Título: Evaluación Comparativa de Modelos de Lenguaje de Inteligencia Artificial vs Evaluación Humana en Trabajos Académicos: Un Caso de Estudio en la Materia "Proyecto de Innovación"

Mónica Cobián Alvarado

1. Resumen

La integración de la inteligencia artificial (IA) en la evaluación educativa representa uno de los avances más significativos en la transformación de los procesos académicos contemporáneos. Este estudio presenta una evaluación comparativa entre tres modelos de lenguaje de inteligencia artificial (ChatGPT4o, Claude Sonnet 4, y DeepSeek-V3) y la evaluación humana tradicional en el contexto de trabajos académicos de estudiantes de sexto semestre de Ingeniería de Software en la materia de Proyecto de Innovación. La investigación empleó un diseño comparativo descriptivo con una muestra de cinco equipos estudiantiles, evaluando cinco criterios específicos mediante una rúbrica estandarizada: hoja de portada, índice, desarrollo del contenido, conclusiones y referencias consultadas. Los resultados revelan diferencias significativas en los patrones de evaluación entre los modelos de IA y el evaluador humano, con la evaluación humana mostrando la media más alta (3.88), seguida de ChatGPT-4o (3.67), mientras que Claude Sonnet y DeepSeek-V3 obtuvieron medias idénticas (3.12). Los modelos de IA demostraron correlaciones altas entre sí, especialmente DeepSeek-V3 y la evaluación humana (r = 0.987), sugiriendo alineación en la identificación de patrones de calidad. Los hallazgos indican que los modelos de IA pueden complementar efectivamente la evaluación realizada por el profesor, proporcionando retroalimentación consistente, aunque requieren ajustes específicos dependiendo del contexto pedagógico.

Palabras clave: inteligencia artificial, evaluación educativa, modelos de lenguaje, educación superior, evaluación automatizada, ChatGPT, Claude, DeepSeek

Contexto y Problemática

La evaluación académica constituye uno de los pilares fundamentales del proceso educativo, funcionando como mecanismo de retroalimentación, medición del aprendizaje y herramienta de mejora continua tanto para los estudiantes como para las instituciones educativas (Magistrum University, 2024). En el contexto de la educación superior, particularmente en carreras como Ingeniería de Software, la evaluación de trabajos académicos requiere un análisis detallado y multidimensional que considere aspectos técnicos, metodológicos y de presentación. Sin embargo, los métodos tradicionales de evaluación enfrentan desafíos significativos relacionados con la consistencia, objetividad, tiempo requerido y escalabilidad (Analytikus, 2024).

La evaluación tradicional realizada por el profesor, aunque rica en contexto y comprensión pedagógica, presenta limitaciones inherentes que han sido ampliamente documentadas en la literatura educativa. Entre estas limitaciones se encuentran la variabilidad inter-evaluador, la subjetividad en la aplicación de criterios, la fatiga del evaluador que puede afectar la consistencia, y el considerable tiempo requerido para proporcionar retroalimentación detallada a cada estudiante (Eniversy, 2024).

En este contexto, la aparición de la inteligencia artificial, específicamente los modelos de lenguaje de gran escala, ha abierto nuevas posibilidades para la automatización y mejora de los procesos de evaluación educativa (Megaprofe, 2025). Los avances recientes en procesamiento de lenguaje natural han permitido el desarrollo de sistemas capaces de analizar, comprender y evaluar textos académicos con niveles de sofisticación anteriormente inalcanzables.

Estado del arte

La investigación sobre el uso de inteligencia artificial en evaluación educativa ha experimentado un crecimiento exponencial en los últimos años, reflejando tanto el avance tecnológico como la necesidad práctica de soluciones escalables para la evaluación académica. La aplicación de IA en la evalaucippon permite precisión y detalle, retroalimentación instantánea a los estudiantes, ahorro de tiempo para el docente, entre otros más beneficios (Bustamante, 2024).

Almegren et al. (2024) realizaron un estudio comparativo entre herramientas de IA y evaluadores humanos en la evaluación de ensayos de estudiantes de inglés como lengua extranjera, encontrando que las herramientas de IA proporcionaron retroalimentación de alta calidad pero calificaron consistentemente más bajo que los evaluadores humanos. Este hallazgo sugiere patrones sistemáticos en las diferencias entre evaluación automatizada y humana que requieren investigación adicional.

González Fernández et al. (2025) destacan la eficiencia y rapidez como beneficios primarios, señalando que la IA puede evaluar grandes cantidades de trabajos en una fracción del tiempo requerido por evaluadores humanos. Adicionalmente, la objetividad y consistencia emergen como características distintivas, ya que los algoritmos de IA aplican criterios de manera uniforme, eliminando la variabilidad inter-evaluador que caracteriza la evaluación humana tradicional.

Justificación y Objetivos

La necesidad de investigación empírica sobre la efectividad comparativa de modelos de IA versus evaluación humana en contextos educativos específicos es evidente y urgente. La materia de Proyecto de Innovación en la carrera de Ingeniería de Software presenta un caso de estudio particularmente relevante debido a la naturaleza multidisciplinaria de los trabajos estudiantiles.

El objetivo general de esta investigación es evaluar la efectividad comparativa de tres modelos de inteligencia artificial versus evaluación humana en la evaluación de trabajos académicos de estudiantes de Ingeniería de Software en la materia de Proyecto de Innovación.

Los objetivos específicos son:

- Analizar la concordancia entre las evaluaciones realizadas por los diferentes modelos de IA y el evaluador humano.
- Identificar las fortalezas y debilidades específicas de cada modelo de IA.
- Evaluar la calidad de la retroalimentación proporcionada.
- Proponer recomendaciones basadas en evidencia para la implementación efectiva de sistemas de evaluación automatizada.

Diseño del Estudio

Esta investigación empleó un diseño comparativo descriptivo de corte transversal para evaluar la concordancia y diferencias entre la evaluación automatizada realizada por modelos de inteligencia artificial y la evaluación humana tradicional realizada por el profesor en trabajos académicos. El enfoque metodológico se fundamenta en el paradigma cuantitativo con elementos cualitativos, permitiendo tanto el análisis estadístico de las puntuaciones como la interpretación contextual de las observaciones proporcionadas por los evaluadores (Creswell & Creswell, 2018).

El diseño comparativo se seleccionó como el más apropiado para este estudio debido a su capacidad para examinar simultáneamente múltiples enfoques de evaluación aplicados a los mismos objetos de estudio, minimizando así las variables confusoras relacionadas con diferencias en el contenido o calidad de los trabajos evaluados (Campbell & Stanley, 2015).

Participantes y Contexto

El estudio se desarrolló en la Facultad de Telemática, una institución reconocida por su enfoque innovador en la formación de profesionales en tecnologías de la información y comunicación. La población objetivo consistió en estudiantes de sexto semestre de la carrera de Ingeniería de Software. La muestra del estudio incluyó cinco equipos de estudiantes, cada uno compuesto por múltiples integrantes que colaboraron en el desarrollo de sus respectivos proyectos de innovación.

Instrumentos de Evaluación

El instrumento principal de evaluación consistió en una rúbrica estandarizada específicamente diseñada para la evaluación del estado del arte en proyectos de innovación. La rúbrica comprende cinco criterios fundamentales de evaluación, cada uno con un valor máximo de cinco puntos, para un total de 25 puntos posibles, los cuales se mencionan a continuación: hoja de portada, índice, desarrollo del contenido, conclusiones y referencias consultadas.

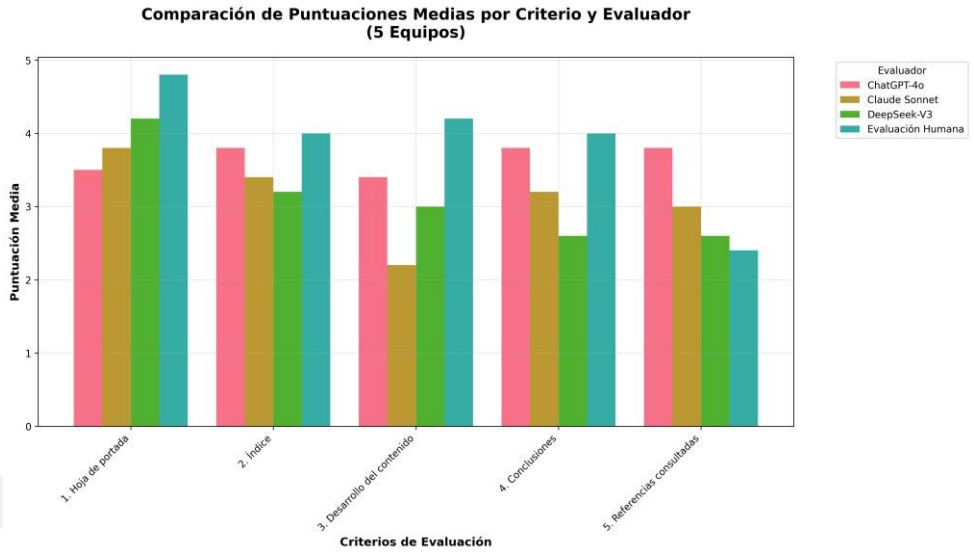
Modelos de Inteligencia Artificial Evaluados

La selección de los modelos de inteligencia artificial se basó en criterios de representatividad, disponibilidad y capacidades técnicas relevantes. ChatGPT-4o, desarrollado por OpenAI (2024), representa la evolución más reciente de la familia GPT de modelos de lenguaje. Claude Sonnet 4, desarrollado por Anthropic (2024), se distingue por su enfoque en seguridad, alineación y razonamiento ético. DeepSeek-V3, desarrollado por DeepSeek (2024), representa un enfoque de código abierto hacia el desarrollo de modelos de lenguaje de gran escala.

Resultados

El análisis de los datos recolectados de cinco equipos estudiantiles revela patrones distintivos en las evaluaciones realizadas por los diferentes modelos. La evaluación humana demostró la puntuación media más alta (M = 3.88, SD = 1.62), seguida por ChatGPT-4o (M = 3.67, SD = 1.40), mientras que Claude Sonnet y DeepSeek-V3 obtuvieron medias idénticas (M = 3.12, SD = 1.59 y 1.48 respectivamente). Estos resultados sugieren que la evaluación humana mantiene una tendencia hacia puntuaciones más altas (ver Figura 1), posiblemente reflejando una mayor consideración de factores contextuales y una aplicación más flexible de los criterios de evaluación (Brookhart, 2013).

Figura 1. Comparación de Puntuaciones Medias por Criterio y Evaluador (5 Equipos)



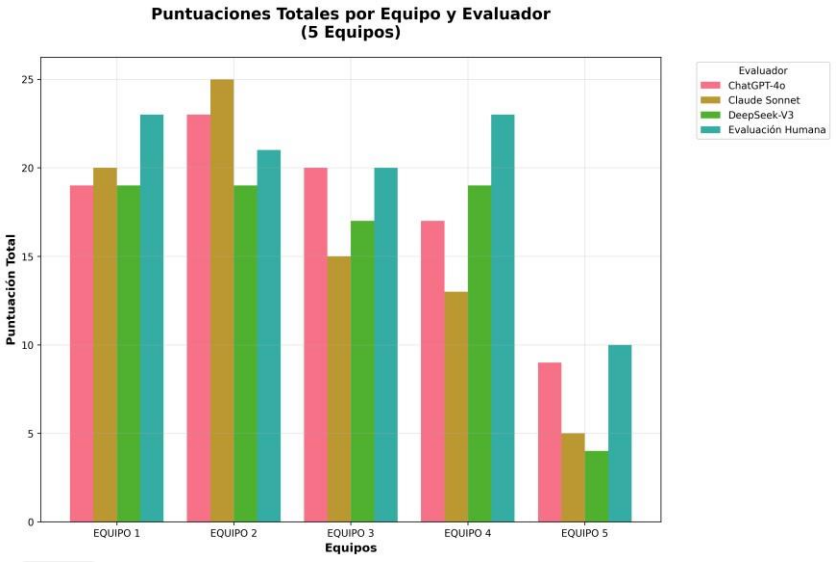
En la Tabla 1 se muestra la media y la desviación estándar para cada uno de los criterios que conforman la rúbrica (filas) según el modelo utilizado (columnas).

Tabla 1. Estadísticas Descriptivas por Criterio (5 Equipos)

Criterio	ChatGPT-4o (M±SD)	Claude Sonnet (M±SD)	DeepSeek-V3 (M±SD)	Evaluación Humana (M±SD)
1. Hoja de portada	3.50±1.00	3.80±1.79	4.20±1.79	4.80±0.45
2. Índice	3.80±1.79	3.40±1.67	3.20±2.05	4.00±2.24
3. Desarrollo del contenido	3.40±1.67	2.20±1.79	3.00±1.41	4.20±0.84
4. Conclusiones	3.80±1.10	3.20±1.48	2.60±0.89	4.00±0.71
5. Referencias consultadas	3.80±1.79	3.00±1.41	2.60±0.89	2.40±2.30

Las puntuaciones totales que obtuvo cada equipo se presentan en la figura 2.

Figura 2. Puntuaciones Totales por Equipo y Evaluador (5 Equipos)

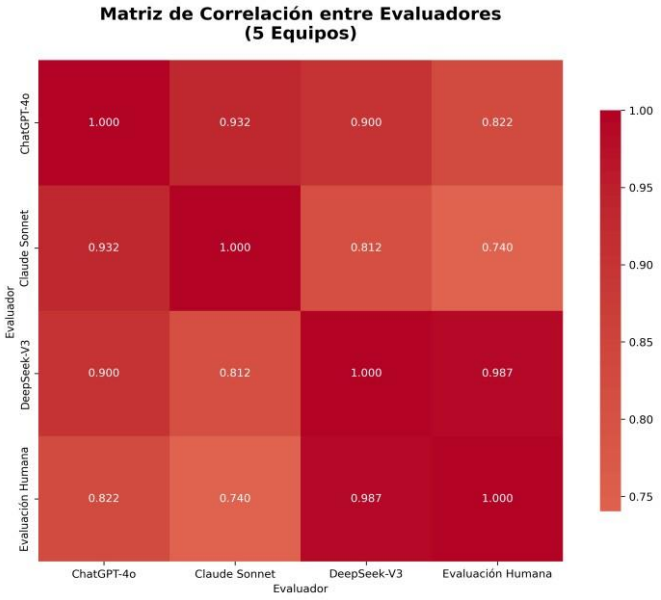


Análisis de Concordancia entre Evaluadores

El análisis de correlación entre evaluadores revela patrones de concordancia notablemente diferentes a los esperados inicialmente. La correlación más alta se observó entre DeepSeek-V3 y la evaluación humana (r = 0.987), indicando una alineación casi perfecta en la identificación de patrones de calidad relativa entre

los trabajos evaluados (ver Figura 3). ChatGPT-4o y Claude Sonnet también mostraron una correlación alta (r = 0.932), sugiriendo convergencia en sus enfoques de evaluación. Estas correlaciones altas indican que, a pesar de las diferencias en puntuaciones absolutas, los evaluadores tienden a identificar patrones similares de calidad relativa entre los trabajos estudiantiles.

Figura 3. Matriz de Correlación entre Evaluadores (5 Equipos)



Análisis de Puntuaciones Totales por Equipo

El examen de las puntuaciones totales por equipo (ver Tabla 2) revela patrones distintivos que resaltan las diferencias sistemáticas entre evaluadores. Del análisis realizado se observa que el EQUIPO 2 obtuvo las puntuaciones más altas de Claude Sonnet (25 puntos) y ChatGPT-4o (23 puntos), mientras que los EQUIPOS 1 y 4 recibieron las puntuaciones más altas de la evaluación humana (23 puntos cada uno). Notablemente, el EQUIPO 5 mostró puntuaciones consistentemente bajas en todas las evaluaciones, sugiriendo deficiencias objetivas en la calidad del trabajo que fueron identificadas uniformemente por todos los sistemas de evaluación.

Tabla 2. Puntuaciones Totales por Equipo (5 Equipos)

Equipo	ChatGPT-4o	Claude Sonnet	DeepSeek-V3	Evaluación Humana
EQUIPO 1	19	20	19	23
EQUIPO 2	23	25	19	21
EQUIPO 3	20	15	17	20
EQUIPO 4	17	13	19	23
EQUIPO 5	9	5	4	10

Fortalezas y Debilidades por Modelo

ChatGPT-4o demostró fortalezas particulares en consistencia y equilibrio en la evaluación, con la segunda media más alta (M = 3.67) y una desviación estándar moderada (SD = 1.40). Claude Sonnet 4 se distinguió por su rigor académico, especialmente en la evaluación del desarrollo del contenido (M = 2.20), reflejando estándares particularmente exigentes. DeepSeek-V3 mostró la alineación más alta con la evaluación humana (r = 0.987), sugiriendo capacidad para detectar aspectos de la calidad académica acordes con el juicio experto. La evaluación humana demostró fortalezas distintivas en flexibilidad contextual, manteniendo la media más alta (M = 3.88) mientras mostraba mayor variabilidad (SD = 1.62), indicando consideración de factores contextuales específicos (Popham, 2017).

2. Conclusiones

Este estudio comparativo entre modelos de inteligencia artificial y evaluación humana en trabajos académicos de cinco equipos de estudiantes de Ingeniería de Software ha proporcionado información valiosa sobre las capacidades, limitaciones y potencial de implementación de sistemas de evaluación automatizada en contextos de educación superior.

Los resultados confirman que los modelos de IA evaluados poseen capacidades significativas para la evaluación de trabajos académicos, aunque con

características y enfoques diferenciados. La evaluación humana mantuvo la puntuación media más alta (M = 3.88), reflejando una consideración holística que incluye factores contextuales y pedagógicos. Sin embargo, la correlación casi perfecta entre DeepSeekV3 y la evaluación humana (r = 0.987) sugiere que este modelo identifica efectivamente los patrones de calidad de los evaluadores expertos.

Las diferencias sistemáticas observadas entre evaluadores reflejan diversas filosofías y prioridades en la evaluación académica. Claude Sonnet demostró el enfoque más riguroso, particularmente en criterios de contenido, mientras que ChatGPT-4o mostró un balance efectivo entre rigor y reconocimiento del progreso estudiantil.

Las limitaciones identificadas incluyen el tamaño de la muestra (cinco equipos) y la naturaleza específica del contexto. Las directrices para investigación futura incluyen la expansión a muestras más grandes y la evaluación longitudinal de la implementación de sistemas automatizados.

En términos de implicaciones prácticas, este estudio sugiere que la implementación exitosa de evaluación automatizada requiere un enfoque contextualizado que considere las fortalezas específicas de cada modelo. Los hallazgos indican que DeepSeek-V3 podría ser particularmente efectivo como complemento a la evaluación humana, mientras que Claude Sonnet podría ser valioso para mantener estándares académicos rigurosos (Black & Wiliam, 2009).

3. Referencias

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of
AI feedback: A comparative study of AI and human essay grading. Innovations in Education and
Teaching International. https://doi.org/10.1080/14703297.2024.2437122

Analytikus. (2024, junio 2). El Futuro de la Educación y la IA: Evaluación y Retroalimentación Automatizadas. https://es.analytikus.com/post/el-futuro-de-la-educaci%C3%B3n-y-la-iaevaluaci%C3%B3n-y-retroalimentaci%C3%B3n-automatizadas

Anthropic. (2024). Claude 3.5 Sonnet: Constitutional AI for Helpful, Harmless, and Honest AI. Anthropic
Research. https://www.anthropic.com/claude

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5-31.

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. ASCD.

Bustamante, P. (2024, enero 21). Inteligencia artificial en evaluación educativa. https://aulasimple.ai/blog/inteligencia-artificial-en-evaluacion-educativa/

Campbell, D. T., & Stanley, J. C. (2015). Experimental and quasi-experimental designs for research.
Ravenio Books.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications.

DeepSeek. (2024). DeepSeek-V3: An Open-Source Large Language Model. DeepSeek Research. https://deepseek.com/

Eniversy. (2024, noviembre 27). Evaluación automatizada: ventajas y desventajas de utilizar inteligencia artificial en la corrección de exámenes y

pruebas. https://eniversy.com/articulos/articulo-evaluacionautomatizada-ventajas-y-desventajas-de-utilizar-inteligencia-artificial-en-la-correccion-de-examenes-ypruebas-5546

González Fernández, M. O., Romero-López, M. A., Sgreccia, N. F., & Latorre Medina, M. J. (2025).
Marcos normativos para una IA ética y confiable en la educación superior: estado de la cuestión. RIED-
Revista Iberoamericana de Educación a Distancia, 28(2). https://doi.org/10.5944/ried.28.2.43511

Magistrum University. (2024, junio 24). Evaluación Automatizada: El Rol de la IA en la Evaluación de Estudiantes. https://magistrum.university/evaluacion-automatizada-el-rol-de-la-ia-en-la-evaluacion-deestudiantes/

Megaprofe. (2025, febrero 25). 7 herramientas de IA para mejorar la evaluación de los estudiantes. https://megaprofe.es/7-herramientas-de-ia-para-mejorar-la-evaluacion-de-los-estudiantes/

OpenAI. (2024). GPT-4o Technical Report. OpenAI Research. https://cdn.openai.com/gpt-4o-systemcard.pdf?utm_source=chatgpt.com

|Popham, W. J. (2017). Classroom assessment: What teachers need to know. Pearson Education.