Universidad de Vigo

Revista de Investigación en Educación, 2026, 24(2), 309-326
DOI: https://doi.org/10.35869/reined.v24i2.6942
https://revistas.webs.uvigo.es/index.php/reined
ISSN 1697-5200 | e-ISSN 2172-3427

Evaluación académica con inteligencia artificial: comparación entre docentes y ChatGPT en Trabajos de Fin de Máster

Academic assessment with artificial intelligence: comparison between teachers and ChatGPT in Master's Theses

Sonia Eusebio-Hermira[1], Rocío Cuberos-Vicente[2], Anna Doquin de Saint-Preux[3]

[1] Universidad Autónoma de Madrid *soniaeusebio@uam.es*
[2] Universidad Complutense de Madrid *marcuber@ucm.es*
[3] Universidad Complutense de Madrid *adoquind@ucm.es*

Recibido: 31/10/2025
Aceptado: 23/4/2026

Dirección de contacto:
Rocío Cuberos Vicente
Dpto. Lengua Española y Teoría de la Literatura
Facultad de Filología, Edif. D, despacho 01.363
Universidad Complutense de Madrid
C/Profesor Aranguren s/n
28040 Madrid

Resumen

La expansión de la inteligencia artificial ha reconfigurado los modos de enseñar, aprender y evaluar, planteando nuevos retos. Este estudio examina la correspondencia entre las evaluaciones realizadas por docentes universitarios y las generadas por un sistema de inteligencia artificial (ChatGPT) en el contexto de Trabajos de Fin de Máster. Se adopta una metodología mixta, cuantitativa y cualitativa. A partir del análisis de 45 trabajos evaluados por ambos agentes utilizando una rúbrica común, se analizan las diferencias. Los resultados muestran una correlación global moderada y estadísticamente significativa entre ambas formas de evaluación, aunque se detectan discrepancias en los ítems relacionados con la estructura formal y la redacción académica, donde la IA tiende a ser más rigurosa. El análisis cualitativo revela diferencias sustanciales en el enfoque y el contenido de los comentarios: mientras que los docentes priorizan la profundidad conceptual y ofrecen retroalimentación contextualizada, la IA se enfoca en aspectos formales con observaciones más genéricas. Estos hallazgos sugieren que la IA puede ser una herramienta útil en procesos evaluativos, siempre que se complemente con el juicio pedagógico humano. Se discuten las implicaciones de estos resultados para el diseño de sistemas de evaluación mixtos y la formación docente en el uso de tecnologías emergentes.

Palabras clave

Evaluación Académica, Inteligencia Artificial, Rúbricas de Evaluación, Educación Superior, Trabajos de Fin de Máster

Abstract

The increasing use of artificial intelligence in education has transformed the way we teach, learn and assess, creating new challenges for academic assessment processes. This study examines the correspondence between academic assessments carried out by university teachers and those generated by an artificial intelligence system

(ChatGPT) in the context of Master's Theses. Based on an analysis of 45 master's theses evaluated by both agents using a common 9-item rubric, this study analyses the differences in the scores awarded. The results reveal a moderate and statistically significant overall correlation between the two forms of assessment. However, discrepancies are evident in items related to formal structure and academic writing, where AI tends to be more rigorous. Qualitative analysis reveals substantial differences in the focus and content of the comments: teachers prioritise conceptual depth and offer contextualised feedback, whereas AI focuses on formal aspects and makes more generic observations. These findings suggest that, when complemented by human pedagogical judgement, AI can be a useful tool to support assessment processes. The implications of these results for the design of mixed assessment systems and teacher training in critically using emerging technologies are discussed.

Key Words

Academic Assessment, Artificial Intelligence, Scoring Rubrics, Higher Education, Masters' Theses

1. INTRODUCCIÓN

En los últimos años, la inteligencia artificial (IA) ha empezado a abrirse paso en áreas que antes dependían exclusivamente del criterio humano, como la evaluación de trabajos académicos. La incorporación de sistemas de inteligencia artificial en este tipo de tareas podría traer ventajas importantes, como agilizar el proceso, minimizar sesgos personales y asegurar una aplicación más uniforme de los criterios de evaluación.

Sin embargo, esta tendencia también plantea interrogantes importantes sobre la fiabilidad y validez de dichas evaluaciones. No está claro hasta qué punto las calificaciones generadas por herramientas automatizadas pueden sustituir o complementar de manera efectiva las emitidas por evaluadores humanos por lo que resulta fundamental analizar la correspondencia entre ambas formas de evaluación e identificar posibles similitudes y discrepancias, así como sus consecuencias en el ámbito educativo. Además, la bibliografía consultada muestra que la relación entre ambas modalidades no ha sido aún explorada en el contexto de los Trabajos de Fin de Máster (TFM), que constituye el ámbito de nuestro trabajo.

A partir de este marco, en este estudio nos planteamos analizar la relación entre ambos tipos de evaluación en TFM. En concreto, nos planteamos determinar qué grado de concordancia existe entre las evaluaciones académicas realizadas por sistemas de inteligencia artificial y evaluadores humanos y qué factores explican sus discrepancias. Para ello, se han comparado las evaluaciones humanas (docentes) y las generadas por ChatGPT-4 de 45 TFM en el ámbito de la enseñanza de lenguas, procedentes de distintas titulaciones impartidas en varias universidades españolas: la Universidad Complutense de Madrid (UCM), la Universidad Autónoma de Madrid (UAM), la Universidad de Alcalá (UAH), la Universidad Nebrija y la Universidad de Castilla-La Mancha (UCLM). Tanto la evaluación realizada por los jueces humanos como la generada por la IA se basó en la misma rúbrica oficial que la Facultad de Educación de la UCM proporciona a sus estudiantes, lo que garantiza la coherencia de los criterios aplicados y permite una comparación más fiable entre ambos tipos de evaluación.

El artículo se estructura en cuatro apartados principales. En primer lugar, esta introducción, donde se contextualiza el problema y se presentan los objetivos del estudio. En el segundo apartado se analiza el estado de la cuestión. En el tercero, se indican los objetivos y las preguntas de investigación. El cuarto se dedica a indicar la metodología empleada. El quinto a exponer y analizar los resultados. Y el sexto a la discusión y conclusiones, con las limitaciones del estudio y las futuras líneas de investigación.

2. ESTADO DE LA CUESTIÓN

La evaluación de trabajos académicos, como es el caso de los TFM, ha sido realizada tradicionalmente por evaluadores humanos, quienes valoran no solo la corrección formal, sino también la calidad conceptual, la coherencia expositiva y argumentativa, la originalidad y la creatividad de los textos. Con la llegada de la IA, se ha empezado a explorar la posibilidad de automatizar al menos parte de este proceso, con el objetivo de ganar eficiencia, rapidez y mayor consistencia en la evaluación.

En la revisión de la literatura sobre el estado de la cuestión se han encontrado algunos estudios que pueden considerarse antecedentes del uso de la IA en la evaluación de textos académicos. Se trata de investigaciones centradas en la puntuación automatizada de ensayos cortos, exámenes o pruebas estandarizadas que representan los primeros intentos de aplicar algoritmos para calificar escritos (Bridgeman et al., 2014; Dikli, 2006; Perelman, 2020).

Los resultados de estos estudios muestran que, aunque los sistemas de puntuación automatizada pueden proporcionar evaluaciones rápidas consistentes y con alta fiabilidad en aspectos formales de la escritura, como la gramática, la ortografía y la estructura, presentan limitaciones importantes cuando se trata de valorar dimensiones cualitativas del texto (Dikli, 2006). Perelman (2020) demuestra que ciertos sistemas de puntuación automatizada pueden otorgar calificaciones altas a ensayos generados sin coherencia, lo que pone de manifiesto que estas herramientas no comprenden el contenido de los textos evaluados. Por su parte, Bridgeman et al. (2014) comparan la puntuación humana y automática en ensayos de exámenes estandarizados y encuentran que, si bien la correlación es de moderada a alta en aspectos formales, existen discrepancias sistemáticas en la evaluación de la coherencia argumentativa, el estilo y la creatividad y que estas diferencias pueden variar según el sexo, la etnia y el contexto educativo de los estudiantes.

Estas limitaciones evidencian que, aunque los sistemas automatizados de puntuación de ensayos y formatos similares constituyen un antecedente importante en el uso de la IA para la evaluación, su aplicación a trabajos académicos complejos, como los TFM, sigue siendo limitada. La evaluación de estos trabajos no solo implica revisar aspectos formales, sino también valorar la calidad conceptual, la argumentación crítica, la originalidad y la profundidad del análisis.

En este contexto, investigaciones recientes han explorado el uso de IA generativa (IAG), como ChatGPT, para complementar o realizar evaluaciones de trabajos académicos más complejos y compararlo con evaluadores humanos (Bouziane y Bouziane, 2024; García-Varela et al., 2025; Impey et al., 2025; Kincl et al., 2024; Kooly y Yusuf, 2025; López de Ramos et al., 2025; Lu et al. 2024; Roberts et al., 2023; Saborido-Fernández et al., 2024; Wetzler et al. 2025). Todos estos estudios coinciden en

que, aunque la IAG puede complementar la evaluación humana, no reemplaza completamente la valoración de aspectos cualitativos en trabajos académicos que requieren juicio crítico y un desarrollo más profundo del contenido.

Por ejemplo, López de Ramos et al. (2025) realizaron un análisis comparativo de la evaluación de resúmenes científicos, comparando la puntuación otorgada por evaluadores humanos con la generada por la IAG (ChatGPT-4 modelo gratuito). Los evaluadores humanos aplicaron una rúbrica detallada que consideraba varios criterios: la profundidad conceptual, evaluando la capacidad del resumen para sintetizar ideas complejas y extraer conclusiones fundamentadas; la calidad argumentativa, incluyendo la coherencia, la lógica y la solidez de los razonamientos; y el cumplimiento de la estructura académica, asegurando que la introducción, la metodología, los resultados y las conclusiones estuvieran presentados de manera clara y ordenada. Cada criterio se calificaba mediante puntuaciones numéricas, con el fin de asegurar la consistencia entre los distintos evaluadores. Esta rúbrica fue la que se le facilitó a la IAG. La IAG demostró consistencia y rapidez al evaluar aspectos formales, como la gramática, el vocabulario y la estructura del texto, pero presentó discrepancias significativas con los evaluadores humanos en la valoración de la profundidad conceptual y la calidad argumentativa, es decir, en aquellos criterios que requieren juicio crítico y comprensión del significado. Por ejemplo, podía otorgar puntuaciones altas a resúmenes correctamente redactados, pero conceptualmente superficiales, mientras que los humanos penalizaban la falta de análisis crítico.

De manera similar, García-Varela et al. (2025) evaluaron la capacidad de ChatGPT-4 (modelo gratuito) para proporcionar calificaciones coherentes en los ensayos (preguntas cualitativas abiertas) de los estudiantes en dos cursos de marketing de una escuela de ingeniería en Chile, con una muestra de 40 participantes. El estudio se desarrolló en cuatro etapas. En la primera no se utilizó rúbrica y se pedía a ChatGPT-4 evaluar una respuesta y luego identificar los criterios usados. En una segunda fase, se proporcionó la rúbrica original del docente. En la tercera, se diseñó un *prompt* enriquecido que instruía a la IA a usar únicamente los criterios de la rúbrica del docente y seguir un formato estandarizado de salida: criterio, puntuación, observación, justificación, nota final y comentarios. En la última, la etapa algorítmica, se usó un *prompt* más complejo, configurado con "*chain-of-thought prompting*" que pedía a ChatGPT-4 razonar paso a paso sus decisiones antes de dar la respuesta final. Los resultados mostraron que, con la rúbrica humana que se proporcionó a la IA, esta no ofrecía evaluaciones consistentes. Solo después de aplicar la rúbrica enriquecida con ejemplos concretos de desempeño, un formato estandarizado que guiara la presentación de la evaluación y descriptores que especificaran cómo asignar calificaciones según cada criterio, ChatGPT-4 pudo realizar evaluaciones más coherentes y justas. Esta mejora proporcionaba a la máquina indicaciones claras y detalladas sobre cómo interpretar los criterios para reducir la ambigüedad que presentaba la rúbrica original. Además, se ajustaron los parámetros del modelo para reducir la creatividad (lo que en términos de IA se conoce como "bajar la temperatura"), de manera que generara respuestas más predecibles y consistentes. Ahora bien, dado que el estudio se realizó con ensayos cortos, no se sabe cómo se comportaría la IA al evaluar trabajos más complejos, como TFM, tesis o artículos científicos, que requieren un análisis más profundo y criterios multidimensionales.

Por su parte, Saborido-Fernández et al. (2024) analizaron la capacidad de ChatGPT-3.5 para evaluar exámenes de programación en español en el grado en Ciencias de la Computación, comparando sus calificaciones con las de evaluadores humanos. Los

exámenes medían tanto las habilidades de los estudiantes para programar como sus conocimientos teóricos, con lo que tenían que redactar respuestas explicando conceptos relacionados con algoritmos, estructuras de datos, etc. Para guiar la evaluación, se utilizó una rúbrica humana que incluía criterios como la corrección sintáctica, la eficiencia y el funcionamiento de algoritmos, la originalidad en la resolución de problemas y la claridad en la explicación de la lógica utilizada. La IA fue entrenada para simular esta rúbrica. Para analizar cada muestra, se utilizaron dos tipos de *prompts*: un primer *prompt* simple, sin contexto y un segundo *prompt* complejo, que incluía el contexto, instrucciones del formato de salida y el rol que debía adoptar la IA. Los resultados mostraron que, aunque la máquina fue capaz de identificar errores sintácticos, reconocer patrones estructurales, evaluar la organización del código y ofrecer comentarios claros y precisos sobre su análisis, esta presentó limitaciones a la hora de valorar la originalidad y la complejidad de los escritos y en ocasiones otorgó puntuaciones altas a textos que eran funcionalmente correctos, pero conceptualmente simples, mientras los evaluadores humanos penalizaban este aspecto. Además, el *prompt* complejo no ayudó al modelo a mejorar la evaluación. En este aspecto, los autores incluso sugieren que la especificación del rol, la indicación de que se tomara tu tiempo para la respuesta y los ejemplos proporcionados pudieron introducir algo de confusión.

Finalmente, Roberts et al. (2023) realizaron un estudio comparativo sobre la evaluación de resúmenes científicos en inglés mediante ChatGPT-3 frente a expertos humanos. La rúbrica utilizada evaluaba la claridad y coherencia del texto, la precisión y relevancia del contenido, así como el cumplimiento de las normas de presentación, de manera que el trabajo resultara correctamente estructurado, documentado y presentado de manera clara y académica. Los resultados mostraron que la IA podía evaluar de manera muy precisa la claridad y coherencia de los textos, pero presentaba limitaciones para juzgar la precisión y la relevancia de la información, especialmente en contextos científicos complejos. Los evaluadores humanos, en cambio, mostraron una mayor capacidad para analizar correctamente los datos y valorar la importancia del contenido según el contexto de la disciplina.

En contraste con lo anterior, Impey et al. (2025) exploraron el uso de GPT-4 para evaluar la escritura científica en inglés en cursos masivos en línea (MOOCs) de astronomía, astrobiología e historia y filosofía de la astronomía, con una muestra de 120 respuestas de estudiantes a 12 preguntas (trabajos cortos de escritura). El diseño incluyó tres fuentes de calificación: la del tutor, la de pares (ambas con rúbricas predefinidas) y la de ChatGPT-4 en tres condiciones: 1. Con la respuesta modelo del instructor; sin rúbrica; 2. Con respuesta modelo y la rúbrica predefinida; 3. Con respuesta modelo y rúbrica generada por el propio GPT-4. Los resultados mostraron que ChatGPT-4 fue significativamente más fiable que la evaluación entre pares, y cuando se le proporcionó la rúbrica y la respuesta modelo, sus calificaciones no diferían de las del tutor. Esto sugiere que la IA puede calificar y generar criterios de evaluación útiles. Sin embargo, los autores advierten que el desempeño fue mejor en tareas más estructuradas (astronomía y astrobiología) que en aquellas de carácter abierto y especulativo (historia y filosofía), donde las discrepancias entre tutor e IA fueron mayores. Estos hallazgos indican que la IA puede superar la fiabilidad del sistema de evaluación entre pares (estudiantes) y acercarse al nivel de los docentes, siempre que cuente con guías claras; sin embargo, puede presentar limitaciones al evaluar tareas más subjetivas o creativas.

De manera complementaria, Kooly y Yusuf (2025) analizaron el uso de ChatGPT-3.5 en la corrección de exámenes universitarios en inglés en un curso de segundo año de ciencias sociales. El estudio se focalizó en una pregunta abierta del examen, que requería identificar y explicar dos tipos de diversificación de programas del gobierno canadiense vinculados a los Objetivos de Desarrollo Sostenible y justificar su utilidad con tres razones. La muestra estuvo compuesta por 25 respuestas de estudiantes, evaluadas con una rúbrica detallada por un corrector humano (doctorando con experiencia en el área) y por ChatGPT-3.5, ambos con idénticos criterios de corrección. Los resultados mostraron que las calificaciones de ChatGPT fueron en promedio más bajas y variables, aunque se observó una correlación moderada y significativa con las del corrector humano. Esto indica que, si bien ChatGPT fue algo más conservador y disperso en sus puntuaciones, logró una alineación estadísticamente válida con la evaluación humana. Los autores concluyen que la IA puede complementar la labor docente al aportar rapidez y retroalimentación inmediata, pero advierten que su efectividad podría reducirse al enfrentarse a tareas más complejas o que requieran un juicio más creativo y profundo.

En conjunto, estos estudios muestran que, aunque la IAG puede complementar la evaluación humana y ofrecer consistencia en aspectos formales y estructurales, presenta ciertas limitaciones para valorar criterios cualitativos, como la profundidad conceptual, la creatividad, la argumentación crítica y la relevancia contextual. Por su parte, las evaluaciones de los humanos resultan especialmente valiosas en cuestiones relacionadas con el contenido de los textos académicos, sin embargo, no están exenta de sesgos e inconsistencias (Sokolov, 2014). En contraste, los sistemas de IAG destacan por un análisis rápido y uniforme (Atasoy y Moslemi, 2025). La mayoría de las investigaciones se han centrado en ensayos cortos, resúmenes científicos o exámenes específicos, con muestras relativamente pequeñas y contextos limitados, pero todavía no se sabe cómo respondería al evaluar trabajos académicos más complejos, como TFM, tesis o artículos científicos. Una importante excepción es el trabajo de Kincl et al. (2024) quienes compararon la evaluación humana y la realizada por dos sistemas de IA (ChatGPT-4, versión gratuita, y Claude) en ensayos escritos en checo por estudiantes de máster que incluían revisión crítica de literatura, definición y análisis de constructos teóricos, y reflexión sobre su validez práctica en el ámbito del marketing con una extensión de 7 a 10 páginas. Sus resultados, en línea con los anteriores, mostraron correlaciones moderadas entre ambos tipos de evaluación. En términos cualitativos, ChatGPT ofreció comentarios generalmente positivos, pero poco específicos, con escasas orientaciones para la mejora, mientras que Claude y los evaluadores humanos proporcionaron observaciones más críticas y focalizadas en las debilidades de cada texto. En el caso de los humanos, además, detectaron plagio. Los autores subrayan, asimismo, que el uso del checo, una lengua con escasa representación en los corpus de entrenamiento de los LLM (*Large Language Model*) constituye un desafío adicional para la fiabilidad de la evaluación automatizada. En nuestro caso, el análisis se realizará en español, lengua que, si bien no alcanza el grado de representación del inglés, cuenta con una presencia significativamente mayor que el checo, lo que previsiblemente podría reducir parte de estas limitaciones. En este sentido, la presente investigación busca ampliar y profundizar estos hallazgos, comparando la evaluación humana y automatizada en trabajos académicos de mayor complejidad, identificando las discrepancias entre ambos enfoques y los posibles factores que las explican. En concreto, este estudio se propone analizar las puntuaciones otorgadas a los TFM y también profundizar en las posibles diferencias

mediante el análisis de las retroalimentaciones cualitativas aportadas por los evaluadores humanos y la IAG. De esta manera, se pretende aportar evidencia sobre la fiabilidad, consistencia y posibles limitaciones de la IA en la corrección de trabajos complejos con el objetivo de proporcionar información que ayude a integrarla de manera adecuada en la Educación Superior.

3. OBJETIVO Y PREGUNTAS DE INVESTIGACIÓN

El objetivo general de la investigación es examinar la relación entre las evaluaciones realizadas por humanos y por IA de TFM, con el fin de examinar el grado de concordancia que existe entre las evaluaciones académicas realizadas por sistemas de inteligencia artificial y evaluadores humanos, así como los factores que explican sus discrepancias. Como resultado, se espera que los hallazgos de este estudio sirvan como base para proponer descriptores más específicos, en los casos donde hay más diferencias, de manera que mejoren la rúbrica general para que sirvan de apoyo al tribunal evaluador en los procesos de automatización con ayuda de la IA.

Con base en lo anterior nos planteamos las siguientes preguntas de investigación:

PI1: ¿En qué medida se da una correlación entre las evaluaciones de trabajos académicos realizadas por evaluadores humanos y un sistema de IA?

PI2: ¿Se observan diferencias significativas en las puntuaciones otorgadas por humanos e IA?

PI3: ¿Qué patrones cualitativos emergen en los casos donde hay discrepancias entre evaluadores humanos e IA?

4. MÉTODO

Este estudio adopta un diseño comparativo con el objetivo de analizar las coincidencias y divergencias entre evaluaciones académicas realizadas por docentes universitarios y por un sistema de inteligencia artificial (ChatGPT). La investigación se centró en una muestra de 45 TFM correspondientes a programas de máster en el ámbito de la Educación, abarcando distintas especialidades: Lengua Castellana, enseñanza de lenguas extranjeras (inglés, francés, español) y Atención Temprana. Los trabajos proceden de cinco universidades de la Comunidad de Madrid, tanto públicas como privadas: la Universidad Complutense de Madrid (UCM), la Universidad Autónoma de Madrid (UAM), la Universidad de Alcalá (UAH), la Universidad Nebrija y la Universidad de Castilla-La Mancha (UCLM). De acuerdo con el protocolo establecido por la UCM, los participantes otorgaron su consentimiento informado por escrito, en el cual se especificaba el uso de sus TFM tanto para la evaluación por parte de docentes como para su análisis automatizado mediante inteligencia artificial. Además, los TFM fueron debidamente anonimizados, garantizando la eliminación de cualquier información que pudiera permitir la identificación de los autores, como nombres, referencias institucionales o cualquier dato sensible relacionado con el contexto educativo de los mismos.

Los evaluadores humanos estuvieron conformados por un panel de 8 docentes universitarias con experiencia en la dirección y corrección de TFM. La evaluación automatizada fue realizada por ChatGPT-4 (versión de pago), configurado previamente

para aplicar criterios académicos estandarizados. En el Anexo 1, se puede consultar el *prompt* facilitado al programa.

Ambos tipos de evaluadores aplicaron una rúbrica común utilizada institucionalmente en la evaluación de TFM en la Facultad de Educación de la UCM. Esta rúbrica incluye nueve ítems cuantitativos, cada uno valorado en una escala de 1 a 10. En la Tabla 1 se incluyen los descriptores de cada criterio de evaluación.

Número de ítem	Descriptor
Ítem 1	Pertinencia y relevancia del tema o de la propuesta de intervención educativa.
Ítem 2	Precisión y fundamentación bibliográfica del marco teórico.
Ítem 3	Claridad en la formulación de objetivos, hipótesis o descripción del problema.
Ítem 4	Adecuación metodológica en función del problema planteado.
Ítem 5	Justificación y argumentación de las valoraciones y juicios emitidos.
Ítem 6	Claridad y desarrollo de las conclusiones derivadas de los resultados.
Ítem 7	Pertinencia y actualización de las referencias bibliográficas.
Ítem 8	Estructura y presentación formal del documento (formato, apartados, bibliografía, etc.).
Ítem 9	Precisión terminológica, claridad expositiva y corrección lingüística.

Tabla 1. Descripción de los criterios de evaluación de la plantilla

Además, cada evaluación incluyó un comentario cualitativo que identificaba los puntos fuertes del trabajo y los aspectos susceptibles de mejora.

Se recopilaron datos cuantitativos correspondientes a las calificaciones asignadas por ambos tipos de evaluadores, junto con información contextual relevante (rama disciplinar, universidad de origen, criterios aplicados). El análisis estadístico se llevó a cabo mediante el software SPSS (versión 26). Se emplearon técnicas de estadística descriptiva para caracterizar las evaluaciones, y se realizaron análisis inferenciales no paramétricos, en particular la correlación bivariada de *Spearman* para examinar el grado de asociación entre las puntuaciones humanas y automatizadas y la prueba *U de Mann-Whitney* para la comparación de puntuaciones. Adicionalmente, se llevó a cabo un análisis cualitativo comparativo de los comentarios proporcionados por docentes y por la IA en aquellos casos en los que se identificaron diferencias significativas en las puntuaciones, con el fin de identificar patrones de discrepancia en los enfoques evaluativos y en los tipos de retroalimentación ofrecida.

5. RESULTADOS

5.1. Descripción y comparación de las evaluaciones entre IA y humanos

En el análisis descriptivo global de las evaluaciones realizadas por docentes humanos y por el sistema de IA (ChatGPT), se observa que ambos grupos utilizaron de forma consistente la rúbrica común de 9 ítems, valorados en una escala de 1 a 10. El promedio general de las calificaciones otorgadas por los evaluadores humanos (M=8) tiende a ser ligeramente superior al otorgado por la IA (M=7,6). Esta última presenta además unas puntuaciones más homogéneas como se puede observar en el Gráfico 1.

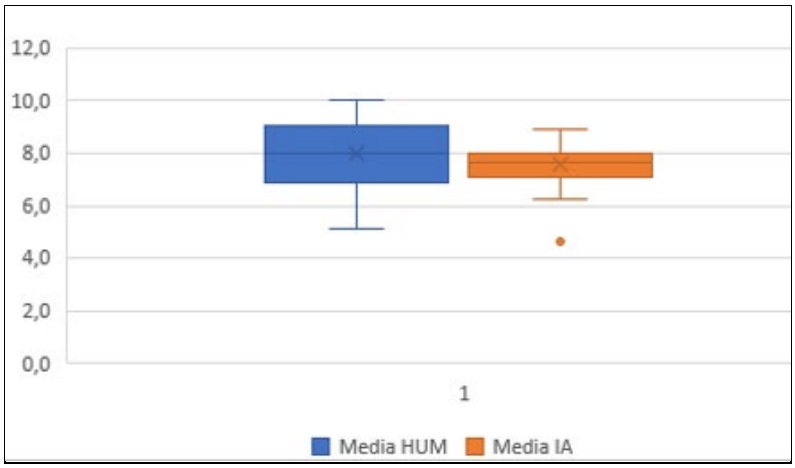

Gráfico 1. Puntuaciones medias de los docentes (Media HUM) y de la IA (Media IA)

El gráfico de barras agrupadas (Gráfico 2) representa las medias obtenidas por cada ítem de la rúbrica de evaluación, diferenciando las puntuaciones asignadas por evaluadores humanos (color azul) y por el sistema de inteligencia artificial (color rojo). En total, se visualizan las medias de los nueve ítems aplicados a los TFM, etiquetados de Ítem1 a Ítem9, que indica que corresponden a la muestra humana. En términos generales, se aprecia que los evaluadores humanos tienden a otorgar puntuaciones ligeramente superiores a las asignadas por la IA en la mayoría de los ítems. Esta diferencia se hace especialmente evidente en los ítems 8 y 9, donde la discrepancia entre ambas medias es más pronunciada visualmente. En estos dos criterios –relacionados con la estructura formal del documento y la calidad de la redacción académica– las barras correspondientes a la IA se encuentran notablemente por debajo de las humanas. En cambio, hay ítems como el 3, que se refiere a la claridad en la formulación de objetivos, hipótesis o problemas de investigación, en los que la IA otorga, en promedio, una puntuación ligeramente superior a la de los evaluadores humanos. Esta excepción sugiere un patrón divergente en la forma de aplicar los criterios evaluativos, aunque sin adelantar interpretaciones estadísticas. En los ítems 1, 2, 4, 5, 6 y 7, las diferencias entre evaluadores humanos e IA son menos marcadas, observándose una cierta proximidad entre ambas puntuaciones medias, lo que sugiere una cierta convergencia en la interpretación de estos aspectos del TFM. Sin embargo, incluso en estos casos, los evaluadores humanos tienden a ubicarse sistemáticamente por encima de la IA, aunque con márgenes más reducidos.

Este patrón gráfico permite observar de forma preliminar tanto los puntos de coincidencia como de divergencia en la aplicación de la rúbrica, configurando una base visual clara para los análisis inferenciales que se presentan en los apartados siguientes.

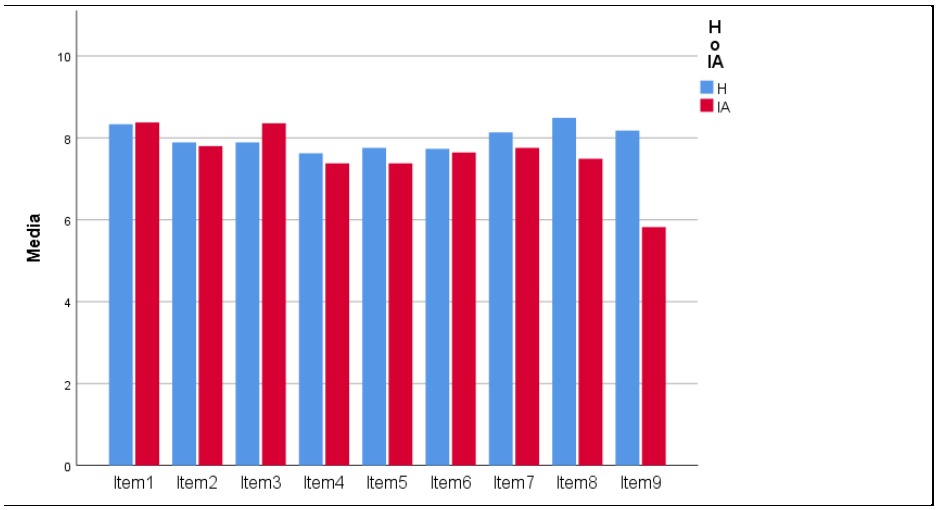

Gráfico 2. Comparación de las medias obtenidas por cada ítem de la rúbrica de evaluación

5.2. Pregunta de investigación 1. ¿Existe una correlación significativa entre las evaluaciones de trabajos académicos realizadas por evaluadores humanos y un sistema de IA?

Para responder a esta pregunta, se aplicó un análisis de correlación bivariada de Spearman entre las puntuaciones otorgadas por evaluadores humanos y las generadas por el sistema de inteligencia artificial (ChatGPT). Este análisis se realizó tanto a nivel global (considerando la nota media de cada trabajo) como por ítems individuales de la rúbrica.

Los resultados muestran una correlación positiva moderada y estadísticamente significativa entre las puntuaciones globales de ambos tipos de evaluación (r = 0,438; p = 0,003). Este valor indica que, en términos generales, existe una tendencia compartida entre humanos e IA al valorar los trabajos, aunque no necesariamente coincidan en todos los casos. El coeficiente de correlación sugiere un grado medio de asociación: es decir, a medida que aumenta la puntuación asignada por un evaluador humano, también tiende a aumentar la calificación otorgada por la IA, aunque con cierta variabilidad.

Cuando se desglosa el análisis por ítems, se observa que la mayoría de los criterios presentan también correlaciones positivas y estadísticamente significativas, con coeficientes que oscilan entre 0,4 y 0,6. Estos resultados confirman que tanto la IA como los evaluadores humanos tienden a coincidir en la valoración de diversos aspectos del TFM, particularmente en aquellos relacionados con el contenido disciplinar, la metodología y la argumentación.

No obstante, hay tres ítems en los que no se ha identificado una correlación significativa, lo que indica una mayor divergencia en los criterios de evaluación aplicados:

- Ítem 1: Justificación de la pertinencia y relevancia del tema o intervención.
- Ítem 6: Claridad y desarrollo de las conclusiones.
- Ítem 8: Estructura y presentación formal del documento.

La ausencia de correlación significativa en estos ítems sugiere que humanos e IA interpretan de forma distinta algunos aspectos del trabajo académico. En el caso del ítem

1, podría deberse a que la valoración de la "relevancia" implica un juicio contextual y disciplinar que la IA no siempre es capaz de captar adecuadamente. En el ítem 6, relacionado con las conclusiones, la diferencia puede reflejar las limitaciones del modelo para evaluar procesos inferenciales y niveles de síntesis. En cuanto al ítem 8, la falta de correlación podría estar asociada a un enfoque más normativo y rígido por parte de la IA al aplicar criterios de formato y presentación, frente a una mayor tolerancia o flexibilidad de los evaluadores humanos.

En conjunto, estos resultados muestran que, si bien existe un grado moderado de concordancia global entre humanos e IA, persisten discrepancias relevantes en la forma en que ciertos criterios son interpretados y aplicados.

5.3. Pregunta de investigación 2. ¿Se observan diferencias significativas en las puntuaciones otorgadas por humanos e IA?

Para analizar esta cuestión, se aplicó la prueba no paramétrica de U de *Mann-Whitney*, que permite contrastar si existen diferencias estadísticamente significativas entre las puntuaciones asignadas por dos grupos independientes –en este caso, evaluadores humanos e IA– en cada uno de los ítems de la rúbrica. La Tabla 2 recoge los resultados de esta prueba estadística. Tal y como queda recogido dicha tabla, los resultados indican que existen diferencias significativas en dos de los nueve ítems:

- Ítem 8: "El documento está bien estructurado y tiene una adecuada presentación (apartados, formato, bibliografía, etc.)".
- Ítem 9: "El documento está redactado con terminología precisa y con una organización clara y sistemática de las ideas, sin errores gramaticales u ortográficos".

Número de ítem	Significación
Ítem 1	,902
Ítem 2	,614
Ítem 3	,349
Ítem 4	,170
Ítem 5	,196
Ítem 6	,879
Ítem 7	,093
Ítem 8	,001
Ítem 9	<,001
Ítem 10	,156

Tabla 2. Resumen de contraste de hipótesis (U de Mann-Whitney)

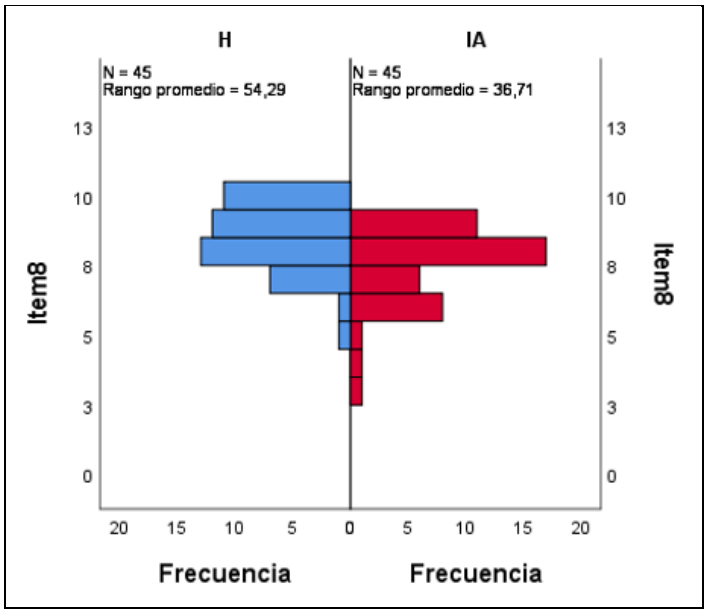

Gráfico 3. Comparación puntuaciones obtenidas para el ítem 8 (Prueba de U de Mann-Whitney)

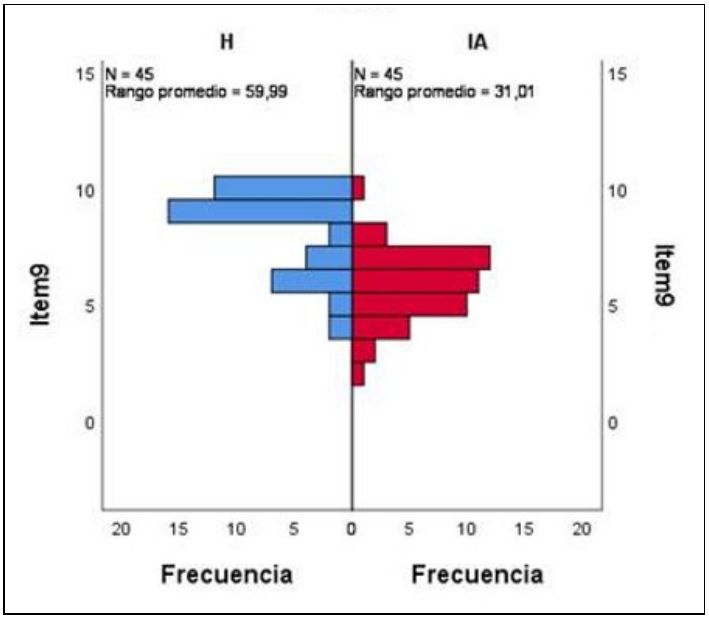

Gráfico 4. Comparación puntuaciones obtenidas para el ítem 9 (Prueba de U de Mann-Whitney)

Como se puede ver en el Gráfico 3 y en el Gráfico 4, en ambos ítems, la IA otorgó puntuaciones significativamente más bajas que los evaluadores humanos. Esta diferencia no solo es estadísticamente significativa, sino que también es visualmente clara en el gráfico de barras agrupadas presentado anteriormente (Gráfico 2), donde se aprecia una caída notable en las medias correspondientes a la IA para estos dos criterios.

Este hallazgo sugiere que la IA aplica estándares más estrictos o inflexibles al valorar aspectos formales y lingüísticos del documento, posiblemente debido a su sensibilidad programada frente a errores gramaticales, faltas de ortografía o desviaciones respecto a normas estructurales. Es decir, mientras que los evaluadores humanos pueden adoptar una

interpretación más holística o contextual de la calidad formal del texto, la IA tiende a penalizar más sistemáticamente cualquier irregularidad detectada.

En los restantes siete ítems, las diferencias entre IA y humanos no alcanzaron significación estadística. Esto no implica necesariamente que las evaluaciones sean idénticas, pero sí que las diferencias observadas pueden atribuirse al azar o a variaciones no sistemáticas dentro del rango esperable de juicios evaluativos.

Este patrón de resultados pone de relieve que, aunque la IA tiende a alinearse con los evaluadores humanos en la mayoría de los aspectos considerados, su comportamiento difiere de forma consistente cuando se trata de aplicar criterios formales, lo que debe tenerse en cuenta en escenarios de evaluación académica compartida o complementaria.

5.4. Pregunta de investigación 3. ¿Qué patrones cualitativos emergen en los casos donde hay discrepancias entre evaluadores humanos e IA?

El análisis cualitativo de los informes redactados por evaluadores humanos e inteligencia artificial (ChatGPT) revela diferencias notables en el estilo, el enfoque y el nivel de profundidad de las valoraciones. Estas diferencias ayudan a explicar varias de las discrepancias detectadas en las puntuaciones cuantitativas.

De manera general, los comentarios de la IA tienden a mantener un tono formal, positivo y diplomático, destacando los aspectos logrados del trabajo, incluso en aquellos casos donde la puntuación otorgada es baja. Las expresiones suelen ser genéricas (“el trabajo está bien estructurado”, “el tema es interesante”, “se aprecia un esfuerzo por organizar las ideas”), lo cual genera valoraciones que, si bien correctas en lo formal, carecen de matices críticos o profundidad analítica.

En contraste, los evaluadores humanos muestran una mayor variedad de registros discursivos, con un lenguaje más directo, matizado e incluso evaluativo en sentido negativo. No es infrecuente encontrar frases como “faltan conexiones claras entre los objetivos y la metodología”, “las conclusiones no aportan nada nuevo”, o “el uso de fuentes es muy pobre”, lo que denota una mayor disposición a señalar debilidades concretas y emitir juicios críticos.

Una de las principales diferencias radica en la capacidad de concreción. Los comentarios humanos suelen incluir referencias precisas a apartados del trabajo, ejemplos textuales o sugerencias para la mejora (“sería recomendable definir claramente los conceptos clave en la introducción”, “las citas del marco teórico están desactualizadas”). Estas observaciones están generalmente contextualizadas y conectadas con la experiencia disciplinar del evaluador.

En cambio, los comentarios de la IA, aunque coherentes y bien redactados, carecen de referencias específicas al contenido del TFM. Las sugerencias tienden a ser genéricas, como “sería beneficioso profundizar en el análisis de los resultados” o “podría mejorar la claridad argumentativa”, sin vinculación directa con aspectos concretos del texto. Esto limita su utilidad como retroalimentación formativa.

Otro patrón detectado es la asimetría en la focalización de los comentarios. La IA otorga mayor peso a criterios formales –estructura, redacción, ortografía, formato APA– y tiende a ignorar o tratar superficialmente los aspectos más sustantivos del contenido, como la originalidad, el marco teórico o la adecuación metodológica. Esto se corresponde con las diferencias observadas en los ítems 8 y 9 del análisis cuantitativo, donde la IA se muestra más exigente.

Por el contrario, los evaluadores humanos enfatizan con más frecuencia la coherencia interna del trabajo, la consistencia argumentativa y la adecuación metodológica, incluso cuando existen deficiencias formales. Esto sugiere que, en caso de conflicto entre fondo y forma, los humanos tienden a priorizar la dimensión conceptual del TFM.

En los casos donde se observaron discrepancias de más de un punto entre las calificaciones de IA y humanos, el análisis de los comentarios muestra que la IA otorga bajas puntuaciones por fallos formales, aunque el contenido sea adecuado, mientras que los humanos tienden a compensar errores formales si el trabajo presenta una aportación sólida. Esta diferencia de enfoque podría explicar parte de la variabilidad en los ítems con menor correlación (1, 6 y 8) y en aquellos con diferencias estadísticamente significativas (8 y 9).

Uno de los patrones más recurrentes es el siguiente: en trabajos donde el contenido es sólido desde el punto de vista conceptual –por ejemplo, con una adecuada selección del tema, una fundamentación bibliográfica pertinente y un planteamiento metodológico razonable–, los evaluadores humanos tienden a valorar positivamente estos aspectos, incluso si el documento presenta deficiencias formales (errores de redacción, estructura imperfecta o presentación descuidada). En estos casos, los comentarios humanos suelen destacar frases como: “El marco teórico está bien trabajado, aunque convendría revisar el estilo de redacción” o “Se aprecia un buen nivel de reflexión y profundidad en el análisis, pese a ciertas limitaciones formales”.

En cambio, la IA en estos mismos trabajos asigna puntuaciones más bajas debido a su énfasis en los elementos formales. Sus comentarios típicos incluyen observaciones del tipo: “El texto presenta problemas de redacción y uso impreciso del lenguaje académico” o “La estructura del documento podría mejorarse para una mejor comprensión del contenido”.

Este sesgo a favor de los aspectos formales por parte de la IA se refleja claramente en los ítems 8 y 9, donde tiende a penalizar con mayor severidad cuestiones como errores gramaticales, problemas en el formato o una organización textual no completamente alineada con los estándares académicos. En contraste, los docentes humanos parecen adoptar una visión más integral del trabajo, equilibrando los criterios formales con los sustantivos.

En otros casos, aunque menos frecuentes, se observa la situación inversa: la IA otorga una calificación relativamente alta debido a la buena presentación del texto, mientras que el docente humano asigna una nota baja justificando su decisión con observaciones como: “El trabajo está bien redactado, pero carece de profundidad en el análisis” o “No hay una reflexión crítica ni una relación clara entre objetivos y resultados”.

En estos escenarios, la IA parece dejarse llevar por la forma superficial del texto (fluidez, estructura, ortografía), sin detectar las carencias más profundas relacionadas con la solidez del contenido, la coherencia argumentativa o la conexión entre secciones clave del TFM. Este patrón podría deberse a la limitación del modelo para captar aspectos de tipo inferencial, crítico o epistémico, que exigen un juicio contextual y disciplinar.

En resumen, las discrepancias significativas entre IA y humanos en los casos analizados responden a una divergencia en los criterios de ponderación: la IA muestra una inclinación técnica y normativa, mientras que el juicio humano se basa en una comprensión más situada del texto académico, sus propósitos, limitaciones y aportaciones. Además, los docentes muestran una mayor disposición a explicar y justificar sus valoraciones con ejemplos concretos o sugerencias de mejora, lo cual

refuerza su rol pedagógico y orientador, algo que la IA no logra replicar plenamente, al menos con los *prompts* empleados en este estudio.

Los patrones cualitativos observados permiten afirmar que las discrepancias entre IA y humanos no son aleatorias, sino que responden a diferencias sistemáticas en los enfoques evaluativos. La IA muestra mayor consistencia y severidad en aspectos formales, mientras que los docentes priorizan el contenido conceptual y son más críticos en términos argumentativos. La IA tiende a ofrecer retroalimentaciones genéricas, menos útiles para la mejora del estudiante, frente a comentarios humanos más específicos y pedagógicos.

Este hallazgo refuerza la necesidad de considerar ambos tipos de evaluación como complementarios, más que equivalentes o sustitutos, y sugiere la conveniencia de enriquecer los sistemas automatizados con guías contextuales y criterios cualitativos más elaborados.

6. DISCUSIÓN Y CONCLUSIONES

Los resultados de esta investigación confirman que, si bien existe una correlación moderada y significativa entre las evaluaciones realizadas por docentes humanos y por el sistema de inteligencia artificial (ChatGPT), persisten discrepancias relevantes en la aplicación de ciertos criterios, especialmente aquellos vinculados a aspectos formales (estructura, redacción, presentación) y a dimensiones más interpretativas (relevancia del tema, formulación de conclusiones).

Este hallazgo se alinea con estudios previos que han señalado las limitaciones de la IA para captar aspectos cualitativos complejos, como la argumentación crítica, la originalidad o la pertinencia contextual. Por ejemplo, Dikli (2006) y Perelman (2020) advertían que los sistemas automatizados tienden a favorecer la corrección superficial de los textos, otorgando puntuaciones elevadas incluso a contenidos incoherentes si estos están bien redactados. En contraste, en nuestro estudio los docentes penalizaron con claridad la falta de profundidad analítica, incluso en trabajos formalmente correctos.

Asimismo, la correlación observada entre IA y evaluadores humanos coincide con los resultados de López de Ramos et al. (2025) y Kincl et al. (2024), quienes identificaron niveles de acuerdo moderado al evaluar tareas escritas académicas complejas, especialmente cuando se usaban rúbricas estandarizadas. En nuestro caso, el uso de una rúbrica común pareció favorecer la convergencia parcial en las puntuaciones, aunque no logró eliminar completamente las diferencias de enfoque entre evaluadores.

En cuanto a los criterios formales, la IA resultó más estricta que los evaluadores humanos, en tanto que penalizó con más severidad deficiencias en la redacción y estructura. Este resultado coincide con los hallazgos encontrados en los trabajos de Kooly y Yusuf (2025) y Saborido-Fernández et al. (2024), quienes señalaron que ChatGPT tiende a aplicar criterios de corrección formal de forma más estricta que los evaluadores humanos. En línea con ello, en nuestro estudio la IA asignó puntuaciones significativamente más bajas en los ítems 8 y 9, penalizando deficiencias formales con mayor rigurosidad que los docentes, quienes en algunos casos las relativizaron en función de la calidad conceptual del trabajo.

Por último, la observación de que los comentarios de la IA resultan más genéricos y menos útiles como retroalimentación coincide con lo reportado por García-Varela et al.

(2025) y Kincl et al. (2024), quienes destacaron la necesidad de enriquecer los *prompts* y las guías proporcionadas a los modelos de IA para mejorar la especificidad y relevancia de sus observaciones.

En conjunto, estos resultados refuerzan la idea de que la evaluación automatizada puede complementar, pero no reemplazar, el juicio humano en la valoración de trabajos académicos complejos. Las coincidencias observadas en ciertos criterios sugieren que la IA puede ser útil como herramienta de apoyo, siempre que se reconozcan sus limitaciones y se integre de forma crítica dentro del proceso evaluativo.

Este estudio aporta evidencias relevantes sobre la correspondencia entre las evaluaciones de docentes y de la IA en la valoración de trabajos académicos, sin embargo, presenta también ciertas limitaciones que condicionan el alcance de los resultados.

En primer lugar, en el estudio cada TFM fue evaluado por un solo evaluador y una sola ronda de evaluación por parte de la IA. Esta decisión metodológica permitió mantener la coherencia del procedimiento, pero limita la generalización de los hallazgos, ya que no se exploran posibles variaciones entre diferentes evaluadores humanos ni entre distintas respuestas del sistema de IA. Siguiendo los argumentos planteados por Kroli y Yusuf (2025), se optó por no incluir múltiples evaluadores ni rondas adicionales de IA porque todo ello podría introducir sesgos adicionales, errores humanos o inconsistencias. Aun así, futuros trabajos deberían considerar la triangulación de evaluaciones con más de un docente y con diversas iteraciones de la IA para analizar la estabilidad y fiabilidad de los resultados.

Otra limitación del trabajo se relaciona con el idioma. La evaluación se realizó en español y la representatividad de este idioma en los corpus utilizados para entrenar a la IA podría ser menor que la del inglés, lo que puede haber afectado parcialmente la capacidad del modelo para evaluar los trabajos con la misma precisión. Esta circunstancia sugiere la necesidad de realizar investigaciones comparativas en diferentes lenguas para examinar si la eficacia de la IA en la evaluación académica varía en función del idioma.

Por último, se decidió utilizar la misma rúbrica para evaluar todos los TFM, independientemente de la disciplina o el tipo de trabajo. Esta decisión garantiza la consistencia en la comparación entre la IA y el evaluador humano, pero ha supuesto simplificar las particularidades de cada área de conocimiento. Tal como señala Hyland (2013), cada disciplina genera TFM con géneros, objetivos y expectativas propios, lo que exige criterios de evaluación ajustados a su especificidad. Por ello, sería relevante diseñar rúbricas específicas para cada disciplina y tipo de TFM en futuras investigaciones.

Además, a partir de los resultados obtenidos, se abren dos líneas de proyección adicionales: (1) perfilar los descriptores de aquellos ítems en los que se observaron discrepancias entre IA y humanos, con el fin de reducir la distancia entre ambos tipos de evaluación; y (2) ampliar el análisis a los ítems en los que no se detectaron diferencias cuantitativas, explorando si los comentarios y la retroalimentación cualitativa revelan divergencias sutiles. Estas aproximaciones permitirían comprender mejor los patrones de convergencia y divergencia y avanzar hacia modelos híbridos de evaluación más precisos.

AGRADECIMIENTOS

Este estudio se enmarca en el Proyecto I+D *Tecnología y Formación pro Máster (TFproM): formación avanzada para optimizar trabajos de Máster con retroalimentación digital supervisada* (acrónimo TfproM, ref. PR17/24-31946) financiado por Comunidad de Madrid y la Universidad

Complutense de Madrid para el fomento de la investigación y la transferencia de tecnología (2023-2026).

BIBLIOGRAFÍA

Atasoy, A. y Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students written texts? *Education and Information Technology 30*, 20.385-20.415. *https://doi-org.bucm.idm.oclc.org/10.1007/s10639-025-13553-1*

Bouziane, K. y Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education, 3*, 201. *https://doi-org.bucm.idm.oclc.org/10.1007/s44217-024-00320-6*

Bridgeman, B., Trapani, C.S. y Attali, Y. (2014). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country NLP. En C. Wendler y B. Bridgeman (Eds.). *The Research Foundation for the GRE Revised General Test: A Compendium of Studies,* pp. 4.8.1-4.8.3). Educational Testing Service. *http://www.ets.org/s/research/pdf/gre_compendium.pdf*

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-35. *https://ejournals.bc.edu/index.php/jtla/article/view/1640*

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C. y Bekerman, Z. (2025). ChatGPT como una herramienta estable y justa para la evaluación automatizada de ensayos. *Education Sciences*, *15*(8), 946. *https://doi.org/10.3390/educsci15080946*

Hyland, K. (2013). *Genre and Second Language Writing* (The Michigan Series on Teaching Multilingual Writers). University of Michigan Press ELT.

Impey, C., Wenger, M., Garuda, N., Golchin, S. y Stamer, S. (2025). Using large language models for automated grading of student writing about science. *International Journal of Artificial Intelligence in Education*, *35*(1), 1-22. *https://doi.org/10.1007/s40593-024-00453-7*

Kincl, T., Gunina, D. y Pospíšil, J. (2024). Comparing human and AI-based essay evaluation in Czech higher education: challenges and limitations. *Business Trends, 14*(2), 25-34. *https://doi.org/10.24132/jbt.2024.14.2.25_34*

Kooli, C. y Yusuf, N. (2025). Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading. *International Journal of Human–Computer Interaction, 41*(5), 3.388-3.399. *https://doi.org/10.1080/10447318.2024.2338330*

López de Ramos, A.L., Bonnett-Bogallo, B., Concepción, D., Quintero-Barreto, G., Durán, J., Meléndez, N. y Esteves, Y. (2025). Análisis comparativo de la evaluación humana y la evaluación basada en inteligencia artificial generativa de resúmenes científicos. *EDUCA. Revista Internacional para a Calidad Educativa, 5*(2), 1-21. *https://doi.org/10.55040/q8sgtr65*

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J. y Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*(5), 616-633. *https://doi.org/10.1080/02602938.2024.2301722*

Perelman, L. (2020). The BABEL Generator and e-rater: 21st Century Writing Constructs and Automated Essay Scoring (AES). *The Journal of Writing Assessment, 13*(1), 1-8. *https://escholarship.org/uc/item/263565cq*

Roberts, R.H., Ali, S.R., Hutchings, H.A., Dobbs, T.D. y Whitaker, I.S. (2023). Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health & Care Informatics, 30*(1), e100830. *https://doi.org/10.1136/bmjhci-2023-100830*

Saborido-Fernández, P., Fernández-Pichel, M. y Losada, D.E. (2024). ChatGPT as a Solver and Grader of Programming Exams written in Spanish. [Preprint]. *arXiv:2409.15112v2. https://doi.org/10.48550/arXiv.2409.15112*

Sokolov, C. (2014). Self-evaluation of rater bias in written composition assessment. *Lingüística, 54*, 261-275. *https://doi.org/10.4312/linguistica.54.1.261-275*

Wetzler, E.L., Cassidy, K.S., Jones, M., Frazier, C., Korbut, N.A., Sims, C.M., Bowen, S.S. y Wood, M. (2025). Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology*, *52*(3), 298-304. *https://doi-org.bucm.idm.oclc.org/10.1177/00986283241282696*

ANEXO 1

Prompt utilizado en la interacción con el programa

Se redactó el siguiente *prompt* como instrucciones al programa:

Debes repasar la rúbrica que está cargada en un documento. Según la rúbrica, debes evaluar el texto que se te cargue en cada conversación. En tu feedback comentarás aspectos relacionados con la rúbrica. Para cada ensayo, es esencial:

- Comprensión crítica: El ensayo debe reflejar una comprensión profunda y crítica de los textos asignados por el/la docente.
- Relación de conceptos: Se debe identificar y establecer conexiones claras entre diferentes conceptos presentados por los autores de los textos.
- Postura personal: A lo largo del ensayo, se debe presentar y justificar una postura propia en relación con la problemática abordada en los textos.
- Reconocimiento de ideas principales: Es crucial identificar y destacar las ideas principales y posturas de los autores, así como evaluar y relacionar los aportes de las diferentes fuentes consultadas.
- Organización: Las ideas deben presentarse de manera estructurada y coherente, siguiendo una lógica argumentativa clara.
- Formato y convenciones: El ensayo debe ser redactado siguiendo las convenciones propias del contexto académico, respetando la normativa APA 7 y garantizando el cumplimiento de todos los indicadores estipulados en la rúbrica proporcionada.

Sin embargo, los ítems que debes valorar y retroalimentar son los de la rúbrica.