Evaluación de Textos Narrativos Basada en Rúbricas con Colaboración Humano-IA: Enfoque Especializado de Modelo

Rubric-Based Assessment of Narrative Texts via Human-AI Collaboration: A Specialized GPT Model Approach

https://doi.org/10.4438/1988-592X-RE-2025-411-730

Tolga Demir
https://orcid.org/0000-0002-1812-3397
Republic of Türkiye Ministry of National Education

Sait Çüm
https://orcid.org/0000-0002-0428-5088
Universidad Dokuz Eylul

Resumen

Este estudio analiza si los textos narrativos pueden ser evaluados con precisión, si es posible mantener calificaciones estables a lo largo del tiempo, y si puede proporcionarse retroalimentación formativa efectiva para estos textos gracias a la colaboración humano-IA. Para ello, se utilizaron dos modelos: la versión estándar de ChatGPT y la Herramienta de Evaluación de Textos (TAT), un modelo GPT específicamente entrenado mediante un proceso de seis pasos diseñado para esta investigación. Se evaluaron 114 textos narrativos en tres ocasiones según los criterios establecidos en una rúbrica, utilizando tanto el modelo estándar como el modelo especialmente entrenado. Se analizaron los niveles de concordancia entre las calificaciones otorgadas por TAT y por ChatGPT estándar con respecto a las calificaciones reales, así como la estabilidad de estas calificaciones a lo largo del tiempo. Los resultados mostraron que, en comparación con el desempeño del ChatGPT estándar, las calificaciones de TAT presentaron altos niveles de concordancia con las calificaciones reales y mantuvieron

su estabilidad a lo largo del tiempo en todas las categorías de la rúbrica, superando de forma constante el umbral mínimo e indicando con frecuencia una alta fiabilidad. Además, se observó que la mayor parte de la retroalimentación generada por TAT cumplía con los criterios de retroalimentación efectiva. La evidencia estadística presentada en este estudio demuestra que los modelos de lenguaje a gran escala, cuando son entrenados específicamente, pueden desempeñarse de manera excelente tanto en la evaluación de textos mediante una rúbrica como en la provisión de retroalimentación formativa. Esto es particularmente alentador para lograr una educación más equitativa, en particular en aulas numerosas y en contextos donde los evaluadores se encuentran sobrecargados.

Palabras clave: evaluación educativa, colaboración humano-IA, entrenamiento de modelos GPT.

Abstract

This study investigates whether narrative texts can be accurately and stably scored over time and whether effective formative feedback can subsequently be provided for these texts through human-AI collaboration. To this end, two models were employed: the default version of ChatGPT and the Text Assessment Tool (TAT), a GPT model specifically trained through a six-step process for this research purpose. 114 narrative texts were scored three times according to criteria in a rubric by both the specially trained and default models. The agreement levels of the scores given by TAT and default ChatGPT with the actual scores, as well as the stability of these scores over time, were examined. The results indicated that, in contrast to the performance of default ChatGPT, TAT's scores demonstrated high levels of agreement with the actual scores and maintained stability over time across all rubric categories, consistently surpassing the threshold and frequently indicating high reliability. Additionally, it was observed that the majority of the feedback generated by TAT met the criteria for effective feedback. Additionally, the feedback provided by TAT for the texts exceeded an 83% success rate in meeting effective feedback criteria across all categories. The statistical evidence presented in this study underscores that large language models, when specifically trained, can perform very well in scoring texts using a rubric and providing feedback. This is particularly promising for achieving fairer education, especially in large classes and situations where evaluators are overburdened.

Keywords: educational assessment, human-AI collaboration, GPT training

Introducción

"Para acortar nuestro camino, necesitábamos un caballo.

Encontramos uno salvaje, indómito y fuerte.

Tuvimos que domarlo, porque un caballo salvaje no nos serviría.

Este artículo es la historia de esa doma."

Sait Çüm & Tolga Demir

En la evaluación educativa, los tipos de preguntas como opción múltiple, completación de oraciones, emparejamiento y verdadero/falso se emplean con frecuencia tanto en entornos de aula como en exámenes a gran escala, particularmente con fines de evaluación sumativa, debido a su capacidad para ofrecer calificaciones objetivas. Sin embargo, para las evaluaciones formativas —que tienen como objetivo identificar y abordar las brechas de aprendizaje de los estudiantes, monitorear su desarrollo y mejorar los procesos de enseñanza— es esencial utilizar preguntas abiertas, así como presentaciones orales y actividades de desempeño basadas en productos o procesos, que brinden al docente datos más detallados. Estos enfoques de evaluación permiten identificar con mayor claridad las deficiencias y conceptos erróneos del alumnado, y además facilitan la medición de habilidades cognitivas de orden superior desde una perspectiva taxonómica. A pesar de estas ventajas, el proceso que implica leer, calificar y proporcionar retroalimentación sobre este tipo de evaluaciones, especialmente en aulas numerosas, suele ser tan demandante en términos de tiempo que muchos docentes optan por no utilizarlas.

Los avances recientes y revolucionarios en la tecnología de inteligencia artificial (IA) han generado discusiones sobre el papel y la relevancia futura de la IA en nuestras vidas. Hoy en día, resulta evidente que los seres humanos ya no son los únicos actores inteligentes en el planeta, lo que hace que la colaboración humano-IA sea inevitable en las organizaciones contemporáneas (Kolbjørnsrud, 2024). No es difícil prever que esta colaboración continuará reduciendo los costos relacionados con el tiempo y el trabajo en diversos ámbitos.

Recibido: 09/10/2024 Aceptado: 12/09/2025

Inteligencia artificial

La inteligencia artificial (IA) abarca sistemas informatizados que ejecutan tareas y responden de formas típicamente asociadas con la inteligencia humana, tales como el aprendizaje, la resolución de problemas y el logro de objetivos en condiciones inciertas y variables. Desde la resolución de problemas en los años 50, pasando por la simulación del razonamiento humano en los años 60, hasta los primeros proyectos de mapeo en los 70 y la aparición de los asistentes inteligentes en la década de 2000, la IA ha logrado avances notables (Dalton, 2024; Fell Kurban & Şahin, 2024). Dentro de este amplio campo, la IA generativa destaca como una subdisciplina especializada en la creación de contenido nuevo que imita datos preexistentes. A día de hoy, los campos del aprendizaje automático y las redes neuronales artificiales han evolucionado significativamente, lo que ha permitido el desarrollo de arquitecturas generativas sofisticadas y algoritmos de aprendizaje profundo. Ejemplos notables incluyen las redes generativas antagónicas (GAN), los autoencoders variacionales (VAE) y los modelos basados en transformadores (Alto, 2023; Chan & Colloton, 2024; Johannesson, 2024), los cuales son fundamentales para el progreso y las aplicaciones de la IA generativa.

El lanzamiento de ChatGPT —un modelo de IA generativa— por parte de OpenAI a finales de 2022 tuvo un impacto global significativo, atrayendo una gran atención a nivel mundial. Este desarrollo actuó como catalizador, motivando a numerosas grandes empresas tecnológicas a incursionar en el competitivo campo de los modelos de IA generativa (Holmes & Miao, 2023). Aunque han surgido competidores sólidos como Gemini, DeepSeek y Llama, ChatGPT mantiene una ligera ventaja en cuanto a popularidad y adopción generalizada por parte de los usuarios.

ChatGPT, un modelo de lenguaje de gran escala (LLM) preentrenado, utiliza una arquitectura lingüística basada en transformadores, un tipo de red neuronal profunda altamente eficaz para las tareas de procesamiento del lenguaje natural (PLN). Es capaz de comprender y generar texto con características humanas a partir de las entradas que recibe. Entrenado con una vasta cantidad de datos, ChatGPT ha aprendido los patrones, estilos y complejidades del lenguaje humano, lo que lo convierte en una herramienta

Recibido: 09/10/2024 Aceptado: 12/09/2025

excepcional para la comunicación. Sus capacidades han transformado el ámbito educativo al ofrecer conversaciones dinámicas de tipo humano, información instantánea, recomendaciones personalizadas y apoyo académico continuo (Chan & Colloton, 2024; Fell Kurban & Şahin, 2024).

Educación y Modelos de Lenguaje de Gran Escala (LLMs)

La popularidad de los modelos de lenguaje de gran escala (LLMs), como ChatGPT, tanto entre docentes como entre estudiantes, exige investigaciones que analicen la concordancia entre sus capacidades y salidas con las expectativas u objetivos definidos. Consideramos crucial explorar el potencial de los LLMs para facilitar y apoyar tareas que resultan agotadoras y que consumen mucho tiempo para los profesores, en lugar de limitarse a asistir en tareas rutinarias o de baja complejidad. Esto es especialmente importante para asegurar la continuidad de procesos educativos críticos, como la evaluación formativa, en aulas numerosas o durante períodos de alta carga laboral docente, garantizando así la calidad educativa.

Al revisar la literatura sobre la integración de los LLMs en las prácticas educativas, en particular en las habilidades de escritura y evaluación, se identificaron dos estudios que destacaban la eficacia de la IA para generar materiales de lectura (Fitria, 2023; Xiao et al., 2023). Además, se encontraron tres estudios alineados con nuestros objetivos que analizaron la capacidad de la IA para proporcionar retroalimentación o calificar ensayos estudiantiles (Steiss et al., 2024; Wang, 2022; Yavuz et al., 2024). Sin embargo, no se empleó un proceso de entrenamiento basado en rúbricas diseñado específicamente para evaluar habilidades de escritura narrativa en ninguno de estos estudios.

Por otro lado, el uso de LLMs en el ámbito educativo puede generar información ilusoria (hallucination), lo que conlleva problemas de precisión y fiabilidad que podrían afectar negativamente el aprendizaje de los estudiantes y sus habilidades de pensamiento crítico (Elsayed, 2024). Algunos estudios han mostrado que los LLMs, especialmente durante los procesos de retroalimentación, pueden no comprender completamente el trabajo del estudiante y, en ocasiones, ofrecer comentarios fuera de

contexto o superficiales (Venter et al., 2024; Jia et al., 2024). Esta situación puede perjudicar la confianza que docentes y estudiantes depositan en la retroalimentación generada por IA, y provocar actitudes más cautelosas o incluso escépticas frente a dichos resultados (Ziqi, 2024). Por estas razones, se considera importante entrenar un modelo especializado basado en rúbricas para un propósito específico dentro del marco de este estudio, ya que tiene el potencial de superar algunos de estos desafíos.

El presente estudio

Este estudio tiene como objetivo determinar si los textos narrativos pueden ser evaluados con precisión y si es posible proporcionar retroalimentación formativa efectiva mediante la colaboración humano-IA. Además, el estudio compara la precisión de las evaluaciones y la estabilidad a lo largo del tiempo entre un modelo GPT entrenado, denominado Herramienta de Evaluación de Textos (TAT, por sus siglas en inglés), y el modelo ChatGPT por defecto. TAT fue entrenado mediante un enfoque de múltiples instrucciones iterativas (many-shot iterative prompting).

Uno de los principales desafíos de este estudio es la subjetividad inherente a la evaluación de textos narrativos, incluso cuando se emplea una rúbrica. El estudio plantea que la colaboración entre humanos e IA puede mejorar la objetividad y la estabilidad en la calificación. Por ejemplo, determinar qué constituye un "título atractivo" implica un juicio personal, lo cual también representa una dificultad para la IA. En lugar de simplificar la rúbrica para minimizar la subjetividad, esta investigación busca demostrar cómo la colaboración humano-IA puede generar soluciones confiables en contextos que requieren evaluaciones subjetivas. Si tiene éxito, este enfoque podría conducir a resultados más justos, incluso en procesos de evaluación a gran escala o en procedimientos de selección de personal.

El estudio también explora el potencial de la colaboración con IA para proporcionar retroalimentación dentro de los procesos de evaluación formativa con el fin de apoyar el desarrollo del estudiantado. Se postula que dicha colaboración puede reducir la carga de trabajo del profesorado al

Recibido: 09/10/2024 Aceptado: 12/09/2025

proporcionar retroalimentación sobre los textos estudiantiles. Los criterios de retroalimentación efectiva utilizados para evaluar dicha efectividad (Brookhart, 2008; Burke & Pieterick, 2010; Irons, 2008; Juwah et al., 2004) se detallan en el Apéndice I. Aunque el proceso de retroalimentación efectiva puede implicar diálogo e interacción cara a cara, este estudio se enfoca exclusivamente en la retroalimentación escrita debido a la naturaleza de los materiales utilizados.

Las siguientes hipótesis orientan el proceso de investigación y los análisis:

- H1: El modelo de lenguaje de gran escala entrenado generará evaluaciones más precisas de textos narrativos que ChatGPT por defecto.
- H2: El modelo de lenguaje de gran escala entrenado demostrará una mayor estabilidad y reproducibilidad en la evaluación de textos narrativos a lo largo del tiempo, en comparación con ChatGPT por defecto.
- H3: El modelo de lenguaje de gran escala entrenado proporcionará retroalimentación escrita más efectiva, que favorezca el desarrollo de las habilidades narrativas del estudiantado, en comparación con ChatGPT por defecto.

Metodología

La metodología de este estudio comprende dos fases principales. La primera fase consiste en entrenar un modelo GPT específicamente con el objetivo de la investigación. La segunda fase evalúa la efectividad del modelo entrenado comparándolo con la versión por defecto de ChatGPT.

Para ambas fases se utilizó una rúbrica del currículo 2024 del Ministerio de Educación Nacional de Turquía, que incluye ocho categorías: "estructura de la página", "título", "estructura del texto", "personaje", "ambientación", "trama", "lenguaje y estilo" y "ortografía y puntuación". Cada categoría se evalúa en tres niveles (véase el Apéndice II para más detalles). La herramienta TAT fue entrenada mediante la aplicación GPT Builder de OpenAI, con el fin

Recibido: 09/10/2024 Aceptado: 12/09/2025

de calificar textos y proporcionar retroalimentación basada en esta rúbrica.

En este estudio, no se emplearon evaluadores humanos como referencia para medir la precisión de la IA, ya que los errores humanos podrían introducir sesgos adicionales. En su lugar, los textos narrativos incluidos en el conjunto de datos fueron elaborados por los investigadores de acuerdo con la rúbrica, incorporando omisiones o errores intencionados. Para garantizar la fidelidad del diseño previsto del conjunto de datos, se llevó a cabo un estudio de fiabilidad con expertos independientes a partir de una muestra de los textos, cuyos resultados se detallan en esta sección. Tras este estudio, las calificaciones predefinidas de los textos —denominadas "calificaciones reales" — se utilizaron como patrón de referencia para las comparaciones.

Proceso de Entrenamiento del GPT

La herramienta TAT fue sometida a un riguroso proceso de entrenamiento de seis etapas, resumido a continuación, utilizando la aplicación GPT Builder. En el Apéndice III se presentan ejemplos de las instrucciones (prompts) utilizadas en cada etapa.

- Definición de objetivos y evaluación inicial

Al modelo GPT se le asigna un rol específico. Los archivos necesarios para desempeñar este rol (preferentemente en formato PDF) se cargan en el sistema y se verifica su comprensibilidad. En el contexto de este estudio, dichos archivos contienen los criterios especificados en la rúbrica, así como los criterios para la provisión de retroalimentación efectiva.

- Introducción a los criterios

Se lleva a cabo una sesión de preguntas y respuestas con GPT sobre los criterios de evaluación. El objetivo de esta sesión es determinar en qué medida GPT comprende dichos criterios e identificar posibles dificultades que podría enfrentar durante el proceso de evaluación. Las instrucciones proporcionadas en respuesta a las respuestas del modelo ayudan a aclarar cómo deben aplicarse los criterios de evaluación.

- Análisis de Ejemplos

Se suministran ejemplos (datos de entrenamiento) a GPT mediante

Recibido: 09/10/2024 Aceptado: 12/09/2025

un enfoque de instrucciones iterativas con múltiples ejemplos (*many-shot prompting*). Para las tareas de calificación, se presentan oraciones representativas de textos que podrían recibir una puntuación de 1, 2 o 3 puntos. En el caso de la retroalimentación, se ofrecen ejemplos de oraciones que ilustran retroalimentación efectiva.

- Carga de Archivos de Muestra

Al evaluar textos narrativos, también es necesario considerar ciertos aspectos formales —como la sangría de los párrafos y la colocación del título— aunque no estén directamente relacionados con el contenido. Dado que no era posible ilustrar eficazmente estos aspectos mediante ejemplos textuales convencionales, se requirieron muestras visuales. Por ello, se creó un conjunto de textos con errores de formato intencionales y se proporcionaron a GPT en formato PDF. Estos archivos sirvieron como puntos de referencia, permitiendo a GPT reconocer y evaluar con precisión los elementos formales como parte del proceso de evaluación.

- Práctica Estructurada

A diferencia de las etapas anteriores, que se centran en partes específicas, esta fase tiene como objetivo observar el proceso de manera integral. Es decir, mientras que las fases previas se enfocan en criterios individuales dentro de una rúbrica o en aspectos particulares de la retroalimentación efectiva, esta etapa examina cómo GPT califica un texto completo y proporciona una retroalimentación global y coherente.

Durante el proceso de entrenamiento, las fases de análisis de ejemplos, carga de archivos de muestra y práctica estructurada pueden repetirse de manera iterativa para asegurar respuestas más precisas. En los casos en que no se logran los resultados deseados, el proceso se repite con nuevos ejemplos y se reevalúan los textos estructurados para identificar áreas problemáticas.

- Evaluación Final y Confirmación

En esta etapa, el investigador verifica los temas acordados con GPT hasta ese momento. Se realizan los ajustes finales a las instrucciones si es necesario y se confirman los archivos fuente que fueron previamente cargados.

Recibido: 09/10/2024 Aceptado: 12/09/2025

FIGURA I. Proceso de entrenamiento del GPT para la evaluación basada en rúbricas utilizando GPT Builder

La Figura I ilustra el proceso de entrenamiento del GPT, el cual es aplicable a tareas similares. Haga clic aquí para acceder a TAT.

Desafíos y Soluciones Encontrados Durante el Entrenamiento de la IA

En esta sección se resumen los desafíos más relevantes enfrentados durante el proceso de entrenamiento de la inteligencia artificial, así como las soluciones implementadas, con el objetivo de orientar a otros investigadores y profesionales que puedan encontrarse con situaciones similares.

- Compatibilidad de tipos de archivo

Uno de los desafíos clave fue la diversidad de tipos de archivo utilizados en el proceso de entrenamiento. Considerando su aplicación práctica, los evaluadores podrían almacenar textos en distintos formatos, como

Recibido: 09/10/2024 Aceptado: 12/09/2025

fotografías de trabajos manuscritos de estudiantes. Durante el entrenamiento, se experimentó con varios tipos de archivo para observar posibles variaciones en el rendimiento. El uso de archivos .png generó más errores, probablemente debido al proceso de reconocimiento óptico de caracteres (OCR) utilizado por el modelo de lenguaje, que alteraba el contenido antes del análisis. Con los archivos .doc, el modelo a veces modificaba los textos, por ejemplo, añadiendo y calificando títulos que originalmente no estaban presentes. No obstante, el uso de archivos .pdf minimizó estos problemas, al reducir significativamente la frecuencia de tales intervenciones.

- Procesamiento por lotes vs. procesamiento individual

Otro problema estuvo relacionado con el modo de envío de los textos: de forma colectiva o individual. El procesamiento por lotes generó errores significativos durante las fases de entrenamiento y prueba, ya que el modelo mostró una automatización no deseada en la calificación y empezó a generar retroalimentación uniforme tras los primeros textos. El envío secuencial y la evaluación individual permitieron mitigar eficazmente estos inconvenientes.

- Extensión del texto

También surgieron dificultades relacionadas con la longitud de los datos de entrenamiento. Los textos extensos o las instrucciones demasiado largas dentro del conjunto de entrenamiento podían provocar confusión en el modelo. Estos problemas se resolvieron organizando los datos en segmentos más breves, claros y concisos.

- Problemas de conexión y generalización

En algunas ocasiones, el modelo entrenado presentó dificultades para establecer conexiones adecuadas con los datos de entrenamiento previos, lo que resultó en una creatividad no deseada. Esto puede deberse a la dificultad del modelo para conectar con los datos anteriores, así como a la tendencia a realizar generalizaciones incorrectas al exponerse a datos más variados. Se observó que estos problemas de conexión tendían a resolverse con el tiempo, sin necesidad de nuevas intervenciones, y que no estaban asociados de manera constante con datos específicos (lo cual indica un comportamiento aleatorio). Esta situación subraya la importancia de la colaboración humano-IA, sugiriendo que algunos procesos no deben ser delegados completamente a la inteligencia artificial. La supervisión humana puede eliminar eficazmente

este tipo de errores.

Recopilación de Datos

Se elaboraron un total de 114 textos narrativos para este estudio, todos redactados en turco. Estos textos presentan una amplia diversidad en relación con los criterios de evaluación. Por ejemplo, algunos consisten en un solo párrafo pero son impecables en cuanto a gramática y puntuación. Otros, aunque tienen una estructura ideal de tres párrafos y un diseño de página adecuado, carecen de título. Algunos omiten intencionadamente elementos de la ambientación. Cada historia tiene un título único, presenta personajes distintos y utiliza diferentes componentes de ambientación, lo que origina tramas diferenciadas. En esencia, cada texto es original y no guarda relación con los demás. Este enfoque busca minimizar el factor de aleatoriedad en la calificación y en la provisión de retroalimentación efectiva por parte de TAT.

Los textos narrativos fueron cargados en la Herramienta Automatizada de Análisis de Textos (TAT) y evaluados en tres ocasiones, generando calificaciones correspondientes a tres momentos distintos. El mismo procedimiento de calificación fue aplicado utilizando ChatGPT en su versión por defecto, dando lugar a tres conjuntos de datos: las calificaciones reales, las calificaciones asignadas por TAT en tres sesiones y las calificaciones asignadas por ChatGPT en tres sesiones. Además, 20 textos narrativos seleccionados al azar fueron cargados nuevamente en TAT para recopilar retroalimentación efectiva basada en las evaluaciones automatizadas previas, y los resultados fueron debidamente registrados.

Análisis de Datos

Los niveles de concordancia entre las calificaciones reales de los 114 textos, las calificaciones asignadas por TAT en tres momentos distintos y las otorgadas por ChatGPT en su versión por defecto también en tres momentos distintos, se calcularon utilizando la técnica del coeficiente α de Krippendorff.

Recibido: 09/10/2024 Aceptado: 12/09/2025

El α de Krippendorff es un coeficiente de fiabilidad ampliamente utilizado en disciplinas como las ciencias sociales y el análisis de contenido, con el propósito de medir la consistencia de las evaluaciones —categóricas o continuas— realizadas por múltiples evaluadores o codificadores (Krippendorff, 2004). Los valores obtenidos de α se interpretan según los siguientes criterios: un coeficiente inferior a 0.67 indica baja concordancia y fiabilidad; un valor entre 0.67 y 0.80 indica un nivel aceptable o moderado; y un valor superior a 0.80 indica alta concordancia y fiabilidad.

La eficacia de la retroalimentación proporcionada por TAT, basada en los criterios de retroalimentación efectiva, fue analizada mediante estadísticas descriptivas. En esta fase, los investigadores examinaron individualmente un total de 160 comentarios correspondientes a 20 textos seleccionados aleatoriamente, utilizando las ocho categorías definidas en la rúbrica. La retroalimentación fue codificada por los investigadores como exitosa o no exitosa, de acuerdo con criterios como la adecuación a la categoría, orientación al rendimiento, claridad y comprensibilidad, calidad de desarrollo, carácter constructivo y especificidad de la tarea. Posteriormente, se calculó el porcentaje de retroalimentaciones exitosas en relación con el total de comentarios evaluados.

Fiabilidad de las Calificaciones Reales

Las calificaciones preliminares de los 114 textos narrativos fueron asignadas por los investigadores utilizando una rúbrica. Para evaluar la fiabilidad de dichas calificaciones, se seleccionó aleatoriamente un subconjunto de textos de los 114 y se enviaron a tres expertos, quienes emitieron su juicio sobre la adecuación de las calificaciones asignadas. La similitud entre las valoraciones de los expertos y las calificaciones de los investigadores se calculó utilizando la fórmula $A = C \div (C + a) \times 100$, basada en el modelo de Miles y Huberman (1994). En esta fórmula, A representa el coeficiente de fiabilidad, C el número de ítems o términos coincidentes, y a el número de ítems o términos no coincidentes. Según este modelo, se requiere una proporción mínima del 80 % de coincidencia para alcanzar una consistencia aceptable.

Nuestro estudio de fiabilidad mostró tasas de acuerdo del 95 %, 96.25 % y 91.25 % entre las calificaciones preliminares asignadas por los investigadores y las evaluaciones de los expertos. Esta consistencia justifica considerar las calificaciones preliminares de los investigadores como el estándar de oro (calificaciones reales) para los fines comparativos del estudio.

Resultados

Todos los textos del conjunto de datos fueron evaluados por TAT en tres momentos distintos, y se utilizó el coeficiente α de Krippendorff para determinar el nivel de concordancia entre cada conjunto de puntuaciones y las calificaciones reales (t). Los resultados obtenidos se presentan en la Tabla I.

TABLA I. Niveles de concordancia entre las calificaciones de TAT y las calificaciones reales

Categoría	α_{t-1}	α_{t-2}	α_{t-3}	α_{media}	Interpretación
Estructura de la página	0.899	0.897	0.857	0.884	Alta fiabilidad
Título	0.870	0.802	0.818	0.830	Alta fiabilidad
Estructura del texto	0.921	0.950	0.950	0.940	Alta fiabilidad
Personaje	0.822	0.763	0.807	0.797	Fiabilidad media
Ambientación	0.728	0.787	0.759	0.758	Fiabilidad media
Trama	0.883	0.894	0.901	0.889	Alta fiabilidad
Lenguaje y estilo	0.874	0.894	0.893	0.887	Alta fiabilidad
Ortografía y puntuación	0.780	0.786	0.814	0.793	Fiabilidad media

Al analizar la Tabla I, se observa que la categoría con el menor nivel de concordancia entre las calificaciones de TAT y las calificaciones reales es la de ambientación, que implica examinar la presencia de los elementos de tiempo y lugar en las historias, así como su impacto. Por otro lado, la categoría con el mayor nivel de acuerdo es estructura del texto, la cual evalúa la presencia y calidad de las secciones de introducción, desarrollo y conclusión en los relatos, alcanzando un valor de 0.940. Al revisar los hallazgos sobre los niveles de concordancia en cada una de las tres comparaciones entre las calificaciones de TAT y las calificaciones reales, se constató que todos

Recibido: 09/10/2024 Aceptado: 12/09/2025

los valores de alfa obtenidos, así como sus medias, superaban el umbral considerado como confiable (0.667).

Además del análisis de la concordancia entre las calificaciones de TAT y las calificaciones reales, también se examinó la estabilidad de las calificaciones asignadas por TAT a lo largo de los tres momentos distintos. Los resultados obtenidos y su interpretación se presentan en la Tabla II.

TABLA II. Estabilidad de las Calificaciones de TAT

Categoría	$\alpha_{1\text{-}2\text{-}3}$	**Interpretación**
Estructura de la página	0.905	Alta fiabilidad
Título	0.896	Alta fiabilidad
Estructura del texto	0.957	Alta fiabilidad
Personaje	0.797	Fiabilidad media
Ambientación	0.846	Alta fiabilidad
Trama	0.908	Alta fiabilidad
Lenguaje y estilo	0.954	Alta fiabilidad
Ortografía y puntuación	0.828	Alta fiabilidad

Al examinar la Tabla II, se observa que las calificaciones asignadas por TAT muestran consistencia a lo largo del tiempo en todas las categorías, lo que subraya la reproducibilidad de los resultados de evaluación. Se advierte que la categoría con el mayor nivel de concordancia entre las tres evaluaciones realizadas por TAT es nuevamente la de estructura del texto. En cambio, la categoría de personaje, que evalúa los rasgos personales y psicológicos de los personajes de la historia, presenta el menor nivel de acuerdo. Puede interpretarse que la categoría de personaje, al tener el valor alfa más bajo, refleja una fiabilidad moderada, mientras que las demás categorías evidencian una fiabilidad alta.

Los valores de α de Krippendorff, que indican el grado de concordancia entre las calificaciones de ChatGPT por defecto y las calificaciones reales, se recogen en la Tabla III**.**

TABLA III. Niveles de concordancia entre las calificaciones de ChatGPT por defecto y las calificaciones reales

Categoría	α_{t-1}	α_{t-2}	α_{t-3}	α_{media}	Interpretación
Estructura de la página	-0.032	0.553	0.370	0.297	Baja fiabilidad
Título	0.159	0.422	0.473	0.351	Baja fiabilidad
Estructura del texto	0.261	0.521	0.502	0.428	Baja fiabilidad
Personaje	0.266	0.604	0.477	0.449	Baja fiabilidad
Ambientación	0.384	0.469	0.516	0.456	Baja fiabilidad
Trama	0.214	0.381	0.477	0.357	Baja fiabilidad
Lenguaje y estilo	0.412	0.562	0.550	0.508	Baja fiabilidad
Ortografía y puntuación	0.233	0.415	-0.171	0.159	Baja fiabilidad

La Tabla III muestra que la categoría ortografía y puntuación presenta el menor nivel de concordancia entre las calificaciones de ChatGPT y las calificaciones reales, mientras que la categoría lenguaje y estilo alcanza el mayor nivel de acuerdo relativo. En términos generales, las calificaciones asignadas por ChatGPT en su versión por defecto evidencian una baja concordancia global con las calificaciones reales en todas las categorías.

Los resultados relacionados con la estabilidad de las calificaciones asignadas por ChatGPT, un modelo de procesamiento del lenguaje general que no fue específicamente entrenado para esta investigación, a lo largo de tres momentos distintos, se presentan en la Tabla IV.

TABLA IV. Estabilidad de las calificaciones de ChatGPT por defecto

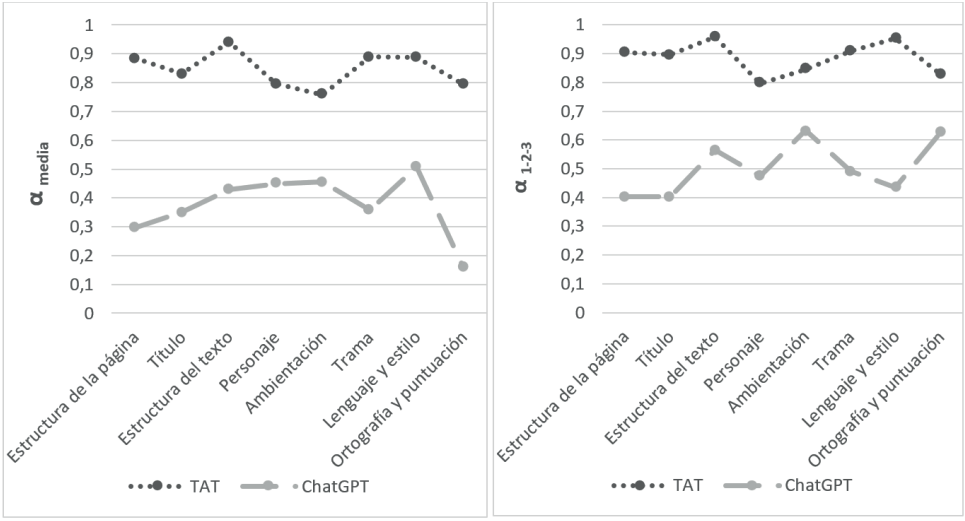
Categoría	α_{1-2-3}	Interpretación
Estructura de la página	0.403	Baja fiabilidad
Título	0.402	Baja fiabilidad
Estructura del texto	0.566	Baja fiabilidad
Personaje	0.475	Baja fiabilidad
Ambientación	0.633	Baja fiabilidad
Trama	0.491	Baja fiabilidad
Lenguaje y estilo	0.436	Baja fiabilidad
Ortografía y puntuación	0.627	Baja fiabilidad

Al analizar la Tabla IV, se observa que la categoría con el mayor nivel

Recibido: 09/10/2024 Aceptado: 12/09/2025

de concordancia entre las calificaciones otorgadas en tres momentos distintos es la de ambientación, mientras que la categoría con el menor nivel de concordancia es la de título. Al interpretar los valores presentados en la Tabla IV, se concluye que los niveles de acuerdo en todas las categorías indican baja fiabilidad.

FIGURA II. Comparación entre TAT y ChatGPT por defecto en términos de concordancia con las calificaciones reales (izquierda) y estabilidad de las calificaciones (derecha)



La Figura II ilustra los valores medios del coeficiente alfa presentados en las Tablas I y III, así como los valores alfa correspondientes a los tres momentos distintos (fiabilidad intraevaluador) detallados en las Tablas II y IV. La figura pone de manifiesto discrepancias significativas entre las calificaciones asignadas por ChatGPT y las otorgadas por TAT, tanto en términos de concordancia con las calificaciones reales (izquierda) como de estabilidad en el tiempo (derecha).

Con base en la evidencia relativa a la concordancia entre la calificación basada en rúbricas realizada por TAT y las calificaciones reales de los textos,

así como la estabilidad de dichas calificaciones a lo largo del tiempo, se concluyó que las hipótesis de investigación H1 y H2 quedaron respaldadas.

Para investigar la hipótesis H3, se analizó la eficacia de la retroalimentación proporcionada por TAT, conforme a la rúbrica utilizada en el estudio, en función de los criterios de retroalimentación efectiva. La retroalimentación fue evaluada por los investigadores utilizando criterios que incluían: orientación al rendimiento, claridad y comprensibilidad, carácter constructivo, valor formativo y especificidad de la tarea, tal como se describe en la introducción del estudio. Durante dichas evaluaciones, se observó que TAT ocasionalmente proporcionaba retroalimentación asociada a una categoría distinta de la que correspondía. Para cuantificar estos casos, se definió un criterio adicional denominado adecuación categorial, que se incorporó junto a los demás criterios de retroalimentación efectiva.

Los resultados, que incluyen el número total de comentarios analizados y sus tasas de éxito, se presentan en la Tabla V.

TABLA V. Rendimiento de la retroalimentación proporcionada por TAT

Categoría	Total de Retroalimentaciones	Retroalimentaciones Exitosas (%)
Adecuación categorial	160	91,88
Orientada al rendimiento	160	100
Clara y comprensible	160	86,25
Formativa (de desarrollo)	160	83,75
Constructiva	160	100
Específica para la tarea	160	89,38

El análisis de la **Tabla V** revela que toda la retroalimentación proporcionada por TAT está **orientada al rendimiento** y tiene un carácter **constructivo**. El criterio con el menor índice de cumplimiento es el de **desarrollo formativo**, con un 83,75 % de éxito. A pesar de la dificultad inherente a redactar retroalimentaciones que fomenten el desarrollo del estudiante, el rendimiento de TAT en esta área sigue siendo notablemente alto, al igual que el observado en los demás criterios.

Cuando se examina secuencialmente la retroalimentación a la luz de los criterios que debe cumplir una retroalimentación efectiva, se observa que el **91,88 %** de los comentarios cumple con el criterio de **adecuación**

Recibido: 09/10/2024 Aceptado: 12/09/2025

categorial, mientras que el **8,12 %** restante no lo hace. A continuación, se presenta un ejemplo de retroalimentación considerada no adecuada según este criterio:

Ejemplo de Retroalimentación 1: Dado que el texto consta de un solo párrafo, la estructura de la página es insuficiente. En tu próxima redacción, intenta utilizar al menos tres párrafos: una introducción, un desarrollo y una conclusión. Al utilizar un párrafo distinto para cada sección, podrás mejorar la fluidez y la comprensibilidad de tu texto. Por ejemplo, en el párrafo de introducción puedes describir cómo Can comienza a buscar a su padre; en el desarrollo, cómo lo encuentra y las experiencias que viven en el laboratorio; y en la conclusión, ofrecer información sobre el futuro de Can.

El Ejemplo de Retroalimentación 1 fue proporcionado por TAT para la categoría estructura de la página del Relato 80. Sin embargo, dicho comentario se centra en sugerencias relativas a la fluidez, claridad y desarrollo narrativo, y no a la estructura física de la página. Aunque estas sugerencias resultan valiosas para el desarrollo del texto, no se ajustan a los criterios establecidos en la rúbrica para la categoría correspondiente**.** En contraste, el siguiente ejemplo sí se considera adecuado:

Ejemplo de Retroalimentación 2: El texto contiene dos párrafos que, actualmente, no están justificados. Para mejorar la organización y legibilidad de tu escritura, se recomienda aplicar sangría al comienzo de cada párrafo en futuras composiciones.

Al examinar si la retroalimentación proporcionada por TAT está orientada al rendimiento, se observa que las 160 observaciones cumplen con este criterio. El Ejemplo de Retroalimentación 3 es una muestra representativa de este tipo de retroalimentación emitida por TAT:

Ejemplo de Retroalimentación 3: Se mencionan el nombre y algunas características del personaje. Sin embargo, puedes añadir rasgos emocionales del personaje para ayudar al lector a comprenderlo mejor y aportar mayor profundidad a tu historia. Por ejemplo, puedes describir la emoción y la curiosidad de Cem durante el descubrimiento mientras avanza por el túnel. "El corazón de Cem latía con fuerza mientras avanzaba por el túnel. Sentía la emoción de un nuevo descubrimiento con cada paso."

Al analizar si la retroalimentación proporcionada por TAT es clara y

Recibido: 09/10/2024 Aceptado: 12/09/2025

comprensible, se observa que el 86,25 % de las 160 observaciones cumple con este criterio. A continuación, se presenta el Ejemplo de Retroalimentación 4:

Ejemplo de Retroalimentación 4: Los eventos progresan de manera lógica, pero puedes aumentar la fluidez de la historia añadiendo más detalles. La historia tiene un comienzo, desarrollo y desenlace claramente definidos. Los preparativos para la fiesta de Kaan, los momentos divertidos durante la celebración y su gratitud por los regalos al final están bien descritos.

El Ejemplo de Retroalimentación 4 fue generado por TAT para la categoría trama del Relato 53. Esta retroalimentación ejemplifica los principios de claridad y comprensibilidad, fundamentales en una retroalimentación efectiva. Evita el uso de sugerencias que puedan confundir al estudiante o palabras que este no comprenda.

En el análisis realizado para evaluar la calidad de la retroalimentación de TAT, se determinó que el 83,75 % de los comentarios fomenta efectivamente el desarrollo del estudiante. Esto se ejemplifica en el siguiente caso:

Ejemplo de Retroalimentación 5:Se mencionan el lugar y el momento en que ocurre el evento, pero no se ofrece información detallada. Puedes reforzar la atmósfera de la historia describiendo el lugar y el tiempo con mayor precisión. Por ejemplo: "En un día de verano, Mehmet encontró un violín silencioso mientras paseaba por la tienda de música. La tienda estaba llena de instrumentos musicales antiguos."

El Ejemplo de Retroalimentación 5 corresponde a la retroalimentación proporcionada por TAT para la categoría ambientación del Relato 15. Al analizar este comentario, se observa que las sugerencias y ejemplos incluidos consisten en expresiones que apoyan el desarrollo del estudiante.

Uno de los principios fundamentales de la retroalimentación efectiva es que debe ser constructiva. Según este criterio, los comentarios dirigidos a los estudiantes deben animarlos y ofrecerles diferentes alternativas, en lugar de imponer órdenes o instrucciones rígidas. Desde esta perspectiva, se observa que todas las retroalimentaciones generadas por TAT se expresan de manera constructiva y motivadora. Esto puede verse en la siguiente retroalimentación proporcionada para la categoría de ambientación en el Relato 22:

Ejemplo de Retroalimentación 6: El lugar y el momento en que se desarrolla el evento están descritos con detalle. El trayecto hacia la biblioteca,

Recibido: 09/10/2024 Aceptado: 12/09/2025

las ruinas, así como los manuscritos y libros dentro de la biblioteca están claramente descritos. La contribución de la ambientación al desarrollo de la historia está bien enfatizada.

Otra de las características que debe tener una retroalimentación efectiva es que sea específica para la tarea y no general. Según este criterio, los comentarios eficaces no deben utilizar expresiones genéricas aplicables a cualquier texto, sino estar personalizados según la producción del estudiante. Al examinar la retroalimentación proporcionada por TAT, se observa que el modelo logra un desempeño bastante alto en este aspecto, con un 89,38 % de los comentarios cumpliendo este criterio.

Ejemplo de Retroalimentación 7: Los eventos progresan de forma lógica y detallada. La historia presenta un inicio, desarrollo y desenlace claramente definidos. Los preparativos para la fiesta de cumpleaños de Selin, los momentos divertidos durante la celebración y, finalmente, la apertura de los regalos y sus agradecimientos están bien descritos.

El Ejemplo de Retroalimentación 7 corresponde a la categoría trama del Relato 97. Al analizar esta retroalimentación, se observa que se trata de un comentario específico y directamente relacionado con el contenido del texto, y no de un comentario genérico.

Conclusión y Discusión

Este estudio tiene como objetivo determinar si los textos narrativos pueden ser evaluados de forma precisa y estable mediante la colaboración entre humanos e inteligencia artificial, así como si puede proporcionarse retroalimentación formativa efectiva. Además, se comparó el rendimiento del modelo GPT entrenado específicamente para este propósito con el de ChatGPT en su versión por defecto, que no fue entrenado con fines evaluativos, a fin de destacar las diferencias de desempeño entre ambos modelos.

Precisión y fiabilidad de la evaluación

Concordancia con las calificaciones reales: TAT evaluó 114 textos narrativos utilizando una rúbrica, y se examinó el nivel de concordancia entre las calificaciones asignadas y las calificaciones reales para cada una de las categorías. Los valores del coeficiente α de Krippendorff indicaron un alto nivel de acuerdo en todos los criterios, superando el umbral de fiabilidad ($\alpha \geq 0.667$). El mayor nivel de concordancia se observó en la categoría estructura del texto ($\alpha = 0.940$), mientras que el más bajo se registró en la categoría ambientación ($\alpha = 0.758$).

Estabilidad a lo largo del tiempo: Al examinar los niveles de concordancia entre las calificaciones asignadas por TAT en tres momentos distintos, se comprobó que los valores de α de Krippendorff se mantuvieron por encima del umbral ($\alpha \geq 0.667$) en todas las categorías. La mayor estabilidad se registró en la categoría estructura del texto ($\alpha = 0.957$), mientras que la menor se observó en personaje ($\alpha = 0.797$).

Tanto en términos de concordancia con las calificaciones reales como en la estabilidad en el tiempo, se identificaron valores de alfa relativamente bajos en las categorías de personaje, ambientación y ortografía y puntuación. En el caso de la categoría de personaje, la rúbrica establece una diferencia sutil entre asignar dos puntos y tres puntos: se otorgan dos puntos cuando se describen los rasgos físicos y psicológicos de los personajes; y tres puntos cuando, además, se identifican las emociones y perspectivas que influyen en el desarrollo narrativo. Determinar qué emoción o punto de vista influye en la narración, o diferenciarlos, puede resultar complejo. Esta dificultad afecta tanto a un evaluador humano como al modelo TAT.

En cuanto a la categoría ambientación, se considera que el reto proviene de las inconsistencias en la representación combinada de los elementos de lugar y tiempo. Por ejemplo, una historia puede ofrecer información detallada sobre el lugar y su impacto en la narración, pero omitir el aspecto temporal, dificultando así la calificación conforme a la rúbrica, que exige evaluar ambos elementos de forma conjunta. Una mayor desagregación de estos criterios en componentes más claros y específicos dentro de la rúbrica podría mejorar el rendimiento del modelo de IA.

Recibido: 09/10/2024 Aceptado: 12/09/2025

En relación con la categoría de ortografía y puntuación, fue necesario incorporar numerosos conjuntos de datos explicativos sobre las reglas ortográficas y de puntuación del turco para mejorar el desempeño de TAT. Esta necesidad resulta paradójica, ya que el uso excesivo de conjuntos de datos puede confundir al modelo durante el entrenamiento. Es probable que, si los textos hubieran estado redactados en inglés, se hubiera requerido una menor cantidad de datos, obteniendo así un rendimiento superior.

En términos generales, todas las evaluaciones superaron el umbral de fiabilidad y se consideran satisfactorias. Las categorías con rendimientos relativamente más bajos podrían mejorarse mediante intervenciones como la revisión de la rúbrica, y no se consideran problemáticas significativas para los procesos de evaluación textual en colaboración con IA.

Rendimiento de ChatGPT por defecto: Las pruebas realizadas con ChatGPT en su versión por defecto revelaron resultados de baja fiabilidad, tanto en términos de concordancia con las calificaciones reales como en la estabilidad interna al evaluar textos narrativos. Esta baja fiabilidad fue evidente incluso en tareas simples, como la evaluación del título de un texto. El modelo por defecto, al no estar entrenado específicamente para la evaluación de textos ni restringido por tareas definidas, suele realizar acciones no deseadas, como introducir correcciones. Por ejemplo, puede añadir un título a un texto que no lo tenía originalmente y luego proceder a calificar ese título añadido por sí mismo.

Al examinar los resultados por categorías, se observaron desempeños particularmente deficientes en algunas de ellas. Por ejemplo, el valor medio de alfa para la categoría de ortografía y puntuación fue de apenas 0.159. El modelo por defecto mostró una debilidad clara al analizar la ortografía y puntuación en textos redactados en turco. Este hallazgo pone de relieve las notables mejoras alcanzadas en las categorías con bajo rendimiento inicial tras un proceso de entrenamiento especializado.

Eficacia de la retroalimentación

Cumplimiento de los criterios: La retroalimentación proporcionada por TAT fue evaluada conforme a los criterios establecidos para una retroalimentación efectiva. La herramienta demostró tasas de éxito superiores al 83 % en todos los criterios, destacándose especialmente en la entrega de retroalimentación orientada al rendimiento, constructiva y específica para la tarea.

Adecuación categorial: Solo alrededor del 8,12 % de las muestras de retroalimentación fueron consideradas inapropiadas para su categoría correspondiente, lo que demuestra el alto rendimiento de TAT al ofrecer comentarios dentro del contexto de cada categoría de la rúbrica, recordando eficazmente al alumnado los criterios pertinentes. Además, cabe destacar que las retroalimentaciones consideradas inapropiadas no se debieron a información alucinada, sino a la confusión generada por matices sutiles entre distintas categorías de la rúbrica.

Análisis comparativo con la literatura existente

En el estudio de Yavuz et al. (2024), se compararon los modelos de lenguaje ChatGPT y Bard para la evaluación de ensayos. ChatGPT se utilizó tanto en su modo por defecto como en una versión ajustada con el nivel de temperatura reducido a 0.2. Las puntuaciones asignadas por la IA se compararon con las de evaluadores humanos. Los resultados indicaron que tanto el ChatGPT por defecto como el ajustado, así como Bard, proporcionaron puntuaciones fiables. En particular, el ChatGPT ajustado mostró una concordancia muy alta con los evaluadores humanos. No obstante, en dicho estudio, los modelos de lenguaje no fueron específicamente entrenados para la tarea. El ajuste se limitó a modificar el valor de temperatura, lo cual restringe la variabilidad en las respuestas del modelo.

En nuestro estudio, por el contrario, no se realizaron ajustes de temperatura, y se utilizó ChatGPT en su versión por defecto para las comparaciones. Los resultados de ambos estudios divergen en cuanto al rendimiento del modelo predeterminado de ChatGPT. Consideramos que

Recibido: 09/10/2024 Aceptado: 12/09/2025

uno de los factores determinantes puede haber sido el idioma de los textos evaluados: mientras que un estudio empleó textos en inglés evaluados con una rúbrica en inglés, el nuestro utilizó textos en turco con una rúbrica en turco. Para respaldar esta hipótesis, se necesitan más investigaciones que comparen el rendimiento de los modelos en diferentes idiomas.

Otro factor que puede haber influido en los resultados dispares es el número de textos evaluados. En el estudio de Yavuz et al. (2024), se evaluaron solo tres textos, mientras que en nuestro estudio se evaluaron 114 textos. Observamos que, al aumentar el número de textos evaluados por ChatGPT, este comenzó a producir respuestas automáticas no deseadas y a aplicar patrones de calificación similares a textos cualitativamente distintos. Por ello, es posible que el otro estudio haya obtenido mejores resultados gracias a la evaluación de un número reducido de textos, con el apoyo de indicaciones específicas y una colaboración humano-IA adecuada. Sin embargo, sostenemos que un modelo entrenado específicamente para una tarea rinde mucho mejor cuando se requiere un trabajo intensivo.

En el estudio de Awidi (2024), se comparó la evaluación de 108 textos realizada por evaluadores humanos y por ChatGPT en su versión predeterminada. El coeficiente de correlación intraclase (ICC) para medidas individuales fue de 0.349, lo que indica baja concordancia, resultado consistente con los hallazgos de nuestro estudio. Awidi también observó que la concordancia mejoraba al considerar medidas promedio y recomendó la colaboración con IA en la evaluación de textos para lograr resultados más consistentes y reducir significativamente la carga de trabajo del profesorado.

En cuanto a la calidad de la retroalimentación proporcionada a los textos, Steiss et al. (2024) compararon los comentarios ofrecidos por humanos y por ChatGPT sobre producciones escritas de estudiantes. El estudio analizó 200 comentarios de evaluadores humanos y 200 de la IA. Los resultados mostraron que los evaluadores humanos fueron más eficaces en la mayoría de las categorías, excepto en la retroalimentación basada en criterios, donde ChatGPT mostró un mejor desempeño. A partir de esto, los autores concluyeron que ChatGPT puede ser útil en ausencia de un educador bien capacitado.

En nuestro estudio, se obtuvieron resultados muy positivos en cuanto

a la calidad de la retroalimentación proporcionada por la IA. La principal diferencia entre ambos estudios radica en si el modelo de lenguaje fue específicamente entrenado para el propósito. Mientras que el otro estudio utilizó un modelo por defecto, nosotros empleamos un modelo entrenado para la evaluación de textos y la provisión de retroalimentación. Nuestro estudio demostró que un modelo de lenguaje entrenado destaca en la entrega de retroalimentación efectiva, la cual se considera un factor que favorece el desarrollo del estudiante.

En esta línea, Escalante et al. (2023) realizaron un estudio para determinar cómo influye la retroalimentación de la IA y la retroalimentación humana en el rendimiento de la escritura estudiantil, así como qué tipo de evaluador prefieren los estudiantes. El estudio reveló que no hubo diferencias significativas en el rendimiento entre los grupos que recibieron retroalimentación de la IA y aquellos que recibieron retroalimentación humana, y que las preferencias del alumnado estaban divididas equitativamente entre ambos evaluadores.

Discusión

Los resultados de este estudio destacan el potencial de la colaboración humano-IA para calificar textos narrativos de manera fiable y objetiva, incluso en contextos que requieren evaluaciones subjetivas. Los altos niveles de concordancia y estabilidad alcanzados por TAT, un modelo GPT desarrollado específicamente para este estudio, demuestran que las herramientas de IA, cuando están suficientemente entrenadas, pueden igualar el desempeño humano tanto en la evaluación de textos como en la provisión de retroalimentación efectiva.

El fuerte potencial de la IA para apoyar los procesos de evaluación formativa resulta especialmente significativo en regiones densamente pobladas y aulas numerosas, ya que puede contribuir a prácticas de evaluación más coherentes y escalables para el alumnado, al tiempo que reduce la carga de trabajo del profesorado en el seguimiento y apoyo del desarrollo individual del estudiante. Esto, a su vez, podría contribuir a un proceso educativo de

Recibido: 09/10/2024 Aceptado: 12/09/2025

mayor calidad.

Impacto del entrenamiento de la IA

El estudio pone de relieve la necesidad de un entrenamiento especializado para mejorar la competencia de los modelos de IA en tareas específicas. Aunque ChatGPT destaca en el procesamiento general del lenguaje, el entrenamiento dirigido es crucial para tareas como la evaluación de textos narrativos. Sin restricciones específicas para la tarea, ChatGPT puede producir resultados inconsistentes, lo cual representa un problema tanto para la investigación científica como para las aplicaciones prácticas. Por lo tanto, los autores desaconsejan el uso de ChatGPT por defecto en tareas críticas y recomiendan emplear un modelo entrenado con fiabilidad demostrada.

No solo IA, sino colaboración humano-IA

La fuerza estadística de los resultados obtenidos por la IA en este estudio ofrece evidencia significativa a favor de su uso. No obstante, durante el proceso tanto de entrenamiento como de utilización del modelo, se observó que la IA podía cometer errores inesperados en áreas imprevistas.

Más allá de las dificultades inherentes a la tarea y de la influencia del juicio subjetivo en la evaluación de textos narrativos, ciertas desviaciones en las tasas de concordancia y el rendimiento temporal de tanto ChatGPT por defecto como TAT pueden explicarse por el fenómeno de la alucinación.

Por lo tanto, sostenemos que un proceso de evaluación completamente automatizado por IA, sin supervisión humana, sería altamente inapropiado. Más allá de la prevención de errores, la colaboración entre humanos e inteligencia artificial es fundamental para el desarrollo de un sistema que pueda mejorar continuamente y abordar eficazmente tareas variadas. La alimentación periódica del modelo con datos adecuados puede mejorar significativamente su rendimiento y hacerlo más competente para enfrentar situaciones diversas.

Limitaciones y recomendaciones para futuras investigaciones

En el presente estudio, los textos narrativos fueron construidos intencionadamente por los investigadores, siguiendo estrictamente una rúbrica predefinida, con omisiones específicas, imprecisiones deliberadas y criterios de puntuación predeterminados. Este enfoque metodológico permitió una evaluación controlada de la competencia del modelo en la interpretación y aplicación de los estándares de evaluación. No obstante, esta elección de diseño introduce limitaciones inherentes. En primer lugar, la ausencia de evaluadores humanos y la dependencia de textos generados artificialmente pueden restringir la autenticidad y variabilidad que suelen caracterizar las composiciones reales de los estudiantes. En consecuencia, los resultados obtenidos mediante este método podrían no reflejar plenamente el rendimiento potencial del modelo en contextos educativos auténticos del mundo real.

En relación con esto, el conjunto de datos estuvo compuesto por 114 textos estandarizados que, si bien promovieron condiciones controladas, podrían no representar adecuadamente la diversidad de perfiles estudiantiles ni las competencias escritas variables que se encuentran en entornos educativos a gran escala. Para superar estas limitaciones, futuras investigaciones podrían beneficiarse de la integración de textos auténticos producidos por estudiantes reales y de la inclusión de evaluadores humanos, con el fin de analizar comparativamente la coherencia de las puntuaciones y la consistencia temporal de modelos GPT personalizados como TAT. Además, ampliar tanto el tamaño de la muestra como la diversidad del conjunto de datos podría mejorar la evaluación de la generalizabilidad y aplicabilidad práctica del modelo.

Asimismo, las variaciones observadas entre este estudio y otros trabajos previos ponen de relieve la importancia de investigar cómo varía el rendimiento de los modelos lingüísticos de IA según el idioma utilizado. Por tanto, sería beneficioso iniciar nuevas investigaciones prácticas y experimentales en esta línea.

Recibido: 09/10/2024 Aceptado: 12/09/2025

Agradecimientos

Deseamos expresar nuestro más sincero agradecimiento a Iria Balayo por la minuciosa revisión lingüística de nuestro manuscrito y por sus valiosas sugerencias, que contribuyeron a mejorar la claridad y la calidad del texto.

Referencias bibliográficas

Alto, V. (2023). *Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation*. Packt Publishing.

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence, 6*, 100226. https://doi.org/10.1016/j.caeai.2024.100226

Brookhart, S. M. (2008). *How to give effective feedback to your students*. ASCD.

Burke, D., & Pieterick, J. (2010). *Giving students effective written feedback*. Open University Press.

Chan, C. K. Y., & Colloton, T. (2024). *Generative AI in higher education: The ChatGPT effect*. Routledge.

Dalton, G. (2024). *Artificial intelligence: Background, risks and policies*. Nova Science Publishers.

Elsayed, H. (2024). The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education. *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity, 9*(8), 11–23.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(57). https://doi.org/10.1186/s41239-023-00425-2

Fell Kurban, C., & Şahin, M. (2024). *The impact of ChatGPT on higher edu-*

cation. Emerald Publishing.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum, 12*(1), 44-58. https://doi.org/10.15294/elt.v12i1.64069

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback*. Routledge.

Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491–499). https://doi.org/10.5281/zenodo.12729868

Johannesson, P. (2024). *Writing your thesis with ChatGPT: Research, scholarship and academic writing in the age of generative AI*. Kindle Direct Publishing. https://writingyourthesiswithchatgpt.wordpress.com/

Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). *Enhancing student learning through effective formative feedback*. The Higher Education Academy.

Kolbjørnsrud, V. (2024). Designing the intelligent organization: Six principles for human-AI collaboration. *California Management Review, 66*(2), 44–64.

Krippendorff, K. H. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Ministerio Nacional de Educación de Turquía. (2024). *Ortaokul Türkçe dersi öğretim programı*. Millî Eğitim Bakanlığı.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech, 40*(1), 80–97. https://doi.org/10.1108/LHT-05-

2020-0113

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 610–625). Association for Computational Linguistics.

Venter, J., Coetzee, S. A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education,* 50(4), 516–536. https://doi.org/10.1080/02602938.2024.2415649

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology,* 56, 150–166. https://doi.org/10.1111/bjet.13494

Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning,* 1–21. https://doi.org/10.1080/09588221.2024.2422478

Dirección de contacto: Universidad Dokuz Eylul, Facultad de Educación, Departamento de Medición y Evaluación Educativa, Esmirna, Turquía. Correo electrónico: sait.cum@deu.edu.tr

APÉNDICE I. Características de la retroalimentación efectiva e inefectiva

Categoría	Características de la retroalimentación efectiva	Características de la retroalimentación inefectiva
Orientada al rendimiento	1. La retroalimentación se dirige hacia el desempeño, no hacia la persona. 2. Se enfoca en aspectos específicos del desempeño en lugar de comentarios generales.	1. Contiene sesgos hacia el estudiante e incluye afirmaciones dirigidas a su personalidad. 2. Utiliza comentarios generales que no son específicos del desempeño.
Clara y comprensible	1. La retroalimentación se expresa con palabras y estructuras gramaticales apropiadas para la edad o nivel de desarrollo del estudiante. 2. Especifica claramente lo que se espera y qué constituye un buen desempeño. 3. Debe ser lo suficientemente detallada y explicativa para evitar confusión.	1. Contiene expresiones técnicas y complejas que dificultan la comprensión del estudiante. 2. Utiliza frases vagas como "puedes hacerlo mejor" en lugar de especificar lo que se espera. 3. Es superficial y aleatoria, lo que dificulta saber qué se espera del estudiante.
Formativa (de desarrollo)	1. Incluye sugerencias para que el estudiante supere deficiencias y logre el desempeño esperado. 2. Puede recomendar tareas similares o estrategias que faciliten el aprendizaje autónomo. 3. Enfatiza lo que el estudiante debería hacer primero para mejorar desempeños futuros.	1. Enfatiza las deficiencias y carencias sin sugerir formas de superarlas.
Constructiva	1. Destaca tanto las fortalezas como las debilidades del desempeño. El buen desempeño también debe recibir retroalimentación. 2. Utiliza un lenguaje que motive al estudiante y apoye su autoestima. 3. Ofrece opciones sobre lo que puede hacer el estudiante, en lugar de imponer órdenes estrictas.	1. Usa un lenguaje condescendiente y afirmaciones que pasivizan al estudiante. 2. Incluye expresiones críticas o amenazantes que desmotivan al estudiante.
Específica para la tarea	1. La retroalimentación no debe contener afirmaciones generales; debe destacar puntos concretos del trabajo del estudiante y estar directamente relacionada con su contenido.	1. Contiene frases genéricas que podrían aplicarse a cualquier tarea similar, haciendo que la retroalimentación parezca repetitiva y estandarizada.

Recibido: 09/10/2024 Aceptado: 12/09/2025

APÉNDICE II. Criterios de la rúbrica y niveles de puntuación (Ministerio Nacional de Educación de Turquía, 2024)

Categoría	1 punto	2 puntos	3 puntos
Estructura de la página	El texto no está escrito en párrafos y está desorganizado visualmente en la página.	El texto está escrito en párrafos, pero las sangrías y/o los saltos de línea no están correctamente alineados.	El texto está escrito en párrafos con sangrías y saltos de línea adecuados, creando una página organizada visualmente.
Título	El texto no tiene título.	El texto tiene un título, pero no refleja el contenido o es un cliché común.	El título es relevante para el tema, refleja el contenido y resulta atractivo.
Estructura del texto	El texto carece de una o más secciones clave: introducción, desarrollo y conclusión.	El texto tiene introducción, desarrollo y conclusión, pero las transiciones entre las secciones son poco fluidas.	El texto contiene introducción, desarrollo y conclusión con relaciones lógicas y transiciones fluidas entre las secciones.
Personaje	Los personajes son mencionados solo por su nombre, sin información adicional.	Los personajes están nombrados, y se describen sus características físicas y/o psicológicas.	Los personajes están nombrados, se describen sus rasgos físicos y psicológicos, y se sugieren o explican sus emociones, perspectivas y actitudes que influyen en la narrativa.
Ambientación	Falta uno de los elementos (espacio o tiempo) o no están claros.	Se mencionan el espacio y el tiempo, pero no se proporcionan detalles.	El lugar está bien descrito con detalles visuales y auditivos, y el tiempo también está especificado, mostrando su impacto en otros elementos narrativos.
Trama	No hay una trama clara.	Existe una trama clara, pero las transiciones entre los eventos son débiles.	La trama es clara y las transiciones entre los eventos son sólidas.
Lenguaje y estilo	La mayoría de las oraciones son poco claras, sin cohesión semántica ni gramatical, y el vocabulario es muy limitado.	Las oraciones son claras y comprensibles, con cierta cohesión semántica y gramatical, aunque el vocabulario es limitado.	Las oraciones son claras y comprensibles, con cohesión semántica y gramatical adecuada, y el texto utiliza un vocabulario rico.
Ortografía y puntuación	El texto contiene 11 o más errores ortográficos y de puntuación.	El texto contiene entre 6 y 10 errores ortográficos y de puntuación.	El texto contiene un máximo de 5 errores ortográficos y de puntuación.

APÉNDICE III. Ejemplos de indicaciones utilizadas en el proceso de entrenamiento

Etapas	Ejemplos de instrucciones (prompts)
Definición del objetivo y evaluación inicial	Como profesor/a de lengua, evaluarás los textos narrativos de tus estudiantes y les proporcionarás retroalimentación formativa efectiva. Para ello, utilizarás la rúbrica y el documento de criterios de retroalimentación que te voy a cargar.
Introducción de los criterios	Vamos a analizar la categoría de ambientación. Al revisar los criterios, ¿ves algún ítem que podría resultarte difícil de evaluar? ¿En qué aspectos podrías encontrar dificultades?
Análisis de ejemplos	Si el evento en la historia ocurre en verano, el tiempo está claro; sin embargo, si no hay información específica, se debe asignar un puntaje de 2. Lo mismo se aplica al espacio: si se menciona que ocurre en una posada sin más detalles, también se asigna 2.
Carga de archivos de muestra	La sangría de párrafo se refiere a que la primera línea comienza más adentro. Ahora te cargaré un ejemplo sin sangría para que lo uses como referencia.
Práctica estructurada	Voy a cargar un texto. De acuerdo con lo que discutimos, quiero que evalúes todas las categorías de la rúbrica para este texto.
Evaluación final y confirmación	Ahora, describe los archivos que te cargué, resume las decisiones tomadas y especifica las reglas que seguirás durante la evaluación.

Recibido: 09/10/2024 Aceptado: 12/09/2025