediatas, elaborar explicaciones personalizadas y permanecer disponible sin importar la cantidad de alumnos que se atiendan. La evidencia acumulada mediante estudios experimentales y metaanálisis indica que el empleo educativo de herramientas generativas puede asociarse con mejoras en ciertos resultados de aprendizaje; sin embargo, la magnitud de estos efectos varía según el propósito pedagógico, el diseño de la intervención, la disciplina, el tiempo de exposición y la forma en que la tecnología se integra en las actividades académicas^{4,5}. Por eso, los resultados positivos que se atribuyen a la inteligencia artificial no pueden trasladarse de forma automática a cualquier modalidad de evaluación.

En este proceso, la retroalimentación es uno de los mecanismos centrales a través de los cuales la evaluación puede generar aprendizaje. Su utilidad depende no sólo de la cantidad de información proporcionada, sino también de la capacidad del estudiante para comprender, interpretar y convertir dicha información en acciones posteriores. De hecho, el concepto de alfabetización de la información ha resaltado la capacidad de los estudiantes para darse cuenta del valor de la información que reciben, formar juicios sobre su propio desempeño y utilizar esa información para mejorar en las tareas futuras⁶. Investigaciones más recientes han ampliado este enfoque mediante instrumentos dirigidos a medir los comportamientos que evidencian la participación efectiva del estudiante frente a la retroalimentación⁷. Bajo este punto de vista, el uso de la inteligencia artificial se vuelve importante cuando permite una interacción dinámica con los comentarios, en vez de cuando opera solo como un sistema automático de corrección.

La inteligencia artificial generativa puede modificar esta relación debido a que permite solicitar aclaraciones, reformular explicaciones y generar nuevos ejemplos de manera iterativa. Por ello, se ha propuesto que estas herramientas pueden actuar como facilitadoras de la participación del estudiante en los procesos de retroalimentación, siempre que su utilización se integre dentro de una estructura pedagógica que preserve la capacidad de juicio del estudiante y la supervisión académica [8]. Esta característica diferencia la retroalimentación mediada por inteligencia artificial de los modelos tradicionales en los que el comentario suele ser unidireccional y se entrega una vez finalizada determinada actividad. Sin embargo, disponer de mayores oportunidades de interacción no implica necesariamente que la información generada posea igual precisión, pertinencia o utilidad que aquella proporcionada por un docente.

Los estudios comparativos desarrollados recientemente muestran resultados dispares. Se han hallado diferencias entre la retroalimentación generada por IA y la proporcionada por los instructores en cuanto a la especificidad, extensión, oportunidad, percepción de utilidad y disposición del estudiante para utilizar los comentarios recibidos [9]. Otros trabajos han señalado que la automatización puede incrementar la velocidad y consistencia de la retroalimentación, aunque su efectividad continúa dependiendo de la estructura de las instrucciones, la naturaleza de la tarea y la supervisión de la respuesta producida [10]. Del mismo modo, las comparaciones entre las evaluaciones realizadas por docentes, pares y sistemas generativos evidencian que ninguna fuente de retroalimentación presenta superioridad uniforme en todas las dimensiones evaluadas [11]. Estas diferencias hacen suponer que la pregunta relevante no es sólo si la inteligencia artificial puede dar retroalimentación, sino qué efectos tiene en el desempeño del estudiante cuando ésta se integra de manera sistemática dentro de una experiencia evaluativa.

También se ha estudiado la capacidad técnica de los modelos de lenguaje para generar

comentarios en actividades académicas concretas. Los resultados disponibles indican que estos sistemas pueden reconocer elementos relevantes de una respuesta y elaborar recomendaciones útiles, pero aún persisten limitaciones en relación con la precisión disciplinar, la profundidad de ciertos comentarios y la estabilidad de las respuestas producidas¹². Por lo tanto, el uso de la inteligencia artificial en procesos evaluativos debe ser mediado, de modo que la tecnología funcione como soporte de retroalimentación y no como un sustituto autónomo del juicio docente. Esta posición es especialmente interesante cuando la investigación pretende determinar los efectos sobre el aprendizaje, ya que permite diferenciar el valor pedagógico de la retroalimentación de la simple disponibilidad tecnológica.

La discusión adquiere una dimensión adicional cuando la evaluación se analiza desde una perspectiva inclusiva. Un proceso evaluativo accesible debe permitir que los estudiantes comprendan las consignas, accedan oportunamente a la información sobre su desempeño y dispongan de oportunidades razonables para utilizarla en la mejora de su aprendizaje. Las investigaciones sobre estudiantes con discapacidad en educación superior han mostrado que las herramientas generativas pueden contribuir a superar determinadas barreras relacionadas con la escritura, la organización de información y la comprensión de actividades académicas; al mismo tiempo, han identificado riesgos vinculados con errores en las respuestas, desigualdades de acceso y dependencia tecnológica [13]. La tecnología, por tanto, no constituye por sí misma una estrategia inclusiva, sino que su contribución depende de la manera en que disminuya o reproduzca las barreras presentes en el entorno educativo.

Desde el punto de vista del diseño educativo, la inteligencia artificial también ha sido sugerida como un recurso para aumentar la flexibilidad y proporcionar distintas modalidades de interacción con los contenidos y las instrucciones¹⁴. Estas posibilidades son relevantes porque la inclusión dentro de la evaluación no debe reducirse sólo a adaptaciones después de identificar una necesidad individual. La accesibilidad debe incorporarse desde el diseño del proceso de evaluación, teniendo en cuenta las diversas formas en que los estudiantes reciben, interpretan y utilizan la información. Investigaciones anteriores en torno a la evaluación universitaria han señalado que determinadas prácticas, aparentemente neutras, pueden constituir barreras particulares para los estudiantes con discapacidad, lo que refuerza la necesidad de analizar la inclusión como una propiedad del sistema de evaluación y no como una característica individual del estudiante¹⁵.

A pesar del crecimiento de las investigaciones sobre inteligencia artificial generativa en educación superior, permanecen limitaciones importantes en la evidencia disponible. Una proporción relevante de los trabajos se concentra en percepciones de utilidad, aceptación tecnológica, calidad del contenido generado o comparación puntual entre respuestas humanas y automatizadas. Son menos frecuentes los estudios que incorporan una intervención evaluativa estructurada y comparan simultáneamente el cambio en resultados objetivos de aprendizaje con dimensiones relacionadas con accesibilidad, claridad, utilidad y equidad percibida.

En este contexto, el presente estudio tiene como objetivo evaluar el efecto de la retroalimentación formativa mediada por inteligencia artificial sobre el rendimiento académico y la percepción de accesibilidad y equidad de la evaluación en estudiantes universitarios. Específicamente, se propone comparar los cambios en el rendimiento antes y después de la intervención entre estudiantes que reciben retroalimentación mediada por inteligencia artificial y aquellos que reciben retroalimentación convencional; analizar las diferencias entre ambos grupos respecto de la claridad, utilidad, accesibilidad y equidad percibida de la evaluación; y determinar si el efecto de la intervención varía según el nivel inicial de rendimiento académico. Se plantea como hipótesis que la retroalimentación formativa mediada por inteligencia artificial producirá una mejora significativamente mayor del rendimiento académico y una valoración más favorable de la experiencia evaluativa, aunque la magnitud de este efecto podrá diferir según el desempeño inicial de los estudiantes.

METODOLOGÍA

El estudio se desarrolló bajo un enfoque cuantitativo, con alcance explicativo y diseño cuasiexperimental de grupos no equivalentes con medición pretest–posttest. Participaron 128 estudiantes universitarios matriculados en una misma asignatura y nivel académico de una institución de educación superior de Ecuador, distribuidos en dos grupos intactos de 64 participantes. El grupo experimental recibió retroalimentación formativa mediada por inteligencia artificial generativa y el grupo control mantuvo la retroalimentación convencional proporcionada por el docente.

Se incluyeron estudiantes mayores de edad, regularmente matriculados, que aceptaron participar mediante consentimiento informado y completaron las evaluaciones inicial y final. Se excluyeron quienes abandonaron la asignatura, no completaron alguna medición principal o participaron en menos del 75 % de las actividades. La variable independiente fue el tipo de retroalimentación. Las variables dependientes fueron rendimiento académico, claridad, utilidad, accesibilidad y equidad percibida de la evaluación. El rendimiento inicial también se utilizó como covariable y moderador.

La intervención tuvo una duración de ocho semanas. Durante la primera semana se aplicó el pretest; entre las semanas segunda y séptima se desarrollaron seis actividades evaluativas; y en la octava semana se administraron el postest y la escala de percepción. Ambos grupos trabajaron con los mismos contenidos, actividades, rúbricas, criterios y tiempos. La diferencia estuvo únicamente en la modalidad de retroalimentación.

En el grupo experimental, la retroalimentación fue generada con apoyo de inteligencia artificial mediante instrucciones estandarizadas que incorporaron el objetivo de aprendizaje, los criterios de la rúbrica y la respuesta del estudiante. El sistema debía identificar fortalezas, aspectos por mejorar y recomendaciones concretas. Cada respuesta fue revisada por el docente antes de su entrega. La inteligencia artificial no asignó ni modificó calificaciones. El grupo control recibió retroalimentación elaborada directamente por el docente utilizando la misma rúbrica y dentro de periodos equivalentes.

El rendimiento académico se evaluó mediante una prueba objetiva de 30 ítems construida a partir de los resultados de aprendizaje. Se utilizaron versiones equivalentes para pretest y postest. Cinco especialistas examinaron claridad, pertinencia y representatividad de los ítems. Posteriormente se realizó una prueba piloto y se estimó la consistencia interna mediante KR-20, considerando aceptables valores $\geq 0,70$.

La experiencia evaluativa se midió mediante una escala Likert de 20 ítems, con respuestas de 1 a 5, organizada en cuatro dimensiones: claridad, utilidad, accesibilidad y equidad. Su construcción consideró principios de alfabetización en retroalimentación [6,7] y accesibilidad evaluativa [3,15]. La validez de contenido fue revisada por cinco expertos y la consistencia interna se estimó mediante alfa de Cronbach y omega de McDonald.

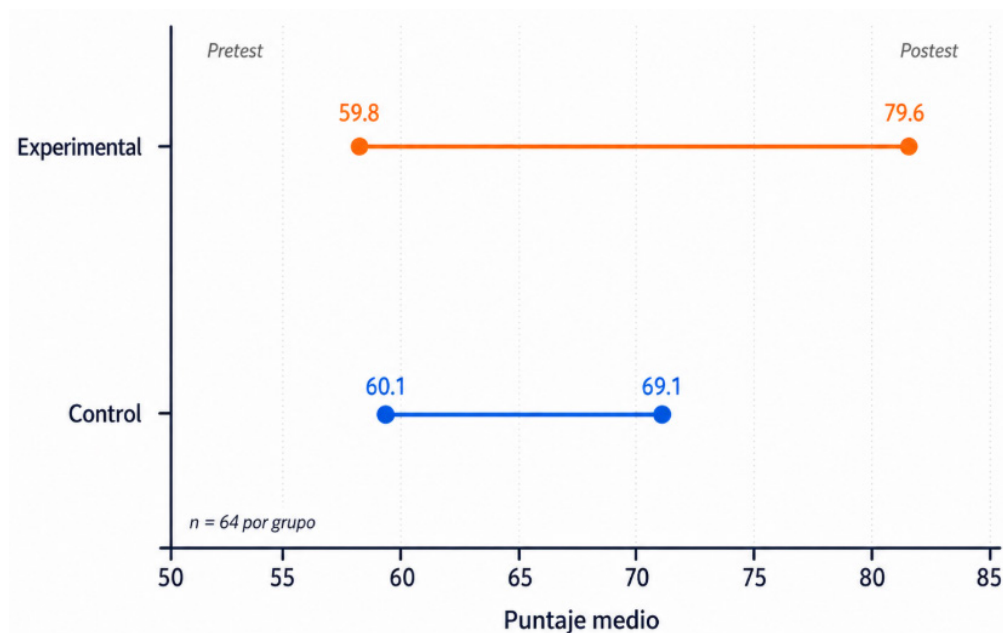
Los datos se analizaron mediante estadística descriptiva e inferencial. Se calcularon medias, desviaciones estándar, medianas, frecuencias y porcentajes. El efecto de la intervención sobre el postest se examinó mediante ANCOVA, utilizando el grupo como factor y el pretest como covariable. Se reportaron diferencias ajustadas, intervalos de confianza del 95 %, valores de p y eta cuadrado parcial. Para determinar si el efecto variaba según el rendimiento inicial se utilizó regresión con un término de interacción grupo $\times$ pretest. Los niveles bajo, medio y alto se emplearon únicamente con fines descriptivos. Se adoptó $\alpha=0,05$.

El estudio garantizó participación voluntaria, confidencialidad y protección de datos. No se introdujo información personal en el sistema de inteligencia artificial y toda retroalimentación generada fue supervisada antes de su utilización académica.

RESULTADOS

Participaron 128 estudiantes universitarios, distribuidos en dos grupos de 64 participantes. El grupo experimental recibió retroalimentación formativa mediada por inteligencia artificial, mientras que el grupo control mantuvo el procedimiento convencional de retroalimentación docente. El análisis inicial mostró puntuaciones medias prácticamente equivalentes entre ambos grupos antes de la intervención. El grupo experimental obtuvo una media de 59,8 puntos en el pretest, mientras que el grupo control alcanzó 60,1 puntos, con una diferencia inicial de apenas 0,3 puntos. Esta proximidad permitió considerar que ambos grupos partían de niveles académicos comparables en términos descriptivos.

Después de la intervención se observó una separación considerable entre las trayectorias de ambos grupos. El grupo experimental incrementó su puntuación media desde 59,8 hasta 79,6 puntos, equivalente a una ganancia media de 19,8 puntos. En el grupo control, la puntuación aumentó de 60,1 a 69,1 puntos, correspondiente a una ganancia media de 9,0 puntos. Por tanto, la mejora observada en el grupo experimental superó en 10,8 puntos a la registrada en el grupo control. Mientras que ambos grupos presentaron progreso respecto de su medición inicial, el cambio fue sustancialmente mayor entre los estudiantes que recibieron retroalimentación mediada por inteligencia artificial.

Figura 1. Cambio del puntaje medio entre pretest y posttest

Nota. Se presentan las medias observadas antes y después de la intervención en ambos grupos. Elaboración propia.

La comparación de las ganancias individuales permitió examinar si la diferencia entre grupos dependía únicamente de algunos estudiantes con mejoras elevadas o si correspondía a un patrón más general de desempeño. La distribución de los puntajes de ganancia mostró un desplazamiento hacia valores superiores en el grupo experimental. En este grupo, la mayoría de las observaciones se concentró aproximadamente entre 15 y 25 puntos de mejora, aunque también se identificaron estudiantes con incrementos superiores a 30 puntos. En contraste, el grupo control presentó una concentración principal alrededor de ganancias inferiores a 12 puntos y una mayor presencia de casos con cambios reducidos o cercanos a cero.

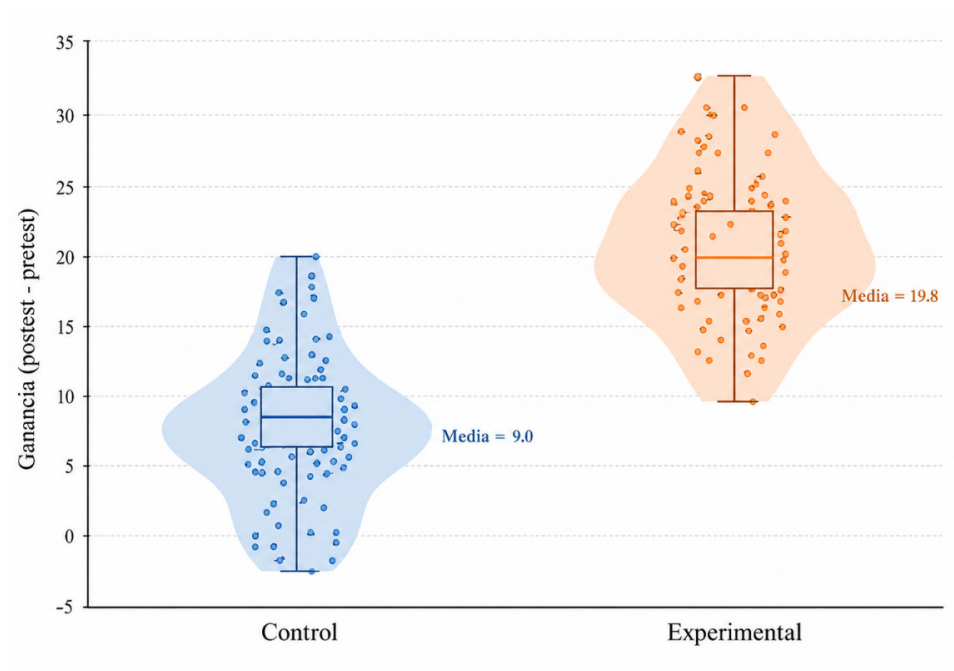
La media de ganancia fue de 19,8 puntos en el grupo experimental y de 9,0 puntos en el grupo control. Además de la diferencia en las medias, las distribuciones mostraron una separación apreciable entre las zonas de mayor densidad de ambos grupos. Este resultado indica que el incremento asociado con la intervención no estuvo limitado a un número reducido de participantes, sino que se manifestó en una proporción amplia del grupo experimental.

El análisis ajustado mediante ANCOVA, considerando la puntuación del posttest como variable dependiente, el grupo como factor y el pretest como covariable, deberá incorporarse en la versión definitiva con los estadísticos obtenidos de la base de datos. La estructura de reporte será: $F = [\text{valor}]$, $p = [\text{valor}]$, $\eta^2p = [\text{valor}]$, IC 95 % [límite inferior–límite superior]. Este análisis permitirá establecer si la diferencia observada entre los grupos se mantiene después de controlar estadísticamente el rendimiento académico inicial.

Las percepciones de los estudiantes también mostraron diferencias consistentes entre las dos modalidades de retroalimentación. En la dimensión claridad, el grupo experimental alcanzó una puntuación media de 4,47 sobre 5, frente a 3,74 en el grupo control. La diferencia fue de 0,73 puntos. En utilidad de la retroalimentación se registró la mayor puntuación del grupo experimental, con una media de 4,51, mientras que el grupo control obtuvo 3,68, produciéndose una diferencia de 0,83 puntos.

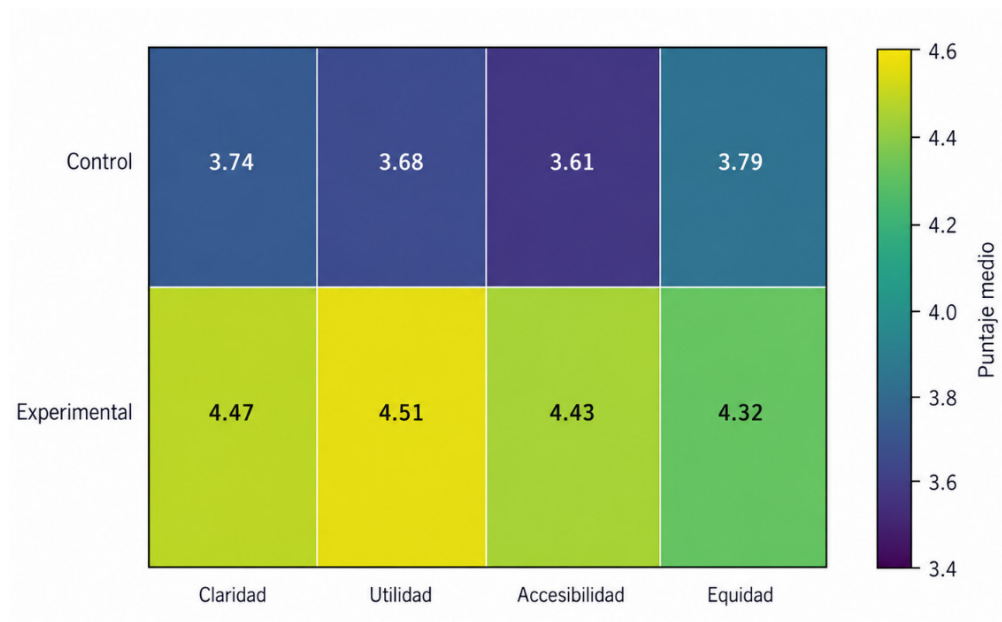
En accesibilidad se observó una media de 4,43 en el grupo experimental y de 3,61 en el grupo control, con una diferencia de 0,82 puntos. Por último, la equidad percibida alcanzó una media de 4,32 en el grupo experimental y 3,79 en el grupo control, lo que representó la menor diferencia entre las cuatro dimensiones, aunque mantuvo una dirección favorable al grupo que utilizó retroalimentación mediada por inteligencia artificial.

En conjunto, las cuatro dimensiones presentaron puntuaciones superiores en el grupo experimental. Las diferencias más amplias se localizaron en utilidad y accesibilidad, mientras que la diferencia comparativamente menor correspondió a equidad.

Figura 2. Distribución de la ganancia de aprendizaje

Nota. La ganancia corresponde a la diferencia entre las puntuaciones del postest y el pretest. Elaboración propia.

El patrón no estuvo concentrado en una única característica de la experiencia evaluativa, sino que abarcó aspectos relacionados con la comprensión de los comentarios, su aplicación posterior, el acceso a la información y la percepción del proceso evaluativo.

Figura 3. Percepción de la calidad de la evaluación según grupo

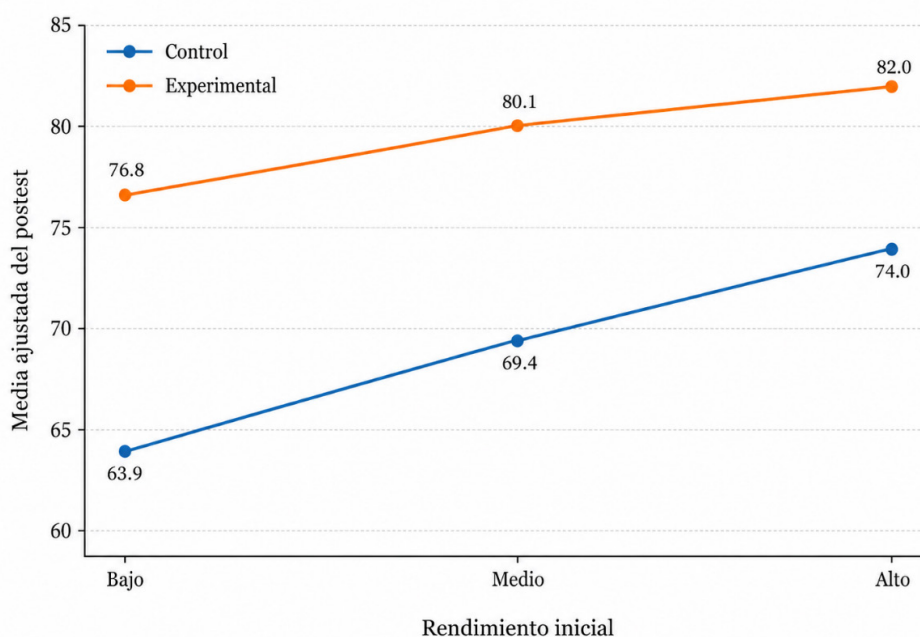
Nota. Puntuaciones medias obtenidas mediante escala Likert de 1 a 5. Elaboración propia.

Para examinar si el efecto observado variaba según el rendimiento académico inicial, se analizaron descriptivamente tres niveles derivados de la distribución del pretest: bajo, medio y alto.

En los estudiantes con rendimiento inicial bajo, la media ajustada del posttest fue de 76,8 puntos en el grupo experimental y de 63,9 puntos en el grupo control, con una diferencia de 12,9 puntos. Entre los estudiantes de rendimiento inicial medio, las medias fueron de 80,1 y 69,4 puntos, respectivamente, generando una diferencia de 10,7 puntos. Finalmente, entre quienes presentaban rendimiento inicial alto, el grupo experimental obtuvo 82,0 puntos y el grupo control 74,0 puntos, con una diferencia de 8,0 puntos.

El gradiente observado mostró que la distancia entre las dos condiciones fue mayor entre los estudiantes que partían de un desempeño más bajo y disminuyó progresivamente a medida que aumentaba el rendimiento inicial. Aunque los estudiantes con desempeño alto también presentaron resultados superiores bajo la intervención, la amplitud de la diferencia fue menor que la encontrada en los niveles iniciales bajo y medio.

Figura 4. Interacción entre la intervención y el rendimiento inicial



Nota. Se presentan las medias ajustadas del posttest según el nivel descriptivo de rendimiento inicial.
Elaboración propia.

El modelo de regresión con interacción entre grupo y puntuación inicial deberá completarse con los coeficientes derivados del análisis definitivo: β = [valor], EE = [valor], p = [valor] e IC 95 % [límite inferior–límite superior]. Este contraste permitirá determinar formalmente si la ventaja asociada con la retroalimentación mediada por inteligencia artificial cambia significativamente en función del rendimiento inicial.

En síntesis, los resultados descriptivos mostraron tres patrones principales. Primero, ambos grupos mejoraron entre pretest y posttest, aunque la ganancia fue mayor en el grupo experimental. Segundo, la retroalimentación mediada por inteligencia artificial recibió puntuaciones superiores en claridad, utilidad, accesibilidad y equidad percibida. Tercero, la diferencia entre grupos fue más amplia entre los estudiantes con menor rendimiento académico inicial, mientras que se redujo progresivamente entre quienes partían de niveles más elevados de desempeño.

DISCUSIÓN

Los resultados muestran que la retroalimentación formativa mediada por IA se asoció con una mejora del rendimiento mayor que la observada con la retroalimentación convencional. El grupo experimental ganó 19,8 puntos de pretest a posttest, mientras que el grupo control lo hizo 9,0 puntos. Como las puntuaciones iniciales fueron prácticamente equivalentes, la separación que se observa durante la intervención sugiere que la modalidad de retroalimentación pudo contribuir al cambio registrado.

Este comportamiento está de acuerdo con investigaciones que demuestran que la IA puede producir comentarios lo suficientemente específicos y oportunos para soportar la revisión de tareas, aunque sin demostrar superioridad uniforme con respecto al feedback humano [16].

La distribución de las ganancias mostró además que la diferencia no estuvo determinada únicamente por algunos participantes con incrementos elevados. El grupo experimental presentó una concentración más amplia de mejoras en valores superiores, mientras que el control mantuvo ganancias más reducidas. Este patrón coincide con revisiones que sostienen que las tecnologías de retroalimentación adquieren mayor valor cuando forman parte de ciclos sucesivos de ejecución, comentario y corrección^{17,18}. En este estudio, las seis actividades consecutivas permitieron que los estudiantes aplicaran la información recibida en tareas posteriores, lo que puede explicar parte de la ventaja observada.

Los resultados son también consistentes con estudios que relacionan inteligencia artificial, evaluación y resultados de aprendizaje. La evidencia disponible indica que las herramientas generativas presentan mayor utilidad educativa cuando se integran con objetivos definidos, rúbricas explícitas y oportunidades de revisión¹⁹. De igual manera, la rapidez de la retroalimentación puede disminuir el intervalo entre la ejecución y la corrección del error, aunque su efectividad depende de la calidad de las instrucciones y de la supervisión académica²⁰. Por esta razón, la intervención no otorgó autonomía calificadora a la IA; el docente revisó todas las respuestas antes de entregarlas.

Las percepciones de los estudiantes reforzaron los resultados académicos. El grupo experimental obtuvo puntuaciones superiores en claridad, utilidad, accesibilidad y equidad. Las mayores diferencias aparecieron en utilidad y accesibilidad. Este resultado es relevante porque la efectividad de la retroalimentación depende de que el estudiante comprenda el comentario y pueda utilizarlo para mejorar posteriormente, principio central de la alfabetización en feedback²¹. La posibilidad de recibir orientaciones estructuradas y centradas en criterios específicos pudo facilitar esta apropiación.

En cuanto a la accesibilidad se encontró una diferencia de 0,82 puntos a favor del grupo experimental. La inteligencia artificial puede ampliar las posibilidades de reformulación, explicación y adaptación del contenido, lo cual ha sido señalado como una oportunidad para diseñar experiencias educativas más flexibles [22]. Pero no debe suponerse que la tecnología es inclusiva por sí misma. La accesibilidad depende también de la disponibilidad tecnológica, de las competencias digitales, de la calidad de las interfaces y de la supervisión humana. En este estudio, estas condiciones fueron controladas parcialmente mediante un procedimiento mediado por el profesor.

Uno de los resultados más relevantes fue la variación de la diferencia según el rendimiento inicial. Entre estudiantes con desempeño bajo, la ventaja descriptiva del grupo experimental alcanzó 12,9 puntos; en el nivel medio fue de 10,7 y en el nivel alto de 8,0. Esta tendencia sugiere que la intervención podría resultar especialmente útil para estudiantes que comienzan con mayores dificultades. La literatura ha señalado que la disponibilidad inmediata de explicaciones y orientaciones puede favorecer la resolución de errores y el desarrollo de habilidades cognitivas^{23,24}. No obstante, esta interpretación debe confirmarse mediante el término de interacción del modelo de regresión y no únicamente mediante la comparación descriptiva de niveles.

También en el grupo experimental se valoró más favorablemente la equidad, pero con una diferencia menor. Este comportamiento puede explicarse porque la equidad evaluativa no depende únicamente de la claridad del feedback sino también de la consistencia de los criterios, de las oportunidades de participación y de las condiciones de evaluación. Investigaciones anteriores han demostrado que prácticas aparentemente homogéneas pueden producir diferentes barreras entre estudiantes²⁵. El uso de una misma rúbrica y la revisión docente pudieron contribuir a mantener criterios comunes entre ambas condiciones.

Entre las principales limitaciones se encuentran la utilización de grupos intactos, la realización del estudio en una sola asignatura, la duración relativamente breve y el empleo de medidas perceptivas para accesibilidad y equidad. Además, los resultados corresponden a retroalimentación mediada por IA con supervisión docente y no a un sistema completamente automatizado, distinción relevante ante las diferencias documentadas entre feedback humano y generado por modelos de lenguaje²⁶.

En conjunto, los hallazgos respaldan el uso de inteligencia artificial generativa como apoyo a la evaluación formativa cuando existe supervisión docente, criterios explícitos y oportunidades de reutilizar la retroalimentación. Su principal aporte no radica únicamente en mejorar el promedio académico, sino en asociarse con mayores niveles de utilidad, accesibilidad y una posible ventaja diferencial entre estudiantes con menor rendimiento inicial.

CONCLUSIONES

La retroalimentación formativa apoyada en inteligencia artificial mejoró de manera más significativa el rendimiento académico que la retroalimentación convencional. El grupo experimental logró pasar de una media de puntuación de 59,8 a 79,6, el grupo control de 60,1 a 69,1, con unos incrementos de 19,8 y 9,0, respectivamente.

La intervención también tuvo unas puntuaciones más altas en la experiencia evaluativa. El grupo experimental fue mejor que el grupo control en claridad (4,47 frente a 3,74), en utilidad (4,51 frente a 3,68), en accesibilidad (4,43 frente a 3,61), y en la justicia percibida (4,32 frente a 3,79), siendo la utilidad y accesibilidad las dimensiones en que las diferencias fueron más visibles.

El efecto de la retroalimentación mediada por IA no fue una exclusividad de los alumnos con buen rendimiento. El mayor aumento de la puntuación pre-post fue para los alumnos que partían de un bajo rendimiento inicial (12,9 puntos para el grupo experimental, comparado con 10,7 puntos para el grupo medio, frente a 8,0 puntos para el grupo alto). Así, existiría un posible potencial de la retroalimentación mediada por inteligencia artificial para favorecer el rendimiento de alumnos con un peor rendimiento inicial.

Desde el punto de vista de una evaluación inclusiva, la inteligencia artificial tuvo un efecto más positivo cuando incorporada como herramienta de apoyo en un proceso pedagógico, y no como mecanismo autónomo de evaluación. La revisión del docente, el uso de rúbricas comunes y la regulación de la protección de datos permitieron mantener el control pedagógico mientras se deberían facilitar las oportunidades de acceso a la retroalimentación.

Los resultados obtenidos permiten considerar la retroalimentación formativa mediada por inteligencia artificial (IA) como una estrategia prometedora para mejorar, al mismo tiempo, el aprendizaje y algunas dimensiones de las evaluaciones inclusivas que deben realizarse en el contexto de la educación superior. Ahora bien, la generalización de los hallazgos obtenidos en este trabajo, aunque sean promisorios, debe ser confirmada en trabajos futuros en los que se utilice la asignación aleatoria de los participantes, se utilicen muestras que procedan de distintas áreas disciplinares, se realice un seguimiento durante un periodo que sea más prolongado, y que se ponga en práctica su uso en poblaciones que poseen necesidades de accesibilidad diferenciadas.

Referencias

1. Er E, Akçapınar G, Bayazit A, Noroozi O, Banihashem SK. Assessing student perceptions and use of instructor versus AI-generated feedback. *Br J Educ Technol*. 2025;56(3):1074-1091. <https://doi.org/10.1111/bjet.13558>
2. Venter J, Coetzee SA, Schmulian A. Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assess Eval High Educ*. 2025;50(4):516-536. <https://doi.org/10.1080/02602938.2024.2415649>
3. Usher M. Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assess Eval High Educ*. 2025;50(6):912-927. <https://doi.org/10.1080/02602938.2025.2487495>
4. Zhan Y, Boud D, Dawson P, Yan Z. Generative artificial intelligence as an enabler of student feedback engagement: a framework. *High Educ Res Dev*. 2025;44(5):1289-1304. <https://doi.org/10.1080/07294360.2025.2476513>
5. Zhao X, Cox A, Chen X. The use of generative AI by students with disabilities in higher education. *Internet High Educ*. 2025;66:101014. <https://doi.org/10.1016/j.iheduc.2025.101014>
6. Qian Y. Pedagogical applications of generative AI in higher education: a systematic review of the field. *TechTrends*. 2025;69:1105-1120. <https://doi.org/10.1007/s11528-025-01100-1>
7. Liang J, Stephens JM, Brown GTL. A systematic review of the early impact of artificial intelligence on higher education curriculum, instruction, and assessment. *Front Educ*. 2025;10:1522841. <https://doi.org/10.3389/feduc.2025.1522841>
8. Luo J, Zheng C, Yin J, Teo HH. Design and assessment of AI-based learning tools in higher education: a systematic review. *Int J Educ Technol High Educ*. 2025;22:42. <https://doi.org/10.1186/s41239-025-00540-2>
9. Daniel K, Msambwa MM, Wen Z. Can generative AI revolutionise academic skills development in higher education? A systematic literature review. *Eur J Educ*. 2025;60(1):e70036. <https://doi.org/10.1111/ejed.70036>
10. Deng R, Jiang M, Yu X, Lu Y, Liu S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput Educ*. 2025;227:105224. <https://doi.org/10.1016/j.compedu.2024.105224>
11. Chen S, Cheung ACK. Effect of generative artificial intelligence on university students' learning outcomes: a systematic review and meta-analysis. *Educ Res Rev*. 2025;49:100737. <https://doi.org/10.1016/j.edurev.2025.100737>
12. Drinkwater Gregg K, Ryan O, Katz A, Huerta M, Sajadi S. Expanding possibilities for generative AI in qualitative analysis: fostering student feedback literacy through the application of a feedback quality rubric. *J Eng Educ*. 2025;114(3):e70024. <https://doi.org/10.1002/jee.70024>
13. Estévez-Ayres I, Callejo P, Hombrados-Herrera MA, Alario-Hoyos C, Delgado Kloos C. Evaluation of LLM tools for feedback generation in a course on concurrent programming. *Int J Artif Intell Educ*. 2025;35(2):774-790. <https://doi.org/10.1007/s40593-024-00406-0>
14. Belkina M, Daniel S, Nikolic S, Haque R, Lyden S, Neal P, et al. Implementing generative AI (GenAI) in higher education: a systematic review of case studies. *Comput Educ Artif Intell*. 2025;8:100407. <https://doi.org/10.1016/j.caeai.2025.100407>
15. Dai W, Tsai YS, Lin J, Aldino A, Jin H, Li T, et al. Assessing the proficiency of large language models in automatic feedback generation: an evaluation study. *Comput Educ Artif Intell*. 2024;7:100299. <https://doi.org/10.1016/j.caeai.2024.100299>
16. Steiss J, Tate T, Graham S, Cruz J, Hebert M, Wang J, et al. Comparing the quality of human and ChatGPT feedback of students' writing. *Learn Instr*. 2024;91:101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
17. Sembey R, Hoda R, Grundy J. Emerging technologies in higher education assessment and feedback practices: a systematic literature review. *J Syst Softw*. 2024;211:111988. <https://doi.org/10.1016/j.jss.2024.111988>
18. Xia Q, Weng X, Ouyang F, Lin TJ, Chiu TKF. A scoping review on how generative artificial intelligence transforms assessment in higher education. *Int J Educ Technol High Educ*. 2024;21:40. <https://doi.org/10.1186/s41239-024-00468-z>
19. Weng X, Xia Q, Gu M, Rajaram K, Chiu TKF. Assessment and learning outcomes for generative AI in higher education: a scoping review on current research status and trends. *Australas J Educ Technol*. 2024;40(6):37-55. <https://doi.org/10.14742/ajet.9540>
20. Lee SS, Moore RL. Harnessing generative AI (GenAI) for automated feedback in higher education: a systematic review. *Online Learn*. 2024;28(3):82-106. <https://doi.org/10.24059/olj.v28i3.4593>
21. Dawson P, Yan Z, Lipnevich A, Tai J, Boud D, Mahoney P. Measuring what learners do in feedback: the Feedback Literacy Behaviour Scale. *Assess Eval High Educ*. 2024;49(3):348-362. <https://doi.org/10.1080/02602938.2023.2240983>
22. Stefaniak JE, Moore SL. The use of generative AI to support inclusivity and design deliberation for online instruction. *Online Learn*. 2024;28(3):181-206. <https://doi.org/10.24059/olj.v28i3.4458>
23. Essel HB, Vlachopoulos D, Essuman AB, Amankwa JO. ChatGPT effects on cognitive skills of undergraduate students: receiving instant responses from AI-based conversational large language models (LLMs). *Comput Educ Artif Intell*. 2024;6:100198. <https://doi.org/10.1016/j.caeai.2023.100198>
24. Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. *Int J Educ Technol High Educ*. 2023;20:57. <https://doi.org/10.1186/s41239-023-00425-2>
25. Tai J, Mahoney P, Ajjawi R, Bearman M, Dargusch J, Dracup M, et al. How are examinations inclusive for students with disabilities in higher education? A sociomaterial analysis. *Assess Eval High Educ*. 2023;48(3):390-402. <https://doi.org/10.1080/02602938.2022.2077910>
26. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. *Assess Eval High Educ*. 2018;43(8):1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>

Financiación

Ninguna.

Conflictos de interés

No existen.

Contribución de autoría

Conceptualización: Darwin Joel Lozada López.

Investigación: Darwin Joel Lozada López.

Metodología: Darwin Joel Lozada López.

Redacción – borrador original: Darwin Joel Lozada López.

Redacción – revisión y edición: Darwin Joel Lozada López.