Comunicar, n. 82, v. XXXIII, 2025 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

Capacidades de la IA en la Evaluación Formativa de TFG: Expertos Humanos vs GPT

AI Capabilities in Formative Assessment of Undergraduate Theses: Human Experts vs GPT

Dr. Cristian Velandia-Mesa*, Director de Investigaciones, Universidad El Bosque (Colombia) (velandiacristian@unbosque.edu.co) (http://orcid.org/0000-0002-7195-3365)
Dra. Ruth Stella Chacón Pinilla, Profesora Titular de la Universidad El Bosque (Colombia) (chaconruth@unbosque.edu.co) (https://orcid.org/0000-0002-5148-7811)
Dra. Erika Fernanda Cortés Ibarra, Profesora Titular de la Universidad El Bosque (Colombia) (corteserika@unbosque.edu.co) (https://orcid.org/0000-0003-3285-1582)
Dr. Carlos Eduardo Rodríguez Muñóz, Profesor Investigador de la Universidad El Bosque (Colombia) (cerodriguezm@unbosque.edu.co) (https://orcid.org/0000-0001-9464-9801)

RESUMEN

El acelerado desarrollo de la inteligencia artificial (IA) en contextos educativos plantea desafíos sustantivos para los modelos tradicionales de evaluación formativa en la Educación Superior. Este estudio tuvo como propósito examinar la capacidad de los modelos Generative Pretrained Transformers (GPT) para desempeñar funciones evaluativas en Trabajos de Fin de Grado (TFG), contrastando sus valoraciones con las emitidas por evaluadores humanos expertos. Se implementó un enfoque mixto de tipo no integrado, articulando un diseño cuasiexperimental de series cronológicas con grupo control, complementado con un análisis cualitativo del corpus de retroalimentaciones. Se seleccionaron intencionalmente 16 TFG como casos de estudio, evaluados en tres momentos sucesivos del proceso formativo. Desde el componente cuantitativo, se analizaron la evolución, consistencia y alineación de las calificaciones; mientras que el componente cualitativo exploró la profundidad crítica, estructura argumentativa y orientación pedagógica del feedback. Los hallazgos evidencian una progresiva convergencia entre las evaluaciones emitidas por GPT y las de los expertos, con altos niveles de correlación y concordancia en la fase final. Asimismo, se observó una mejora sostenida en la riqueza semántica, la precisión argumentativa y la adaptabilidad de la retroalimentación generada por la IA. Se concluye que, bajo condiciones controladas y criterios normativos claramente definidos, los modelos GPT presentan un potencial significativo como agentes complementarios en la evaluación formativa universitaria, aportando eficiencia, coherencia y escalabilidad, aunque con limitaciones en la personalización y el fomento de la reflexión crítica.

ABSTRACT

The accelerated development of artificial intelligence (AI) in educational contexts poses significant challenges to traditional models of formative assessment in Higher Education. This study aimed to examine the capacity of Generative Pretrained Transformer (GPT) models to perform evaluative functions in undergraduate thesis assessment, comparing their judgments with those issued by expert human evaluators. A non-integrated mixed-methods approach was employed, combining a quasi-experimental time-series design with a control group and a qualitative corpus analysis of feedback content. Sixteen undergraduate theses were intentionally selected as case studies and assessed at three successive stages of the formative process. The quantitative component analyzed the evolution, consistency, and alignment of scores, while the qualitative analysis explored the critical depth, argumentative structure, and pedagogical orientation of the feedback. The findings reveal a progressive convergence between GPT and expert evaluations, with high levels of correlation and agreement in the final stage. Additionally, GPT-generated feedback showed sustained improvement in semantic richness, argumentative precision, and adaptive capacity. It is concluded that, under controlled conditions and clearly defined evaluative criteria, GPT models exhibit significant potential as complementary agents in formative assessment within Higher Education, offering efficiency, consistency, and scalability. However, limitations remain regarding the personalization of feedback and the promotion of critical reflection, highlighting the need to enhance their pedagogical and metacognitive capabilities.

PALABRAS CLAVE | KEYWORDS

Inteligencia artificial, Educación Superior, Evaluación, Tecnología Educativa, Sistemas Inteligentes, Innovación Educativa.
Artificial Intelligence, Higher Education, Assessment, Educational Technologies, Intelligent Systems, Educational Innovation.

Received: 2025-03-20 | Reviewed: 2025-04-04 | Accepted: 2025-04-05 | Online First: 2025-07-21 | Published: 2025-07-24
DOI: https://doi.org/10.5281/zenodo.15995749 | Pages: 103-115

1. Introducción y Estado de la Cuestión

El avance vertiginoso de la inteligencia artificial (IA) ha transformado diversos ámbitos del conocimiento y la práctica profesional, incluyendo la Educación Superior. En este contexto, los Transformadores Generativos Preentrenados (GPT) emergen como herramientas innovadoras con la capacidad analizar y evaluar producciones académicas basadas en resultados de aprendizaje con un nivel de sofisticación sin precedentes. Su implementación en la evaluación formativa de Trabajos de Fin de Grado (TFG) plantea un desafío fundamental: ¿pueden estos modelos GPT alcanzar un nivel de criterio y profundidad evaluativa comparable al de los expertos humanos en el ámbito académico?

La evaluación formativa es un componente crucial en la Educación Superior, ya que permite potenciar la calidad de la producción estudiantil mediante retroalimentación continua y orientada al desarrollo de competencias investigativas y argumentativas. Tradicionalmente, este proceso ha sido llevado a cabo por expertos humanos, quienes, con base en su formación, experiencia y juicio crítico, proporcionan valoraciones fundamentadas y personalizadas. Sin embargo, este modelo evaluativo enfrenta limitaciones como la variabilidad en los criterios de evaluación, desafíos en la evaluación de procesos formativos en grupos numerosos y la demanda académica de profesores y tutores. En este sentido, la IA surge como una alternativa que proyecta eficiencia, ausencia de sesgos y escalabilidad en la evaluación formativa. No obstante, a pesar de su potencial, el uso de modelos de lenguaje como los GPT en la evaluación académica no está exento de controversias. Existen preocupaciones legítimas en torno a la capacidad de estas herramientas para interpretar matices argumentativos, evaluar la originalidad del pensamiento crítico y ofrecer retroalimentación constructiva alineada con los estándares académicos. Adicionalmente, la validez y fiabilidad de sus evaluaciones requieren un análisis comparativo riguroso frente a la evaluación humana tradicional.

Diversos estudios advierten que, si bien los modelos GPT pueden generar retroalimentaciones estructuradas y consistentes, su capacidad para identificar errores conceptuales, interpretar el contexto discursivo del autor y valorar de forma crítica los resultados de aprendizaje permanece limitada (García-Peñalvo y Vázquez-Ingelmo, 2023). Asimismo, se ha señalado el riesgo de una dependencia excesiva por parte de los estudiantes hacia estas herramientas, lo cual podría obstaculizar el desarrollo de competencias de autorregulación y pensamiento crítico (Paredes-Marín, Ramírez-Chumbe y Ramírez-Chumbe, 2024). A ello se suma el sesgo inherente a los datos utilizados en su entrenamiento, que puede incidir negativamente en la imparcialidad y contextualización de sus valoraciones. En consecuencia, su implementación en procesos de evaluación formativa requiere una validación empírica rigurosa que permita determinar en qué medida sus desempeños pueden complementar, enriquecer o aproximarse al juicio evaluativo de expertos humanos en contextos académicos complejos.

Esto conllevó a reflexionar sobre la necesidad de un enfoque riguroso y analítico en la implementación de estas tecnologías en la evaluación académica formativa. En primer lugar, es fundamental garantizar la solidez del diseño instruccional del Transformador Generativo. Además, resulta imprescindible evaluar con precisión tanto la eficiencia computacional de la inteligencia artificial como el juicio de los expertos humanos en su rol de evaluadores, con el fin de asegurar la calidad y confiabilidad del proceso educativo. En este sentido, es fundamental establecer criterios de evaluación rigurosos que reflejen con precisión el proceso de formación y aprendizaje en investigación de los estudiantes (Velandia-Mesa, Serrano-Pastor y Martínez-Segura, 2021). Además, en el ámbito de la inteligencia artificial, es crucial diseñar y estructurar de manera precisa los prompts y las instrucciones computacionales utilizadas en la construcción y ajuste de modelos GPT, garantizando así una evaluación alineada con estándares académicos y pedagógicos.

Los modelos GPT utilizan redes neuronales profundas y mecanismos de autoatención para procesar lenguaje natural, identificando dependencias sintácticas y semánticas en textos extensos. Su entrenamiento auto-regresivo y masivo mejora la generalización a distintos contextos lingüísticos. La arquitectura Multi-Head Attention optimiza la detección de patrones y la coherencia textual, favoreciendo su uso en evaluación formativa. No obstante, su rendimiento depende de la calidad y sesgo de los datos, limitando su capacidad para evaluar aspectos metacognitivos y contextuales frente al juicio experto humano (Kasneci et al., 2023).

Tradicionalmente, la evaluación formativa ha dependido del juicio de expertos, quienes no solo analizan aspectos técnicos y estructurales de los TFG, sino que también incorporan su experiencia disciplinaria, su conocimiento del contexto académico y sus capacidades para detectar fortalezas y debilidades en la argumentación investigativa de los estudiantes (Chacón Pinilla, 2014). Así mismo, los modelos GPT ofrecen

una alternativa basada en inteligencia artificial al automatizar este proceso (Guárdia-Ortiz, Bekerman y Zapata-Ros, 2024); esto nos lleva a pensar que esta alternativa podría ofrecer mayor eficacia y eficiencia en la amplitud del análisis de criterios de evaluación, en la búsqueda de cohesión y consistencia interna durante la evaluación formativa. La evaluación formativa es un proceso continuo que mejora el aprendizaje al ofrecer feedback durante el desarrollo de trabajos académicos. A diferencia de la evaluación sumativa, que califica el producto final, busca intervenir en momentos clave para que los estudiantes identifiquen fortalezas, corrijan debilidades y mejoren la argumentación y estructura antes de la entrega.

La evaluación formativa en la formación investigativa y en los TFG permite ajustes progresivos basados en observaciones detalladas (Cortés Ibarra y Martínez Clares, 2020), fortaleciendo la coherencia, el rigor metodológico y la calidad académica. Su carácter iterativo fomenta la autonomía y autorregulación de los estudiantes, facilitando la revisión crítica de sus trabajos y la toma de decisiones informadas. Basada en la interactividad y la personalización del feedback, la evaluación formativa impulsa la mejora continua, adaptando las recomendaciones a las necesidades específicas de cada estudiante y promoviendo la consolidación del conocimiento y la argumentación académica. En este contexto, la comparación entre evaluadores humanos y modelos de inteligencia artificial como GPT adquiere especial relevancia, al posibilitar el análisis de la capacidad de estas herramientas automatizadas para replicar las propiedades esenciales de la evaluación formativa. No obstante, persiste el interrogante sobre la medida en que dichas tecnologías pueden igualar la capacidad analítica, interpretativa y contextual del juicio experto humano, lo que plantea un debate sobre la complementariedad y las limitaciones de la inteligencia artificial en la evaluación académica.

1.1. Revisión de la Literatura

Los estudios recientes sobre modelos GPT en educación evidencian su potencial para complementar la evaluación académica y mejorar el aprendizaje. Paredes-Marín et al. (2024) compararon las evaluaciones de estudiantes de Magisterio con las generadas por ChatGPT, encontrando coincidencias en las calificaciones, lo que sugiere su viabilidad como apoyo en corrección automatizada. García-Peñalvo y Vázquez-Ingelmo (2023) analizaron la retroalimentación de GPT-4 en trabajos universitarios, concluyendo que la IA ofrece feedback rápido y estructurado, aunque limitada en aspectos críticos y reflexivos. Un estudio previo identificó que, identificaron que, aunque la IA es eficiente y coherente, carece de profundidad y contextualización en la interpretación del aprendizaje.

La revisión sistemática de la literatura permitió establecer niveles diferenciados de capacidad evaluativa entre expertos humanos y sistemas de IA, fundamentados en criterios de evaluación académica, procesamiento de lenguaje natural (PLN) y aprendizaje automático, como se refleja en la Tabla 1. El análisis empírico evidenció contrastes en la aplicación de criterios, la profundidad crítica y la precisión del feedback, definiendo parámetros comparativos relevantes para la Educación Superior. Los expertos destacan por su habilidad para interpretar contextos disciplinares, detectar falacias y generar retroalimentación contextualizada (Boud y Falchikov, 2006; Nicol y Macfarlane-Dick, 2006), mientras que los modelos GPT sobresalen por su rapidez y consistencia algorítmica, aunque limitados en comprensión semántica y razonamiento metacognitivo (Floridi y Chiriatti, 2020). Estos hallazgos permitieron estructurar, de manera comparativa y con criterios operacionales, un análisis riguroso sobre las fortalezas y limitaciones de ambos enfoques en el ámbito universitario.

Kalla et al. (2023) evidenciaron que, aunque las narrativas de ChatGPT resultaron estructuralmente correctas, fueron percibidas por los lectores como carentes de inmersión emocional frente a las producciones humanas. Yeadon, Peach y Testrow (2024) identificaron, mediante un análisis computacional de corpus, divergencias estilísticas y patrones léxicos distintivos entre GPT-3.5, GPT-4 y escritores expertos. Estos hallazgos subrayan tanto las posibilidades como las limitaciones de GPT en contextos educativos, y refuerzan la necesidad de su integración estratégica con la evaluación humana para propiciar aprendizajes más profundos y significativos.

En este contexto, el presente estudio tuvo como objetivo examinar la capacidad de los modelos GPT para llevar a cabo procesos de evaluación formativa de Trabajos de Fin de Grado (TFG) en Educación Superior. Para ello, se compararon sus valoraciones con las de evaluadores humanos expertos y se analizó su evolución a lo largo del tiempo, con el fin de identificar patrones de aprendizaje y consistencia en sus juicios. Los hallazgos de esta investigación contribuirán a una mejor comprensión del papel que la inteligencia

artificial puede desempeñar en los procesos de evaluación académica, proporcionando evidencia empírica sobre sus ventajas, limitaciones y potencial complementario en la Educación Superior. Con ello, se espera generar insumos relevantes para la toma de decisiones respecto a la integración de tecnologías avanzadas en la enseñanza y evaluación del aprendizaje en contextos universitarios.

Tabla 1: Niveles de capacidad evaluativa.

Características	Evaluador Experto		Evaluación con IA (GPT)	
Comprensión del contexto disciplinar	↑	Alta	↓	Limitada (no comprende contexto disciplinar más allá de datos entrenados)
Capacidad de interpretación crítica	↑	Alta	↔	Moderada (evalúa estructura, pero puede fallar en análisis profundo)
Consistencia en la evaluación	↔	Moderada (puede variar entre evaluadores)	↑	Alta (mantiene consistencia algorítmica en criterios predefinidos)
Rapidez en la evaluación	↔	Media (requiere tiempo para análisis detallado)	↑	Alta (procesa grandes volúmenes de texto en poco tiempo)
Adaptabilidad del feedback	↑	Alta (personaliza el feedback según el estudiante)	↔	Moderada (adapta respuestas, pero con limitaciones contextuales)
Detección de matices discursivos	↑	Alta (detecta ironías, ambigüedades y argumentación implícita)	↓	Baja (dificultad para interpretar ironías, ambigüedades o intenciones implícitas)
Análisis de originalidad y plagio	↑	Alta (analiza la construcción argumentativa y referencias)	↔	Moderada (detecta coincidencias textuales, pero no interpreta argumentación)
Capacidad de razonamiento ético	↑	Alta (evalúa la ética del contenido en contexto académico)	↓	Baja (no posee razonamiento ético autónomo)
Aplicación de criterios normativos	↑	Alta (ajusta la evaluación según normativas y estándares académicos)	↑	Alta (aplica criterios preprogramados de manera uniforme)
Capacidad de aprendizaje y mejora	↑	Alta (aprende con la experiencia y nuevos conocimientos)	↔	Moderada (aprende de interacciones, pero con sesgos y limitaciones en actualización de conocimientos)

2. Material y Método

La investigación se estructuró bajo un enfoque mixto cuantitativo-cualitativo no integrado (Creswell y Plano Clark, 2017), permitiendo el análisis comparativo de la evaluación formativa de Trabajos de Fin de Grado (TFG) realizada por Transformadores Generativos Preentrenados (GPT) y evaluadores humanos expertos. La elección de este enfoque respondió a la necesidad de examinar la profundidad del proceso evaluativo desde una perspectiva complementaria pero metodológicamente diferenciada, garantizando que cada componente cuantitativo y cualitativo mantuviera autonomía en su análisis e interpretación. Esta aproximación permitió contrastar las valoraciones y el feedback generado por ambos tipos de evaluadores sin fusionar los métodos de análisis, asegurando así una evaluación rigurosa y precisa del desempeño de GPT en comparación con el criterio experto humano.

La investigación empleó un diseño longitudinal comparativo (Menard, 2002), estructurado en tres momentos de evaluación, para contrastar las evaluaciones cuantitativas y valoraciones cualitativas de GPT y evaluadores humanos (O^x). En el primer momento, se evaluaron versiones preliminares de los TFG (X^x), identificando fortalezas y debilidades en estructura, argumentación, claridad y fundamentación. GPT y los expertos emitieron juicios y retroalimentación simultáneamente. En el segundo momento, los estudiantes presentaron versiones revisadas tras aplicar las recomendaciones, permitiendo analizar la consistencia de los criterios evaluativos entre ciclos. En el tercer momento, se evaluaron las versiones finales para examinar la evolución del trabajo y el impacto del feedback, comparando profundidad, precisión y aplicabilidad de los juicios. De forma complementaria, se aplicó un diseño transversal para comparar las evaluaciones en cada fase, identificando patrones de ajuste, alineación progresiva y consistencia entre el modelo de IA y los evaluadores humanos. Ambos diseños se ilustran en la Tabla 2.

Tabla 2: Diseño metodológico.

	Longitudinal comparativo →				
Diseño transversal ↓	Inicial		Ajuste		Final
Experimental - Evaluación GPT	O^1	X^1	O^3	X^3	O^5
Control – Evaluación Expertos Humanos -	O^2	X^2	O^4	X^4	O^6

El componente cuantitativo se sustentó en un diseño cuasiexperimental de series cronológicas con grupo control, en el cual los TFG fueron evaluados en tres etapas sucesivas. El grupo experimental

estuvo compuesto por las valoraciones generadas por GPT, mientras que el grupo control correspondió a las evaluaciones realizadas por expertos humanos. La comparación sistemática entre ambos permitió identificar similitudes y diferencias en la aplicación de criterios, la consistencia y la utilidad de las valoraciones, aportando evidencia empírica sobre la eficacia de la inteligencia artificial en la evaluación académica en Educación Superior. Desde la perspectiva cuantitativa, este estudio realizó un análisis comparativo entre las valoraciones de GPT y las de evaluadores humanos expertos, determinando el grado de coincidencia y coherencia en relación con los criterios normativos aplicados a los TFG. Este procedimiento permitió valorar la capacidad de la IA para reproducir patrones evaluativos expertos e identificar posibles sesgos sistemáticos. Los datos fueron procesados con IBM SPSS Statistics (versión 27).

En paralelo, el componente cualitativo se centró en el análisis del corpus de las dimensiones discursivas y lingüísticas del feedback proporcionado por GPT y por los expertos. A través de un análisis de contenido con enfoque descriptivo-interpretativo, se exploraron categorías clave que permitieron evaluar el nivel de profundidad crítica, estructura y utilidad pedagógica de las retroalimentaciones. El proceso de codificación y triangulación de datos (Flick, 2018) facilitó la identificación de patrones discursivos y divergencias en la orientación y calidad de las observaciones generadas por la IA frente a las de los evaluadores humanos. La codificación y análisis de contenido fueron gestionados con Atlas.ti (versión 23).

Considerando que la pregunta central de esta investigación se centró en analizar la capacidad de los modelos GPT para aproximarse al criterio y a la profundidad evaluativa de los expertos humanos, la adopción de un enfoque mixto no integrado se configuró como una decisión metodológica idónea. Bajo esta lógica de diseño, el estudio logró ofrecer un análisis sistemático y riguroso sobre las fortalezas y limitaciones que presenta la inteligencia artificial en los procesos de evaluación académica universitaria. En conjunto, esta estrategia metodológica aportó evidencia empírica robusta sobre la capacidad de los modelos GPT para emular patrones evaluativos estandarizados, al tiempo que dejó en evidencia sus limitaciones en la replicación de la sensibilidad crítica y contextual propia del juicio experto humano en el ámbito de la Educación Superior.

2.3. Dispositivos y Técnicas de Registro de Información

Para el análisis de los criterios de evaluación, se aplicó la escala acumulativa de Guttman para medir de forma progresiva y estructurada cuatro dimensiones clave de la producción académica en los Trabajos de Fin de Grado (Velandia-Mesa et al., 2021). Cada dimensión permitió identificar niveles de calidad y rigor en la elaboración de los TFG. Este enfoque garantizó una valoración jerárquica y precisa de la estructura, fundamentación y coherencia argumentativa, bajo criterios metodológicos rigurosos. La naturaleza jerárquica y unidimensional de la escala facilitó examinar la calidad y evolución de los TFG, asegurando la coherencia y secuencialidad en su evaluación. Además, permitió identificar el grado de desarrollo de los trabajos y analizar la evolución de los GPT en las cuatro dimensiones evaluadas, propiciando una valoración integral de la estructura, argumentación y fundamentación teórica de cada TFG.

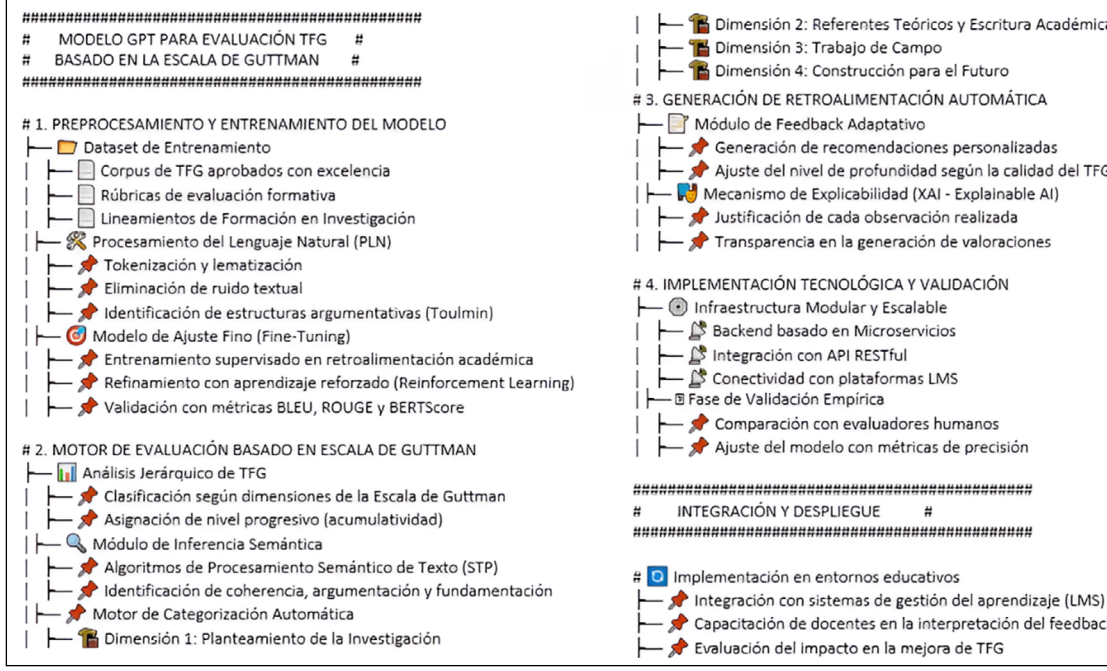
La escala acumulativa de Guttman utilizada en este estudio fue previamente validada en dos investigaciones anteriores, las cuales se enfocaron en su diseño y consistencia para la evaluación de la formación en investigación educativa (Velandia-Mesa, Serrano-Pastor y Martínez-Segura, 2019). Posteriormente, la escala fue aplicada en un estudio adicional (Velandia-Mesa et al., 2021), consolidando su uso como un instrumento riguroso para la valoración del desarrollo investigativo en educación superior. Basada en el modelo de Teoría de Respuesta al Ítem (IRT), su validación se estructuró en cuatro fases: exploratoria (validez de contenido), descriptiva (validez de constructo), analítica (consistencia interna) y explicativa (análisis R&R). La combinación de conglomerados, medidas de concordancia e índices de estabilidad permitió consolidar su rigor metodológico.

El desarrollo del asistente GPT para la evaluación formativa de Trabajos de Fin de Grado (TFG) se fundamentó en un modelo avanzado de procesamiento del lenguaje natural (PLN), optimizado para el análisis académico y la generación de retroalimentación especializada (Figura 1). La arquitectura se estructuró sobre redes neuronales profundas basadas en Transformers (Fúquene Ardila, 2024) y se sometió a un proceso de ajuste fino (fine-tuning) utilizando corpus específicos de producción científica y normativas académicas. Con el fin de garantizar la alineación con los estándares de calidad, se integró un módulo de correspondencia con la escala acumulativa de Guttman, permitiendo que las observaciones

generadas por la IA respondieran a criterios jerárquicos y acumulativos. Asimismo, el sistema incorporó un motor de inferencia semántica, diseñado para evaluar la argumentación, la coherencia discursiva y la estructura metodológica de los TFG, emulando la granularidad del análisis propio de un evaluador experto.

Desde la perspectiva de la ingeniería de software, el asistente fue implementado bajo una arquitectura modular y escalable, con un back-end basado en microservicios, lo que facilitó su integración en entornos de gestión del aprendizaje. Su motor de evaluación combinó algoritmos de procesamiento semántico de texto (STP) con técnicas de aprendizaje supervisado y no supervisado, optimizando la precisión de la retroalimentación. Además, el sistema incorporó principios de interpretabilidad mediante la adopción de mecanismos de explicabilidad en IA (Explainable AI, XAI) que sustentaron cada observación generada. Para fortalecer su funcionalidad, se diseñó un módulo de retroalimentación adaptativa capaz de modular la profundidad del feedback en función de la complejidad del TFG y el perfil del estudiante. Finalmente, el asistente fue sometido a una validación empírica basada en métricas estándar de PLN (BLEU, ROUGE y BERTScore), asegurando que la calidad de su retroalimentación resultara equiparable a la de evaluadores humanos expertos.

Figura 1: Arquitectura modelo GPT para Evaluación de TFG basado en escala de Gutmann.



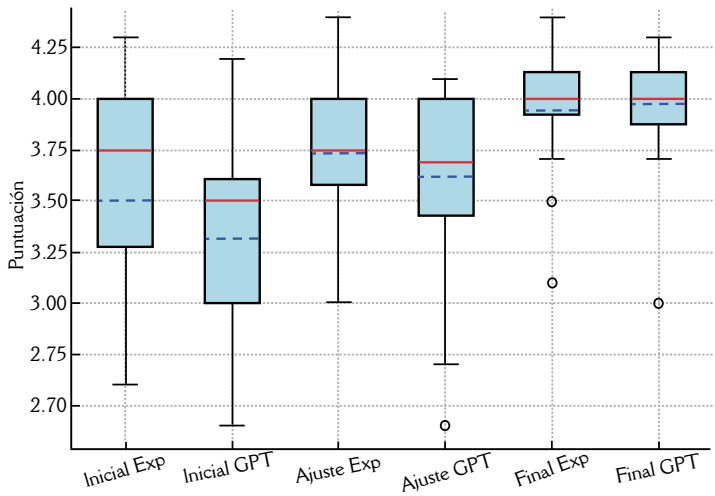
3. Análisis y Resultados

3.1. Componente Cuantitativo

El análisis de las evaluaciones formativas de TFG revela una convergencia progresiva entre los juicios de los expertos humanos y los modelos GPT. En una primera aproximación en el análisis, la tendencia sugiere que los modelos GPT podrían replicar, en términos generales, la coherencia evaluativa de los expertos, especialmente cuando se aplican criterios bien definidos. Además, a lo largo del proceso, ambas fuentes mantienen un patrón estable en sus calificaciones, evidenciando una aplicación sistemática de los criterios de evaluación. No obstante, las diferencias iniciales observadas, con GPT asignando una μ más baja en la fase inicial (3.35 vs. 3.57 en expertos), reflejan una menor flexibilidad en su valoración frente a la adaptación progresiva de los evaluadores humanos. En la fase final del estudio, las medidas de tenencia central de las evaluaciones ($\bar{x} = 3.94$) y ($Md = 4.00$) muestran alta correspondencia, lo que indicaría una posible alineación en la asignación de puntuaciones (Figura 2).

Entre las principales capacidades del modelo GPT en la evaluación formativa destaca su baja dispersión en las puntuaciones, con una desviación estándar (σ) de 0.48 en la fase inicial y 0.31 en la final, lo que demuestra su estabilidad y consistencia en la aplicación de criterios. Esta característica minimiza sesgos individuales y contribuye a una evaluación equitativa y reproducible. Asimismo, GPT ofrece una retroalimentación inmediata, lo que optimiza el proceso de revisión y permite a los estudiantes recibir valoraciones rápidas sobre su desempeño. A pesar de la mejora progresiva en su desempeño a lo largo del proceso evaluativo, los modelos GPT aún enfrentan desafíos en cuanto a su capacidad de aprendizaje autónomo. Si bien pueden ajustar sus valoraciones mediante calibraciones algorítmicas y refinamientos en su programación, este ajuste no implica un aprendizaje auténtico, ya que no incorporan una interpretación crítica ni un razonamiento contextualizado basado en la evolución del trabajo. No obstante, su capacidad para asimilar patrones y replicar criterios evaluativos sugiere un potencial para perfeccionar su desempeño, especialmente si se integran mecanismos de optimización continua que permitan una mayor adaptación a los estándares y dinámicas de la evaluación formativa en educación superior. A pesar de su eficiencia, el modelo GPT presenta limitaciones en la evaluación formativa. Su menor σ en comparación con los expertos (0.53 vs. 0.36 en la fase de ajuste) indica una menor variabilidad en sus valoraciones, lo que podría traducirse en una evaluación más rígida y menos adaptativa. En contraste, los expertos humanos ajustan sus criterios conforme avanza la evaluación, permitiendo una apreciación más matizada del trabajo en función de su evolución.

Figura 2: Distribución de evaluaciones – GPT vs Expertos.



	Inicial	Ajuste	Final
M Exp	3.57	3.79	3.94
Md Exp	3.57	3.8	4.0
ó Exp	0.54	0.36	0.31
M GPT	3.35	3.57	3.94
Md GPT	3.5	3.7	4.0
ó GPT	0.48	0.53	0.31

Los resultados de las pruebas de normalidad indican que GPT y los expertos presentan un comportamiento evaluativo diferenciado a lo largo del proceso formativo, con una tendencia inicial más estructurada que se vuelve progresivamente más heterogénea. En la fase inicial, ambos modelos presentan distribuciones normales, lo que sugiere que, en un primer momento, sus criterios de evaluación son relativamente consistentes y predecibles dentro de un patrón estadístico estable. Sin embargo, en la fase de ajuste, GPT deja de cumplir con la normalidad ($p = 0.019$), mientras que los expertos mantienen una distribución aceptablemente normal ($p = 0.476$), lo que indica que la IA comienza a manifestar una variabilidad en su asignación de puntuaciones que no sigue un patrón convencional. En la fase final, tanto GPT como los expertos muestran distribuciones no normales ($p = 0.016$ y $= 0.009$, respectivamente), lo que refleja una mayor heterogeneidad en las evaluaciones en este punto del proceso. Esto sugiere que, a medida que avanza la evaluación formativa, los juicios de ambos evaluadores dejan de seguir una estructura uniforme, posiblemente debido a la complejidad creciente de los ajustes que realizan los expertos y al comportamiento adaptativo de GPT en la asignación de calificaciones. La prueba de Anderson-Darling refuerza esta interpretación, mostrando una mayor desviación de la normalidad en la fase final para ambos modelos (Expertos = 1.052, GPT = 0.747).

Los resultados sugieren que, si bien GPT y los expertos mantienen una correspondencia inicial en sus valoraciones, su evolución a lo largo del proceso evaluativo refleja diferencias en la forma en que ajustan sus calificaciones. Mientras que los expertos pueden estar aplicando criterios más flexibles y adaptativos en función del contenido y la retroalimentación formativa, GPT sigue un patrón de ajuste que, aunque dinámico, responde a una lógica distinta en la asignación de puntuaciones. Este comportamiento evidencia la necesidad de emplear enfoques estadísticos no paramétricos para analizar de manera más precisa la relación entre ambos evaluadores y determinar en qué medida la IA es capaz de replicar la complejidad del juicio humano en la evaluación formativa.

A través del coeficiente de correlación de Spearman (ρ) se determinó una relación monótona entre las evaluaciones emitidas por los expertos humanos y el modelo GPT. Los resultados indican una correlación positiva alta y estadísticamente significativa en la fase final de la evaluación ($\rho = 0.92$, $p < 0.001$), lo que sugiere que ambos modelos tienden a otorgar puntuaciones siguiendo un patrón similar. Esta alta correspondencia implica que, a medida que avanza el proceso de evaluación formativa, GPT se alinea con los criterios de los evaluadores expertos, reflejando una estabilidad en la tendencia de calificación. No obstante, esta relación debe ser interpretada considerando que la alta correlación no implica una completa equivalencia en la asignación de puntuaciones, sino una consistencia en la forma en que ambos modelos distribuyen sus evaluaciones.

Para examinar si los evaluadores humanos y el modelo GPT establecen un orden jerárquico equivalente en la calidad de los trabajos, se aplicó el coeficiente de concordancia de Kendall (τ). Los resultados muestran una concordancia alta y significativa ($\tau = 0.85$, $p < 0.001$), lo que indica que ambos evaluadores clasifican de manera similar los trabajos en términos de su calidad relativa. Esto sugiere que, aunque puedan existir diferencias en la magnitud de las calificaciones, la priorización relativa de los trabajos mejor y peor evaluados sigue un esquema consistente entre GPT y los expertos. Este hallazgo es relevante en el contexto de la evaluación formativa, pues evidencia que el modelo de IA es capaz de replicar la estructura de ordenamiento utilizada por los expertos, alineándose con la lógica subyacente a sus criterios de evaluación.

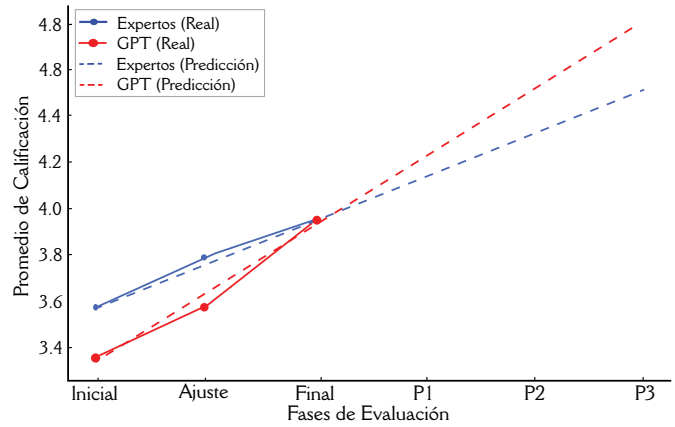
Se utilizó la prueba de Wilcoxon para muestras pareadas con el fin de evaluar si existen diferencias sistemáticas en las puntuaciones otorgadas por GPT y los expertos en cada una de las fases de evaluación. Los resultados revelan diferencias estadísticamente significativas en la fase inicial ($p = 0.019$) y en la fase de ajuste ($p = 0.006$), lo que indica que GPT y los expertos presentan discrepancias notables en las primeras etapas del proceso evaluativo. Sin embargo, en la fase final, la diferencia entre ambos no resulta significativa ($p = 0.367$), sugiriendo que, a medida que avanza la evaluación formativa, GPT ajusta su criterio y converge progresivamente con las calificaciones otorgadas por los expertos humanos. Este ajuste progresivo es indicativo de un patrón de alineación con los evaluadores humanos, lo que respalda la viabilidad del uso de modelos de IA en procesos de evaluación educativa, siempre que se implementen mecanismos que refuercen la calibración de sus criterios (Martínez-Comesaña et al., 2023).

La prueba de Wilcoxon permitió observar que las diferencias iniciales y en la fase de ajuste tienden a reducirse en la etapa final, lo que respalda la hipótesis de que GPT mejora su alineación con los expertos a medida que avanza el proceso evaluativo. Esto sugiere que, aunque el modelo presenta discrepancias significativas en sus primeras evaluaciones, su desempeño en la fase final tiende a ser más acorde con el juicio humano, lo que refuerza su potencial en la evaluación formativa, especialmente si se ajustan sus parámetros en función de criterios específicos previamente definidos. Este hallazgo refuerza la hipótesis de que, aunque GPT y los expertos pueden diferir en la asignación de puntuaciones en las primeras fases, el modelo de IA presenta una tendencia de alineación con los criterios humanos en la etapa final, lo que podría explicarse por la naturaleza programada de su ajuste y la coherencia de sus evaluaciones a lo largo del tiempo. No obstante, es necesario complementar este análisis con otras pruebas, como la prueba de Cochran, para determinar si esta convergencia sigue un patrón de ajuste similar al de los evaluadores humanos.

Para analizar la evolución del ajuste en las evaluaciones realizadas por los modelos GPT y los expertos a lo largo del tiempo, se transformaron los datos en una estructura binaria, diferenciando entre coincidencias y discrepancias en las calificaciones en las distintas fases de evaluación. Posteriormente, se aplicaron la prueba de Cochran y la prueba de McNemar, cuyos resultados revelaron diferencias estadísticamente significativas en la evolución de los juicios de ambos evaluadores ($\chi^2 = 7.8$, $p = 0.020$ en Cochran; $p = 0.039$ en McNemar). Estos hallazgos indican que, aunque GPT ajusta progresivamente sus calificaciones a lo largo del tiempo, su patrón de modificación no es completamente equivalente al de los expertos humanos, quienes pueden ajustar sus juicios en función de criterios adicionales y elementos cualitativos más complejos.

Si bien GPT modifica sus calificaciones en función del progreso de los trabajos, el análisis sugiere que los expertos humanos integran dimensiones evaluativas que van más allá de la simple consistencia en la aplicación de criterios. Mientras que el modelo de IA tiende a realizar ajustes de manera sistemática y predecible, los expertos pueden modificar sus calificaciones atendiendo a aspectos pedagógicos, interpretativos y contextuales que no son capturados por GPT. Esto explica la diferencia significativa en la evolución de las evaluaciones, ya que la IA no muestra un proceso de ajuste completamente alineado con el juicio humano, sino que sigue un esquema más rígido basado en patrones predefinidos. Estos resultados refuerzan la necesidad de una calibración más precisa de los modelos de IA en procesos de evaluación formativa (Delso Vicente, Carvajal Camperos y Corral De La Mata, 2024), asegurando que sus ajustes sean más alineados con las dinámicas de evaluación humana en educación superior. Para lograrlo, sería necesario integrar mecanismos de adaptación más avanzados en la IA, permitiendo que su evaluación no solo refleje cambios en la estructura del texto o en los elementos formales del trabajo, sino que también incorpore criterios pedagógicos y cognitivos más cercanos a los empleados por los expertos en la valoración de trabajos académicos.

Figura 3: Evolución predictiva extendida de GPT vs Expertos.



Para evaluar la evolución de las calificaciones otorgadas por el modelo GPT y los evaluadores humanos expertos a lo largo del tiempo, se implementó un modelo de regresión lineal simple. Se tomaron como variables independientes las fases de evaluación (Fase Inicial, Fase de Ajuste y Fase Final), y como variables dependientes los promedios de las calificaciones en cada fase. Se ajustaron modelos de regresión independientes para los expertos (R^2=0.89) y GPT (R^2=0.94), lo que sugiere un ajuste adecuado a los datos observados. Posteriormente, se utilizó la ecuación estimada de cada modelo para proyectar valores en tres fases adicionales, permitiendo visualizar la tendencia evolutiva de ambas evaluaciones (Figura 3). Este enfoque se basa en estudios previos que emplean regresión lineal para modelar la evolución de sistemas de evaluación automatizados en educación superior.

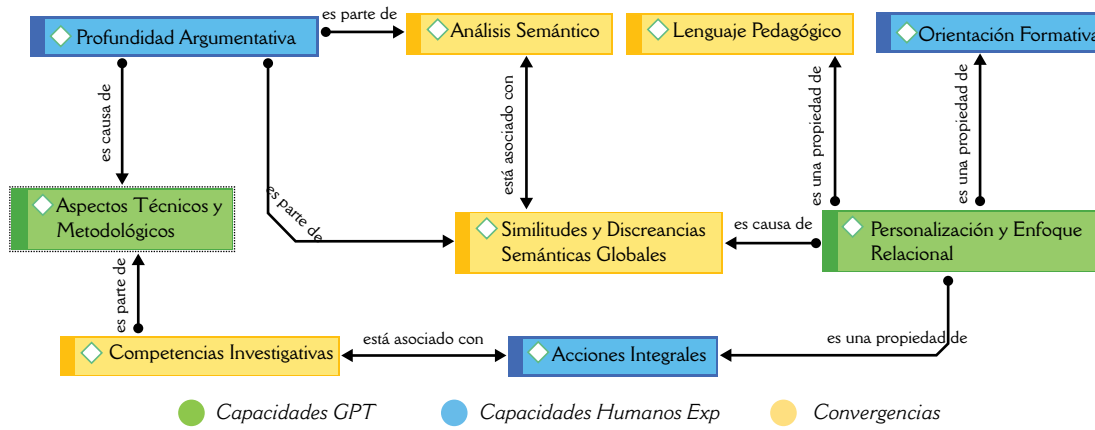
Los resultados obtenidos indican que, bajo un modelo de tendencia lineal, las calificaciones de GPT continuarían ajustándose progresivamente hacia las de los expertos, aunque con una progresión más rígida y sistemática. La predicción muestra que, en fases futuras, la diferencia entre las calificaciones de GPT y los expertos tendería a reducirse, aunque no desaparecería por completo. Este hallazgo sugiere que, aunque GPT puede aproximarse a la lógica evaluativa de los expertos, su proceso de ajuste sigue patrones algorítmicos más predecibles en comparación con la flexibilidad y la capacidad de razonamiento crítico de los evaluadores humanos. La necesidad de modelos más sofisticados que incorporen calibraciones dinámicas o aprendizaje contextualizado se plantea como una línea de investigación futura para mejorar la replicabilidad del juicio humano en la evaluación formativa de Trabajos de Fin de Grado.

3.2. Componente Cualitativo

El análisis semántico del corpus (Figura 4) revela que los evaluadores humanos desarrollan una

mayor Profundidad Argumentativa y Análisis Semántico basado en el modelo argumentativo de Toulmin, con juicios críticos que trascienden la mera descripción de hechos. Este hallazgo se obtuvo mediante un análisis de contenido inductivo con codificación abierta (Cáceres, 2008), segmentando y categorizando las unidades de análisis para identificar patrones de profundidad argumentativa y riqueza semántica. Las narrativas humanas conectan aspectos técnicos con Acciones Integrales de mejora, añadiendo matices pedagógicos que fortalecen la formación investigativa del estudiante. En contraste, la narrativa de GPT, aunque estructurada y coherente, se centró inicialmente en aspectos formales y técnicos, con un enfoque correctivo y menos crítico. Sin embargo, a lo largo de los tres ciclos de feedback se evidenció una evolución cualitativa significativa: GPT mejoró en la producción de retroalimentaciones con mayor riqueza semántica, incrementando la variedad léxica, la precisión argumentativa y la capacidad para formular recomendaciones más elaboradas y contextualizadas. Esta progresión también implicó una mejora en la incorporación de matices pedagógicos y en la orientación de las sugerencias hacia el desarrollo de competencias investigativas.

Figura 4: Capacidades evaluativas GPT vs Humanos Expertos. (análisis cualitativo del corpus).



El corpus también muestra que la Personalización y el Enfoque Relacional son distintivos en los evaluadores humanos, quienes adaptan el mensaje al perfil del estudiante y al contexto específico del TFG. Mediante el análisis temático reflexivo de Braun y Clarke (2006), se identificaron patrones de personalización y tono relacional en la narrativa humana. Aunque GPT comenzó con un feedback **más** homogéneo y estandarizado, en las fases posteriores mostró un avance relevante hacia una retroalimentación más adaptativa, ajustando mejor sus recomendaciones según las características del documento y las necesidades detectadas. Desde una perspectiva discursiva, la evolución cualitativa de GPT se manifestó en el mejor uso de conectores lógicos, estructuras narrativas más flexibles y un mayor equilibrio entre el enfoque técnico y la orientación pedagógica. Mientras que en el primer ciclo la retroalimentación de GPT era predominantemente formal y menos formativa, en los ciclos posteriores se observó una capacidad creciente para generar recomendaciones más específicas, formular preguntas abiertas y promover una reflexión más activa, aunque sin alcanzar la profundidad crítica de los evaluadores humanos.

En la comparativa, los evaluadores humanos sobresalen por el Lenguaje Pedagógico y la Orientación Formativa, incluyendo preguntas abiertas, sugerencias personalizadas y recomendaciones que promueven la reflexión crítica y el aprendizaje autónomo. Si bien GPT mejoró progresivamente en la generación de recomendaciones, aún no reproduce con igual eficacia los patrones discursivos de tipo didáctico observados en la narrativa humana. Esta evolución se identificó tras un análisis de contenido deductivo que permitió detectar un aumento en la complejidad y pertinencia de las sugerencias a lo largo de los tres momentos. Además, los expertos humanos sugieren Acciones Integrales que abordan tanto Aspectos Técnicos como el desarrollo de Competencias Investigativas y la mejora del proceso de aprendizaje (Velandia-Mesa et al., 2019). A través del análisis comparativo constante de Glaser y Strauss (1967), se contrastaron sistemáticamente las recomendaciones de mejora, observándose una evolución en GPT hacia propuestas más completas y

alineadas con las necesidades formativas del estudiante.

El análisis del corpus también detectó Similitudes y Discrepancias Semánticas Globales en la identificación de aspectos como calidad metodológica, coherencia textual y uso adecuado de instrumentos. Mediante un análisis semántico diferencial, se identificaron convergencias en aspectos formales y estructurales, pero también divergencias significativas en la capacidad de fomentar la metacognición y la autorreflexión crítica. Los expertos movilizan estrategias discursivas que favorecen la autonomía del estudiante, mientras que GPT, aunque mejoró en la profundidad de su feedback, sigue presentando limitaciones para desafiar al estudiante a cuestionar críticamente su producción académica. El análisis longitudinal del feedback evidencia una evolución cualitativa notable en los asistentes GPT, que mejoraron su capacidad para generar retroalimentación más rica, profunda y pedagógicamente orientada en cada ciclo de evaluación. No obstante, persisten diferencias sustanciales frente a la capacidad reflexiva, contextual y didáctica de la retroalimentación humana, lo que refuerza la importancia de integrar la IA como un complemento, y no como un sustituto, de la labor evaluativa experta en la educación superior.

4. Discusión y Conclusiones

Los hallazgos de esta investigación aportan evidencia sustantiva sobre el comportamiento evaluativo de los modelos GPT en comparación con evaluadores humanos expertos en la evaluación formativa de Trabajos de Fin de Grado (TFG) en Educación Superior. En línea con la pregunta de investigación planteada, los resultados sugieren que GPT, cuando opera bajo un diseño instruccional calibrado y una estructura de criterios normativos estandarizados, es capaz de aproximarse a la lógica evaluativa humana con niveles de consistencia y estabilidad notables.

El análisis estadístico confirma que los modelos GPT presentan una alta capacidad para reproducir la lógica algorítmica de la evaluación humana en entornos controlados y estructurados. La baja dispersión y la estabilidad de sus puntuaciones sugieren su potencial para ofrecer objetividad y reproducibilidad en contextos de alta demanda evaluativa, donde la consistencia resulta crítica. Sin embargo, la pérdida de normalidad y la menor flexibilidad observadas en la fase de ajuste reflejan limitaciones en la capacidad de GPT para adaptar sus juicios a factores contextuales y pedagógicos, una fortaleza inherente al criterio experto humano (García-Peñalvo y Vázquez-Ingelmo, 2023). A diferencia de los evaluadores humanos, que modulan sus valoraciones de acuerdo con aspectos situacionales y evolutivos, GPT mantiene una lógica de ajuste más lineal y previsible, característica de su programación algorítmica.

Desde la perspectiva cualitativa, el estudio evidencia una evolución notable en la capacidad de los modelos GPT para generar retroalimentación formativa. A lo largo de las tres fases evaluativas, GPT mejoró en riqueza semántica, diversidad léxica y precisión de sus recomendaciones, avanzando de un feedback inicial técnico hacia uno con mayor enfoque pedagógico y formativo. Estos avances, pese a las limitaciones detectadas, destacan su potencial para apoyar la formación investigativa en estudiantes universitarios, en línea con estudios que valoran la IA en procesos de retroalimentación automatizada (Guárdia-Ortiz et al., 2024; Paredes-Marín et al., 2024).

Los resultados permiten reconocer que, aunque los modelos GPT aún no igualan la profundidad crítica ni la sensibilidad didáctica del evaluador humano, presentan un potencial creciente para operar como sistemas complementarios en procesos de evaluación formativa. Su capacidad para ofrecer feedback inmediato, consistente y escalable los posiciona como aliados estratégicos en entornos educativos contemporáneos, donde la eficiencia y la cobertura masiva son cada vez más relevantes. La tendencia de mejora detectada sugiere que, con calibraciones avanzadas y con la integración de modelos de aprendizaje más sofisticados, la IA podría evolucionar hacia desempeños evaluativos cada vez más cercanos a la práctica experta.

Los resultados de este estudio permiten concluir, con base en los hallazgos empíricos, que los modelos GPT, bajo configuraciones cuidadosamente estructuradas y criterios normativos bien definidos, poseen la capacidad de replicar parcialmente el criterio evaluativo humano en procesos de evaluación formativa de TFG. Esta afirmación se sustenta en tres evidencias clave: (1) la consistencia observada en la aplicación de criterios, (2) la estabilidad estadística en las puntuaciones a lo largo de las tres fases evaluativas, y (3) la progresiva alineación con los patrones evaluativos de los expertos humanos, especialmente en la fase final del proceso.

Desde un enfoque holístico, los modelos GPT demuestran un potencial considerable para consolidarse como herramientas autónomas y complementarias en los procesos de evaluación formativa, aportando

eficiencia en la generación de retroalimentación y fortaleciendo la capacidad de los sistemas evaluativos para gestionar entornos educativos de gran escala. La evolución positiva observada en la calidad del feedback —más estructurado, argumentativo y orientado al aprendizaje— respalda su creciente utilidad en la formación investigativa universitaria. No obstante, las limitaciones identificadas, particularmente en la personalización y en la generación de retroalimentación crítica y metacognitiva, evidencian la necesidad de optimizar sus algoritmos de procesamiento discursivo. En consecuencia, este estudio sostiene que la integración de GPT debe concebirse como un recurso complementario, que fortalezca la labor del evaluador experto sin reemplazar su juicio pedagógico.

Los hallazgos de esta investigación abren diversas oportunidades para profundizar en la optimización del uso de modelos GPT en la evaluación formativa de Trabajos de Fin de Grado (TFG). En primer lugar, si bien el desempeño de GPT ha demostrado una progresión significativa en términos de consistencia evaluativa y generación de retroalimentación estructurada, su funcionamiento se encuentra influenciado por el diseño específico de los prompts y la alineación previa con rúbricas normativas. Esto plantea una interesante línea de trabajo para perfeccionar las estrategias de diseño instruccional que permitan maximizar la capacidad adaptativa y contextual de estos modelos en situaciones evaluativas diversas.

Una de las áreas de proyección más prometedoras radica en el desarrollo de mecanismos de calibración dinámica y de aprendizaje adaptativo para los sistemas GPT. Estos avances permitirían que la IA evolucione hacia una mayor sensibilidad frente a la dimensión pedagógica y metacognitiva de los procesos evaluativos, generando feedback que no solo responda a criterios normativos, sino que también promueva la reflexión crítica y el desarrollo de competencias investigativas en los estudiantes. Finalmente, el estudio abre la posibilidad de diseñar ecosistemas híbridos de evaluación donde los modelos GPT actúen en sinergia con el juicio experto humano. Esta integración estratégica podsría fortalecer la eficiencia y la calidad de la evaluación formativa, contribuyendo a la creación de sistemas educativos más personalizados, equitativos y sostenibles en contextos universitarios caracterizados por la alta demanda y la heterogeneidad de los estudiantes.

Apoyos

Universidad El Bosque (Colombia). Rectoría y Vicerrectoría de Investigaciones. Facultad de Educación. Grupo de Investigación: Educación e Investigación UNBOSQUE. Ministerio de Ciencia Tecnología e Innovación: Categoría A1.

Referencias

Boud, D. y Falchikov, N. (2006). Assessment & Evaluation in Higher Education. *Assessment & Evaluation in Higher Education*, *31*(4), 399-413. https://doi.org/10.1080/02602930600679050

Braun, V. y Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Cáceres, P. (2008). Análisis cualitativo de contenido: una alternativa metodológica alcanzable. Psicoperspectivas. *Individuo y sociedad*, *2*(1), 53-82. https://doi.org/10.5027/psicoperspectivas-Vol2-Issue1-fulltext-3

Chacón Pinilla, R. S. (2014). Del maestro como investigador: ¿reto y necesidad? *Itinerario Educativo*, *28*(64), 249-257. https://doi.org/10.21500/01212753.1430

Cortés Ibarra, E. F. y Martínez Clares, P. (2020). Adaptación y validación de un cuestionario para la caracterización de las prácticas evaluativas de los aprendizajes en educación superior PREVAPREDU. *Horizontes pedagógicos*, *22*(2), 25-36. https://doi.org/10.33881/0123-8264.hop.22202

Creswell, J. W. y Plano Clark, V. L. (2017). *Diseño y desarrollo de métodos mixtos de investigación*. SAGE Publications.

Delso Vicente, A. T., Carvajal Camperos, M. y Corral De La Mata, D. Á. (2024). La evolución del procesamiento del lenguaje natural y su influencia en la inteligencia artificial: Una revisión y líneas de investigación futura. *European Public & Social Innovation Review*, *10*, 1-23. https://doi.org/10.31637/epsir-2025-782

Flick, U. (2018). *An Introduction to Qualitative Research* (6ª ed.). SAGE Publications. https://doi.org/10.4135/9781529622737

Floridi, L. y Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

Fúquene Ardila, H. J. (2024). Procesamiento de Lenguaje Natural, los Transformers y los Bots Conversacionales. *XIKUA Boletín Científico de la Escuela Superior de Tlahuelilpan*, *12*(Especial), 151-160. https://doi.org/10.29057/xikua.v12iEspecial.12904

García-Peñalvo, F. y Vázquez-Ingelmo, A. (2023). What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(4), 7-16. https://doi.org/10.9781/ijimai.2023.07.006

Glaser, B. y Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine Press.

Guárdia-Ortiz, L., Bekerman, Z. y Zapata-Ros, M. (2024). Presentación del número especial "IA generativa, ChatGPT y Educación.

Consecuencias para el Aprendizaje Inteligente y la Evaluación Educativa". *Revista de Educación a Distancia (RED)*, *24*(78). https://doi.org/10.6018/red.609801

Kalla, D., Smith, N., Samaah, F. y Kuraku, S. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology*, *8*(3), 827-833. https://ssrn.com/abstract=4402499

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Martínez-Comesaña, M., Rigueira-Díaz, X., Larrañaga-Janeiro, A., Martínez-Torres, J., Ocarranza-Prado, I. y Kreibel, D. (2023). Impacto de la inteligencia artificial en los métodos de evaluación en la educación primaria y secundaria: revisión sistemática de la literatura. *Revista de Psicodidáctica*, *28*(2), 93-103. https://doi.org/10.1016/j.psicod.2023.06.001

Menard, S. W. (2002). *Longitudinal research* (2ª ed.). Sage Publications. https://doi.org/10.4135/9781412984867

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

Paredes-Marín, R. V., Ramírez-Chumbe, I. y Ramírez-Chumbe, C. A. (2024). La competencia digital y desempeño docente en instituciones educativas públicas: estudio bibliométrico en Scopus. *Revista Científica UISRAEL*, *11*(1), 31-48. https://doi.org/10.35290/rcui.v11n1.2023.1066

Velandia-Mesa, C., Serrano-Pastor, F. J. y Martínez-Segura, M. J. (2019). El desafío de la formación en competencias para la investigación educativa: aproximación conceptual. *Actualidades Investigativas en Educación*, *19*(3), 1-27. https://doi.org/10.15517/aie.v19i3.38738

Velandia-Mesa, C., Serrano-Pastor, F. J. y Martínez-Segura, M. J. (2021). Evaluación de la investigación formativa: Diseño y validación de escala. *Revista Electrónica Educare*, *25*(1), 1-20. https://doi.org/10.15359/ree.25-1.3

Yeadon, W., Peach, A. y Testrow, C. (2024). A comparison of human, GPT-3.5, and GPT-4 performance in a university-level coding course. *Scientific Reports*, *14*(1), 23285. https://doi.org/10.1038/s41598-024-73634-y