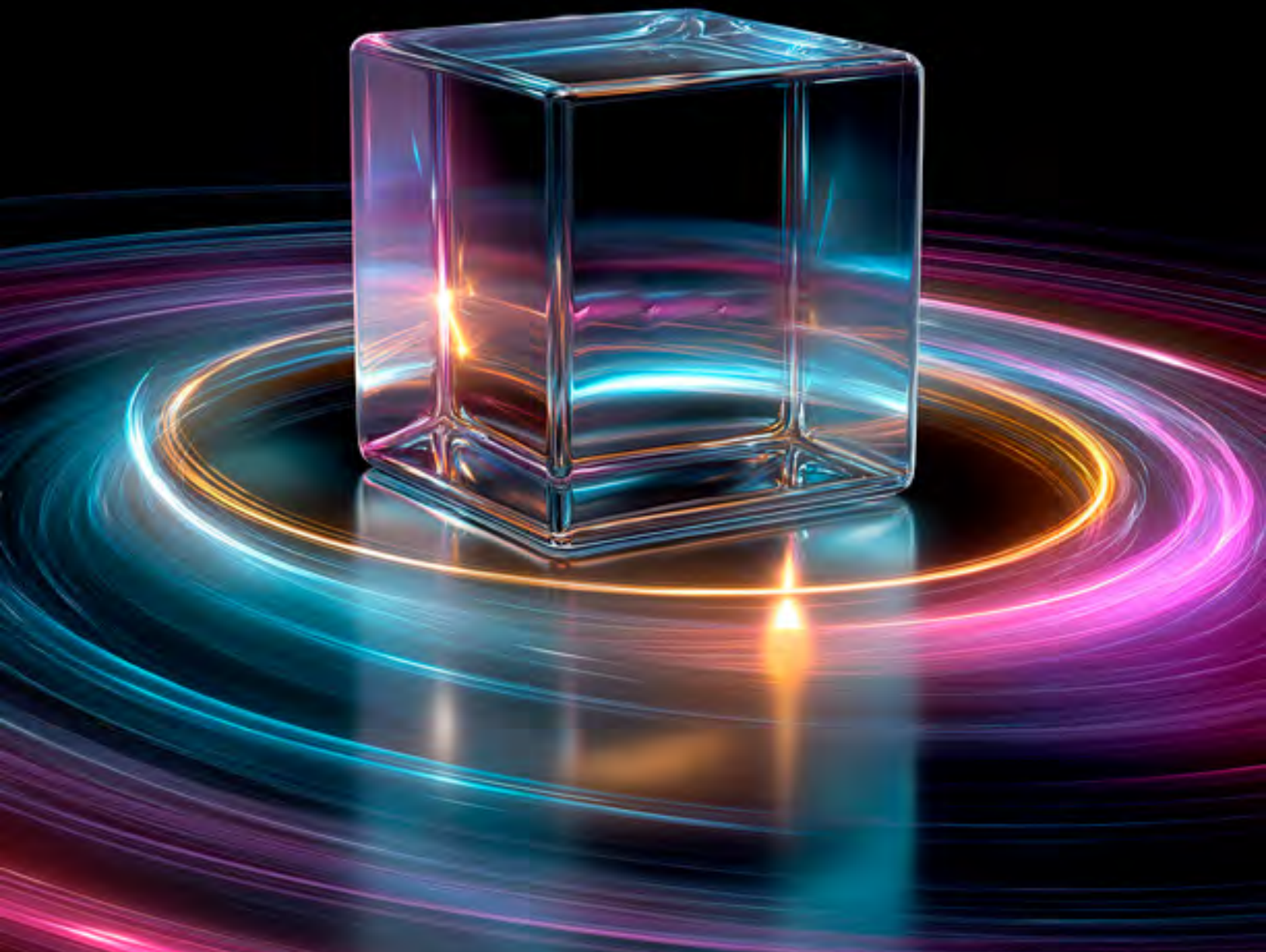
INTELIGENCIA ARTIFICIAL

NUEVAS EXPERIENCIAS ACADÉMICAS



EMMA PATRICIA MERCADO-LÓPEZ
ALEXANDRO ESCUDERO-NAHÓN
(COORDS.)

Transdigital
editorial

INTELIGENCIA ARTIFICIAL

Nuevas experiencias académicas

Emma Patricia Mercado-López

Alexandro Escudero-Nahón

(Coords.).

Nélida Bethel Alcalá Cortés, Guillermo Barrera Gómez, Sandra Luz Canchola-Magdaleno, Ahmed Alejandro Cardona Mesa, Luis Alonso Castañeda Negrete, Patricia Delgadillo Gómez, Sergio Alberto Díaz Alvarado, Alexandro Escudero-Nahón, Víctor Guillermo Flores Rodríguez, Mauricio Hernández Ramírez, Luis Jesús Ibarra Manrique, Fernando Leal Ríos, José Carlos López Hernández, Esperanza Manrique Rojas, Edith Martin-Galindo, David Martínez Cerqueda, Emma Patricia Mercado-López, René Sebastián Mora Ortiz, Georgina del Carmen Mota Valtierra, Emmanuel Munguía Balvanera, Salvador Ortiz Santos, Benito Parra Pacheco, Margarita Ramírez Ramírez, Margarita Ramirez-Torres, Alejandro Guadalupe Rincón Castillo, Cándida Marcela Rodríguez Chávez, Jesús Ángel Rodríguez García, Alma Eloisa Rodríguez Medina, Manuel Ruiz Méndez, Adriana Mercedes Ruiz Reynoso, María del Consuelo Salgado Soto, Nancy Azucena Salgado-Iriarte, Edgar Fabián Torres Hernández, Oralia Zamora Pequeño, Raymundo Said Zamora Pequeño y Santiago Zapata Vargas

Autores y autoras

Transdigital®

editorial

Título original: Inteligencia artificial: nuevas experiencias académicas / Emma Patricia Mercado-López y Alexandro Escudero-Nahón (Coords.) — Ciudad de Querétaro, México: Editorial Transdigital, 2025 — 245 páginas.

International Standard Book Number (ISBN): 978-968-9724-12-4.

Digital Object Identifier (DOI) del libro: https://doi.org/10.56162/transdigitalbc04

Clasificación DEWEY. Materia: 006.3 - Inteligencia artificial. Tipo de Contenido: Libros universitarios. Clasificación thema: JN–Educación. Tipo de soporte: libro digital gratuito descargable. Formato: PDF. Tamaño: 2.7 Mb.

Sociedad de Investigación sobre Estudios Digitales, S. C. Nombre de marca: *Transdigital*. Dirección: Circuito Altos Juriquilla 1132. Colonia Altos Juriquilla. C. P. 76230, Juriquilla, Querétaro, México. +52 (442) 301 32 38. editorial@transdigital.mx www.editorial.transdigital.mx

Registro en el Padrón Nacional de Editores como agente editor Sociedad de Investigación sobre Estudios Digitales, S. C., con el Dígito Identificador 978-607-99594.

Afiliación a la Cámara Nacional de la Industria Editorial Mexicana (CANIEM) con el número 4069, de conformidad con el artículo 17 de la Ley de Cámaras Empresariales y sus Confederaciones en vigor.

Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) de México con el folio: RENIECYT 2400068.

Sugerencia de referencia para el libro en APA 7a. edición:

Mercado-López, E. P., y Escudero-Nahón, A. (2025) (Coords.). *Inteligencia artificial: nuevas experiencias académicas.* Editorial Transdigital. https://doi.org/10.56162/transdigitalbc04

Contenido

12.
Comparación entre la evaluación docente y la realizada por un modelo de lenguaje extenso

Raymundo Said Zamora Pequeño
Universidad Autónoma de Nuevo León, México
ORCID: 0009-0001-3482-701X

Oralia Zamora Pequeño
Universidad Autónoma de Nuevo León, México
ORCID: 0009-0004-7338-1518

DOI del capítulo de libro:
https://doi.org/10.56162/transdigitalbc04.12

12.
Comparación entre la evaluación docente y la realizada por un modelo de lenguaje extenso

Introducción

La digitalización en el ámbito educativo ha facilitado la integración de inteligencia artificial (IA) en procesos como la enseñanza, el aprendizaje y la evaluación. Los modelos de lenguaje extenso (LLM, por sus siglas en inglés) generan y analizan texto con coherencia y precisión cada vez mayores (Stasuik, 2025). Esta tecnología permite automatizar la calificación de trabajos escritos, lo que reduce la carga laboral del personal docente y proporciona retroalimentación al estudiantado.

La evaluación de respuestas abiertas presenta desafíos debido a la diversidad del lenguaje y la subjetividad en su interpretación (Yavuz et al., 2024). Mientras que el profesorado emplea criterios explícitos e implícitos basados en su experiencia, los LLM utilizan patrones lingüísticos y semánticos derivados de grandes conjuntos de datos textuales para calificar. Esta situación genera cuestiones relacionadas con la consistencia, equidad y validez de las calificaciones producidas por IA en comparación con las evaluaciones humanas (Mok et al., 2025).

Este estudio examinó el grado de concordancia y las diferencias entre la evaluación realizada por un docente y por un LLM en reactivos de la unidad de aprendizaje llamada *Temas selectos de optimización*. Dicha unidad incluye, entre otros temas, la optimización mediante métodos heurísticos, un concepto que se presta a diversas formulaciones y que se utiliza para evaluar la sensibilidad y precisión de un evaluador automático.

Objetivo general

Comparar la valoración de un modelo de lenguaje extenso y de un docente en la calificación de ejercicios y preguntas sobre optimización con métodos heurísticos, con el fin identificar patrones de concordancia, sesgo y discrepancia.

Preguntas de investigación

Las principales preguntas que busca responder este proyecto de investigación son:

1. ¿Existe correlación significativa entre las calificaciones del LLM y las del docente?
2. ¿Qué patrones de sesgo o discrepancia se observan en la calificación del LLM?
3. ¿Qué implicaciones tienen estos resultados para la evaluación asistida por IA en educación superior?

Método de investigación

Participantes

Se incluyeron 50 estudiantes inscritos en la asignatura *Temas selectos de optimización* durante el semestre enero-junio 2025.

Instrumentos

Los estudiantes que participaron en el instrumento de evaluación respondieron tres reactivos, mediante una tarea designada en el equipo de la clase en la plataforma de *Microsoft Teams*. El primero de los reactivos corresponde a la pregunta abierta: a) Defina qué es un método heurístico. Los siguientes dos ejercicios fueron: b) Considerando que los nodos de la Tabla 1 tienen la forma (coordenada x, coordenada y demanda), y que cada grupo tiene una capacidad de 70, utilice la heurística de barrido para formar grupos de nodos. Utilizando distancias euclidianas resuelva el problema del agente viajero para el primer grupo utilizando búsqueda tabú.

Tabla 1
Datos del ejercicio b, coordenada x, coordenada y demanda

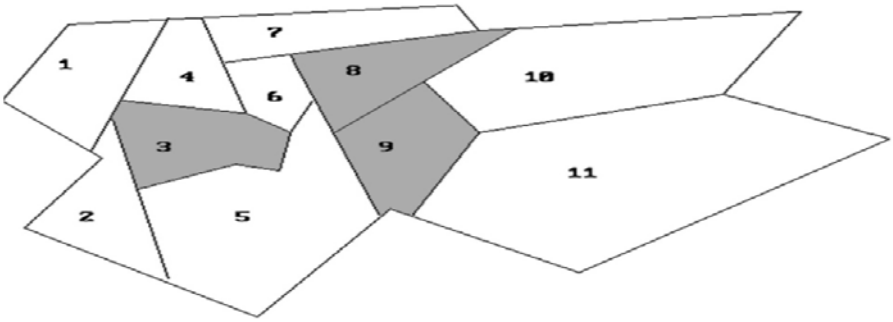
Listado uno de ejercicios	Listado dos de ejercicios	Listado tres de ejercicios	Listado cuatro de ejercicios
(0, -2, 17)	(-3, -2, 16)	(1.6,-2.5, 14)	(0.8, 1.2,16)
(-3, 5, 18)	(4, 4, 22)	(3, 1, 24)	(1.4, 3.1,18)
(4, 1, 21)	(3, 2, 20)	(1, -6, 17)	(6, 8, 11)

c) Existen once localidades en un estado. Se desea construir una mínima cantidad de estaciones de bomberos para asegurar que por lo menos una estación ente dentro de 15 minutos (tiempo de viaje entre ciudades adyacentes) de cada ciudad. Plantee un vector

valido de solución. Analice si existen columnas dominadas. El esquema de resolución se muestra en la Figura 1.

Figura 1

Esquema para la resolución del ejercicio c



Prompt de evaluación

Para el primer reactivo se planteó el siguiente *prompt en ChatGPT 4*: “Si la mejor respuesta (7 puntos) es: ‘método de solución para problemas complejos basado en la experiencia que permite encontrar soluciones aproximadas en un tiempo razonable’, asigna de 0 a 7 puntos a cada respuesta siguiendo la rúbrica proporcionada”.

Para el segundo ejercicio se planteó la siguiente instrucción: “Considerando un máximo de quince puntos, donde cinco se otorgan por realizar las agrupaciones de manera correcta, incluyendo los ángulos para el método de barrido, otros cinco para la tabla de distancias euclidianas y los últimos cinco en la realización correcta de una iteración de búsqueda tabú, asigna puntuaciones adecuadas para cada respuesta siguiendo el archivo de solución guía”.

Para el ejercicio c el *prompt* proporcionado al LLM para la evaluación fue: “Otorgando un máximo de 10 puntos posibles, designe cinco a la verificación de un vector de solución válido y cinco a la revisión de columnas dominadas de acuerdo con las soluciones proporcionadas en el archivo anexado”.

Análisis de datos

El análisis comparativo entre las calificaciones del docente y las emitidas por el LLM se realizó mediante los siguientes estadísticos:

Diferencia absoluta promedio

La diferencia absoluta promedio mide la magnitud promedio de las discrepancias entre las puntuaciones otorgadas por el LLM y las del docente, ignorando el signo de la diferencia.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

donde *xi* son las calificaciones del LLM y *yi* las del docente. Valores bajos indican alta similitud en las puntuaciones (Willmott & Matsuura, 2005).

Sesgo promedio

El sesgo promedio cuantifica la tendencia sistemática del LLM a sobrestimar o subestimar las calificaciones en comparación con el docente.

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$$

Un sesgo positivo indica que el LLM asigna, en promedio, puntuaciones mayores que el docente; un sesgo negativo indica lo contrario (Jolliffe, 2011).

Prueba de normalidad de Shapiro-Wilk

La prueba de Shapiro-Wilk determina si una muestra sigue una distribución normal (Shapiro & Wilk, 1965). Su estadístico de prueba se define como:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

donde:

x(i) son los valores ordenados de la muestra.

$\bar{x}$ representa la media de la muestra.

ai son coeficientes calculados a partir de la varianza y la media de una distribución normal.

n es el tamaño de la muestra.

Si el p-valor asociado a W es menor que un nivel de significancia (α), se rechaza la hipótesis de normalidad.

Si los datos no muestran una distribución normal, para comparar la relación entre dos muestras, se debe emplear el coeficiente de correlación de Spearman.

Coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman (ρ) evalúa la relación monótona entre dos variables con los rangos de los datos.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde *di* es la diferencia entre los rangos de *xi* y *yi* (Spearman, 1904).

Prueba de rangos con signo de Wilcoxon

La prueba de rangos con signo de Wilcoxon se utiliza cuando las diferencias entre dos mediciones relacionadas no siguen una distribución normal. Es una prueba no paramétrica que compara las medianas de dos conjuntos de datos apareados. Consiste en calcular las diferencias entre pares de observaciones, eliminar las que sean cero, ordenar las diferencias absolutas de menor a mayor, asignarles un rango y un signo según la dirección de la diferencia, y finalmente sumar por separado los rangos positivos y negativos. El estadístico de prueba *V* se define como el menor de dichas sumas.

$$V = \min\left(\sum R^{+}, \sum R^{-}\right)$$

Donde *R*+ y *R*– representan la suma de rangos positivos y negativos, respectivamente (Wilcoxon, 1945).

3. Resultados

Puntuaciones coincidentes

En los tres reactivos examinados se tomó en cuenta el número de coincidencias en la puntuación asignada. Se consideró un total de cincuenta respuestas analizadas. Los resultados correspondientes se presentan en la Tabla 2.

Tabla 2
Número de coincidencias de evaluación entre el docente y el modelo de lenguaje extenso

Reactivo a	Reactivo b	Reactivo c
35	31	41

Estadística descriptiva

Se calcularon las medidas de tendencia central y dispersión para las calificaciones otorgadas por el LLM y por el docente en cada uno de los tres reactivos evaluados. Para cada par de calificaciones (LLM–Docente) se obtuvieron la media y la desviación estándar, las cuales pueden ser encontradas en la Tabla 3, con el fin de caracterizar el nivel promedio y la variabilidad de las puntuaciones.

Tabla 3
Estadísticas descriptivas de las evaluaciones

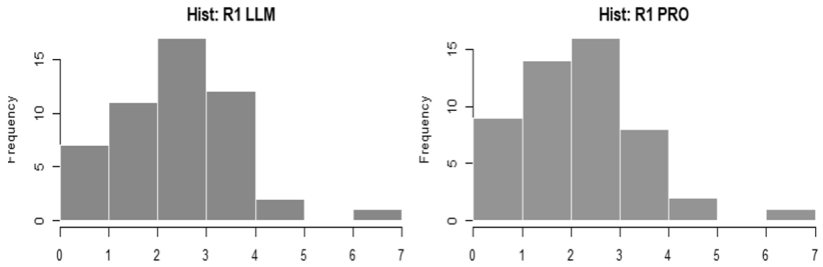
Evaluación	LLM a	Docente a	LLM b	Docente b	LLM c	Docente c
Media	2.9	2.66	8.84	8.58	5.36	5.3
Desviación estándar	1.233	1.287	4.077	4.081	2.529	2.667

3.3. Histogramas

En la pregunta abierta, tanto el LLM como el docente asignaron la mayoría de las calificaciones dentro del rango de 1 a 3 puntos, con una asimetría leve hacia la derecha. El patrón de dispersión es comparable entre ambos evaluadores; sin embargo, el LLM muestra una frecuencia ligeramente mayor de puntajes en el rango de 3 puntos, lo que corresponde al sesgo positivo identificado en el análisis descriptivo y representado en la Figura 2.

Figura 2

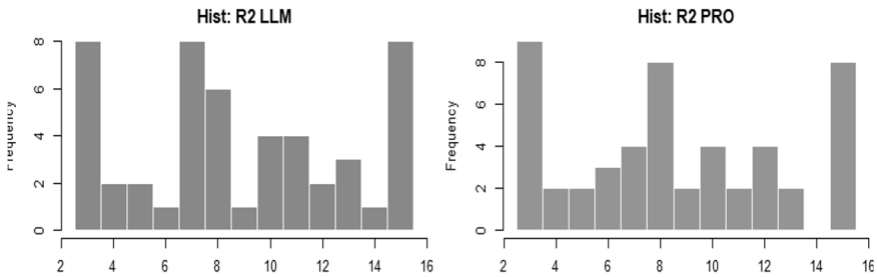
Histograma de la distribución de evaluaciones del ejercicio a



En el ejercicio b, las distribuciones de calificaciones del LLM y del docente son visualmente semejantes, con valores dispersos en todo el rango de 3 a 15 puntos. En la Figura 3, se puede observar los picos en las puntuaciones altas y bajas, lo que indica que ambos evaluadores identificaron claramente tanto respuestas completas como incompletas. No obstante, el LLM muestra una ligera tendencia a asignar calificaciones más elevadas en los extremos superiores, lo que refuerza el sesgo positivo identificado en el análisis estadístico.

Figura 3

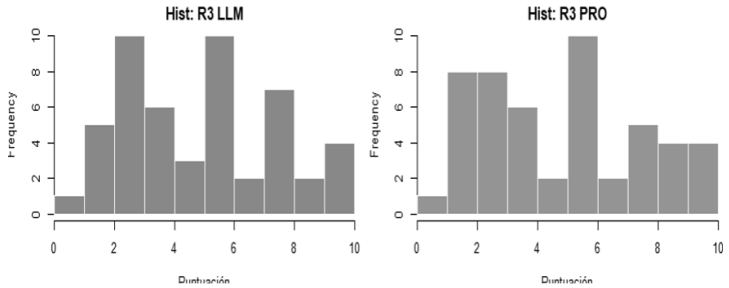
Histograma de la distribución de evaluaciones del ejercicio b



En el ejercicio c, las calificaciones de ambos evaluadores están distribuidas de forma relativamente uniforme en todo el rango de 0 a 10 puntos, sin una concentración clara en valores específicos. La similitud en la forma de ambas distribuciones, se puede observar en la Figura 4, la cual sugiere un alto grado de coincidencia en la evaluación.

Figura 4

Histograma de la distribución de evaluaciones del ejercicio c



Resultados del análisis

Diferencia absoluta promedio

El análisis de la diferencia absoluta promedio entre las calificaciones del LLM y del docente mostró valores bajos en los tres reactivos, con 0.48 en R1, 0.38 en R2 y 0.18 en R3. Estos resultados indican que, en promedio, las discrepancias en magnitud fueron menores a medio punto, lo que evidencia una alta similitud en las puntuaciones emitidas por ambos evaluadores. La reducción progresiva de R1 a R3 sugiere que la coincidencia en las calificaciones fue mayor en los ejercicios prácticos que en la pregunta abierta.

Sesgo promedio

El sesgo promedio fue positivo en todos los casos (+0.24 en R1, +0.26 en R2 y +0.06 en R3), lo que revela una tendencia consistente del LLM al asignar calificaciones ligeramente superiores a las del docente. No obstante, la magnitud de este sesgo es pequeña, lo que indica que la sobreestimación no es sustancial. El sesgo más alto se observó en R2, lo que podría asociarse con una mayor valoración del LLM a las soluciones de optimización en comparación con la evaluación docente.

Prueba de normalidad de Shapiro-Wilk

Se aplicó la prueba de Shapiro–Wilk a las diferencias entre las calificaciones del LLM y las del docente con el fin de evaluar el cumplimiento del supuesto de normalidad requerido por ciertos métodos estadísticos paramétricos. La hipótesis nula de esta prueba plantea que los datos provienen de una distribución normal, mientras que la hipótesis alternativa establece que la distribución es no normal.

En los tres reactivos, los valores p fueron extremadamente bajos (R1: $p \approx 3.6 \times 10^{-8}$; R2: $p \approx 3.0 \times 10^{-8}$; R3: $p \approx 1.0 \times 10^{-10}$), todos muy inferiores al nivel de significancia $\alpha = 0.05$. Esto llevó a rechazar la hipótesis de normalidad con alto grado de certeza, concluyéndose que las diferencias presentan distribuciones no normales. Esta condición puede deberse a asimetría, curtosis elevada o la existencia de valores atípicos que afectan la forma de la distribución.

La ausencia de normalidad afecta la elección de la prueba de comparación de medias (descartando la t de Student) e influye en la selección del coeficiente de correlación más apropiado. Dado que el coeficiente de Pearson asume normalidad y linealidad en los datos, su interpretación bajo distribuciones no normales puede ser menos fiable y más sensible a valores extremos. Por ello, se priorizó el uso del coeficiente de Spearman, que se basa en rangos y no en los valores crudos, lo que lo hace robusto frente a distribuciones no normales y valores atípicos, además de ser adecuado para medir relaciones monótonas que no necesariamente son lineales.

Coeficiente de correlación de Spearman

El análisis de correlación de Spearman (ρ) mostró asociaciones positivas altas en todos los reactivos: $\rho = 0.667$ en R1, $\rho = 0.988$ en R2 y $\rho = 0.983$ en R3, todas con $p < 0.001$. Estos resultados indican que, aunque en la pregunta abierta la coincidencia en el orden de las calificaciones es moderada-alta, en los ejercicios prácticos la correspondencia en la jerarquía de las evaluaciones entre LLM y docente es prácticamente perfecta. Esto sugiere que, a medida que el formato de la tarea se vuelve más estructurado, la consistencia relativa entre evaluadores aumenta notablemente.

Prueba de rangos con signo de Wilcoxon

Dado que la prueba de Shapiro–Wilk evidenció ausencia de normalidad en las diferencias, se utilizó la prueba no paramétrica de rangos con signo de Wilcoxon para comparar

las calificaciones del LLM y del docente en cada reactivo. En la pregunta abierta se obtuvo un estadístico $V = 31.0$ con $p = 0.092$, lo que indica que no hay diferencias estadísticamente significativas entre ambos evaluadores. En el primer ejercicio b de optimización b, el estadístico fue $V = 30.0$ con $p = 0.0029$, mostró diferencias significativas a favor de puntuaciones más altas por parte del LLM. Finalmente, en el segundo ejercicio c el resultado fue $V = 15.0$ con $p = 0.317$, sin diferencias significativas. Estos resultados confirman que, salvo en R2, las calificaciones emitidas por el LLM y el docente son estadísticamente equivalentes, aunque con ligeras variaciones en magnitud.

4. Discusión

Los resultados evidencian una elevada concordancia entre las calificaciones asignadas por el LLM y el docente en los tres reactivos analizados, con diferencias absolutas promedio inferiores a medio punto. Las correlaciones de Spearman se ubicaron entre valores moderados-altos ($\rho = 0.667$ en R1) y muy altos ($\rho > 0.98$ en R2 y R3). Este resultado es consistente con investigaciones previas que indican que los LLM reproducen criterios de evaluación con mayor uniformidad en ejercicios estructurados, mientras que en preguntas abiertas la interpretación semántica y el peso de criterios implícitos tienen mayor relevancia.

Por otro lado, el sesgo positivo identificado en los tres reactivos, aunque de baja magnitud, indica que el LLM presenta una ligera tendencia a sobreestimar las respuestas en comparación con el docente. Este comportamiento es coherente con investigaciones que sugieren que los modelos de lenguaje, al priorizar la coherencia y completitud textual, pueden otorgar valor adicional a elementos formales incluso si el contenido presenta deficiencias sustantivas.

Asimismo, la prueba de Wilcoxon evidenció que, salvo en el ejercicio b, no existen diferencias estadísticamente significativas entre las evaluaciones del LLM y las del docente. La discrepancia en este caso podría atribuirse a la forma en que el LLM interpreta los criterios técnicos del problema de optimización, particularmente en la ponderación de procedimientos parciales correctos, mientras que el docente pudo aplicar criterios más estrictos.

En conjunto, los hallazgos respaldan la viabilidad del uso de LLM como herramienta de apoyo a la evaluación, especialmente en contextos donde las tareas poseen un formato claro y criterios de calificación bien definidos. No obstante, la presencia de sesgo positivo y la menor concordancia en reactivos abiertos sugieren la necesidad de calibrar los *prompts* y

validar periódicamente los resultados frente a la evaluación humana para garantizar equidad y alineación con los estándares académicos.

Conclusiones

El estudio realizó una comparación entre la evaluación de un docente y un modelo de lenguaje en una unidad de aprendizaje sobre optimización por métodos heurísticos. Se observó un alto nivel de coincidencia en las calificaciones, especialmente en ejercicios prácticos con criterios estructurados. El modelo de lenguaje asignó puntuaciones ligeramente superiores a las del docente, un aspecto relevante para considerar en su uso.

Se concluye que los LLM pueden ser incorporados como recurso complementario en procesos de evaluación, reduciendo la carga de trabajo docente con retroalimentación rápida al estudiantado. Sin embargo, su uso debe acompañarse de calibraciones específicas, revisión por parte del profesorado y diseño cuidadoso de las instrucciones de evaluación, especialmente en actividades que involucren respuestas abiertas o interpretaciones conceptuales.

Como líneas futuras, se propone ampliar la muestra a diferentes asignaturas y niveles académicos, así como explorar la combinación de evaluación automática y revisión docente en un esquema híbrido que maximice precisión, eficiencia y equidad en la valoración del aprendizaje.

Referencias

Jolliffe, I. (2011). Principal Component Analysis. En: Lovric, M. (eds), *International Encyclopedia of Statistical Science*. Springer. https://doi.org/10.1007/978-3-642-04898-2_455

Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., & Campanelli, M. (2025). Using large language models for grading in education: An applied test for physics. *Physics Education*, *60*(3), 035006. https://iopscience.iop.org/article/10.1088/1361-6552/adb92b/meta

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Stasuik, N. C. (2025). *Evaluating LLM performance in essay assessment: A comparative analysis of AI grading and feedback systems for university English courses*. [Tesis de licenciatura, University of British Columbia]. https://open.library.ubc.ca/soa/cIRcle/collections/undergraduateresearch/52966/items/1.0448868

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150–166. https://doi.org/10.1111/bjet.13494

INTELIGENCIA ARTIFICIAL

NUEVAS EXPERIENCIAS ACADÉMICAS

ISBN: 978-968-9724-12-4

Trans digital editorial