e-ISSN 1135-9250

EDUTEC. Revista Electrónica de Tecnología Educativa

Número 89 – Septiembre 2024

Sección: Inteligencia artificial en la evaluación y la personalización del aprendizaje

Un análisis experimental de la relación entre las evaluaciones proporcionadas por la inteligencia artificial y las proporcionadas por los docentes en formación

An experimental analysis of the relationship between the evaluations of artificial intelligence and pre-service teachers

Héctor Galindo-Domínguez[1]; hector.galindo@ehu.eus;

Nahia Delgado[1]; nahia.delgado@ehu.eus;

Martín Sainz de la Maza[1]; martin.sainzdelamaza@ehu.eus;

Ernesto Expósito[2]; ernesto.exposito@univ-pau.fr;

Resumen

Uno de los beneficios potenciales de la inteligencia artificial (IA) es que puede permitir la optimización de las tareas de los docentes. Este estudio tuvo como objetivo analizar las posibles diferencias entre las evaluaciones realizadas por docentes en formación y las realizadas por diferentes IA. Participaron un total de 507 docentes en formación, a quienes se les proporcionó una rúbrica para evaluar 12 textos de distintos tipos y calidades. Los resultados mostraron cómo el desempeño de las IA en la evaluación de tareas escritas replicó con bastante precisión el funcionamiento de los docentes en formación, siendo ChatGPT la IA que mejor replicó el comportamiento de los docentes en formación, con una precisión cercana al 70% de la evaluación proporcionada por humanos. Del mismo modo, hubo diferencias mínimas en las evaluaciones realizadas por los docentes en formación según su género y año académico. Asimismo, las evaluaciones realizadas por los docentes en formación con mejor desempeño estuvieron más alineadas con las proporcionadas por la IA en comparación con los estudiantes con menor desempeño. Estos resultados son útiles, al destacar cómo la IA podría ser una herramienta de apoyo que guíe el conocimiento pedagógico de los docentes en formación en tareas de evaluación.

Palabras clave: Evaluación, Inteligencia Artificial, ChatGPT, Formación Docente

Abstract

One of the potential benefits of artificial intelligence (AI) is its ability to optimize teachers' tasks. The aim of this study was to analyze the possible differences between assessments carried out by pre-service teachers and those performed by various AI systems. A total of 507 pre-service teachers participated, and they were provided with a rubric to evaluate 12 texts of different types and quality. The results showed that AI performance in evaluating written tasks closely replicated the functioning of pre-service teachers, with ChatGPT being the AI that most accurately mirrored the teachers' evaluations, achieving approximately 70% precision compared to human assessments. Similarly, there were minimal differences in the assessments made by pre-service teachers based on gender and academic year. Moreover, evaluations conducted by higher-performing pre-service teachers were more aligned with those provided by AI, compared to those from lower-performing students. These findings are valuable, highlighting how AI could serve as a supportive tool to guide the pedagogical knowledge of pre-service teachers in assessment tasks.

Keywords: *Assessment, Artificial Intelligence, ChatGPT, Teacher Training*

[1] Universidad del País Vasco/Euskal Herriko Unibertsitatea (España)

[2] Université de Pau et des Pays de l'Adour (Francia)

DOI: https://doi.org/10.21556/edutec.2024.89.3509

Recibido: 02-08-2024
Aceptado: 06-09-2024

1. INTRODUCCIÓN

1.1. La evaluación en la formación docente: responsabilidades, desafíos y factores influyentes

En la Ley Orgánica 3/2020, la principal ley educativa en España, el artículo 91 detalla las funciones del profesorado, incluyendo la evaluación de los procesos de aprendizaje de los estudiantes. Trabajos previos, como el de Stiggins (2014), han demostrado que los docentes invierten entre un tercio y hasta la mitad de su tiempo profesional en tareas relacionadas con la evaluación y calificación, lo que muestra que es una tarea que consume mucho tiempo, especialmente con el aumento de la ratio de estudiantes por profesor (Ramesh y Kumar, 2022). Sin embargo, a pesar de ser una labor que requiere esfuerzo y dedicación, se ha observado que un aumento en la alfabetización evaluativa de los docentes tiene un impacto directo en los resultados de aprendizaje de los estudiantes (por ejemplo, Mellati y Khademi, 2018; Xu y Brown, 2016).

Considerando esta función que cualquier docente a nivel mundial realiza con sus estudiantes, enseñar a los futuros docentes nuevos conocimientos y habilidades para evaluar a sus futuros alumnos podría ser esencial para hacer la tarea de enseñanza lo más eficiente posible (Atjonen et al., 2022). En esta misma línea, estudios previos muestran importantes limitaciones en la manera en que se enseña a los docentes en formación a evaluar, ya sea porque es excesivamente teórico o porque está desconectado de las tareas diarias de un docente (Atjonen, 2017; DeLuca et al., 2019; Salama y Subahi, 2020). Como resultado, en muchos casos, los futuros docentes aplican estrategias de evaluación que se utilizaron con ellos cuando eran estudiantes (Hill et al., 2017).

Asimismo, la cantidad de conocimientos y habilidades de un docente en formación al evaluar trabajos podría estar condicionada por una serie de variables personales y académicas. Aunque la evidencia existente hasta la fecha es escasa, algunos estudios, como el de Salama y Subahi (2020), observaron que la alfabetización evaluativa de los docentes en formación era relativamente baja y similar, independientemente de su género, rendimiento académico o años de experiencia. Además, Lovorn y Reza (2011) observaron cómo la formación recibida puede influir en la forma en que se lleva a cabo la evaluación mediante rúbricas, y Deneen y Brown (2016) notaron cómo el rendimiento académico de los docentes en formación desempeña un papel determinante en la profundidad de una evaluación. No obstante, debemos considerar que el grado en que se detalla la evaluación de una tarea puede estar influenciado por el rendimiento académico, que a su vez podría deberse al impacto de otros factores socioemocionales, como la motivación de los estudiantes hacia la tarea o su estado emocional actual (Eklöf, 2010). Por esta razón, el grado de detalle en la evaluación de tareas podría ser un fenómeno multidimensional.

1.2. La integración de la inteligencia artificial en la evaluación digital

Con los avances tecnológicos de los últimos años, algunos de los métodos de evaluación que se están utilizando para abordar problemas como la alta ratio de estudiantes por docente, la instrucción personalizada y la reducción del tiempo excesivo, implican el uso de sistemas basados en inteligencia artificial (por ejemplo, Vij et al., 2019).

La inteligencia artificial (IA de aquí en adelante) es la capacidad de una máquina para replicar el comportamiento humano inteligente (Wang, 2019). Hoy en día, existe una abundancia de herramientas basadas en IA. Dentro del campo de la IA, un área en expansión es la IA generativa, que se entiende como una IA enfocada en crear contenido basado en las entradas del usuario. Algunas de las IA generativas más importantes en la actualidad son ChatGPT, de OpenAI, Copilot de Bing, de Microsoft, o Gemini, de Google, por nombrar algunas.

La IA generativa tiene aplicaciones potenciales en la evaluación de tareas estudiantiles, ya que diversos estudios han demostrado su precisión al proporcionar retroalimentación. Estos sistemas de IA, apoyados por el procesamiento de lenguaje natural, ofrecen respuestas personalizadas para complementar el esfuerzo de los docentes (Ocaña-Fernández et al., 2019; González-Calatayud et al., 2021). Jani et al. (2020) destacaron el papel de la IA en la evaluación formativa, utilizando el aprendizaje automático para monitorear el progreso de los estudiantes y mejorar las prácticas clínicas. De manera similar, la IA se ha aplicado en la formación médica (Mirchi et al., 2020), la educación en ingeniería (Samarakou et al., 2016; Liu et al., 2017) y las evaluaciones de programación (Grivokostopoulou et al., 2017), proporcionando retroalimentación automatizada basada en el rendimiento. Otros estudios (Rhienmora et al., 2011; Ouguengay et al., 2015; Ulum, 2020; Choi y McClenen, 2020) demuestran además la capacidad de la IA para calificar y evaluar habilidades en diversos campos.

Asimismo, ciertos estudios han comparado el efecto de utilizar sistemas basados en IA frente a no utilizarlos. Por ejemplo, Grivokostopoulou et al. (2017) realizaron una comparación entre los resultados obtenidos por una IA y las evaluaciones manuales realizadas por docentes para verificar la precisión de esta tecnología. Los resultados mostraron una correlación entre ambos, con solo ligeras diferencias observadas en trabajos excelentes donde los docentes tendían a sobrevalorar la tarea en comparación con las puntuaciones otorgadas por la IA. Estos resultados son coherentes con los obtenidos por Houtao et al. (2022), quienes observaron cómo la retroalimentación de la IA generativa podría ser tan útil como la de los docentes, aunque con algunas diferencias. En particular, en la corrección de textos, mientras que la retroalimentación del docente tendía a centrarse en la estructura y el contenido de la tarea, la retroalimentación de la IA era más detallada en vocabulario y gramática. Estos hallazgos subrayan el valor potencial de integrar ambas formas de retroalimentación para garantizar una evaluación más exhaustiva.

Como comentó Dillenbourg (2016), la transición de la educación tradicional a la digital no significa la obsolescencia de los docentes en el futuro. En lugar de debatir si la IA reemplazará a los docentes, Hrastinski et al. (2019) proponen reconocer los beneficios potenciales de la IA y cómo estas ventajas podrían redefinir su papel en el aula. Por lo tanto, en la evaluación educativa, los docentes continúan desempeñando un papel crucial en garantizar la utilización adecuada de la IA para los objetivos de medición y evaluación. Algunas de estas responsabilidades incluyen la creación de evaluaciones y el establecimiento de objetivos de aprendizaje, la contextualización de las preguntas de evaluación para hacerlas más relevantes y significativas para los estudiantes, la interpretación de los resultados para proporcionar retroalimentación personalizada adaptada a las fortalezas y debilidades de los estudiantes, y el monitoreo del progreso estudiantil, entre otras (Owan et al., 2023).

1.3. Propósito del estudio

Como se ha observado, la gran mayoría de los estudios mencionados anteriormente utilizan sistemas basados en IA (generativa) para proporcionar retroalimentación a los estudiantes, pero no comparan experimentalmente dicha retroalimentación con la que podrían proporcionar los docentes. Asimismo, los estudios que realizan análisis comparativos entre evaluaciones proporcionadas por IA y las realizadas por docentes en activo son mínimos (por ejemplo, Grivokostopoulou et al., 2017; Houtao et al., 2022), pero, según el conocimiento de los autores, no existen estudios que evalúen la relación entre las evaluaciones proporcionadas por IA generativa y las de los docentes en formación. Además, revisiones sistemáticas recientes han demostrado el creciente interés de la comunidad científica en el uso de sistemas de inteligencia artificial y aprendizaje automático para automatizar la puntuación de ensayos, con el fin de abordar el aumento de la proporción de estudiantes y las tareas que consumen mucho tiempo, como la retroalimentación y la calificación (Ramesh y Kumar, 2022). Por esta misma razón, y dado que la literatura sobre el tema es escasa, es imperativo determinar si dichas herramientas están o no preparadas para servir como una ayuda adicional en el conocimiento pedagógico del docente durante sus procesos de evaluación digital.

Basándose en estas necesidades, los objetivos de este estudio son:

- O1: Analizar si existen diferencias estadísticamente significativas entre las evaluaciones realizadas por la IA generativa y las evaluaciones realizadas por los docentes en formación sobre diferentes textos escritos.
- O2: Analizar si las diferencias en las evaluaciones entre la IA generativa y los docentes en formación dependen de su género.
- O3: Analizar si las diferencias en las evaluaciones entre la IA generativa y los docentes en formación dependen de su nivel de formación.
- O4: Analizar si las diferencias en las evaluaciones entre la IA generativa y los docentes en formación dependen de su rendimiento académico.

2. MÉTODO

2.1. Participantes

Un total de 507 estudiantes universitarios participaron en el presente estudio, con una edad promedio de 20.56 años (*DT* = 5.42). De ellos, 155 eran hombres, 348 mujeres y 4 personas que no se identificaron con ninguna categoría de género específica. En cuanto a la titulación académica, 130 estudiantes provenían del Grado en Educación Infantil, 327 del Grado en Educación Primaria y 50 de áreas relacionadas como Pedagogía o Educación Social. En términos de progreso académico, 172 estudiantes estaban en su primer año, 137 en su segundo, 168 en su tercero, 25 en su cuarto y 5 en su quinto año de estudios. Una parte de la muestra fue seleccionada por proximidad, compuesta por estudiantes de los investigadores involucrados en este estudio. Otra parte de la muestra fue seleccionada mediante la difusión de un mensaje institucional que invitaba a la participación de los estudiantes matriculados en los programas de grado ofrecidos por las tres Facultades de Educación de la Universidad del País Vasco. A pesar

de que la muestra no es probabilística, existe evidencia previa que indica que las muestras por conveniencia pueden arrojar resultados similares a los obtenidos de muestras aleatorizadas (por ejemplo, Coppock et al., 2018).

2.2. Instrumentos

Para recopilar todos los datos, se utilizaron dos instrumentos diferentes. En primer lugar, se solicitó a los participantes información sobre variables personales como el género, la edad, el año académico, el grado universitario que estaban cursando y su rendimiento académico. Para el objetivo O3, el nivel de formación se codificó en función del año académico del estudiante. Específicamente, los estudiantes de primer año se ubicaron en formación inicial, los de segundo año en formación intermedia y los de tercer y cuarto año en formación avanzada de grado universitario. Los estudiantes de cuarto año se agruparon con los de tercer año debido a que tienen menos asignaturas. Asimismo, para el objetivo O4, el rendimiento académico se midió a través de la media aritmética de las calificaciones obtenidas por los estudiantes en el año anterior al que estaban matriculados. Este valor se conoce a partir del expediente académico que se les envía al final de cada año. En base a sus puntuaciones, se crearon grupos de bajo (percentil < 33), medio (percentil 34 a 66) y alto (percentil > 67) rendimiento académico.

En segundo lugar, se generaron una serie de textos *ad-hoc* escritos en español utilizando ChatGPT 3.5, y posteriormente fueron supervisados por expertos en educación. Estos textos fueron evaluados por docentes en ejercicio (n = 3). El *prompt* utilizado para generar los textos fue el siguiente:

> *Escribe un texto [tipo de texto] de 5 a 10 líneas escrito por un estudiante de 10 años sobre un tema de libre elección, con contenido, organización, vocabulario, coherencia y cohesión de [tipo de calidad].*

En total, se desarrollaron 12 textos de diferentes tipos (3 textos descriptivos, 3 textos narrativos, 3 textos argumentativos y 3 textos instructivos) y calidades (4 excelentes, 4 normales y 4 significativamente mejorables). Los textos generados se encuentran en el Anexo I. Además, se creó una rúbrica *ad-hoc* con ChatGPT 3.5, también supervisada por docentes en ejercicio (n = 3).

En cuanto a la evaluación de ensayos mediante IA, varios estudios han analizado los principales criterios que los docentes deben utilizar. En este sentido, algunos criterios están relacionados con características estadísticas, como la longitud del ensayo en relación con el número de palabras, la longitud de las oraciones o la longitud promedio de las palabras (Contreras et al., 2018; Kumar et al., 2019; Mathias y Bhattacharyya, 2018). Respecto a los criterios de estilo o basados en la sintaxis, los más relevantes incluyen la estructura de la oración, la puntuación, la gramática, los operadores lógicos y el vocabulario (Cummins et al., 2016; Darwish y Mohamed, 2020; Ke et al., 2019). Asimismo, en cuanto a los criterios basados en el contenido, los más relevantes son la cohesión entre oraciones en un documento, la relevancia de la información, la corrección y la consistencia (Dong et al., 2017).

Basándose en estos criterios, se construyó la herramienta *ad-hoc* con 4 criterios diferentes (Contenido, Organización, Vocabulario, Coherencia y Cohesión) y 4 niveles de logro (excelente, bueno, regular y pobre). Por lo tanto, cada texto podría obtener una puntuación máxima de 16 puntos (4 criterios x 4 niveles de logro), aunque, para fines prácticos y alinearse con el sistema

educativo español, estos 16 puntos se ponderaron en una escala de 10, siendo 10 la máxima calificación para cada texto (equivalente a 16 puntos no ponderados). Los niveles de logro fueron establecidos por ChatGPT basados en criterios derivados de trabajos empíricos previos ya mencionados. Estos niveles fueron revisados por los investigadores, mostrando un alto acuerdo entre los proporcionados por la IA y los encontrados en otras rúbricas de evaluación de trabajos escritos (por ejemplo, Gobierno de Terranova y Labrador, 2014).

Esta rúbrica, recogida en el Anexo II, fue creada para ayudar a los docentes en formación a evaluar los textos generados. Todos los estudiantes participantes habían recibido formación específica sobre cómo usar rúbricas de evaluación, ya que a lo largo de su grado cursan asignaturas de didáctica general y didáctica específica en las que se les enseña a diseñar, interpretar y aplicar rúbricas de evaluación. Este aspecto es importante, ya que estudios previos han demostrado que un docente en formación sin capacitación en el uso de rúbricas emite una evaluación tan subjetiva como un docente en formación sin ninguna herramienta de evaluación (Lovorn y Reza, 2011). La fiabilidad de los diferentes textos usando la rúbrica de evaluación se muestra en la Tabla 1. Como se puede observar, dado que todos los valores están por encima del punto de corte de $\alpha > .70$ (Tavakol y Dennick, 2011), se puede asumir que existe una buena consistencia interna en los criterios utilizados para la evaluación de cada texto.

Tabla 1

Índices de fiabilidad para cada texto

Orden	Texto	Caldiad	Fiabilidad
1	Texto descriptivo	Significativamente mejorable	.792
2	Texto argumentativo	Normal	.801
3	Texto instructivo	Significativamente mejorable	.741
4	Texto narrativo	Excelente	.904
5	Texto descriptivo	Excelente	.904
6	Texto argumentativo	Significativamente mejorable	.781
7	Texto instructivo	Excelente	.922
8	Texto narrativo	Normal	.816
9	Texto descriptivo	Normal	.857
10	Texto argumentativo	Excelente	.875
11	Texto instructivo	Normal	.869
12	Texto narrativo	Significativamente mejorable	.814

Finalmente, se solicitó a cada una de las IAs analizadas que evaluara los diferentes textos utilizando el siguiente *prompt*:

> *El siguiente texto ha sido producido por un estudiante de 10 años. Considerando los criterios de contenido, organización, vocabulario, y coherencia y cohesión, proporciona a cada texto una calificación del 0 al 10: [Texto elaborado por ChatGPT].*

2.3. Procedimiento

El proceso comenzó con la creación de los diferentes textos. Estos textos fueron elaborados utilizando ChatGPT, siguiendo la estructura: Escribe un [tipo de texto] de 5 a 10 líneas escrito por un estudiante de 10 años sobre un tema de libre elección, con contenido, organización, vocabulario, coherencia y cohesión de [tipo de calidad]. Estos textos fueron supervisados por expertos en educación para detectar errores de coherencia, léxicos y gramaticales. Posteriormente, los textos fueron convertidos digitalmente a formularios de Google (*Google Forms*), y la rúbrica para evaluar cada texto también se añadió para permitir a los estudiantes evaluar cada criterio (contenido, organización, vocabulario, y coherencia y cohesión).

Inicialmente, se seleccionó un subconjunto de la muestra de los estudiantes del grupo de los investigadores. Asimismo, el equipo de investigación solicitó permiso a la decanatura para invitar a participar a todos los demás estudiantes que no estaban bajo su instrucción. En este proceso, se siguieron estrictamente todas las medidas éticas aceptadas en la Declaración de Helsinki. En ambos casos, se requería que los participantes revisaran y aceptaran los procedimientos y términos de participación antes de aportar cualquier dato, garantizando el cumplimiento de los estándares éticos. Este acuerdo incluía informar a los participantes sobre los objetivos del estudio, el tiempo de respuesta anticipado, el manejo confidencial y anónimo de los datos, y la naturaleza voluntaria de la participación, que incluía la opción de retirar sus respuestas durante la encuesta. Los datos del grupo de estudiantes de los investigadores se recopilaron durante el horario laboral regular, mientras que las respuestas de los estudiantes externos se aceptaron en cualquier momento. Finalmente, tras completar el análisis del estudio, se envió un informe que resumía los hallazgos clave a aquellos que expresaron su deseo de recibir los resultados.

2.4. Análisis de datos

El proceso de análisis de datos se llevó a cabo completamente utilizando el software estadístico *SPSS Statistics 27*. Inicialmente, considerando las respuestas de todos los participantes, se calcularon las puntuaciones para cada texto. Como se mencionó anteriormente, aunque cada texto podía recibir un máximo de 16 puntos (4 criterios x 4 niveles de logro), para facilitar la interpretación, estos 16 puntos se ponderaron a 10 (una puntuación de 10 equivale a 16 puntos en la rúbrica), dado que, en el sistema educativo español, un 10 representa la máxima calificación.

Luego, para abordar el primer objetivo, se calcularon las medias aritméticas y las desviaciones estándar para cada grupo (IA vs. docentes en formación). Posteriormente, para identificar posibles diferencias significativas, se realizó una prueba t de Student complementada con la d de Cohen para conocer el tamaño del efecto. Además, se calculó el porcentaje de precisión, que se define como el número de textos que se desviaron en menos de 1 punto de las medias aritméticas de los docentes en formación, dividido por el número total de textos. Si la precisión fuera del 100%, significaría que las evaluaciones proporcionadas por la IA coincidirían estrechamente (en todos los casos, con una desviación de menos de 1 punto) con las evaluaciones dadas por los docentes en formación.

Para abordar el segundo objetivo, se recalcularon las medias aritméticas y las desviaciones estándar para cada grupo, seguidas de un ANOVA de un factor para examinar posibles

diferencias entre los grupos. Este análisis se acompañó de una prueba post-hoc de Tukey para identificar entre qué grupos se hallaban diferencias significativas. Para el tercer objetivo, se siguió el mismo procedimiento que para el segundo.

Finalmente, para el cuarto objetivo, dado el rendimiento académico de los docentes en formación en una escala, se formaron tres grupos diferentes: bajo rendimiento (percentil 1 a 33), rendimiento medio (percentil 34 a 66) y alto rendimiento (percentil 67 a 99). Una vez clasificados, se realizó otro ANOVA de un solo factor para identificar posibles diferencias significativas. Este análisis se complementó con un análisis post-hoc utilizando la prueba de Tukey para determinar entre qué grupos se encontraron diferencias significativas.

3. RESULTADOS

Primero, en relación con el objetivo 1, para determinar si había diferencias significativas entre las puntuaciones asignadas a los diferentes textos utilizando la rúbrica de evaluación y las puntuaciones asignadas por las IAs, se realizaron diversas pruebas T.

Como se observa en la Tabla 2, de los 12 textos, se encontraron diferencias estadísticamente significativas en solo 4 de ellos entre las calificaciones proporcionadas por los docentes en formación y las IAs. Sin embargo, vale la pena señalar que, aunque no había diferencias significativas en la mayoría de los casos, en aquellos donde sí se encontraron diferencias, el tamaño del efecto fue grande [Argumentativo Excelente, $p = .036$, $d = 1.25$; Argumentativo Normal, $p = .029$, $d = 1.56$; Narrativo Excelente, $p < .001$, $d = .70$; Narrativo Normal, $p = .046$, $d = 1.43$].

También se analizó la precisión de varias IAs en replicar la capacidad evaluativa de los docentes en formación. El método utilizado para calcular la precisión de la IA consistió en contar el número de textos evaluados por la IA con una desviación de menos de 1 punto de la evaluación dada por los docentes en formación y dividirlo por el número total de textos (12). La desviación de 1 punto se seleccionó de manera arbitraria, ya que se entiende que puede haber un buen acuerdo entre la evaluación proporcionada por la IA y la evaluación dada por los docentes en formación cuando existe una desviación de menos de 1 punto de 10 entre las dos puntuaciones. Los resultados revelaron que ChatGPT fue la IA entre las analizadas que mejor replicó las calificaciones proporcionadas por los docentes en formación (8/12 = 66.66% de precisión), seguida de Gemini (7/12 = 58.33% de precisión), y Copilot de Bing en la peor posición (6/12 = 50% de precisión).

Tabla 2

Resultados principales de las pruebas T

Tipo	Calidad	Inteligencia Artificial				Docentes en Formación	p	d
		ChatGPT	Bing	Gemini	Total IA			
Descriptivo	Excelente	8+	9.25+	9+	8.75 (.661)+	8.95 (1.46)	(ns)	-
	Normal	6+	8.5	7+	7.16 (1.25)+	6.75 (1.77)	(ns)	-
	Bajo	4+	4.5+	6	4.83 (1.04)+	4.95 (1.43)	(ns)	-
Argumentativo	Excelente	9	9	8+	8.66 (.577)	7.08 (1.68)	*	1.25
	Normal	7	8.5	7	7.5 (.866)	5.52 (1.56)	*	1.56
	Bajo	6	4+	7	5.66 (1.52)	4.52 (1.32)	(ns)	-
Instructivo	Excelente	9+	8.75+	9+	8.91 (.144)+	9.08 (1.37)	(ns)	-
	Normal	5	8.75	8+	7.25 (1.98)+	7.11 (1.63)	(ns)	-
	Bajo	4+	3.5	5+	4.16 (.763)+	4.61 (1.44)	(ns)	-
Narrativo	Excelente	9+	8.75+	9+	8.91 (.144)+	8.08 (1.65)	***	.70
	Normal	7+	8.75	8	7.91 (.877)	6.07 (1.59)	*	1.43
	Bajo	3+	2.75+	5	3.58 (1.23)+	3.18 (1.09)	(ns)	-
Precisión		66.66%	50%	58.33%	66.66%			

Nota. +, La puntuación de las IAs está por debajo a 1 punto de desviación de la media de la evaluación proporcionada por los docentes en formación; * p < .05; ** p < .01; *** p < .001; (ns), no signitivativo.

Posteriormente, en relación con el objetivo 2, se realizaron diversas pruebas ANOVA de un factor, considerando tres grupos diferentes: evaluaciones proporcionadas por la IA, evaluaciones proporcionadas por docentes en formación masculinos y evaluaciones proporcionadas por docentes en formación femeninas. Como se observa en la Tabla 3, las diferencias entre géneros fueron mínimas, ya que solo en 3 de los 12 textos se encontraron diferencias significativas basadas en el género. De estas 3 diferencias por género, en 2 de ellas, las mujeres tuvieron una puntuación más similar a la proporcionada por la IA, mientras que en solo un caso los hombres tuvieron una puntuación más similar a la proporcionada por la IA. En los casos restantes, no hubo diferencias estadísticamente significativas basadas en el género.

A partir de este análisis, se calculó la precisión como el número de textos de los docentes en formación que tuvieron una media aritmética inferior a 1 punto en comparación con la media aritmética proporcionada por las IAs generativas. En este caso, la precisión para ambos géneros respecto a las evaluaciones proporcionadas por las IAs generativas fue del 66.6% (8/12 textos).

Tabla 3

Principales resultados de la ANOVA de un factor en base al género de los docentes en formación

Tipo	Calidad	Total IA (3)	Género		p	Post-Hoc
			Masculino (1)	Femenino (2)		
Descriptivo	Excelente	8.75 (.661)	8.72 (1.54)+	9.04 (1.42)+	(ns)	-
	Normal	7.16 (1.25)	6.62 (1.64)+	6.81 (1.82)+	(ns)	-
	Bajo	4.83 (1.04)	5.06 (1.49)+	4.90 (1.41)+	(ns)	-
Discutidor	Excelente	8.66 (.577)	6.87 (1.50)	7.17 (1.75)	(ns)	-
	Normal	7.5 (.866)	5.78 (1.63)	5.41 (1.51)	**	2<1; 1<3; 2<3
	Bajo	5.66 (1.52)	4.56 (1.37)+	4.50 (1.30)	(ns)	-
Instructivo	Excelente	8.91 (.144)	8.98 (1.36)+	9.13 (1.38)+	(ns)	-
	Normal	7.25 (1.98)	6.89 (1.34)+	7.21 (1.74)+	(ns)	-
	Bajo	4.16 (.763)	4.56 (1.43)+	4.63 (1.45)+	(ns)	-
Narrativo	Excelente	8.91 (.144)	7.81 (1.55)	8.20 (1.68)+	*	1<2
	Normal	7.91 (.877)	5.89 (1.47)	6.15 (1.64)	*	(ns)
	Bajo	3.58 (1.23)	3.29 (1.21)+	3.13 (1.04)+	(ns)	-
	Precisión		8/12 (66.6%)	8/12 (66.6%)		

Nota. +, La puntuación de las IAs está por debajo a 1 punto de desviación de la media de la evaluación proporcionada por los docentes en formación; * p < .05; ** p < .01; *** p < .001; (ns), no significativo. Post-Hoc realizado a través de la prueba de Tukey.

A continuación, en relación con el objetivo 3, se realizaron varias pruebas ANOVA de un solo factor, considerando 4 grupos: las IAs, docentes en formación que comienzan su carrera universitaria (año 1), docentes en formación en medio de su carrera universitaria (año 2) y docentes en formación que concluyen su carrera universitaria (años 3 y 4). En términos generales, los análisis recogidos en la Tabla 4 revelan que las diferencias observadas entre los diversos grupos son mínimas, siendo en todos los casos 8 de las 12 evaluaciones (66.6% de precisión) similares a las proporcionadas por las IAs generativas.

Estos resultados pueden confirmar que, independientemente de la etapa en la que se encuentre el docente en formación, su capacidad para calificar y evaluar no será muy diferente. Las únicas excepciones fueron los textos descriptivos de baja calidad, donde los docentes en formación que concluían sus estudios sobrestimaron la puntuación proporcionada por sus compañeros y por la IA (p < .001), así como en los textos instruccionales excelentes, donde también sobreestimaron la puntuación proporcionada (p = .004).

Tabla 4

Principales resultados de la ANOVA de un factor en base al nivel de formación de los docentes en formación

Tipo	Calidad	IA Total (4)	Nivel de formación			p	Post-Hoc
			Inicial (1)	Medio (2)	Finalizando (3)		
Descriptivo	Excelente	8.75 (.661)	8.82 (1.53)+	9.01 (1.20)+	9.01 (1.56)+	(ns)	-
	Normal	7.16 (1.25)	6.52 (1.73)+	6.78 (1.75)+	6.92 (1.81)+	(ns)	-
	Bajo	4.83 (1.04)	4.76 (1.35)+	4.61 (1.19)+	5.34 (1.57)+	***	1<3; 2<3
Argumentativo	Excelente	8.66 (.577)	7.04 (1.65)	7.18 (1.63)	7.06 (1.76)	(ns)	-
	Normal	7.5 (.866)	5.69 (1.54)	5.48 (1.48)	5.41 (1.62)	*	(ns)
	Bajo	5.66 (1.52)	4.44 (1.28)	4.47 (1.23)	4.62 (1.40)	(ns)	-
Instructivo	Excelente	8.91 (.144)	8.79 (1.57)+	9.34 (1.00)+	9.16 (1.37)+	**	1<2; 1<3
	Normal	7.25 (1.98)	7.16 (1.74)+	7.07 (1.57)+	7.11 (1.59)+	(ns)	-
	Bajo	4.16 (.763)	4.81 (1.47)+	4.42 (1.23)+	4.57 (1.53)+	(ns)	-
Narrativo	Excelente	8.91 (.144)	8.04 (1.61)+	8.10 (1.70)+	8.10 (1.70)+	(ns)	-
	Normal	7.91 (.877)	6.16 (1.63)	5.96 (1.46)	6.07 (1.64)	(ns)	-
	Bajo	3.58 (1.23)	3.16 (1.10)+	3.07 (1.05)+	3.26 (1.11)+	(ns)	-
Precisión			8/12 (66.6%)	8/12 (66.6%)	8/12 (66.6%)		

Nota. +, La puntuación de las IAs está por debajo a 1 punto de desviación de la media de la evaluación proporcionada por los docentes en formación; * p < .05; ** p < .01; *** p < .001; (ns), no significativo. Post-Hoc realizado a través de la prueba de Tukey.

Finalmente, en relación con el objetivo 4, se realizaron varias pruebas ANOVA de un solo factor, considerando 4 grupos: las IAs, docentes en formación con bajo rendimiento académico (percentil < 33), docentes en formación con rendimiento académico medio (percentil 34 a 66) y docentes en formación con alto rendimiento académico (percentil > 67). Los resultados, como se observa en la Tabla 5, mostraron que los docentes en formación con un rendimiento académico más alto fueron evaluadores más precisos (9 de 12 textos, 75%) de las calificaciones proporcionadas por la IA, en contraste con los docentes en formación con rendimiento medio (8 de 12 textos, 66.6%) o bajo rendimiento académico (7 de 12 textos, 58.3%).

Asimismo, como se puede observar en las diferencias de medias presentadas en la Tabla 5, el número de textos con puntuaciones más altas aumenta con un mayor rendimiento académico. Por lo tanto, se puede observar que los estudiantes con un rendimiento más alto tienden a evaluar los textos escritos de manera más favorable en comparación con los estudiantes con un rendimiento académico más bajo.

Además, se puede ver que las IAs analizadas tendieron a sobreestimar las puntuaciones en la mayoría de los textos, independientemente del rendimiento académico de los estudiantes. Así, el número de textos sobreestimados por la IA en comparación con las evaluaciones de los estudiantes con bajo rendimiento fue de 10 de 12 textos (5 de ellos con más de 1 punto de

desviación de la media), en el caso de estudiantes con rendimiento medio fue de 8 de 12 textos (4 de ellos con más de 1 punto de desviación de la media), y en el caso de estudiantes con alto rendimiento fue de 7 de 12 textos (3 de ellos con más de 1 punto de desviación de la media). Estos datos demuestran cómo el rendimiento académico podría ser un factor importante en la generación de evaluaciones más alineadas con las proporcionadas por la IA.

Tabla 5

Principales resultados de la ANOVA de un factor en base al rendimiento académico de los docentes en formación

Tipo	Calidad	IA Total (4)	Rendimiento académico						p	Post-Hoc
			Bajo P<33 (1)		Medio P 34-66 (2)		Alto P>67 (3)			
			M (*DT*)	Dif M[1]	M (*DT*)	Dif M[1]	M (*DT*)	Dif M[1]		
Descriptivo	Excelente	8.75 (.661)	8.72 (1.52)+	.03	8.88 (1.50)+	-.13	9.35 (1.24)+	-.60	**	1<3; 2<3
	Normal	7.16 (1.25)	6.41 (1.73)+	.75	6.84 (1.81)+	.32	7.08 (1.69)+	.08	**	1<3
	Bajo	4.83 (1.04)	4.74 (1.39)+	.09	4.84 (1.39)+	-.01	5.39 (1.48)+	-.56	***	1<3; 2<3
Argumentativo	Excelente	8.66 (.577)	6.87 (1.64)	1.79	7.11 (1.75)	1.55	7.34 (1.61)	1.32	*	(ns)
	Normal	7.5 (.866)	5.48 (1.54)	2.02	5.43 (1.56)	2.07	5.71 (1.56)	1.79	(ns)	-
	Bajo	5.66 (1.52)	4.40 (1.36)	1.26	4.44 (1.19)	1.22	4.80 (1.40)+	.86	*	1<3
Instructivo	Excelente	8.91 (.144)	8.94 (1.40)+	-.03	9.06 (1.43)+	-.15	9.31 (1.24)+	-.40	(ns)	-
	Normal	7.25 (1.98)	7.05 (1.59)+	.20	6.99 (1.68)+	.26	7.37 (1.60)+	-.12	(ns)	-
	Bajo	4.16 (.763)	4.48 (1.47)+	-.32	4.60 (1.37)+	-.44	4.79 (1.49)+	-.63	(ns)	-
Narrativo	Excelente	8.91 (.144)	7.68 (1.68)	1.23	8.07 (1.64)+	.84	8.65 (1.46)+	.26	***	1<3; 2<3
	Normal	7.91 (.877)	5.78 (1.50)	2.13	6.11 (1.64)	1.8	6.40 (1.57)	1.51	***	1<3
	Bajo	3.58 (1.23)	3.20 (1.11)+	.38	3.10 (1.04)+	.48	3.25 (1.15)+	.33	(ns)	-
	Precisión		7/12 (58.3%)		8/12 (66.6%)		9/12 (75%)			

Nota. +, La puntuación de las IAs está por debajo a 1 punto de desviación de la media de la evaluación proporcionada por los docentes en formación; Dif M, La diferencia entre la media aritmética de la IA y la media aritmética del grupo. Un valor positivo más alto indica una mayor subestimación por parte del grupo en comparación con la evaluación de la IA, mientras que un valor negativo más alto indica una mayor sobreestimación por parte del grupo en comparación con la evaluación de la IA.; * p < .05; ** p < .01; *** p < .001; (ns), no significativo. Post-Hoc realizado a través de la prueba de Tukey.

4. DISCUSIÓN Y CONCLUSIONES

La evaluación es una de las principales tareas que cualquier docente realiza en sus funciones profesionales (Ley Orgánica 3/2020). Esta tarea, a pesar de ser compleja y consumir mucho tiempo debido a las ratios en el aula (por ejemplo, Ramesh y Kumar, 2022), puede aportar mejoras significativas en los procesos de aprendizaje de los estudiantes (Mellati y Khademi, 2018; Xu y Brown, 2016). Ante este dilema y con la introducción de nuevas tecnologías en las aulas, una alternativa posible para abordar este problema es emplear herramientas basadas en IA que permitan una replicación válida y confiable de las evaluaciones realizadas por los docentes.

Tomando esta idea como punto de partida, el objetivo principal de este estudio ha sido determinar si las evaluaciones proporcionadas por diferentes IAs generativas son fieles a las dadas por los docentes en formación, así como entender si existe alguna variable, como el género, el nivel de formación o el rendimiento académico, que pueda influir en la fidelidad de una evaluación en comparación con la proporcionada por una IA generativa.

Los resultados mostraron que las diferentes IAs analizadas son capaces de replicar bastante bien los patrones de los docentes en formación al evaluar tareas escritas, siendo ChatGPT la IA que obtuvo la mayor precisión (cerca del 70% de acuerdo con la evaluación de los docentes en formación) y Copilot de Bing la que presentó la menor precisión (50% de acuerdo con la evaluación de los docentes en formación). Estos resultados son consistentes con la limitada literatura previa sobre este tema, que generalmente revela una concordancia entre la retroalimentación proporcionada por la IA y la retroalimentación dada por los docentes en servicio, encontrando pequeñas diferencias entre ambos grupos (Grivokostopoulou et al., 2017; Houtao et al., 2022).

Además, los resultados han mostrado que este grado de acuerdo, en términos generales, es idéntico a lo que indica la literatura limitada sobre este tema, independientemente del género y del nivel de formación del docente en formación. Como excepción, se encontraron diferencias significativas claras solo en las evaluaciones dadas según el rendimiento académico de los docentes en formación, donde aquellos con un rendimiento académico más alto proporcionaron evaluaciones más alineadas con las ofrecidas por la IA en comparación con aquellos con un rendimiento académico más bajo. Estos resultados también son parcialmente consistentes con la limitada literatura sobre este tema. Específicamente, de acuerdo con el presente estudio, Salama y Subahi (2020) también observaron que el género y el nivel de formación eran variables que tenían poca influencia en el conocimiento y habilidades de evaluación, mientras que trabajos como los de Deneen y Brown (2016) mostraron cómo el rendimiento académico de los docentes en formación influía significativamente en la profundidad de la evaluación realizada. Sin embargo, los hallazgos del presente estudio contradicen los de Salama y Subahi (2020), quienes observaron que el rendimiento académico no era una variable influyente en la evaluación de conocimientos y habilidades.

4.1. Implicación teóricas y prácticas

Estos resultados tienen importantes implicaciones teóricas y prácticas que deben ser discutidas. En primer lugar, los hallazgos de este estudio pueden ser valiosos para la comunidad científica,

ya que podrían contribuir a ampliar la comprensión actual sobre el grado de concordancia entre las evaluaciones proporcionadas por educadores y las realizadas por sistemas basados en IA.

En segundo lugar, estos resultados pueden ser relevantes para los docentes universitarios en áreas relacionadas con la educación, ya que subrayan el interés potencial en capacitar a los futuros docentes en tecnologías digitales, como el uso de inteligencia artificial, big data o analíticas de aprendizaje, para optimizar los recursos temporales y proporcionar asistencia y monitoreo personalizados a sus estudiantes.

En Europa, se han realizado esfuerzos significativos para desarrollar un marco para la conceptualización y desarrollo de la competencia digital de los docentes, siendo el modelo DigCompEdu el principal referente (Redecker, 2017). Este marco destaca la relevancia de emplear tecnologías digitales en la formación docente para llevar a cabo tareas evaluativas dentro de la cuarta competencia "Evaluación". Esta competencia está orientada a mejorar las estrategias de evaluación, analizar la evidencia de aprendizaje y proporcionar retroalimentación mediante tecnologías digitales. En este sentido, el uso de tecnologías basadas en IA podría contribuir parcialmente a abordar esta competencia. Asimismo, estudios previos muestran cómo la mejora en la competencia digital docente podría tener efectos importantes en la gestión del tiempo y la autoeficacia docente, variables clave para reducir el estrés y la incomodidad generados por las tareas profesionales (Galindo-Domínguez y Bezanilla, 2021).

Finalmente, estos resultados pueden ser útiles para los docentes en ejercicio, ya que destacan el creciente potencial de los sistemas basados en inteligencia artificial como herramientas para evaluar el trabajo escrito de los estudiantes. Si bien estudios previos han demostrado que los docentes de educación primaria y secundaria actualmente utilizan herramientas basadas en IA principalmente para fines de creación de contenido, como textos o imágenes (Galindo-Domínguez et al., 2024), podría ser interesante añadir formación específica sobre cómo emplear la IA para la evaluación y el monitoreo de estudiantes como parte del proceso de desarrollo profesional continuo en las instituciones educativas. Como mencionan diferentes autores (por ejemplo, Kasneci et al., 2023; Owan et al., 2023), el uso de la IA en el ámbito educativo requiere que tanto docentes como estudiantes desarrollen un conjunto de competencias necesarias para comprender la tecnología, aprovechar sus potencialidades y reconocer sus limitaciones.

4.2. Limitaciones y prospectiva

El presente estudio tiene varias limitaciones que deben tenerse en cuenta al interpretar los resultados. La primera limitación se refiere a la muestra. Aunque esta es relativamente grande, los resultados podrían variar si los mismos participantes se evaluaran en unos años cuando estén en activo. Por ello, futuros estudios podrían replicar la metodología utilizada en este estudio, empleando a docentes en servicio de diferentes etapas educativas como participantes, para observar si los resultados obtenidos son similares o no. Además, podría ser interesante comparar las evaluaciones de docentes novatos, educadores de media carrera y docentes veteranos, ya que esto podría revelar importantes perspectivas sobre cómo la experiencia docente puede influir en los procedimientos de evaluación. Esta idea se justifica en que, mientras algunos estudios sugieren que la cantidad de experiencia docente podría influir en la alfabetización evaluativa (por ejemplo, Spear-Swerling et al., 2005), otros apuntan en la dirección opuesta (por ejemplo, Bagsao y Peckley, 2020; Salama y Subahi, 2020).

Asimismo, el presente estudio solo consideró la evaluación de trabajos escritos, lo que imposibilita determinar si el nivel de acuerdo entre las evaluaciones proporcionadas por la IA y los docentes en formación también podría ocurrir en trabajos presentados en otros formatos, como audio, video, imágenes o ecuaciones matemáticas. Aunque es más complejo para el estado actual de los sistemas de IA, futuros estudios podrían intentar replicar la metodología utilizada, pero evaluando tareas en formatos diferentes al escrito.

Finalmente, los textos escritos fueron generados por la IA bajo la instrucción de asumir el papel de un estudiante de 10 años. Sin embargo, las diferencias en los resultados obtenidos podrían surgir si se utilizaran textos reales escritos por estudiantes de 10 años. Por esta razón, futuros estudios podrían replicar la metodología del presente trabajo, pero utilizando textos redactados por estudiantes reales.

5. REFERENCIAS

Atjonen, P. (2017). Development of teacher assessment literacy in comprehensive schools – Views from the curriculum analysis. *Kriteerit Puntarissa, 74*, 132–169.

Atjonen, P., Pöntinen, S., Kontkanen, S., & Ruotsalainen, P. (2022). In Enhancing Preservice Teachers' Assessment Literacy: Focus on Knowledge Base, Conceptions of Assessment, and Teacher Learning. *Frontiers in Education, 7,* 1-12. https://doi.org/10.3389/feduc.2022.891391

Bagsao, J., & Peckley, M.K. (2020). Assessment Literacy of Public Elementary School Teachers in the Indigenous Communities in Northern Philippines. *Universal Journal of Educational Research, 8*(11b), 5693-5703. http://dx.doi.org/10.13189/ujer.2020.082203

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 8196. https://www.mdpi.com/2076-3417/10/22/8196#

Contreras, J.O., Hilles, S.M., & Abubakar, Z.B. (2018) Automated essay scoring with ontology based on text mining and NLTK tools. In I. Zen (Pres.), *2018 International Conference on Smart Computing and Electronic Enterprise* (pp. 1-6). IEEExplore.

Coppock, A., Leeper, T.J., Mullinix, K.J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *PNAS, 115*(49), 12441-12446. http://www.pnas.org/cgi/doi/10.1073/pnas.1808083115

Cummins, R., Zhang, M., & Briscoe, E. (2016). *Constrained multi-task learning for automated essay scoring.* Association for Computational Linguistics.

Darwish, S.M., & Mohamed, S.K. (2019) Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In A.E. Hassanien, A.T. Azar, T. Gaber, R. Bhatnagar, & M.F. Tolba (Eds.), *The International Conference on Advanced Machine Learning Technologies and Applications* (pp. 566-575). Springer.

DeLuca, D., Willis, J., Cowie, B., Harrison, C., Coombs, A., Gibson, A., et al. (2019). Policies, programs, and practices: exploring the complex dynamics of assessment education in teacher education across four countries. *Frontiers in Education, 4,* 1-19. https://doi.org/10.3389/feduc.2019.00132

Deneen, C.C., & Brown, G.T.L (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education, 3*(1), 1225380. https://doi.org/10.1080/2331186X.2016.1225380

Dillenbourg, P. (2016). The evolution of research on digital education. International *Journal of Artificial Intelligence in Education, 26*(2), 544-560. https://doi.org/10.1007/s40593-016-0106-z

Dong, F., Zhang, Y., Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In R. Levy & L. Specia (Eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 153–162). Association for Computational Linguistics.

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345-356. https://doi.org/10.1080/0969594X.2010.516569

Galindo-Domínguez, H., & Bezanilla, M.J. (2021). Promoting Time Management and Self-Efficacy Through Digital Competence in University Students: A Mediational Model. *Contemporary Educational Technology, 13*(2), ep294. https://doi.org/10.30935/cedtech/9607

Galindo-Domínguez, H., Delgado, N., Losada, D., & Etxabe, J.M. (2024). An analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education, 40*(1), 41-56. https://doi.org/10.1080/21532974.2023.2284726

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for student assessment: a systematic review. *Applied Sciences,* 11, 5467. https://doi.org/10.3390/app11125467

Government of Newfoundland and Labrador (2014). *English Language Arts Grade 6. Appendix D: Sample Elementary Classroom Rubrics and Checklists.* Department of Education of the Government of Newfoundland and Labrador. https://www.gov.nl.ca/education/files/k12_curriculum_guides_english_grade6_300614_g6_ela.pdf

Grivokostopoulou, F., Perikos, I., Hatzilygeroudis, I. (2017). An Educational System for Learning Search Algorithms and Automatically Assessing Student Performance. *International Journal of Artificial Intelligence in Education*, 27, 207–240. http://dx.doi.org/10.1007/s40593-016-0116-x

Hill, M., Ell, F., & Eyers, G. (2017). Assessment capability and student self-regulation: the challenge of preparing teachers. *Frontiers in Education, 2,* 1-15. https://doi.org/10.3389/feduc.2017.00021

Houtao, L., Wenjia, M., Tingting, W., & Chuanhua, X. (2022). The Study of Feedback in Writing from College English Teachers and Artificial Intelligence Platform Based on Mixed Method Teaching. *Pacific International Journal, 5*(4), 147-154. https://doi.org/10.55014/pij.v5i4.270

Hrastinski, S., Olofsson, A. D., Arkenback, C., Ekström, S., Ericsson, E., Fransson, G., Jaldemark, J., Ryberg, T., Öberg, L.-M., Fuentes, A., Gustafsson, U., Humble, N., Mozelius, P., Sundgren, M., & Utterberg, M. (2019). Critical imaginaries and reflections on artificial intelligence and robots in post-digital K-12 education. *Post-Digital Science and Education, 1*(2), 427-445. https://doi.org/10.1007/s42438-019-00046-x

Jani, K.H., Jones, K.A., Jones, G.W., Amiel, J., Barron, B., & Elhadad, N. (2020). Machine learning to extract communication and historytaking skills in OSCE transcripts. *Medical Education, 54,* 1159–1170. https://doi.org/10.1111/medu.14347

Kasneci, E., Sessler, K., Küchemann, S., …, Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences,* 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Ke, Z., Inamdar, H., Lin, H., & Ng, V. (2019). Give me more feedback II: Annotating thesis strength and related attributes in student essays. In A. Korhonen, D. Traum & L. Márquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3994-4004). Association for Computational Linguistics.

Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get it scored using autosas—an automated system for scoring short answers. In B. Williams, Y. Chen, & J. Neville (Eds.), *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 9662–9669). AAAI Press.

Liu, M., Wang, Y., Xu, W., & Liu, L. (2017). Automated Scoring of Chinese Engineering Students' English Essays. *International Journal of Distance Education Technologies, 15*(1), 52–68.

Lovorn, M.G., Reza, A. (2011). Assessing the Assessment: Rubrics Training for Pre-service and New In-service Teachers. *Practical Assessment, Research, and Evaluation, 16*(1), 16. https://doi.org/10.7275/sjt6-5k13

Mathias, S., & Bhattacharyya, P. (2018). Thank "Goodness"! A Way to Measure Style in Student Essays. In Y. Tseng, H. Chen, V. Ng. & M. Komachi (Eds.), *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 35–41). Association for Computational Linguistics.

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education, 43*(6), 1-18. http://dx.doi.org/10.14221/ajte.2018v43n6.1

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R.F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE 15,* e0229596. https://doi.org/10.1371/journal.pone.0229596

Ocaña-Fernández, Y., Valenzuela-Fernández, L.A., & Garro-Aburto, L.L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones, 7*(2), 536-568. https://doi.org/10.20511/pyr2019.v7n2.274

Organic Law 3/2020, of December 29th, amending Organic Law 2/2006, of May 3rd, on Education. *Official State Gazette,* 340, 122868-122953. https://www.boe.es/eli/es/lo/2020/12/29/3

Ouguengay, Y.A., El Faddouli, N.-E., & Bennani, S. (2015). A neuro-fuzzy inference system for the evaluation of reading/writing competencies acquisition in an e-learning environnement. *Journal of Theoretical and Applied Information Technology, 81*(3), 600–608.

Owan, V.J., Bekom, K., Emoji, D., Onor, E., & Asuquo, B. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. Modestum. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Ramesh, D., & Kumar, S. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review,* 55, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Redecker, C. (2017). *European Framework for the Digital Competence of Educators: DigCompEdu.* Joint Research Centre. http://dx.doi.org/10.2760/159770

Rhienmora, P., Haddawy, P., Suebnukarn, S., Dailey, M.N. (2011). Intelligent dental training simulator with objective skill assessment and feedback. *Artificial Intelligence in Medicine, 52*(2), 115–121. https://doi.org/10.1016/j.artmed.2011.04.003

Salama, S., & Subahi, A. M. (2020). The Impact of Specialty, Sex, Qualification, and Experience on Teachers' Assessment Literacy at Saudi Higher Education. *International Journal of Learning, Teaching and Educational Research, 19*(5), 200-216. https://doi.org/10.26803/ijlter.19.5.12

Samarakou, M., Fylladitakis, E.D., Karolidis, D., Früh, W.-G., Hatziapostolou, A., Athinaios, S.S., & Grigoriadou, M. (2016). Evaluation of an intelligent open learning system for engineering education. *Knowledge Management & E-Learning: An International Journal, 8*(3), 496–513.

Spear-Swerling, L., Owen, P., & Alfano, M.P. (2005). Teachers' literacy-related knowledge and self-perceptions in relation to preparation and experience. *Annals of Dyslexia, 55,* 266-296. https://doi.org/10.1007/s11881-005-0014-7

Stiggins, R. (2014). Improve assessment literacy outside of schools too. Phi Delta Kappan, 96, 65–72.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Ulum, Ö.G. (2020). A critical deconstruction of computer-based test application in Turkish State University. *Education and Information Technologies, 25,* 4883–4896. https://doi.org/10.1007/s10639-020-10199-z

Vij, S., Tayal, D., & Jain, A. (2020). A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. *Wireless* Personal Communications, 111(2), 1271–1282. https://doi.org/10.1007/s11277-019-06913-x

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence, 10*(2), 1-37. https://doi.org/10.2478/jagi-2019-0002

Xu, Y., & Brown, G.T.L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teaching and Teacher Education, 58,* 149-162. http://dx.doi.org/10.1016/j.tate.2016.05.010

Para citar este artículo:

Galindo-Domínguez, H., Delgado, N., Sainz de la Maza, M., y Expósito, E. Un análisis experimental de la relación entre las evaluaciones proporcionadas por la inteligencia artificial y las proporcionadas por los docentes en formación [An experimental analysis of the relationship between the evaluations of artificial intelligence and pre-service teachers]. *Edutec, Revista Electrónica de Tecnología Educativa*, (89), 84-104. https://doi.org/10.21556/edutec.2024.89.3509

6. ANEXOS

6.1. ANEXO 1: Textos generados

Tipo	Nivel	Texto
Texto descriptivo	Alto	En el tranquilo bosque detrás de mi casa, el sol de la tarde se filtra a través de las hojas de los altos árboles, creando un mágico juego de luces y sombras en el suelo. Los pájaros cantan melodías alegres mientras exploran entre las ramas, y el aire está lleno del aroma fresco de la naturaleza. A medida que camino por el sendero cubierto de hojas crujientes, el murmullo del arroyo cercano me guía hacia su orilla, donde el agua fluye suavemente sobre las piedras redondas y pulidas. Cada paso que doy me sumerge aún más en este misterioso mundo, donde la paz y la belleza se entrelazan en perfecta armonía."
	Medio	Me gusta mucho el parque cerca de mi casa. Cuando llego allí, siento una emoción especial. El parque es un lugar dibertido y emocionante para mí. Hay tantas cosas que hacer y explorar. Los colunpios son mi parte favorita. Me balanceo alto y siento como si volara. También hay un tobogán grande y emocionante en el que deslizo rápidamente. Mis amigos y yo nos divertimos mucho compitiendo a ver quién llega primero. Además de los juegos, el parque está yeno de árboles altos que proporcionan sonbra en los días calurosos. Me encanta sentarme bajo un árbol con mis amigos y hablar de nuestras aventuras. También hay un pequeño estanque en el que a veces vemos patitos nadando. Cuando tenemos hambre, a veces traemos helado y nos sentamos en los bancos del parque para disfrutarlo. Es tan refrescante y delicioso.
	Bajo	Aier fui a un lugar bonito. Estaba afuera y hacia mucha calor. Me gustó mucho por qué era un día ermoso. Había columpios, un tovogán grande y un par de vancos para sentarse. Jugamos en los colunpios, subimos al tovogán y reímos mucho. Yo y mis amigos pasamos un buen rato. Después de jugar, fuimos a una tienda cercana y comimos elado. Yo pedí un elado de fresa por qué es el más rico. La parte más dibertida fue ver cómo se deretía bajo el sol caliente. En general, el día estuvo guay, aunque no recuerdo el nombre esacto del lugar. Haber si volvemos otro dia.
Texto argumentativo	Alto	Creo que debemos cuidar a los animales. Los animales son nuestros amigos y necesitan amor y protección. Si cuidamos de ellos, podemos ayudar a que estén felices y saludables. También es importante no tirar basura en el suelo porque puede lastimar a los animales y contaminar su hogar. Debemos mantener limpio nuestro entorno para que los animales vivan en un lugar seguro. En resumen, debemos ser buenos amigos de los animales y cuidar de nuestro planeta para que todos podamos vivir en armonía.
	Medio	Creo que los videojuegos son divertidos. Muchos niños juegan a los videojuegos y se divierten mucho. A veces, es mejor jugar afuera, pero los videojuegos son buenos para cuando hace frío o llueve. También es importante hacer la tarea antes de jugar a los videojuegos. La tarea es aburrida, pero es importante hacerla para aprender cosas. En resumen, los videojuegos son divertidos, pero la tarea es importante. Los dos pueden ser buenos si los hacemos en el momento adecuado.
	Bajo	Los bideojuegos son guais. Me gusta jugar porque es dibertido. Tengo muchos juegos en mi consola. Algunos son de carreras y otros de abenturas. Me emociona mucho cuando gano. Mis amigos tanbién juegan a bideojuegos. Jugamos juntos y nos dibertimos mucho. Abeces, nos retamos haber quién gana más partidas. Son guais, pero tanvién se que es inportante acer otras cosas, como leer o acer los deveres. Asique, aunque me gusten mucho los bideojuegos, sé que no puedo jugar todo el tienpo.
Texto instructivo	Alto	Cómo Hacer una Cometa Materiales: Papel de seda o papel de envolver Palos de madera Tijeras Hilo fuerte Pegamento

		Cinta adhesiva Pinturas y pinceles (opcional) Instrucciones: Primero, corta cuatro palos de madera de igual longitud para hacer el armazón de la cometa. Dos de ellos serán los lados verticales y los otros dos, los lados horizontales. Une los palos para formar un marco en forma de cruz. Utiliza cinta adhesiva o hilo fuerte para asegurar las uniones. A continuación, corta el papel de seda o papel de envolver en forma de diamante, dejando unos centímetros extra en los bordes para doblarlos y pegarlos al armazón. Pega con cuidado el papel al marco de madera, asegurándote de que quede estirado y sin arrugas. Ahora, corta un trozo de hilo fuerte de aproximadamente un metro de largo. Amarra un extremo al centro del palo horizontal y deja el otro extremo libre. Decorar la cometa con pinturas y pinceles si lo deseas. Puedes hacer diseños creativos en el papel. Cuando hayas terminado de decorar la cometa y el pegamento esté seco, ¡está lista para volar! Encuentra un día con viento suave, ve a un campo abierto y suelta tu cometa al viento sujetando el extremo del hilo. ¡Disfruta viendo cómo se eleva y vuela alto en el cielo!”
	Medio	Cómo Hacer un Sándwich de PByJ Materiales: Pan de molde Mantequilla de maní Mermelada de fresa Cuchillo Instrucciones: Coge dos rebanadas de pan de molde. Abre la mantequilla de maní y la mermelada de fresa. Unta mantequilla de maní en una de las rebanadas de pan. Luego, pon mermelada de fresa en la otra rebanada. Presiona las dos rebanadas juntas para que la mantequilla de maní y la mermelada se mezclen.
	Bajo	Como Hacer Un Pastel: Conprar una mescla para pastel. Agregar uevos y leche. Mezclarlo todo. Poner en un molde. Meter el molde en el orno. Sacarlo cuando este listo.
Texto Narrativo	Alto	Ayer, junto a mis amigos, pasé un emocionante día en el parque. Juntos, construimos un inmenso castillo de arena y nos sumergimos en un emocionante juego de escondidas. Posteriormente, comimos deliciosos helados mientras admirábamos el colorido arco iris que se formó en el cielo. Sin duda, fue uno de los días más sorprendentes que he vivido.
	Medio	Un día soleado, fui al parque con mis amigos. Corrimos y jugamos en los columpios. Después, decidimos explorar el bosque cercano. Seguido, encontramos un arrollo y lanzamos piedras al agua. Después, nos sentamos bajo un árvol a comer sándwiches. Para acabar el día, regresamos a casa, cansados pero felices.
	Bajo	Un día, fui al parke. Jugamos mucho y comimos elado. Luego, fuimos a casa. Fin.

6.2. Anexo II: Rúbrica empleada por los docentes en formación para evaluar los diferentes textos.

Criterio	Excelente (4)	Bueno (3)	Regular (2)	Pobre (1)
Contenido	El texto presenta información detallada y rica del tema, con una variedad de detalles y ejemplos relevantes que enriquecen la comprensión del lector.	El texto presenta una cantidad de información adecuada del tema, con detalles y ejemplos que hacen que el tema sea claro para el lector.	El texto presenta una cantidad de información limitada del tema, con detalles insuficientes o poco claros que dificultan la comprensión del lector.	El texto tiene una cantidad de información tan pobre que no se puede entender el tema.
Organización	El texto tiene una estructura clara y lógica, con una introducción, desarrollo y conclusión bien definidos. Las ideas están organizadas de manera efectiva.	El texto tiene una estructura generalmente clara y lógica, aunque la organización podría mejorar en algunos lugares. Las ideas están organizadas de manera adecuada.	El texto tiene una estructura poco clara o desorganizada, lo que dificulta la comprensión de las ideas.	El texto carece de estructura y organización, lo que hace que sea difícil seguir las ideas presentadas.
Vocabulario y Lenguaje	El texto utiliza un vocabulario variado y preciso. Las oraciones son complejas y están bien construidas.	El texto utiliza un vocabulario adecuado, aunque podría incorporar más variedad y precisión. Las oraciones son en su mayoría correctas.	El texto utiliza un vocabulario limitado y repetitivo, lo que afecta la calidad de la producción. Las oraciones son simples y pueden contener errores gramaticales.	El texto utiliza un vocabulario muy limitado y/o inadecuado, lo que dificulta la comprensión. Las oraciones son incorrectas y confusas.
Coherencia y Cohesión	El texto muestra una alta coherencia y cohesión, con conexiones claras entre las ideas y un uso efectivo de conectores y referencias.	El texto muestra coherencia y cohesión en general, aunque algunas conexiones entre las ideas pueden ser más claras. Se utilizan algunos conectores y referencias.	El texto carece de coherencia y cohesión, lo que dificulta la transición entre las ideas. Los conectores y referencias son escasos o inapropiados.	El texto es incoherente y carece de cualquier forma de cohesión, lo que hace que sea difícil seguir la narración.