

Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios

Feedback on learning with generative artificial intelligence in university students

-----  -----

ID María Verónica Leiva-Guerrero - Pontificia Universidad Católica de Valparaíso, PUCV (Chile)
ID Ignacio Araya Zamorano - Pontificia Universidad Católica de Valparaíso, PUCV (Chile)
ID Rafael Escobar Collins - Pontificia Universidad Católica de Valparaíso, PUCV (Chile)
ID Francisca Silva Castro - Pontificia Universidad Católica de Valparaíso, PUCV (Chile)

RESUMEN

La evaluación para el aprendizaje ha adquirido creciente relevancia en la docencia universitaria, especialmente el proceso de retroalimentación. Sin embargo, persiste una percepción de insatisfacción en estudiantes sobre la calidad de retroalimentación brindada por el profesorado, lo que evidencia la necesidad de innovar en sus estrategias. Esta investigación exploró la pertinencia pedagógica y tecnológica de integrar la Escalera de Retroalimentación de Wilson con inteligencia artificial generativa, específicamente GPT-4o, para fortalecer la retroalimentación formativa en estudiantes universitarios. El estudio se desarrolló bajo un enfoque cualitativo y exploratorio en dos fases. En primer lugar, se diseñó y validó un prompt mediante el método Delphi con la participación de ocho expertos en evaluación e inteligencia artificial, aplicándolo a siete modelos de lenguaje de última generación. En la segunda fase, el prompt validado se implementó en dos cursos universitarios de distinta naturaleza, Evaluación para el Aprendizaje y Estructura de Datos, integrando la retroalimentación automática en la plataforma Moodle. Los resultados mostraron que los expertos coincidieron en la idoneidad de la Escalera de Wilson mediada por IA y destacaron el desempeño superior de GPT-4o. A nivel de aula, los estudiantes valoraron la claridad, utilidad e inmediatez de la retroalimentación, aunque identificaron limitaciones en la falta de contextualización y tono impersonal de la herramienta. Se concluye que la integración de la Escalera de Wilson con inteligencia artificial generativa representa una innovación prometedora, pero que requiere ajustes disciplinares, supervisión docente y resguardo de la dimensión humana en los procesos de retroalimentación en entornos abiertos y a distancia.

Palabras clave: retroalimentación; evaluación formativa; inteligencia artificial generativa; ChatGPT; Escalera de Wilson; estudiantes universitarios.

ABSTRACT

Assessment for learning has become increasingly important in university teaching, particularly regarding the feedback process. However, there is still a perception of student dissatisfaction with the quality of feedback provided by faculty, highlighting the need to innovate in feedback strategies. This study aimed to explore the pedagogical and technological relevance of integrating Wilson's Feedback Ladder with generative artificial intelligence, specifically GPT-4o, to strengthen formative feedback in university students. The study was conducted using a qualitative and exploratory approach in two phases. First, a prompt was designed and validated using the Delphi method with the participation of eight experts in assessment and artificial intelligence, applying it to seven state-of-the-art language models. In the second phase, the validated prompt was implemented in two university courses of different nature, Assessment for Learning and Data Structures, integrating automatic feedback into the Moodle platform. The results showed that the experts agreed on the suitability of AI-mediated Wilson's Ladder and highlighted the superior performance of GPT-4o. At the classroom level, students valued the clarity, usefulness, and immediacy of the feedback, although they identified limitations in the tool's lack of contextualization and impersonal tone. It is concluded that the integration of Wilson's Ladder with generative artificial intelligence represents a promising innovation, but one that requires disciplinary adjustments, teacher supervision, and careful attention to the human dimension of feedback in e-learning contexts.

Keywords: feedback; formative evaluation; generative artificial intelligence; ChatGPT; Ladder of Feedback; university students.

INTRODUCCIÓN

La evaluación para el aprendizaje con un enfoque formativo ha adquirido creciente relevancia a nivel global, tanto en contextos escolares como universitarios, al centrarse en el acompañamiento y mejora continua del proceso educativo (Black y Wiliam, 2018; Andrade y Brookhart, 2019). En este marco, la retroalimentación se posiciona como un componente clave, al utilizar evidencia del desempeño estudiantil para orientar la mejora de los aprendizajes (Carless y Boud, 2018).

Sin embargo, diversos estudios reportan una persistente insatisfacción del estudiantado universitario respecto a la calidad de la retroalimentación proporcionada por el profesorado, especialmente en lo que respecta a su oportunidad, especificidad y utilidad (Quezada y Salinas, 2021; Galindo-Domínguez et al., 2023). Esta situación ha sido evidenciada en investigaciones realizadas en países como Australia, Reino Unido, México y Chile, lo que sugiere una problemática de alcance internacional.

En varios sistemas educativos, las prácticas de retroalimentación no han evolucionado significativamente, lo que limita su impacto formativo. Estrategias como la Escalera de Retroalimentación de Wilson (2013) o el Panel de Retroalimentación (Booth et al., 2008) han sido propuestas como alternativas efectivas para mejorar los procesos evaluativos, aunque su incorporación en la práctica docente sigue siendo limitada.

Frente a este escenario, se vuelve necesario renovar las estrategias de evaluación universitaria, incorporando tecnologías digitales emergentes que puedan apoyar la retroalimentación formativa (Puertas y Cano, 2024). En particular, la Inteligencia Artificial Generativa (IAGen) ofrece oportunidades relevantes para transformar la retroalimentación en tareas de escritura y pensamiento complejo (Dai et al., 2023; Al-Azawei et al., 2023).

Entre las herramientas basadas en esta tecnología, destaca ChatGPT (OpenAI, 2024), que integra modelos como GPT-4o, GPT-4.1 o GPT-4.5 y se caracteriza por generar textos coherentes y contextualizados a partir de prompts diseñados con criterios pedagógicos. Aunque su incorporación al ámbito educativo es reciente, estudios preliminares evidencian un uso creciente en enseñanza, aprendizaje y evaluación, lo que abre nuevas líneas de investigación sobre su aplicabilidad y eficacia (Galindo-Domínguez et al., 2023; García-Peñalvo, 2023).

En este marco, el presente estudio se plantea tres preguntas centrales: ¿es adecuada la Escalera de Retroalimentación de Wilson para su aplicación mediante inteligencia artificial?, ¿qué herramienta de IAGen, programada con prompts según estos criterios, resulta más pertinente para retroalimentar aprendizajes universitarios?, y ¿cómo perciben los estudiantes la retroalimentación recibida a través de esta estrategia aplicada con GPT-4o?

Inteligencia artificial en educación

La integración de la inteligencia artificial (IA) en el campo educativo ha redefinido los marcos tradicionales de enseñanza, constituyéndose en un eje fundamental de la innovación pedagógica contemporánea. El rol docente, históricamente centrado en la transmisión de contenidos, ha experimentado una transición sustantiva hacia funciones de facilitación del aprendizaje, gracias a las potencialidades de la IA para automatizar procesos, analizar datos y personalizar trayectorias formativas (Bonales-Daimiel et al., 2025). Esta reconfiguración permite a los educadores concentrarse en tareas de mayor valor pedagógico, tales como la retroalimentación cualitativa y el acompañamiento diferenciado (García Peñalvo et al., 2024).

Además, la IA ha demostrado ser una aliada estratégica en la mejora de la experiencia educativa, tanto en contextos presenciales como virtuales, mediante la personalización de recursos, la evaluación automatizada, la retroalimentación inmediata y el monitoreo en tiempo real del desempeño estudiantil (López Regalado et al., 2024; Romero Alonso et al., 2025). La implementación de sistemas adaptativos y chatbots educativos ha permitido no solo optimizar la eficiencia docente, sino también enriquecer los procesos de aprendizaje con enfoques centrados en el estudiante.

Inteligencia Artificial Generativa

La inteligencia artificial generativa (IAGen) abarca técnicas y modelos diseñados para crear contenido nuevo (texto, imágenes, audio, video) que imita la distribución estadística de grandes conjuntos de datos. En el núcleo de muchas soluciones de IAGen están los transformers, arquitecturas de atención que, tras un pre-entrenamiento masivo, aprenden a prever la siguiente unidad (palabra, píxel, nota) y así generan salidas coherentes y contextualizadas (Vaswani et al., 2017; OpenAI, 2023). Gracias a algoritmos de ajuste fino y a la incorporación de múltiples modalidades de entrada y salida estos sistemas pueden adaptarse a tareas diversas desde redacción creativa hasta simulaciones conversacionales manteniendo fluidez y sentido en sus respuestas (Weng et al., 2024).

Desde 2024, el ecosistema de LLMs se ha diversificado con propuestas comerciales GPT-4o y GPT-4.5 (OpenAI, 2024), Claude 3.7 Sonnet (Anthropic, 2024) y Gemini 2.5 (Google DeepMind, 2024) y de código abierto LLaMa 4 (LLaMa, 2024), Mistral (Mistral AI, 2024). Además, algunos modelos, como DeepSeek R1 (DeepSeek, 2024), destacan por exponer su razonamiento interno frente a problemas complejos, y xAI Grok 3 (xAI, 2024) dedica fases de “reflexión” profundas para matemáticas y lógica avanzadas.

En educación, estas herramientas no solo sirven para automatizar contenidos, sino que ayudan a personalizar la enseñanza y la retroalimentación, adaptándose a cada estudiante y aliviando tareas repetitivas (corrección de trabajos), siempre que

se consideren aspectos éticos y metodológicos. Su integración responsable puede potenciar significativamente el aprendizaje (Weng et al., 2024).

Prompts y técnicas de prompting

En el contexto de la inteligencia artificial generativa, un prompt es la entrada textual o estructurada que se entrega a un modelo de lenguaje para guiar su respuesta. Su diseño no es trivial: la forma, el contenido y el formato del prompt influyen directamente en la calidad, precisión y utilidad de la salida generada. De ahí surge el campo del prompt engineering, dedicado a la elaboración de estrategias específicas para construir prompts efectivos según la tarea o el contexto educativo (Liu et al., 2023; Reynolds y McDonell, 2021).

Entre las principales técnicas de prompting se encuentran el zero-shot prompting, en el que el modelo responde únicamente a una instrucción directa sin ejemplos; y el few-shot prompting, que incorpora ejemplos representativos antes de la tarea para modelar el tipo de respuesta esperada (Brown et al., 2020). Estas técnicas permiten modular el comportamiento del modelo y anticipar mejor el tipo de contenido que se generará.

Otra técnica fundamental es chain-of-thought, que guía al modelo a razonar paso a paso antes de emitir una respuesta final, mejorando su desempeño en tareas complejas de lógica o evaluación estructurada (Wei et al., 2022). A esto se suma el role prompting, que asigna explícitamente un rol al modelo —como “evaluador experto” o “docente universitario”— para controlar el tono, la profundidad y el estilo de la retroalimentación (Sahoo et al., 2024). Finalmente, el structured prompting especifica el formato de salida requerido, como un diccionario o una tabla, lo que resulta especialmente útil cuando la información generada debe ser procesada automáticamente por sistemas externos, como plataformas educativas (Schulhoff et al., 2025).

En educación, las técnicas de prompting permiten estructurar retroalimentación automatizada de forma precisa y pedagógicamente alineada. El uso combinado de role prompting, chain-of-thought y structured prompting guía al modelo a razonar como un evaluador experto, generando respuestas organizadas y coherentes. Esto facilita su integración en plataformas educativas y potencia una retroalimentación más útil, trazable y centrada en el aprendizaje.

Retroalimentación formativa y su mediación con inteligencia artificial

La retroalimentación es un pilar del aprendizaje formativo, concebida no solo como corrección de errores, sino como un proceso interactivo que promueve reflexión, ajuste de estrategias y construcción activa del conocimiento. Winstone et al. (2022) la describen como una interacción dialógica entre docente y estudiante, o entre pares, que favorece la comprensión profunda siempre que sea oportuna,

específica y con metas claras. En la misma línea, Carless y Boud (2018) subrayan que debe entenderse como un proceso de influencia mutua y de agencia estudiantil, mientras que Molloy et al. (2020) destacan la necesidad de que los estudiantes desarrollen competencias para interpretarla y aplicarla en contextos autónomos.

No obstante, en la práctica persisten barreras estructurales. En Chile, el Ministerio de Educación (2019) advierte que, pese al reconocimiento teórico de la retroalimentación, las prácticas evaluativas siguen siendo mayoritariamente unidireccionales y correctivas, con poco énfasis en la metacognición y la autorregulación. Para responder a esta brecha, se requieren enfoques estructurados que potencien la dimensión formativa. En este marco, la Escalera de Retroalimentación de Wilson (Goodrich, 2011) ofrece una guía clara con cuatro niveles progresivos (Aclarar: formular preguntas que apunten a resolver ambigüedades o aspectos omitidos en el trabajo del estudiante, fomentando la comprensión conceptual; Valorar: destacar logros y avances, reconociendo de manera auténtica los aspectos positivos del desempeño; Expresar inquietudes: identificar dificultades observadas con un tono respetuoso y constructivo, evitando juicios descalificadores; y, Sugerir mejoras: ofrecer orientaciones específicas que orienten al estudiante hacia una comprensión más profunda y acciones de mejora viables) orientados al pensamiento crítico y la autorregulación. Su valor radica en transformar la retroalimentación en una experiencia empática, participativa y adaptable a diversos contextos educativos, presenciales o virtuales.

En este estudio, la Escalera de Wilson se adopta como base conceptual para diseñar retroalimentación automatizada mediada por inteligencia artificial. Desde esta perspectiva, la IA se plantea como recurso para personalizar y ampliar la cobertura, en línea con lo señalado por Wiliam (2011) y Shute y Rahimi (2017) respecto al potencial de las tecnologías emergentes en la evaluación para el aprendizaje.

Inteligencia artificial para la retroalimentación educativa

La incorporación de IA en los procesos de retroalimentación educativa responde a la necesidad de transformar prácticas tradicionales hacia modelos más interactivos, personalizados y efectivos. Diversas investigaciones han propuesto innovar en esta dimensión mediante el uso de IA, destacando su potencial para mejorar la calidad del aprendizaje y reducir la carga evaluativa docente (Puertas y Cano, 2024; Ossa y Willatt, 2023). El informe *Teaching with AI – Assessment, Feedback and Personalisation*, elaborado por European Commission (2023), profundiza en cómo la IA puede intervenir positivamente en los procesos de evaluación y retroalimentación desde cuatro niveles complementarios: social, institucional, docente y estudiantil.

En este contexto, la IA ofrece herramientas capaces de adaptar la retroalimentación a las necesidades individuales del estudiantado mediante la generación automatizada de comentarios, la evaluación en tiempo real y el seguimiento del progreso académico (Ayeni et al., 2024; Holmes et al., 2019). Este

tipo de apoyo permite a los docentes identificar con mayor precisión las dificultades específicas de cada estudiante y ofrecer una enseñanza más focalizada y diferenciada (Luckin y Holmes, 2016). Asimismo, la retroalimentación inmediata proporcionada por sistemas inteligentes incide positivamente en la motivación estudiantil, ya que entrega respuestas oportunas que facilitan la comprensión de errores y refuerzan aprendizajes claves.

En particular, las herramientas de chat basadas en modelos IAGen han cobrado relevancia como asistentes pedagógicos eficaces. Estos sistemas permiten automatizar la corrección y la retroalimentación a partir de criterios previamente establecidos, incrementando así la eficiencia y coherencia del proceso evaluativo (Carless y Winstone, 2023). Para el profesorado, esto significa una reducción en la carga operativa; para los estudiantes, una oportunidad de recibir orientación formativa inmediata, promoviendo la autoevaluación y el pensamiento crítico.

La efectividad de estas herramientas ha sido documentada en diversos estudios empíricos. Zhang et al. (2024) evaluaron el uso de ChatGPT para generar retroalimentación en tareas de programación con estudiantes de un curso introductorio de Ciencias de la Computación. La mayoría de los participantes valoró positivamente la claridad y utilidad de los comentarios generados por la IA, aunque algunos expresaron que habrían preferido una retroalimentación más rica en ejemplos específicos. Por su parte, Jauhiainen y Garagorry Guerra (2024) analizaron la aplicación de ChatGPT-4 en la evaluación de respuestas abiertas escritas por estudiantes universitarios. Utilizando una rúbrica con cinco criterios, relevancia, exactitud, exhaustividad, coherencia y corrección lingüística, observaron que el modelo generaba evaluaciones detalladas y consistentes, aunque insistieron en la necesidad de supervisión docente para garantizar su precisión final.

Complementariamente, Baral et al. (2024) compararon ChatGPT-4 con otros modelos (como LLaMa y SBERT-Canberra) en la corrección de respuestas matemáticas de nivel secundario. A partir de 500 respuestas evaluadas con una rúbrica común, encontraron que ChatGPT-4 alcanzó un 92 % de concordancia con las calificaciones de docentes humanos, además de generar retroalimentación cualitativa de alta calidad, evaluada mediante métricas lingüísticas y juicios de expertos. No obstante, los autores advierten que, aunque la IA puede aproximarse al criterio humano, su uso debe enmarcarse en un sistema que contemple revisión experta y reflexión pedagógica.

METODOLOGÍA

Diseño

El estudio adoptó un enfoque cualitativo de carácter exploratorio, orientado a diseñar, validar e implementar un prompt de retroalimentación formativa automatizada mediante inteligencia artificial generativa (IAGen), sustentado en la

Escalera de Retroalimentación de Wilson (2013). El proceso se desarrolló en dos fases.

En la Fase 1, se elaboró el prompt a partir de una revisión documental sobre estrategias de retroalimentación y se validó mediante el método Delphi (Landeta, 1999). El procedimiento incluyó dos rondas sucesivas de juicio experto. En la primera, se evaluó la adecuación del prompt y de los cuatro peldaños de la Escalera (aclarar, valorar, expresar inquietudes, sugerir), aplicados a siete modelos de lenguaje (LLM) representativos. En la segunda, se devolvió una síntesis de resultados y se solicitó re-juicio para convergencia. Se utilizó una escala Likert 0–3 anclada en desempeño y se definió como criterio de consenso una desviación intercuartílica ≤ 1 . Además, se calculó el coeficiente W de Kendall como medida de consistencia global (criterio de aceptabilidad $\geq 0,70$).

En la Fase 2, el prompt validado se integró en la plataforma Moodle¹ en dos asignaturas universitarias. Las respuestas abiertas de los estudiantes fueron procesadas automáticamente con GPT-4o para generar retroalimentación inmediata estructurada según la Escalera de Wilson. Finalmente, se recogieron percepciones estudiantiles mediante entrevistas semiestructuradas para complementar los datos técnicos con la experiencia del alumnado.

Procedimiento

En la primera fase, el objetivo fue asegurar que el prompt diseñado generara retroalimentación estructurada y coherente con los cuatro niveles de la Escalera de Wilson (2013) antes de ser aplicado con estudiantes. Este proceso incluyó la selección de modelos de lenguaje, la elaboración iterativa del prompt y su validación. Se eligieron siete modelos de lenguaje representativos por su diversidad técnica y accesibilidad: GPT-4o², GPT-4o-mini³ (OpenAI, 2024), Microsoft Copilot, Google Gemini (2024), Claude (Anthropic, 2024), Perplexity AI y LLaMA (2024), accedidos desde sus canales oficiales.

El prompt fue diseñado como plantilla estructurada con: (i) contexto teórico, (ii) enunciado de la pregunta, (iii) criterios de evaluación y/o respuesta esperada, (iv) respuesta del estudiante y (v) instrucciones explícitas para orientar la retroalimentación siguiendo los peldaños de Wilson. En la evaluación piloto se utilizaron dos actividades: una pregunta abierta basada en un caso de estudio y una tarea de elaboración de rúbrica. Las respuestas de cada modelo fueron analizadas y posteriormente evaluadas por un panel de ocho expertos —tres en evaluación, tres en IA y dos docentes universitarios— usando una escala de 0 a 3 y comentarios cualitativos, en tres rondas según el enfoque Delphi (Steurer, 2011).

La segunda fase tuvo como propósito implementar el prompt validado en dos asignaturas de pregrado: Evaluación para el aprendizaje y Estructura de Datos. La evaluación (prueba escrita) se realizó en Moodle mediante el plugin Coderunner⁴, automatizando el flujo entre respuestas estudiantiles y GPT-4o. Cada estudiante

respondió preguntas abiertas que, una vez enviadas, eran procesadas en tiempo real generando retroalimentación inmediata estructurada según la Escalera de Wilson, junto con un puntaje preliminar de 0 a 10 basado en rúbricas.

El estudiantado fue informado sobre la naturaleza experimental del sistema y se ofreció una opción de apelación, revisada manualmente por el equipo docente, como medida de equidad. Además, se realizó un análisis cualitativo posterior mediante entrevistas semiestructuradas a estudiantes voluntarios, indagando en la claridad, utilidad y confiabilidad de la retroalimentación automatizada. Las entrevistas fueron analizadas utilizando categorías deductivas (basadas en los peldaños de Wilson) y subcategorías emergentes Mayring (2000). ChatGPT asistió en la sistematización del análisis, apoyando la codificación temática y la extracción de patrones significativos en los testimonios estudiantiles.

Instrumentos

Se utilizaron instrumentos y técnicas que se detallan a continuación:

1. *Protocolo de Validación del Prompt*

Se elaboró un protocolo para validar el desempeño del prompt en siete modelos de lenguaje (LLM). Este contenía las retroalimentaciones generadas por cada modelo, organizadas según los cuatro peldaños de la Escalera de Wilson (2013): Aclarar, Valorar, Expresar inquietudes y Sugerir mejoras. El panel de expertas evaluó cada peldaño con escala ordinal (0–3) y emitió observaciones cualitativas cuando la puntuación fue igual o inferior a 1.

La validación se llevó a cabo mediante el método Delphi con un panel interdisciplinario de ocho expertos. El proceso contempló dos rondas sucesivas, utilizando una escala Likert 0–3 y criterios de consenso basados en desviación intercuartílica (≤ 1) y coeficiente de Kendall W. Esta dinámica permitió ajustar el instrumento antes de su aplicación, confirmando su coherencia con los peldaños de la Escalera de Wilson y respaldando la selección de GPT-4o como modelo preferente para la fase piloto.

2. *Prueba Escrita para Estudiantes (Evaluación Sumativa)*

Las evaluaciones aplicadas en ambas asignaturas incluyeron preguntas abiertas diseñadas para activar habilidades cognitivas de orden superior. En ‘Estructura de Datos’, se abordaron explicaciones, análisis de estructuras algorítmicas y justificaciones técnicas. En ‘Evaluación del y para el aprendizaje’, las preguntas demandaban respuestas argumentadas sobre conceptos pedagógicos a partir de

clases observadas. Las preguntas fueron acompañadas de rúbricas específicas, lo cual permitió aplicar el prompt de manera coherente.

3. *Prompt Estructurado*

El instrumento central del estudio fue un prompt diseñado conforme a principios de prompt engineering (Brown et al., 2020; Wei et al., 2022; Hao et al., 2022; Liu et al., 2023; Schulhoff et al., 2025), integrado por seis componentes: (i) Contexto teórico: sintetiza los fundamentos conceptuales y metodológicos que enmarcan la tarea, permitiendo situar la evaluación dentro de un marco disciplinar y pedagógico sólido. (ii) Enunciado de la pregunta: plantea la tarea evaluativa con claridad, asegurando coherencia con los objetivos de aprendizaje. (iii) Criterios de evaluación/respuesta esperada: establecen los parámetros de calidad y pertinencia de las respuestas, descritos mediante una rúbrica analítica con todos sus niveles de desempeño. (iv) Respuesta del estudiante: contiene la producción a evaluar. (v) Instrucciones para retroalimentar: orientan la formulación de juicios y comentarios siguiendo los peldaños de la Escalera de Wilson (valoraciones, aclaraciones, inquietudes, sugerencias), con límites de extensión y tono constructivo. (vi) Formato de salida en diccionario Python: organiza los resultados en los siguientes campos: análisis y puntaje por cada criterio de evaluación, más los cuatro peldaños de la Escalera de Wilson. Se escribió en español, empleando formato markdown para favorecer la legibilidad y se asignó el rol de “evaluador experto” para asegurar aspectos técnicos. Tras su validación, se decidió incluir en el punto (iii) la descripción completa de todos los niveles de la rúbrica analítica y agregar el punto (vi) para facilitar la automatización de la respuesta en la plataforma Moodle. La versión validada se adjunta como anexo para su replicabilidad. Es importante resaltar que los puntos (v) y (vi) pueden permanecer sin modificación, mientras que los apartados anteriores se deben adecuar a la evaluación correspondiente.

4. *Sistema Automatizado de Retroalimentación*

Se desarrolló un sistema integrado en Moodle mediante el plugin Coderunner, que conecta los datos con el modelo GPT-4o. Cada vez que una respuesta era enviada por un o una estudiante, se generaba un prompt estructurado que era procesado sincrónicamente, devolviendo retroalimentación inmediata. La salida era formateada automáticamente en Moodle junto al puntaje preliminar. Además, se almacenaron los datos para su posterior análisis. El modelo se configuró con una temperatura de 0.1, priorizando así respuestas más deterministas y consistentes con la rúbrica de evaluación establecida.

Técnica

Entrevista Semiestructurada

Con el fin de explorar la percepción estudiantil sobre el sistema, se diseñó una entrevista con preguntas abiertas organizadas en torno a seis dimensiones: los cuatro peldaños de Wilson, la usabilidad del sistema y el impacto percibido en el aprendizaje. Antes de su aplicación, cada participante firmó un consentimiento informado, resguardando los principios de voluntariedad, confidencialidad y anonimato. Las entrevistas fueron grabadas en audio con autorización de los estudiantes y transcritas íntegramente para su análisis. Se garantizó en todo momento que la información sería utilizada únicamente con fines académicos, cumpliendo las normas éticas institucionales vigentes.

El análisis se realizó mediante el enfoque de contenido cualitativo de Mayring (2000), combinando categorías deductivas (derivadas de los peldaños de la Escalera de Wilson y de las dimensiones complementarias de usabilidad, impacto en el aprendizaje y juicio crítico con recomendaciones) e inductivas, surgidas de la propia información. Dos codificadores trabajaron de manera independiente en la lectura, segmentación y codificación de las entrevistas, aplicando la matriz previamente consensuada. Posteriormente, revisaron conjuntamente los resultados para resolver discrepancias y consolidar la codificación definitiva, lo que permitió asegurar la fiabilidad intercodificador y dotar de mayor solidez a los hallazgos.

Las categorías se enriquecieron con subcategorías emergentes basadas en las experiencias de los participantes, diferenciando entre valoraciones favorables (utilidad, claridad, motivación, guía formativa) y problemáticas (rigidez, descontextualización, automatismo, falta de empatía). El proceso contó además con el apoyo de ChatGPT como herramienta metodológica complementaria, utilizada para organizar y contrastar fragmentos textuales, explorar patrones temáticos y verificar la coherencia del sistema categorial. Su uso se limitó a facilitar tareas operativas, sin sustituir el juicio analítico de los investigadores, y contribuyó a fortalecer la trazabilidad y consistencia del proceso. La Tabla 1 resume las categorías y subcategorías empleadas en el análisis.

Tabla 1
Categorías y subcategorías usadas para el análisis

Categoría	Valoraciones Favorables	Valoraciones Problemáticas
I. Experiencia general	• Valoración de la innovación. • Rapidez y precisión del sistema.	• Extrañeza ante evaluación automatizada. • Ausencia de vínculo humano.
II. Aclaraciones (Escalera)	• Señala omisiones útiles. • Ejemplos clarificadores.	• Correcciones innecesarias. • Sugerencias confusas.

Categoría	Valoraciones Favorables	Valoraciones Problemáticas
III. Valoración (Escalera)	• Refuerzo positivo genera confianza. • Estimula el aprendizaje.	• Impacto emocional limitado por ser de IA. • Tono mecánico o forzado.
IV. Inquietudes (Escalera)	• Críticas claras y constructivas. • Impulso a la mejora.	• Redundancia o irrelevancia en algunas críticas.
V. Sugerencias (Escalera)	• Claridad y aplicabilidad. • Mejora futura.	• Sugerencias fuera del contenido visto. • Poca pertinencia contextual.
VI. Usabilidad y confianza	• Confianza en criterios. técnicos estables.	• Falta de transparencia evaluativa. • Dudas sobre sensibilidad pedagógica.
VII. Impacto en el aprendizaje	• Mejora de comprensión. • Retroalimentación útil para el estudio.	• Exigencias percibidas como excesivas
VIII. Juicio crítico y recomendaciones	• Disposición a seguir usando la IA. • Aporte al desarrollo del sistema.	• Necesidad de personalización. • Críticas al ajuste de nivel.
IX. Autonomía y control percibido (emergente)	• Claridad sobre expectativas. • Retroalimentación como guía. • Ajuste estratégico a la IA(sin conflicto explícito).	• Evaluación rígida. • Penalización por estilo personal o creatividad. • Desmotivación por falta de reconocimiento a respuestas válidas alternativas.

Fuente: Elaboración propia.

RESULTADOS

A continuación, se presentan los principales resultados en torno a las preguntas de investigación: ¿Es adecuada la Escalera de Retroalimentación de Wilson para ser aplicada mediante herramientas de inteligencia artificial?; ¿Cuál de las herramientas de IAGen, con prompts diseñados según escalera Wilson, resulta más pertinente para retroalimentar aprendizajes en estudiantes universitarios?; y ¿Cómo percibe el estudiantado universitario la retroalimentación recibida a través de GPT-4o?

Escalera de Retroalimentación de Wilson con IA

El 100 % de las valoraciones del panel de expertos concordó en que la estrategia de la Escalera de Wilson para retroalimentar aprendizajes con el uso de IA era apropiada, ya que brinda al estudiante claridad, valor, reflexión y sugerencias para

impulsar su aprendizaje, fortaleciendo la relación entre profesor(a) y alumno(a) al estar claramente definidos los pasos de retroalimentación. Al respecto, los expertos señalaron:

Entrega un nivel de descripción y detalle, que no solo se centra en señalar aspectos positivos y negativos, sino que va más allá, ofreciendo más posibilidades para la reflexión, corrección y mejora de aprendizajes a través de los pasos aclarar, valorar, señalar inquietudes y entregar sugerencias. (Experto 6).

Es interesante y apropiada la estrategia de retroalimentación Escala de Wilson, proporciona un marco para que el *feedback* sea empático, descriptivo y con foco en el aporte para la construcción de aprendizajes, en lugar de la crítica. (Experto 2).

Herramientas de IA más apropiadas de acuerdo con la Escalera de Retroalimentación de Wilson

La Tabla 2 presenta los promedios de valoraciones entregadas por expertos sobre retroalimentaciones generadas por distintas herramientas de IA, evaluadas en una escala de 0 (muy bajo) a 3 (muy alto) según los cuatro peldaños de la Escalera de Wilson. Cada herramienta fue evaluada en términos de claridad, valoración, expresión de inquietudes y sugerencias, incluyendo además su media y desviación estándar. De acuerdo con estos datos, GPT-4o destaca como la herramienta con mejor desempeño general, al obtener puntajes consistentemente altos en todos los criterios y una media de 2,87, junto con una baja desviación estándar (0,18), lo que refleja tanto efectividad como estabilidad en la calidad de sus retroalimentaciones.

Tabla 2
Promedios de valoraciones expertos de retroalimentaciones con IA

Herramienta IA	Promedio Retroalimentación Escalera de Wilson Ejemplos 1 y 2				Media	Desviación estándar
	Aclarar	Valorar	Expresar inquietudes	Ofrecer Sugerencias		
GPT-4o	2,9	2,75	2,9	2,95	2,87	0,18
GPT-4o mini	2,2	2,1	2,6	2,35	2,30	0,35
Microsoft Copilot	1,55	1,6	1,75	1,95	1,70	0,49
Gemini	1,65	1,45	1,7	1,95	1,67	0,46
Claude	1,6	1,5	1,1	1,35	1,36	0,62
Perplexity	2,45	1,95	2,3	2,5	2,28	0,25
LlaMa 3.2	2	1,9	2	2,3	2,04	0,27

Fuente: Elaboración propia.

La concordancia entre los ocho jueces al evaluar las siete herramientas (GPT-4, GPT Gratuito, Copilot, Gemini, Claude, Perplexity y LLaMa 3.2) fue medida mediante el coeficiente W de Kendall, que alcanzó un valor de 0,43, lo cual indica un nivel moderado de acuerdo entre los jueces. La prueba de significancia asociada arrojó un valor de $\chi^2 = 20,67$ con 6 grados de libertad y un $p < 0,01$, lo que permite concluir que la concordancia observada es estadísticamente significativa, indicando que los jueces mostraron un consenso consistente y validable estadísticamente respecto a la clasificación de las herramientas según los criterios de retroalimentación de Wilson.

Aplicación de la herramienta en asignatura Estructura de Datos

La herramienta fue aplicada a un total de 107 estudiantes de segundo año. Durante la prueba, se procesaron 846 respuestas abiertas, cada una evaluada automáticamente mediante el modelo GPT-4o, utilizando el prompt estructurado previamente validado.

Se habilitó un canal de apelación voluntaria a través del correo institucional, permitiendo a los estudiantes solicitar una revisión de la retroalimentación o del puntaje asignado. Se recibieron 34 apelaciones, de las cuales 27 fueron consideradas válidas, lo que representa un 3,19 % del total de respuestas, tasa relativamente baja de disconformidad con el sistema automatizado.

La Tabla 3 presenta un análisis de los principales patrones de error detectados en el modelo GPT-4o, sobre la base del total de respuestas procesadas ($N = 846$). Para cada patrón identificado, se incluye un ejemplo representativo de retroalimentación generada, un análisis interpretativo del tipo de error, y la frecuencia (Frec.) con que este fue observado, expresada tanto en número de casos como en porcentaje del total.

Tabla 3

Patrones de error observado en respuestas del modelo

Patrón de error	Ejemplo retroalimentación GPT-4o	Análisis del error	Casos	Frec.
La retroalimentación del modelo sugiere incluir detalles que ya estaban presentes en la respuesta del estudiante.	“Sería útil considerar una descripción más clara de las operaciones específicas que se realizan en la lista, como la adición de elementos al final y la eliminación del primer elemento”.	La retroalimentación sugiere que faltan detalles sobre las operaciones de agregar y eliminar elementos en la lista. Sin embargo, el estudiante ya lo había mencionado en su respuesta.	31	3,7 %

Patrón de error	Ejemplo retroalimentación GPT-4o	Análisis del error	Casos	Frec.
La retroalimentación del modelo pide detalles o aspectos que no estaban solicitados en la pregunta.	“Sería útil considerar la importancia de la función hash y cómo afecta la eficiencia de la tabla”.	La retroalimentación sugiere incluir información adicional (como la importancia de la función hash) que no se pedía en la pregunta original. Esto puede confundir al estudiante porque extiende el alcance de la pregunta y lo hace sentir que su respuesta estaba incompleta aunque en realidad respondió adecuadamente a lo solicitado.	127	15 %
La retroalimentación es poco clara o ambigua limitando la orientación a la mejora del estudiante	“tu respuesta no aborda cómo manejar el rango específico de búsqueda ni cómo recolectar los elementos dentro de ese rango”.	La retroalimentación de la IA sugiere que el estudiante no explicó nada sobre la recolección de elementos, cuando en realidad sí lo hizo, pero su enfoque no era el más adecuado. Esto puede desorientar al estudiante porque no queda claro si lo que escribió estaba mal o si simplemente necesitaba mejorarlo.	47	5,6 %

Fuente: Elaboración propia.

Aplicación de la herramienta en asignatura Evaluación para el Aprendizaje

La herramienta fue aplicada a un total de 37 estudiantes. Durante la prueba, se procesaron 592 respuestas abiertas, cada una evaluada automáticamente mediante el modelo GPT-4o, utilizando el mismo prompt estructurado previamente validado.

Al igual que en la experiencia anterior, se habilitó un canal de apelación voluntaria a través del correo institucional, con el fin de que los estudiantes pudieran solicitar una revisión de la retroalimentación o del puntaje asignado. En este caso, no se recibieron apelaciones, lo que sugiere una alta aceptación de las evaluaciones generadas automáticamente.

La Tabla 4 presenta un análisis de los principales patrones de error detectados en este segundo experimento, sobre la base del total de respuestas procesadas ($N = 592$).

Tabla 4

Patrones de error observado en respuestas del modelo

Patrón de error	Ejemplo retroalimentación GPT-4o	Análisis del error	Casos	Frec.
La retroalimentación del modelo pide detalles o aspectos que no estaban solicitados en la pregunta.	“¿Qué ejemplos específicos de evaluación formativa observaste durante la clase?”	La retroalimentación no resulta pertinente respecto de la respuesta de la o el estudiante ni del caso presentado en la pregunta, el cual no refleja una aplicación explícita de los principios de la evaluación formativa.	3	0,5 %
La retroalimentación contiene errores conceptuales o afirmaciones incorrectas.	La respuesta destaca correctamente la función pedagógica de la evaluación y su enfoque formativo, lo cual es fundamental en un contexto de aprendizaje de idiomas. Además, se identifica adecuadamente el uso de la heteroevaluación, lo que muestra una comprensión de los roles de los agentes evaluativos en la clase.	La retroalimentación sugiere que la respuesta de la o el estudiante es correcta y adecuada; sin embargo, esta no aborda todos los aspectos requeridos por la pregunta. Esta incongruencia puede generar confusión, especialmente si no se asigna el puntaje máximo y la retroalimentación no especifica claramente qué elementos deben ser mejorados.	4	0,7 %

Patrón de error	Ejemplo retroalimentación GPT-4o	Análisis del error	Casos	Frec.
La retroalimentación ocupa conceptos negativos o rígidos que no se condicen con una retroalimentación constructiva.	La respuesta carece de una discusión sobre cómo las pruebas estandarizadas podrían influir...	Se reemplazó la palabra “carece” por considerarse poco apropiada en el contexto de una retroalimentación constructiva, ya que podría resultar rígida o negativa. En su lugar, se utilizó una formulación más formativa y orientada a la mejora, como “faltaría incorporar...” o “aún no se aborda...”.	26	4,4 %
La retroalimentación es poco clara o ambigua limitando la orientación a la mejora del estudiante.	La respuesta destaca correctamente la naturaleza formativa de la evaluación en la clase.	Se emplea una doble valoración al utilizar términos como “destaca” y “correctamente”; sin embargo, no se entrega una retroalimentación constructiva desde la perspectiva del logro alcanzado. La respuesta no especifica qué aspecto de la evaluación formativa se valora en el contexto de la clase, lo que limita su utilidad para orientar mejoras.	12	2,0 %

Fuente: Elaboración propia.

Percepción de estudiantes sobre retroalimentación recibida

A partir de este sistema de categorías, se procedió al análisis sistemático de las entrevistas, codificando las respuestas de los estudiantes en función de las categorías y subcategorías definidas. Esta codificación permitió detectar patrones de sentido, recurrencias temáticas y contrastes en las valoraciones individuales, así como matices relevantes en la forma en que los participantes interpretaron la retroalimentación automatizada. El análisis se realizó combinando una lectura inductiva del discurso con la clasificación estructurada por dimensiones, lo que facilitó una comprensión

holística y matizada de la experiencia estudiantil. Cada testimonio fue interpretado a la luz de su coherencia interna, su relación con las categorías del marco analítico y su conexión con los demás casos.

En términos generales, los estudiantes valoran positivamente la experiencia de recibir retroalimentación automatizada, destacando especialmente su rapidez, claridad e innovación en comparación con la retroalimentación tradicional de los docentes. La sorpresa ante el uso de inteligencia artificial es recurrente: “es sorprendente que esta te revise la evaluación, ya que uno como estudiante está acostumbrado a que esta retroalimentación la realice el profesor” (estudiante 3). Esta innovación es vista como un cambio significativo en la práctica evaluativa, percibido como algo novedoso y prometedor. No obstante, esta valoración técnica convive con matices emocionales: algunos describen la experiencia como “un poco estresante” (estudiantes 1 y 8), porque “te hace pensar que tienes que estar mucho más preparada y ser más precisa” (Estudiantes, 1, 4, 6). En conjunto, los estudiantes reconocen el valor funcional del sistema, pero también notan la ausencia del “factor humano” (estudiantes 1, 2, 5, 7 y 8) y una cierta distancia afectiva en el estilo de la retroalimentación.

Respecto al peldaño de aclaraciones, los participantes aprecian que la IA sea capaz de identificar omisiones o ambigüedades que ellos mismos no habían detectado. Se valora, por ejemplo, cuando la IA indica que “falta detallar el proceso completo de ajuste de punteros para que el nuevo nodo se integre correctamente en la lista” (estudiante 7). Estas observaciones son percibidas como útiles y formativas, pero no están exentas de crítica. Algunos estudiantes cuestionan que “me corrige y me explica algo que yo ya puse en mi respuesta” (estudiante 10), lo cual puede generar confusión y una sensación de evaluación injusta. Esta tensión entre estructura y subjetividad se refuerza cuando la “IA se vuelve más crítica cuando uno da una respuesta con libertad” (estudiante 9), lo que pone en cuestión la flexibilidad interpretativa del sistema frente a respuestas no literales pero válidas.

En cuanto al peldaño de valoración, muchos estudiantes reconocen el refuerzo positivo como un estímulo para continuar aprendiendo. Afirmaciones como “resalta lo que hiciste bien y te invita a seguir estudiando” (estudiante 6) o “me confirma que lo que sé no es tan erróneo” (estudiante 2) muestran que el reconocimiento, incluso automatizado, puede cumplir una función pedagógica significativa. Sin embargo, su impacto emocional es limitado para algunos: “si los comentarios vinieran de un profesor, probablemente me sentiría más reconocida” (estudiante 4). Esta ambivalencia revela que, aunque el refuerzo valida el conocimiento y orienta al estudiante, carece de la carga afectiva y la empatía que normalmente aporta un docente humano, lo cual plantea consideraciones éticas sobre la calidad del vínculo e interacción pedagógico mediado por IA.

El peldaño de inquietudes muestra que las críticas de la IA son, en general, bien recibidas gracias a su tono claro y respetuoso. El estudiante 1 señala que “estas críticas son constructivas, lo cual refuerza tus errores”, mientras otro comenta que

“me sentí bien criticado y al grano, sin dar vueltas” (estudiante 3). No obstante, aparecen cuestionamientos cuando la IA exige explicaciones adicionales que no fueron requeridas por la consigna. Por ejemplo, “aunque me dio todo el puntaje, creo que la parte donde menciona incluir ejemplos es un poco innecesaria” (estudiante 9). Estas observaciones sugieren que la retroalimentación sería más efectiva si ajustara su nivel de exigencia al contenido realmente trabajado y a los criterios explícitos de la evaluación.

En relación con el peldaño de sugerencias, los estudiantes consideran que la IA ofrece recomendaciones “claras, concisas y fáciles de aplicar”, lo que contribuye a su aprendizaje autónomo. “Me ayudó a tenerlo en cuenta para futuras ocasiones” (estudiante 10), comenta uno, mientras otro destaca que “me recalco el orden de redacción al explicar mi algoritmo” (estudiante 3). Sin embargo, la utilidad de estas sugerencias depende del grado de alineación con el contexto del curso. Algunos critican que “no habíamos visto complejidad temporal, por lo cual confundía la retroalimentación” (estudiante 5), evidenciando que el sistema puede perder eficacia si opera desconectado del temario. A pesar de ello, la mayoría valora la estructura y claridad de las sugerencias como un recurso para mejorar progresivamente.

Finalmente, en cuanto al impacto en el aprendizaje, los estudiantes coinciden en que el uso de inteligencia artificial ha fortalecido su proceso formativo, contribuyendo significativamente a la reflexión sobre sus propios aprendizajes. La retroalimentación inmediata fue valorada por su capacidad de activar conocimientos previos, “recordar detalles que se me olvidaban” (estudiante 6) y permitir una revisión oportuna: “retroalimentar al momento” (estudiante 2); además, fue descrita como “de utilidad para poder reflexionar sobre mi proceso de enseñanza-aprendizaje” (estudiante 4).

Algunos participantes reconocen el aporte de esta herramienta como un recurso estratégico: “me hizo mejorar en temas que no tenía muy bien profundizados” (estudiante 8) y “me ayudó a identificar elementos importantes no considerados antes” (estudiante 1). Estas percepciones reflejan una apropiación activa de la retroalimentación automatizada como apoyo al desarrollo de competencias reflexivas y metacognitivas.

Por otra parte, en la categoría de juicio crítico y recomendaciones, se observa una actitud proactiva y analítica. Los estudiantes no solo valoran el uso de la IA, sino que expresan propuestas concretas para mejorar su funcionamiento. Entre ellas destacan el reorganizar los pasos de la Escalera de Wilson colocando valorar antes que aclarar, incorporar una sección de discusión para profundizar contenidos, entrenar previamente la IA con materiales del curso, y evitar penalizaciones injustificadas no bajar puntaje por falta de ejemplos si no son requeridos o están en los criterios de evaluación.

Estas propuestas evidencian que el estudiantado no se posiciona como receptor pasivo del sistema, sino como agente reflexivo y comprometido con el perfeccionamiento de una herramienta pedagógica que consideran útil, pero que desean ver más contextualizada, justa y alineada con los objetivos formativos del curso.

DISCUSIÓN Y CONCLUSIONES

Los resultados de este estudio evidencian el potencial pedagógico de integrar retroalimentación automatizada mediada por inteligencia artificial generativa, en particular a través de GPT-4o estructurado mediante la Escalera de Wilson. Esta propuesta se alinea con la necesidad de transformar las prácticas de retroalimentación en la educación universitaria, que los estudiantes suelen percibir como tardías, superficiales y poco específicas (Carless y Boud, 2018; Winstone et al., 2022; Quezada y Salinas, 2021). En entornos abiertos y a distancia, donde la inmediatez y la claridad son factores críticos, la IAGen demuestra ser capaz de reducir tiempos de respuesta y de entregar comentarios estructurados y focalizados, potenciando la dimensión formativa de la evaluación (Puertas y Cano, 2024; Holmes et al., 2019).

La Escalera de Wilson ofreció un andamiaje útil para organizar la retroalimentación en los cuatro peldaños (aclarar, valorar, expresar inquietudes y sugerir), lo que permitió generar observaciones más comprensibles, específicas y orientadas a la mejora, incluso en un formato automatizado. Los estudiantes destacaron esta experiencia por su rapidez, claridad e innovación, reconociendo en ella una herramienta capaz de complementar el rol docente. No obstante, también señalaron tensiones asociadas a sentimientos de estrés, mayor exigencia en la preparación de sus respuestas y la ausencia del factor humano que confiere cercanía y calidez al proceso de retroalimentación. Estos matices refuerzan que la tecnología no puede sustituir al profesor, sino que debe concebirse como un apoyo complementario a la interacción pedagógica. Esta interpretación coincide con lo planteado por Valenzuela y Pérez (2025), quienes subrayan el carácter insustituible del profesorado en términos de interacción humana y apoyo personalizado, recordando que la inteligencia artificial puede constituir un recurso valioso, pero siempre subordinado a la dimensión relacional y afectiva propia de la enseñanza.

Un hallazgo relevante fue la observación estudiantil respecto del peldaño Aclarar. En la práctica actual, la herramienta no permite al estudiante aportar información adicional antes de recibir comentarios, lo que reduce el alcance formativo de este paso. Incluir un mecanismo interactivo, como un chat o ventana de aclaración, podría enriquecer la personalización de la retroalimentación, alineando los comentarios con las necesidades reales de los estudiantes (Luckin y Holmes, 2016; López Regalado et al., 2024). Esta posibilidad adquiere mayor importancia en contextos masivos o de educación a distancia, donde los espacios de interacción suelen ser limitados.

En el plano tecnológico, la validación por expertos y la percepción positiva del estudiantado coinciden con lo planteado por Zhang et al. (2024) y Baral et al. (2024), quienes destacan que GPT puede generar retroalimentación coherente y funcional siempre que los prompts estén cuidadosamente diseñados. Este aspecto resulta crítico en escenarios que requieren estandarización y escalabilidad (Carless y Winstone, 2023). No obstante, como advierten Jauhiainen y Garagorry Guerra (2024), persisten limitaciones vinculadas a la empatía, la contextualización y la

sensibilidad afectiva, lo que reafirma la necesidad de supervisión docente para asegurar la pertinencia pedagógica y el impacto formativo.

El estudio también muestra que estudiantes de distintas áreas y niveles académicos (segundo y cuarto año) valoraron de manera similar la retroalimentación automatizada, lo que sugiere que la Escalera de Wilson, aplicada con IA generativa, constituye un marco adaptable y replicable en diversos contextos. Sin embargo, su implementación requiere incorporar criterios éticos, revisión docente y un diseño riguroso de prompts que orienten adecuadamente la producción de la IA.

Finalmente, emergen desafíos que invitan a profundizar en investigaciones futuras, particularmente en el análisis de sesgos algorítmicos, la evaluación de riesgos éticos vinculados a la privacidad y la equidad, así como en la necesidad de garantizar la sostenibilidad técnica de estas herramientas. Resulta igualmente clave explorar su adaptación a diversas disciplinas académicas, asegurando la pertinencia pedagógica y la alineación con estándares de calidad en educación superior. En esta línea, se recomienda avanzar hacia una integración responsable y contextualizada de la inteligencia artificial en los procesos evaluativos universitarios, de manera que sus beneficios técnicos se complementen con la mediación pedagógica docente. A futuro, resulta pertinente ampliar el alcance de los estudios a distintas disciplinas y explorar nuevas dimensiones de su impacto en la educación, considerando variables como la calidad argumentativa, la resolución de problemas, el pensamiento crítico, la autonomía estudiantil y la transformación de las prácticas evaluativas. Asimismo, se sugiere desarrollar investigaciones longitudinales que permitan analizar los efectos sostenidos del uso de IAGen en el aprendizaje y en la práctica docente. En paralelo, se destaca la necesidad de fortalecer la formación del profesorado en el diseño de prompts, la reflexión ética y la evaluación crítica del uso de estas tecnologías, con el fin de asegurar una implementación más consciente y formativa.

NOTAS

- ¹ Moodle es una plataforma de gestión del aprendizaje (LMS, por sus siglas en inglés) de código abierto ampliamente utilizada en contextos educativos para la creación, entrega y seguimiento de cursos en línea.
- ² Versión gpt-4o-2024-08-06
- ³ Versión gpt-4o-mini-2024-07-18
- ⁴ CodeRunner es un plugin para Moodle que permite la evaluación automática de código informático. Se integra con un motor de ejecución (por ejemplo, Python) que procesa las respuestas en tiempo real, generando retroalimentación y puntajes inmediatos de acuerdo con criterios de evaluación preestablecidos.

REFERENCIAS

- Al-Azawei, A., Abdullah, A. A., Mohammed, M. K. y Abod, Z. A. (2023). Predicting online learning success based on learners' perceptions: The integration of the information system success model and the security triangle framework. *International Review of Research in Open and Distributed Learning*, 24(2), 7295. <https://doi.org/10.19173/irrodl.v24i2.6895>
- Andrade, H. L. y Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, 26(1), 103-117. <https://doi.org/10.1080/0969594X.2019.1571992>
- Anthropic. (2024). *Claude 3.7* [Modelo de lenguaje de gran tamaño]. <https://www.anthropic.com>
- Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B. y Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, 18(2), 261-271. <https://doi.org/10.30574/gscarr.2024.18.2.0062>
- Baral, S., Worden, E., Lim, W.-C., Luo, Z., Santorelli, C., Gurung, A. y Heffernan, N. (2024). *Automated feedback in math education: A comparative analysis of LLMs for open-ended responses*. arXiv. <https://doi.org/10.48550/arXiv.2411.08910>
- Black, P. y Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551-575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Bonales-Daimiel, G., Martínez-Estrella, E. C. y Sierra-Sánchez, J. (2025). Evolución del perfil docente y surgimiento de nuevos roles profesionales en la era de la inteligencia artificial (IA). *Pixel-Bit. Revista de Medios y Educación*, 73, art. 3. <https://doi.org/10.12795/pixelbit.109085>
- Booth, W. C., Colomb, G. G. y Williams, J. M. (2008). *The craft of research* (3ª ed.). University of Chicago Press. <https://doi.org/10.7208/chicago/9780226062648.001.0001>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
- Carless, D. y Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Carless, D. y Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150-163. <https://doi.org/10.1080/13562517.2020.1782372>
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gašević, D. y Chen, G. (2023). *Can large language models provide feedback to students? A case study on ChatGPT*. EdArXiv. <https://doi.org/10.35542/osf.io/hcgzj>
- DeepSeek. (2024). *DeepSeek R1* [Modelo de IA generativa]. <https://www.deepseek.com/>
- European Commission. (2023). *Teaching with AI – Assessment, feedback and personalisation. Briefing report No. 7 (European Digital Education Hub)*. Erasmus+ Programme. https://resitve.sio.si/wp-content/uploads/sites/7/2023/11/AI-squad-output_briefing-report-7.pdf
- Galindo-Domínguez, H., Delgado, N., Losada, D. y Etxabe, J. M. (2023). An

- analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education*, 40(1), 41-56. <https://doi.org/10.1080/21532974.2023.2284726>
- García-Peñalvo, F. J. (2023). La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: ¿Disrupción o pánico? *Education in the Knowledge Society (EKS)*, 24, e31279-e31279. <https://doi.org/10.14201/eks.31279>
- García Peñalvo, F. J., Llorens-Largo, F. y Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39. <https://doi.org/10.5944/ried.27.1.37716>
- Goodrich, H. (2011). Escalera de la retroalimentación [Adaptación basada en la Escalera de Retroalimentación de Ron Berger]. En A. Goodrich (Coord.), *Protocols in the classroom*. Harvard Project Zero.
- Google DeepMind. (2024). *Gemini 2.5* [Multimodal AI model]. <https://deepmind.google/>
- Hao, Y., Sun, Y., Dong, L., Han, Z., Gu, Y. y Wei, F. (2022). *Structured prompting: Scaling in-context learning to 1,000 examples*. arXiv. <https://doi.org/10.48550/arXiv.2212.06713>
- Holmes, W., Bialik, M. y Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Jauhainen, J. S. y Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 1-18. <https://doi.org/10.1080/14703297.2024.2422337>
- Landeta, J. (1999). *El método Delphi: Una técnica de previsión para la incertidumbre*. Ariel.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. y Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35. <https://doi.org/10.1145/3560815>
- LLaMa. (2024). *LLaMA4* [Large language model]. <https://www.llama.com/>
- López Regalado, O., Núñez-Rojas, N., López Gil, O. R. y Sánchez-Rodríguez, J. (2024). El análisis del uso de la inteligencia artificial en la educación universitaria: Una revisión sistemática. *Pixel-Bit. Revista de Medios y Educación*, 70, 97-122. <https://doi.org/10.12795/pixelbit.106336>
- Luckin, R. y Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson. <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/IntelligenceUnleashedSPANISH.pdf>
- Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research*, 1(2), 1-10. <https://doi.org/10.17169/fqs-1.2.1089>
- Ministerio de Educación de Chile. (2019). *Orientaciones para la implementación del Decreto 67/2018 de evaluación, calificación y promoción*. MINEDUC. <https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/14279/orientaciones%20decreto%2067.pdf>
- Mistral AI. (2024). *Mistral* [Large language model]. <https://mistral.ai/>
- Molloy, E., Boud, D. y Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, 45(4), 527-540. <https://doi.org/10.1080/02602938.2019.1667955>
- OpenAI. (2023). *Best practices for prompt engineering with the OpenAI API*. <https://>

- help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api
- OpenAI. (2024). *GPT-4o, GPT-4.5 model cards*. <https://openai.com>
- Ossa, C. y Willatt, C. (2023). Uso de inteligencia artificial generativa para retroalimentar escritura académica en procesos de formación inicial docente. *European Journal of Education and Psychology*, 16(2), 1-16. <https://doi.org/10.32457/ejep.v16i2.2412>
- Puertas, E. y Cano, E. (2024). ¿Puede la inteligencia artificial proporcionar un feedback más sostenible? *Digital Education Review*, 45(1), 50-58.
- Quezada, S. y 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.
- Reynolds, L. y McDonell, K. (2021, mayo). Prompt programming for large language models: Beyond the few-shot paradigm. En *Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). <https://doi.org/10.1145/3411763.3451760>
- Romero Alonso, R., Araya Carvajal, K. y Reyes Acevedo, N. (2025). Rol de la inteligencia artificial en la personalización de la educación a distancia: Una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(1), 9-36. <https://doi.org/10.5944/ried.28.1.41538>
- Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö., Cohen, T., Yetisgen, M. y Wang, Y. (2024). Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association*, 31(9), 2114-2124. <https://doi.org/10.1093/jamia/ocae074>
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025). *The prompt report: A systematic survey of prompt engineering techniques*. arXiv. <https://doi.org/10.48550/arXiv.2406.06608>
- Shute, V. J. y Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1), 1-19. <https://doi.org/10.1111/jcal.12172>
- Steurer, J. (2011). The Delphi method: An efficient procedure to generate knowledge. *Skeletal Radiology*, 40, 959-961. <https://doi.org/10.1007/s00256-011-1145-z>
- Valenzuela Caico, R. y Pérez Carvajal, 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/ried.vi9.221>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. y Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. <https://doi.org/10.48550/arXiv.2201.11903>
- Weng, X., Xia, Q., Gu, M., Rajaram, K. y Chiu, T. K. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, 40(6), 37-55. <https://doi.org/10.14742/ajet.9540>
- William, D. (2011). What is assessment for learning? *Studies in Educational*

- Evaluation*, 37(1), 3-14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
- Wilson, D. (2013). *Ladder of feedback* [Documento de trabajo]. Project Zero, Harvard Graduate School of Education. <https://pz.harvard.edu/resources/ladder-of-feedback>
- Winstone, N. E., Boud, D., Dawson, P. y Heron, M. (2022). From feedback-as-information to feedback-as-process: A linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education*, 47(2), 213-230. <https://doi.org/10.1080/02602938.2021.1902467>
- xAI. (2024). *Grok 3* [Large language model]. <https://grok.com/>
- Zhang, Z., Dong, Z., Shi, Y., Price, T., Matsuda, N. y Xu, D. (2024). Students' perceptions and preferences of generative artificial intelligence feedback for programming. En *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23250-23258. <https://doi.org/10.1609/aaai.v38i21.30372>

ANEXO

Prompt replicable

i. Contexto teórico

En este apartado se debe describir de manera breve y precisa el instrumento de evaluación utilizado. La descripción debe incluir: propósito general, público objetivo, objetivos de aprendizaje evaluados, criterios de evaluación, puntaje total, condiciones de aprobación, formato de aplicación y lineamientos para responder. También debe detallar la estructura general (tipos de ítems y cantidad aproximada), el tipo de habilidades o conocimientos que busca medir y el modo en que se espera que el estudiante desarrolle sus respuestas. La redacción debe ser clara, suficiente y comprensible para permitir la replicabilidad del instrumento. (Borrar este texto después de completar el contexto)

ii. Enunciado de la pregunta:

Aquí se coloca el enunciado de la pregunta.

iii. Criterios de evaluación/respuesta esperada:

En este espacio se debe incluir una rúbrica que contenga los criterios de evaluación, los niveles de desempeño esperados y las descripciones correspondientes a cada nivel. Los criterios deben estar alineados con los objetivos de aprendizaje y permitir una valoración objetiva de la calidad, pertinencia y profundidad de la respuesta.

La escala de puntuación debe ser explícita y consistente, y debe incluirse el rango de puntajes para cada nivel. (Borrar este texto después de completar la rúbrica)

iv. Respuesta del estudiante:

Aquí se coloca la respuesta del estudiante.

v. Instrucciones para retroalimentar:

Asigna puntuación a la Respuesta del Estudiante utilizando la rúbrica (puntuación para cada criterio).

Justifica cada uno de los puntajes de la rúbrica.

Luego, redacta una retroalimentación breve en español utilizando el marco de la Escalera de Wilson (valoraciones, aclaraciones, inquietudes, sugerencias).

La retroalimentación debe tener los 4 componentes de 100 palabras cada uno (máximo), enfocados en cada componente de la Escalera:

VALORACIONES: Reconoce de forma genuina los aspectos positivos de la respuesta. Destaca fortalezas, ideas correctas o puntos bien expresados para reforzar buenas prácticas.

ACLARACIONES: Invita a precisar o desarrollar aspectos faltantes de la respuesta para mejorar su claridad y profundidad, sin emitir juicio.

INQUIETUDES: Señala errores en la respuesta utilizando un tono respetuoso y constructivo. Prefiere expresiones sugerentes y amables como “sería útil considerar...”, “podrías revisar...”, o “tal vez convenga precisar...”.

SUGERENCIAS: Ofrece sugerencias específicas de mejora para futuras respuestas. No pidas ejemplos ni códigos.

vi. Formato de salida en diccionario Python:

Como evaluador, sé generoso: si la respuesta cumple parcialmente con el criterio, asigna el puntaje máximo para ese nivel.

Se espera un diccionario en Python como respuesta: {"analisis1": "(nombre criterio 1) análisis criterio 1", "score1": score1, "analisis2": "(nombre criterio 2) análisis criterio 2", "score2": score2, ..., "valoraciones": "valoraciones", "aclaraciones": "aclaraciones", "inquietudes": "inquietudes", "sugerencias": "sugerencias"}

Fecha de recepción del artículo: 1 de junio de 2025

Fecha de aceptación del artículo: 2 de septiembre de 2025

Fecha de aprobación para maquetación: 1 de octubre de 2025

Fecha de publicación en OnlineFirst: 17 de octubre de 2025

Fecha de publicación: 1 de enero de 2026