Comparación del Desempeño Humano y de la Inteligencia Artificial en la Revisión de Tesis

Comparison of Human Performance and Artificial Intelligence in Thesis Evaluation

Juan Guillermo Muñoz Tapia
Universidad de Aconcagua | Pedro de Villagra, 2265, 8540000, Vitacura | Chile
Juan.munoz@uac.cl
https://orcid.org/0000-0002-7260-516X

Recepción 00/00/2025 · Aceptación 00/00/2025 · Publicación 00/00/2025

Resumen
El aumento en la matrícula universitaria, impulsado por la implementación de clases en modalidad online, ha generado un incremento significativo en el número de tesis de pregrado presentadas para su evaluación. En una universidad con presencia regional, la disponibilidad de aproximadamente veinte docentes para revisar más de trescientas tesis anualmente dificulta garantizar un proceso de evaluación exhaustivo y homogéneo. Este estudio tuvo como objetivo comparar el desempeño del juicio humano y de la inteligencia artificial en la revisión de tesis de pregrado pertenecientes a diversas áreas disciplinares. Se adoptó un enfoque positivista, con un diseño cuantitativo y comparativo, evaluando un total de 118 tesis mediante ambas modalidades de revisión. Los resultados iniciales evidenciaron que la inteligencia artificial otorgó calificaciones más benevolentes que las evaluaciones humanas. Sin embargo, tras la implementación de ajustes metodológicos, se observó una tendencia de la inteligencia artificial a emitir evaluaciones ligeramente más estrictas que el juicio humano. Se concluye que, aunque la inteligencia artificial no reproduce de manera exacta la calificación otorgada por un docente, puede constituir una herramienta valiosa para el análisis y retroalimentación de tesis, favoreciendo la eficiencia del proceso evaluativo y contribuyendo al aseguramiento de la calidad académica. Este aporte resulta relevante para instituciones con alta carga de revisiones, al proponer un recurso complementario que optimiza tiempos sin comprometer la rigurosidad del proceso.
Palabras clave: inteligencia artificial, educación superior, tesis, análisis cuantitativo.

Abstract
The increase in university enrollment driven by the implementation of online classes has led to a significant rise in the number of undergraduate theses submitted for evaluation. At a regional university, the availability of approximately twenty faculty members to review more than three hundred theses annually poses a challenge to ensuring a thorough and consistent evaluation process. This study aimed to compare the performance of human judgment and artificial intelligence in the review of undergraduate theses from various disciplinary areas. A positivist approach was adopted, employing a quantitative and comparative design, and a total of 118 theses were evaluated through both review modalities. Initial results indicated that artificial intelligence assigned more lenient grades than human evaluations. However, after methodological adjustments were applied, artificial intelligence tended to issue slightly stricter evaluations compared to human judgment. It is concluded that, although artificial intelligence does not exactly replicate the grades assigned by faculty members, it can serve as a valuable tool for thesis analysis and feedback, benefiting both faculty and students. This contribution is particularly relevant for institutions with a high review workload, as it offers a complementary resource that optimizes time without compromising the rigor of the evaluation process.
Keywords: Artificial intelligence, Higher education, Thesis, Quantitative analysis.

1. Introducción

La revisión de trabajos de titulación constituye un proceso esencial en la formación universitaria, pues asegura que los egresados demuestren competencias investigativas y disciplinares acordes con los estándares académicos. En las últimas décadas, la masificación del acceso a la educación superior, potenciada por la implementación de modalidades de enseñanza en línea, ha generado un incremento sostenido en la matrícula y, en consecuencia, en el número de tesis de pregrado

presentadas para su evaluación. Este fenómeno, aunque positivo en términos de democratización educativa, ha impuesto una carga considerable sobre los comités evaluadores, especialmente en instituciones con presencia regional y recursos humanos limitados. La situación se agrava cuando la relación entre el número de revisores disponibles y la cantidad de tesis a evaluar es desproporcionada, lo que compromete la oportunidad, exhaustividad y uniformidad del proceso evaluativo.

En este contexto, la literatura reciente ha explorado alternativas tecnológicas que contribuyan a optimizar procesos académicos sin detrimento de la calidad. Entre estas, la inteligencia artificial (IA) ha emergido como una herramienta con potencial para asistir o complementar la labor docente en diversas áreas, desde la retroalimentación en tareas escritas hasta la evaluación automatizada de contenidos académicos. Estudios previos han evidenciado que los modelos de IA pueden ofrecer criterios de análisis coherentes, rápidos y replicables, aunque persisten interrogantes sobre su precisión, sesgos y capacidad para interpretar matices disciplinares. A pesar del creciente interés, la comparación sistemática entre el juicio humano y la evaluación realizada por IA en el contexto específico de tesis de pregrado continúa siendo limitada, lo que abre un campo de investigación relevante tanto para la educación superior como para la innovación tecnológica.

La necesidad de explorar este ámbito se hace particularmente evidente en universidades que enfrentan altos volúmenes de revisiones con dotaciones reducidas de personal académico. En dichas condiciones, el uso de IA podría no solo aliviar la carga de trabajo, sino también contribuir a homogeneizar criterios y detectar áreas de mejora en los trabajos estudiantiles. Sin embargo, su implementación requiere una validación rigurosa que considere la variabilidad entre disciplinas, la naturaleza de los criterios de evaluación y las posibles divergencias con la valoración humana. El presente estudio se inserta en este escenario, planteando la necesidad de establecer evidencias empíricas que permitan dimensionar la concordancia y discrepancia entre ambos enfoques evaluativos.

El problema de investigación que guía este trabajo se centra en determinar en qué medida el juicio emitido por la inteligencia artificial difiere o coincide con el juicio humano en la revisión de tesis de pregrado pertenecientes a diversas áreas del conocimiento, considerando que el volumen anual de trabajos supera ampliamente la capacidad operativa de los revisores disponibles. El objetivo general consiste en comparar, mediante un análisis estadístico, las evaluaciones de tesis realizadas por docentes y por IA, identificando patrones de coincidencia y divergencia, así como posibles ajustes para optimizar la herramienta tecnológica.

Metodológicamente, el estudio adopta un paradigma positivista, con un enfoque cuantitativo y diseño comparativo, evaluando un conjunto de 118 tesis a través de ambas modalidades de revisión. Los datos obtenidos se someten a pruebas estadísticas que permiten contrastar las diferencias y similitudes entre los dos métodos, buscando no solo establecer niveles de concordancia, sino también identificar tendencias hacia evaluaciones más benevolentes o más estrictas por parte de la IA.

El aporte potencial de esta investigación radica en ofrecer evidencia empírica que permita fundamentar la integración de herramientas de IA en procesos de evaluación académica de alta exigencia, optimizando tiempos y recursos sin comprometer la rigurosidad. Asimismo, sus resultados pueden servir de referencia para el diseño de políticas institucionales que incorporen la tecnología como apoyo a la labor docente, fortaleciendo la calidad de la educación superior en entornos con alta demanda de revisión y limitaciones de personal. Con ello, se contribuye a un debate global sobre el papel de la inteligencia artificial en la educación, no como sustituto del juicio humano, sino como un complemento estratégico para la mejora continua de los procesos formativos y evaluativos.

2. Marco teórico

La transformación digital en la educación superior ha generado un escenario en el que la inteligencia artificial generativa (GenAI) se presenta como un recurso estratégico para optimizar procesos académicos, entre ellos la evaluación de trabajos de titulación. La literatura reciente reconoce que la irrupción de herramientas como ChatGPT ha ampliado las posibilidades de apoyo a docentes y estudiantes, ofreciendo asistencia en redacción, revisión de contenidos y generación de retroalimentación (Baidoo-Anu & Owusu Ansah, 2023; Farrelly & Baker, 2023). Sin embargo, la integración efectiva de estas tecnologías en el contexto universitario requiere comprender sus alcances, limitaciones y condiciones óptimas de uso.

En este sentido, estudios como los de Adeyele y Ramnarain (2024a, 2024b) han explorado la incorporación de ChatGPT en entornos de aprendizaje basado en la indagación, destacando su potencial para estimular el pensamiento crítico, pero también subrayando la necesidad de guías pedagógicas claras. A nivel institucional, la transformación digital implica no solo la adopción de nuevas tecnologías, sino también el rediseño de marcos normativos, competencias docentes y modelos de interacción educativa (Adiwijaya et al., 2025; Rahmadi, 2024). El desarrollo de habilidades de alfabetización digital en el profesorado se convierte, por tanto, en un requisito fundamental para garantizar la pertinencia de la implementación (Mardiana, 2024).

Diversos enfoques han documentado aplicaciones concretas de la IA en la educación superior, abarcando desde la asistencia en la enseñanza de programación (Xue et al., 2024) hasta su uso en entornos clínicos y de formación profesional (Durmuş Sarıkahya et al., 2025). Michel-Villarreal et al. (2023) y Kazimova et al. (2025) señalan que la IA no debe concebirse como sustituto del juicio humano, sino como un complemento que, adecuadamente calibrado, puede mejorar la consistencia y eficiencia de la evaluación académica. En esta línea, Romero-Rodríguez et al. (2023) y Robles y Ek (2024) destacan que la percepción de utilidad por parte de los estudiantes está vinculada a la calidad y pertinencia de la retroalimentación generada por la herramienta.

Las investigaciones recientes identifican tanto beneficios como riesgos asociados al uso de IA generativa en contextos educativos. Entre los beneficios se incluyen la estandarización de criterios, la reducción de carga de trabajo docente y el apoyo a la escritura académica (del Mar Sánchez Vera, 2024; Memon & Kwan, 2025). No obstante, se advierten riesgos relacionados con sesgos algorítmicos, errores de contenido y sobredependencia tecnológica (Chiappe et al., 2025; Reyes et al., 2025). La confianza en las evaluaciones generadas por IA depende, en gran medida, de la transparencia de los procesos y de la presentación de la incertidumbre asociada a sus resultados (Reyes et al., 2025).

En el ámbito de la gestión educativa, Sposato (2025) propone una taxonomía para la integración de la IA en liderazgo educativo, mientras que Pycka y Zastempowski (2025) evidencian cómo las técnicas de aprendizaje automático pueden optimizar auditorías y procesos de control académico. Estas perspectivas refuerzan la idea de que la incorporación de IA en la revisión de tesis requiere no solo un enfoque técnico, sino también una estrategia de gobernanza institucional que asegure calidad y equidad.

Modelos como el Stimulus-Organism-Behavior-Consequence (Biswas et al., 2025) han sido aplicados para explicar la intención de uso y recomendación de ChatGPT, destacando que la adopción de la herramienta depende tanto de factores tecnológicos como de la experiencia del usuario. Por otra parte, Rahiman y Kodikal (2024) subrayan que el éxito de la IA en educación radica en su integración en un ecosistema de aprendizaje más amplio, donde se articule con metodologías activas y objetivos formativos claros.

En términos macro, organismos internacionales como la OCDE (2021) han señalado que la transformación tecnológica en educación debe ir acompañada de políticas que fortalezcan la equidad en el acceso y el desarrollo de competencias digitales. Esto es coherente con el planteamiento de Zahari et al. (2018), quienes proponen un modelo de “universidad del futuro” sustentado en la convergencia entre innovación tecnológica, pedagogía y gestión institucional.

En síntesis, la literatura revisada converge en que la IA generativa, y en particular ChatGPT, ofrece un potencial significativo para optimizar procesos de evaluación en educación superior, siempre que su implementación esté respaldada por un marco pedagógico sólido, políticas institucionales claras y mecanismos de validación que garanticen la fiabilidad de sus resultados. Este cuerpo de evidencia justifica la pertinencia de investigaciones que comparen, de manera empírica y rigurosa, la concordancia y divergencia entre el juicio humano y la IA en la revisión de tesis de pregrado.

3. Metodología

El presente estudio se enmarcó en el paradigma positivista, adoptando un enfoque cuantitativo orientado a la medición objetiva y la comparación de resultados entre dos modalidades de evaluación de tesis de pregrado: juicio humano e inteligencia artificial. El diseño fue de tipo comparativo no experimental y de corte transversal, dado que se analizaron datos recolectados en un único momento temporal, sin manipulación de variables independientes, con el propósito de identificar patrones de coincidencia y divergencia entre ambos métodos evaluativos.

La investigación se desarrolló en una universidad con presencia regional, caracterizada por un alto volumen anual de tesis de pregrado y una disponibilidad limitada de docentes revisores. La muestra estuvo compuesta por 118 tesis defendidas en distintas áreas disciplinares, seleccionadas mediante un muestreo no probabilístico por conveniencia (se utilizaron las tesis que la universidad proporciono). Se descartaron aquellas tesis que estaban incompletas y aquellas que no contaban con nota final. Los criterios de inclusión consideraron trabajos completos que hubiesen sido evaluados formalmente por al menos dos docentes bajo criterios institucionales vigentes. Se excluyeron tesis con evaluaciones incompletas o sin registro de la calificación final emitida por el comité evaluador. Las tesis proporcionadas por la universidad se encontraban revisadas y calificadas por los revisores humanos, los cuales fueron un total de 20 (en total de todas las diciplinas). Se desconoce el número de revisores por diciplina.

Para la recolección de datos se emplearon dos instrumentos principales: (1) un conjunto de *prompts* estandarizados diseñados para la plataforma ChatGPT-5 según la actualización realizada en el 2025, estructurados a partir de la rúbrica institucional de evaluación de tesis, con el fin de garantizar consistencia en la retroalimentación y en la asignación de calificaciones por parte de la IA; y (2) una tabla de comparación que registró las calificaciones otorgadas por los revisores humanos y por la IA para cada tesis, organizada por criterios evaluativos y calificación final. La discrepancia de las notas la universidad las considero como información complementaria, las calificaciones finales fueron las consideradas por los revisores humanos. El primer prompts utilizando para calificar las evaluaciones consistía en dos etapas primero se cargaba un textos que traía una serie de instrucciones (como se muestra en la figura 1) y posterior a eso se cargaba un archivo que contenía la rúbrica de evaluación. El primer prompts utilizado es el que se presenta en la figura 1)

Actúa como un profesor universitario con experiencia en dirección y evaluación de tesis de pregrado en [nombre del programa, por ejemplo: Derecho, Psicología, Ingeniería eléctrica, Ingeniería en minas, etec].

Tu tarea es revisar el siguiente capítulo de una tesis de pregrado. Usa los criterios de la rúbrica que te entregaré y adapta tu retroalimentación considerando el contexto de un estudiante que está en su proceso formativo final.

Evalúa el texto según los siguientes criterios y la siguiente rubrica.
1 Criterios:
Claridad y coherencia del planteamiento del problema
- Pertinencia del marco teórico y uso de fuentes académicas
- Adecuación de la metodología al problema de investigación
- Análisis de resultados: profundidad, interpretación y uso de evidencias
- Conclusiones: vínculo con objetivos, aportes y limitaciones
- Redacción académica: estilo, gramática y citación

2. Asigna una calificación cualitativa a cada criterio (Excelente, Bueno, Regular, Débil) y cuantitativa (según el puntaje de la rúbrica), para finalmente presentarle una calificación en relación al puntaje.
3. Entrega retroalimentación específica por cada criterio evaluado. Sé claro, pedagógico y fomenta la mejora del texto.
4. Considera que la misión, visión y valores de la universidad son [*aquí se debe ingresar la misión, visión y valores de la universidad*]
5. Considera el perfil de egreso de la carrera del estudiante la cual es [*Insertar aquí el perfil de egreso*]
6. Finalmente, haz una síntesis general con recomendaciones finales.
7. La evaluación debe ser en base a esta rúbrica [*Insertar o adjuntar la rubrica*]

[*Inserta aquí el texto del capítulo o sección de tesis*]

Figura 1: Primer prompt utilizado para evaluar las tesis. Elaboración propia.

Se debe considerar que el procedimiento a seguir consistía en pegar el prompt llenar los datos y adjuntar la tesis y la rúbrica de evaluación y luego se daba “Enter”.
Después del ajuste empelado, el cual consistía en una sola serie de ordenes ajustada a la nota según se muestra en la figura 2.

Actúa como un **profesor universitario experimentado y muy riguroso** en la dirección y evaluación de tesis de pregrado en **[nombre de la carrera]**.
Debes evaluar un capítulo de tesis siguiendo los **mismos estándares, tendencia de calificación y exigencia** que usan los docentes evaluadores de la Universidad de Aconcagua.
Calibración según evaluaciones humanas reales
Basado en datos históricos de evaluaciones, las notas humanas siguen este patrón:

- Trabajos con errores graves (planteamiento confuso, fallas metodológicas, citas incompletas) suelen obtener **entre 3,5 y 4,8**.
- Trabajos aceptables pero con mejoras necesarias obtienen **entre 4,9 y 5,5**.
- Trabajos sólidos, bien fundamentados pero no excepcionales, obtienen **entre 5,6 y 6,2**.
- Solo los trabajos excepcionales, sin errores y con aportes originales obtienen **6,3 a 7,0**.

Aplica este patrón al asignar tus calificaciones para **cada criterio**. No otorgues notas máximas si hay faltas relevantes, aunque el resto esté bien logrado.
Instrucciones estrictas

1. Usa la **rúbrica oficial** que te proporcionaré y no flexibilices criterios.
2. Aplica deducciones mínimas de **0,5 puntos por cada error grave** en:
    - Coherencia del planteamiento del problema.
    - Uso incorrecto o insuficiente de fuentes académicas.
    - Metodología mal definida o inconsistente.
    - Análisis superficial o sin evidencias.
    - Citación incorrecta o incompleta.
3. Si un criterio no está presente o está desarrollado de forma insuficiente, asígnale la calificación mínima posible según la rúbrica.

Criterios de evaluación
Evalúa y comenta cada uno de los siguientes criterios:

- Claridad y coherencia del planteamiento del problema.
- Pertinencia del marco teórico y uso de fuentes académicas.
- Adecuación de la metodología al problema de investigación.
- Análisis de resultados: profundidad, interpretación y uso de evidencias.
- Conclusiones: vínculo con objetivos, aportes y limitaciones.
- Redacción académica: estilo, gramática y citación.

Formato de respuesta obligatorio
Para cada criterio, responde así:
yaml
CopiarEditar
Criterio: [nombre del criterio]
Calificación cualitativa: [Excelente / Bueno / Regular / Débil]
Calificación cuantitativa: [número con un decimal, escala 1,0 a 7,0]
Retroalimentación: [comentario pedagógico detallado, explicando fortalezas, debilidades y justificando por qué la nota no es mayor]
Síntesis final
Al final, incluye:

- **Promedio final** (promedio aritmético de las calificaciones cuantitativas).
- **Evaluación global** del capítulo, manteniendo el nivel de exigencia.
- **Recomendaciones específicas** para alcanzar un nivel de excelencia.

Considera siempre:

- La misión, visión y valores de la universidad.
- El perfil de egreso de la carrera correspondiente.

[Aquí insertar la rúbrica oficial y la tesis a evaluar]

Figura 2: Segundo prompt ajustado utilizado para evaluar las tesis. Elaboración propia.

El prompt fue completado con los datos requeridos como “el perfil de egreso” o “valores, misión y visión” para que el prompt se adapte a las características particulares de la universidad y de la carrera. Junto con lo anterior se adjunta la rúbrica con la que se evaluar la tesis, la cual es particular a la facultad que corresponde. Por ejemplo para la facultad de salud, se utilizó una rubrica particular para la carrera de enfermería o kinesiología y para la facultad de ingeniería, se utilizó una rubrica correspondiente a carreras como ingeniería eléctrica e ingeniería en minas. A continuación se presenta un extracto de la rúbrica utilizada para ingeniería. El fracmento de la rúbrica que se presenta corresponde a la presentación del problema.

Criterios de evaluación: Planteamiento del problema

Dimensión evaluada: Claridad, contextualización y justificación del problema

Indicador	**Puntaje máximo**	**Descripción del nivel excelente (100%)**
1. **Claridad en la redacción del problema**	2 puntos	El problema está redactado de forma precisa, directa y comprensible; no presenta ambigüedades.
2. **Contextualización (temporal, geográfica y disciplinar)**	2 puntos	Se sitúa adecuadamente el problema en un contexto académico, social o profesional, delimitando espacio y tiempo si corresponde.
3. **Vinculación con una necesidad real o vacíos del conocimiento**	2 puntos	Se evidencia que el problema responde a una situación concreta, relevante y/o a un vacío teórico o práctico identificable.
4. **Relación clara con los objetivos y preguntas de investigación**	2 puntos	El planteamiento se vincula lógicamente con los objetivos generales y específicos; existe coherencia en el enfoque.
5. **Justificación del problema con respaldo teórico o empírico**	2 puntos	La relevancia del problema se argumenta con apoyo en literatura científica o datos empíricos; incluye referencias pertinentes y actualizadas.

Nivel Excelente (9–10 puntos)
El problema está **claramente formulado, contextualizado y justificado con base teórica o empírica**. Muestra una comprensión profunda del área de estudio y su relevancia investigativa.

Figura 3: Rubrica de la presentación del problema. Elaboración propia.

El procedimiento de recolección se llevó a cabo en tres fases. En la primera, se recopilaron las versiones digitales de las tesis y las calificaciones humanas correspondientes, asegurando que los datos estuvieran completos y anonimizados. En la segunda, cada tesis fue procesada en ChatGPT mediante los *prompts* previamente diseñados, obteniendo un informe evaluativo y una calificación

global según la rúbrica. En la tercera, se consolidaron los resultados en la tabla comparativa, generando un único archivo de datos para el análisis estadístico.
El análisis de datos se efectuó combinando herramientas de inteligencia artificial y software estadístico. Inicialmente, ChatGPT-5 se empleó como asistente de análisis para generar observaciones cualitativas estructuradas que complementaran la calificación cuantitativa. Posteriormente, los datos fueron procesados en *Jupyter Notebook* utilizando bibliotecas como *pandas* y *matplotlib* para la depuración, organización y visualización de resultados. Asimismo, se utilizó RStudio para realizar pruebas estadísticas de comparación, incluyendo medidas de concordancia interevaluador (índice Kappa) y análisis correlacional (coeficientes de Pearson y Spearman), con un nivel de significancia de 0,05.
En cuanto a las consideraciones éticas, el estudio cumplió con los principios establecidos en la Declaración de Helsinki para investigaciones con datos de carácter académico. Todos los participantes implicados en la entrega de tesis y revisión otorgaron su consentimiento informado para el uso de la información con fines de investigación. Se garantizó el anonimato de los autores de las tesis y de los revisores humanos, así como la confidencialidad de las calificaciones. El protocolo de investigación fue evaluado y aprobado por el comité de ética de la institución, asegurando el cumplimiento de las normativas nacionales e internacionales aplicables a estudios en educación superior.
La evaluación y revisión de las tesis se dividió en cinco fases donde la primera fase concite en la revisión humana de las tesis, en esta etapa el docente guía le da una nota al trabajo de la tesis del alumno; la segunda fase corresponde a en la recolección de la tesis y la nota dada por el profesor para el estudio; la tercera fase corresponde a la evaluación de la tesis por medio de inteligencia artificial, donde se incorpora en la interfaz de la IA la tesis el prompts y la rubrica correspondiente ala carrera; la cuarta fase corresponde a comparación de las notas por medio de un análisis estadístico y la quinta fase corresponde a la elaboración de conclusiones. Se debe mencionar que una vez evaluada las 118 tesis se ajustó el prompts y se realizó nuevamente la revisión de las 118 tesis.

4. Resultados

El análisis incluyó un total de 118 tesis de pregrado correspondientes a diversas áreas disciplinares. Para cada documento se dispuso de dos evaluaciones independientes: una emitida por docentes humanos siguiendo la rúbrica institucional y otra generada por el sistema de inteligencia artificial (IA) mediante prompts estandarizados. Los datos fueron organizados en una matriz de comparación que permitió contrastar las calificaciones globales y por criterio. Para la revisión de las tesis se realizó la elaboración de un promps, el cual fue ajustado después de la primera revisión y se evaluó nuevamente considerando los ajustes al promps
La presente tabla realiza un Análisis descriptivo de la comparación del juicio humano en la revisión de tesis comparado con el juicio de la IA. En la tabla se observa los datos obtenidos de una primera revisión de las tesis (Ant.) y una segunda revisión después de un ajuste en el promps (Nuev.) La columna “Variable” representa las características de estadística descriptiva de las notas obtenidas por humanos (Nota) y la nota obtenida por la inteligencia artificial (Nota IA). La columna “media (Ant.)” corresponde a la nota promedio obtenida por la versión del primer prompts y la columna “Media (Nuev.)” es la nota promedio del prompts ajustado. La columna “DE (Ant.)” es la desviación estándar de la nota obtenida con el primer prompts y ¨DE (Nuev.)¨ es la desviación estándar de la nota del segundo prompts. La columna “CV%(Ant.)” es el porcentaje coeficiente de varianza correspondiente a la nota del primer prompts, “CV% (Nuev.)” es el porcentaje de coeficiente de variación del segundo prompt.

Variable	Media (Ant.)	Media (Nuev.)	DE (Ant.)	DE (Nuev.)	CV% (Ant.)	CV% (Nuev.)
Nota	5.70	4.91	0.76	0.75	13.38	15.38
Nota IA	5.33	5.09	0.97	0.37	18.26	7.17

Tabla 1: Análisis descriptivo de la comparación de evaluación de tesis. Elaboración propia

La tabla muestra que en las **Notas** la media pasó de 5.70 a 4.91, con desviación estándar casi igual (0.76 a 0.75) y CV % de 13.38 a 15.38. En **Notas IA**, la media bajó de 5.33 a 5.09, la desviación estándar de 0.97 a 0.37 y el CV % de 18.26 a 7.17, reflejando mayor consistencia.

En la gráfica 1 se observa la comparación de la media, desviación estándar y coeficiente de variación de la nota obtenida por la revisión humana comparada con la nota obtenida por la nota obtenida por la IA (Nota IA). La barra azul corresponde a la nota del primer prompts y la barra azul al de segundo prompts.

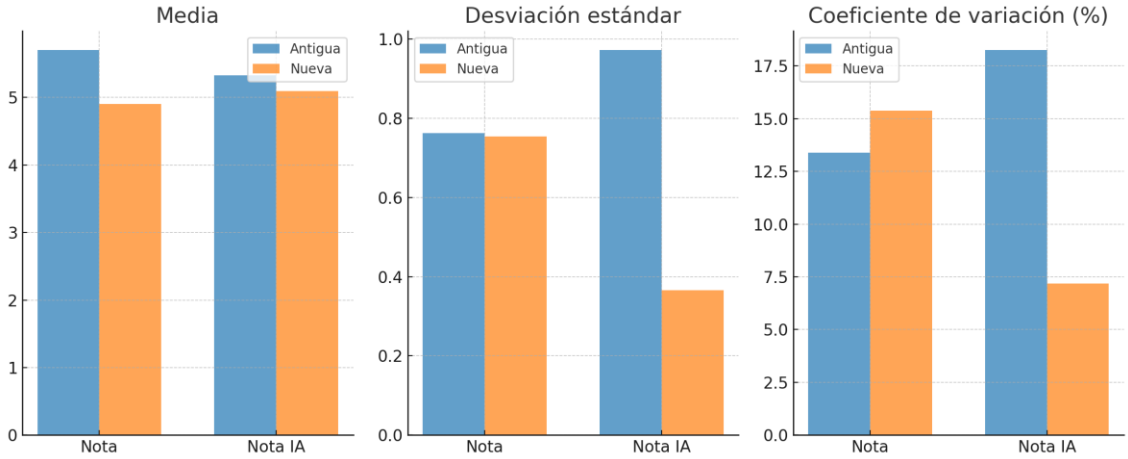

Figura 4: datos descriptivos entre la primera revisión y la segunda revisión. Elaboración propia.

El análisis comparativo presentado en el Gráfico 1 evidencia cambios relevantes entre la primera revisión (versión antigua) y la segunda revisión (versión nueva), tanto en la evaluación humana como en la evaluación realizada por inteligencia artificial. En términos de tendencia central (media), se observa una reducción en las calificaciones humanas al pasar de la revisión antigua a la nueva, mientras que las calificaciones otorgadas por la IA muestran un aumento leve, lo que sugiere un comportamiento más conservador inicial de la IA y un ajuste posterior hacia una valoración más estricta. En relación con la dispersión de los datos, la desviación estándar disminuye de manera significativa en la evaluación realizada por IA en la segunda revisión, lo que indica una mayor consistencia del modelo respecto de su primera iteración. Esta tendencia se refuerza al analizar el coeficiente de variación, donde la IA reduce su variabilidad porcentual desde valores altos en la primera revisión hasta el nivel más bajo entre todas las condiciones evaluadas, reflejando un comportamiento más estable y homogéneo. En contraste, la evaluación humana mantiene niveles similares de dispersión entre ambos momentos. En síntesis, los resultados muestran que los ajustes metodológicos aplicados al modelo de IA no solo modificaron la tendencia de las calificaciones, sino que también mejoraron su consistencia, acercando su comportamiento al estándar evaluativo esperado en un proceso académico formal.

En la tabla 2 se presentan los resultados de la correlación entre el juicio humano comparado con el juicio de la IA en la revisión de tesis. La tabla expone la correlación en la primera revisión y los resultados obtenidos en la segunda revisión. para este análisis inferencia se consideró la prueba de correlación de Pearson y la de Spearman debido a que en la nueva revisión de la IA los datos no se mostraban estadísticamente normalizados. La columna “Versión” indica si es la revisicon con el romper prompts (antigua) o con el segundo prompt (Nueva). La segunda columna “Tipo de correlación” Me indica si la correlación es la de Pearson o si es la correlación de Spearman. La columna “Correlación” indica el nivel de correlación. Finalmente en la columna “p-valor” corresponde al valor p de los intervalos.

Versión	Tipo de correlación	Coeficiente	p-valor
Antigua	Pearson	-0.1204	0.6691
Antigua	Spearman	-0.3065	0.2665
Nueva	Pearson	0.4406	0.1002
Nueva	Spearman	0.4390	0.1016

Tabla 2: correlación entre el juicio humano con el de la IA. Elaboración propia.

La tabla 2 muestra que en la versión Antigua la correlación entre Nota (Humana) y Nota (IA) fue negativa: Pearson -0.1204 (p=0.6691) y Spearman -0.3065 (p=0.2665). En la Nueva, ambas fueron positivas moderadas: Pearson 0.4406 (p=0.1002) y Spearman 0.4390 (p=0.1016), sin significancia estadística.

A continuación, se muestra los resultados obtenidos por medio de la prueba t de Studen,

Versión	Prueba estadística	t	p-valor
Antigua	t-test pareado	-0.3880	0.7038
Nueva	t-test pareado	-1.0662	0.3044

Tabla 3: Prueba de t de Student para comparar las medias entre los ambos juicios. Elaboración propia.

Se evaluó la asociación entre las calificaciones emitidas por modelos de inteligencia artificial y la calificación humana utilizando el coeficiente de correlación de Pearson, acompañado del p-value e intervalo de confianza del 95%. Ninguna de las IA presentó correlación significativa con la calificación humana (r entre −0.12 y 0.12; p > .05 en todos los casos). Los intervalos de confianza incluyeron el cero en todos los modelos, indicando ausencia de evidencia estadística que respalde una asociación entre ambas evaluaciones. Por lo tanto, las calificaciones generadas por IA no replican el criterio evaluativo humano.

Debido a que el análisis general no mostro evidencia estadística, se realizó el análisis por carrera, como se muestra en la figura 2. Es en esta figura que se muestra en el eje vertical las carreras analizadas, en el eje orizontal se muestran el nivel del intervalo de confianza al 95%

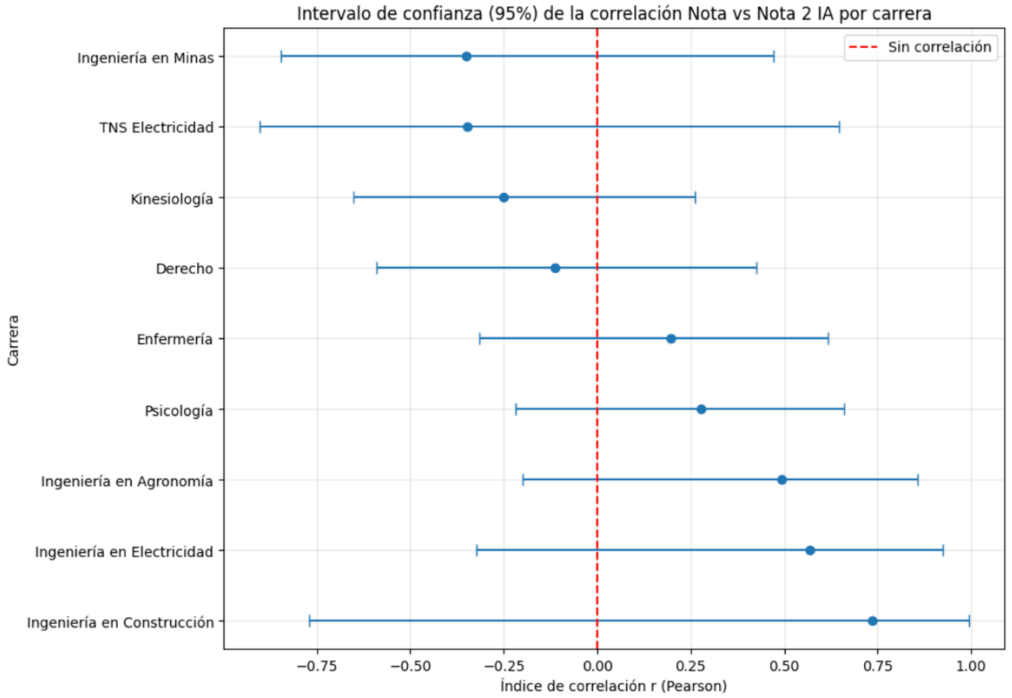

Figura 2: Intervalos de confianza por carrera. Elaboración propia.

La Figura 2 presenta los intervalos de confianza al 95% de los coeficientes de correlación entre la calificación otorgada por el docente y la calificación generada por la inteligencia artificial, desagregados por carrera. La línea punteada vertical ubicada en cero representa el punto de ausencia de correlación. Se observa que, en varias carreras, los intervalos de confianza atraviesan esta línea, lo que indica que no existe evidencia estadística suficiente para afirmar una relación consistente entre ambos tipos de evaluación en dichos programas académicos. Esto se aprecia con mayor claridad en carreras como Derecho, Enfermería, Psicología y TNS Electricidad, donde la amplitud de los intervalos sugiere una alta variabilidad en los puntajes y, por ende, una baja estabilidad en la relación entre el juicio humano y el de la IA. Por el contrario, algunas carreras muestran correlaciones más definidas: Ingeniería en Construcción e Ingeniería en Electridad presentan intervalos más desplazados hacia el lado positivo, lo que indica una mayor concordancia entre ambos evaluadores. De forma inversa, Kinesiología evidencia un valor de correlación central negativo, lo que implica una tendencia opuesta entre las calificaciones del docente y de la IA. En síntesis, los resultados sugieren que la relación entre la evaluación humana y la generada por IA no es uniforme entre carreras, sino que depende del contexto disciplinar y probablemente de la estructura metodológica de cada tesis. Esto respalda la hipótesis de que la IA puede comportarse de manera diferenciada según las características del contenido evaluado, lo que refuerza la necesidad de ajustar sus parámetros de uso para cada disciplina.

En la tabla 4 se observa se observa en la columna “Carrera” cada una de las carreras analizadas, la columna “r (Correlación)” presenta la correlación de Pearson que existe entre el la nota humana y la nota obtenida por el algoritmo en la segunda versión del prompt. La columna “IC 95% (Mín)” representa el intervalo de confianza en su límite por el extremo izquierdo y la columna “IC 95% (Máx)” re presenta el intervalo de confianza por el lado de la derecha. Las columna “n” representa al número de entrevistados

Carrera	r (Correlación)	IC 95% (Mín)	IC 95% (Máx)	n
Ingeniería en Construcción	0,736	-0,77	0,994	4

Ingeniería en Electricidad	0,567	-0,325	0,925	7
Ingeniería en Agronomía	0,492	-0,199	0,856	10
Psicología	0,277	-0,218	0,659	18
Enfermería	0,195	-0,315	0,618	17
Derecho	-0,112	-0,59	0,425	15
Kinesiología	-0,251	-0,653	0,261	17
TNS Electricidad	-0,347	-0,904	0,647	6
Ingeniería en Minas	-0,351	-0,846	0,47	8

Tabla 4: Correlación e intervalo de confianza por carrera. Elaboración propia.

La Tabla 4 muestra los valores del coeficiente de correlación de Pearson entre la calificación otorgada por los docentes y la asignada por la inteligencia artificial (versión ajustada del prompt), desagregados por carrera, junto con sus respectivos intervalos de confianza al 95% y el tamaño de muestra utilizado en cada caso. Los resultados revelan una alta variabilidad entre carreras, lo que sugiere que la relación entre el juicio humano y el algoritmo de IA no es homogénea en todos los contextos disciplinares. Se observan correlaciones positivas moderadas a altas en Ingeniería en Construcción (r = 0,736) e Ingeniería en Electricidad (r = 0,567), lo que indica una concordancia considerable entre las evaluaciones humanas y las generadas por IA en estas áreas. No obstante, los intervalos de confianza siguen siendo amplios debido al número reducido de tesis evaluadas, lo que limita la capacidad de generalización.

En contraste, otras carreras muestran correlaciones débiles o incluso negativas, como Derecho (r = –0,112), Kinesiología (r = –0,251), TNS Electricidad (r = –0,347) e Ingeniería en Minas (r = –0,351). Estas correlaciones negativas sugieren que, en dichas carreras, a medida que la calificación humana aumenta, la otorgada por IA tiende a disminuir, lo cual refleja un patrón evaluativo divergente. Además, en la mayoría de las carreras los intervalos de confianza incluyen el valor cero, lo que implica ausencia de evidencia estadística suficiente para afirmar una correlación significativa entre ambas formas de evaluación.

En conjunto, los datos indican que la relación entre la evaluación humana y la evaluación por IA depende del contexto disciplinar y posiblemente de las características propias de las tesis (estructura metodológica, nivel de redacción, tipo de análisis realizado). Estos resultados refuerzan la necesidad de ajustar el uso de IA como herramienta de apoyo considerando las particularidades de cada campo profesional, más que asumir un desempeño uniforme del modelo en toda la institución.

En síntesis, los resultados muestran que la IA se aproxima mejor al criterio humano en carreras donde el producto evaluado es más cuantitativo y estructurado, mientras que su desempeño disminuye en disciplinas donde la evaluación depende del análisis interpretativo. Esto evidencia que la efectividad de la IA como herramienta evaluativa no es universal, sino dependiente del contexto disciplinar.

5. Discusión y conclusiones

Los hallazgos de este estudio permiten establecer que, en su configuración inicial, la inteligencia artificial (IA) mostró una tendencia a asignar calificaciones más altas que las emitidas por evaluadores humanos, mientras que, tras ajustes metodológicos y de calibración, se observó una leve inversión de la tendencia, con evaluaciones ligeramente más estrictas que el juicio humano. Esta variación confirma la sensibilidad de los sistemas generativos a los parámetros de diseño y a la formulación de prompts, lo que coincide con lo planteado por Adeyele y Ramnarain (2024a, 2024b), quienes subrayan que la efectividad de ChatGPT en contextos educativos depende en gran medida de su adaptación a marcos pedagógicos específicos.

En términos de concordancia, la mejora observada tras la calibración refuerza la hipótesis de que la IA puede alinearse de manera más precisa con criterios evaluativos humanos si se cuenta con ajustes iterativos, como ya se ha documentado en trabajos de Michel-Villarreal et al. (2023) y

Kazimova et al. (2025), quienes destacan la necesidad de procesos de entrenamiento contextualizados. Este alineamiento progresivo también está en sintonía con lo reportado por Romero-Rodríguez et al. (2023), quienes indican que la utilidad percibida de ChatGPT por parte de los estudiantes aumenta cuando la retroalimentación es coherente con las expectativas académicas y los criterios de evaluación institucionales.

La tendencia inicial a la benevolencia puede interpretarse como una consecuencia del enfoque generativo de modelos como ChatGPT, que prioriza la producción de contenido coherente y constructivo, reduciendo la probabilidad de emitir valoraciones negativas sin un contexto explícito de evaluación (Chiappe et al., 2025; del Mar Sánchez Vera, 2024). Esta característica ha sido señalada como una limitación en entornos donde la rigurosidad evaluativa es esencial, dado que la sobreestimación de la calidad puede afectar la validez de las decisiones académicas (Reyes et al., 2025).

Desde una perspectiva práctica, los resultados respaldan lo señalado por Farrelly y Baker (2023) y Rahiman y Kodikal (2024) sobre la potencialidad de la IA como herramienta de apoyo a la docencia, particularmente en contextos de sobrecarga evaluativa, como el caso de universidades con alta matrícula y limitados recursos humanos. La capacidad de la IA para reducir tiempos de revisión y estandarizar criterios es coherente con las proyecciones de Memon y Kwan (2025), quienes proponen modelos colaborativos entre docentes y sistemas de IA generativa para optimizar procesos académicos.

No obstante, los hallazgos también sugieren que la implementación de IA en la revisión de tesis no debe concebirse como un reemplazo del juicio humano, sino como un complemento que aporte rapidez y consistencia, siempre bajo supervisión académica. Este planteamiento es consistente con las advertencias de Baidoo-Anu y Owusu Ansah (2023) y Robles y Ek (2024) respecto a los riesgos de sobredependencia tecnológica, que incluyen la pérdida de matices disciplinares y la aceptación acrítica de resultados generados por sistemas automáticos.

En el plano teórico, el estudio aporta evidencia que dialoga con el marco Stimulus-Organism-Behavior-Consequence (Biswas et al., 2025), al mostrar que la experiencia de uso y los ajustes técnicos influyen directamente en la recomendación y aceptación de la IA en procesos evaluativos. Además, reafirma la visión de Adiwijaya et al. (2025) y Rahmadi (2024) sobre la importancia de integrar la transformación digital en la educación superior mediante marcos estratégicos que incluyan formación docente, protocolos de calidad y evaluación continua del impacto de las tecnologías emergentes.

Entre las implicancias metodológicas, se confirma que la calidad de las evaluaciones generadas por IA está fuertemente condicionada por la precisión de los prompts y la claridad de los criterios de la rúbrica, lo que coincide con los hallazgos de Jahani et al. (2024) sobre el papel de la ingeniería de prompts en la efectividad del uso de ChatGPT. Asimismo, se evidencia que la calibración basada en retroalimentación humana incrementa los niveles de concordancia, alineándose con lo planteado por Sposato (2025) respecto a la necesidad de modelos de liderazgo académico que faciliten la integración de tecnologías disruptivas con prácticas docentes consolidadas.

Entre las limitaciones del estudio se reconoce que el análisis se centró exclusivamente en la nota final de la tesis, sin considerar las evaluaciones parciales por capítulos, lo que podría ocultar discrepancias más finas en criterios específicos. Además, la muestra estuvo limitada a una única institución con presencia regional, lo que condiciona la generalización de los resultados a otros contextos.

A partir de estas limitaciones, se abren líneas de investigación futuras orientadas a: (i) aplicar el prompt definitivo a un número representativo y diverso de tesis para validar su eficacia; (ii) desarrollar un chatbot académico que permita interactuar con el contenido de las tesis y generar retroalimentación detallada; (iii) realizar análisis por capítulos para detectar patrones de discrepancia específicos; y (iv) comparar el rendimiento de diferentes sistemas de IA generativa en la revisión académica, considerando variables como área disciplinar, complejidad metodológica y calidad argumentativa.

Se analizaron 118 tesis de pregrado, seleccionadas como muestra representativa del universo de más de 300 egresados que genera anualmente la universidad. El volumen total de tesis supera la capacidad de revisión exhaustiva por parte del cuerpo académico, compuesto por un número acotado de docentes evaluadores. Por ello, trabajar con esta muestra permitió desarrollar un análisis estadísticamente manejable, manteniendo diversidad disciplinar y variabilidad en los resultados.

Aunque no se evaluaron todas las tesis, el tamaño muestral fue suficiente para identificar tendencias relevantes, comparar el desempeño entre evaluación humana e inteligencia artificial y obtener conclusiones generalizables al contexto institucional.

En conjunto, los resultados de este estudio contribuyen a la discusión internacional sobre el uso de IA generativa en educación superior, aportando evidencia empírica sobre su potencial como herramienta de apoyo en la revisión de tesis y ofreciendo orientaciones para su implementación responsable, calibrada y pedagógicamente pertinente.

El presente estudio tuvo como propósito comparar la revisión de tesis de pregrado realizada por juicio humano con aquella efectuada mediante inteligencia artificial, evaluando el grado de concordancia, las diferencias de tendencia y el potencial de la IA como herramienta de apoyo en procesos académicos. Los hallazgos evidenciaron que, en su configuración inicial, la IA tendió a asignar calificaciones más elevadas que las otorgadas por evaluadores humanos, mostrando una clara inclinación hacia la benevolencia. No obstante, tras la implementación de ajustes metodológicos y la calibración de los prompts en función de la rúbrica institucional, esta tendencia se revirtió ligeramente, resultando en evaluaciones marginalmente más estrictas que las humanas y en un aumento sustancial de la concordancia entre ambas modalidades.

Estos resultados permiten afirmar que la IA, correctamente configurada y supervisada, puede aproximarse de manera significativa a los criterios de evaluación humana, contribuyendo a la estandarización y eficiencia del proceso de revisión. El aporte principal de este trabajo radica en ofrecer evidencia empírica que respalda la viabilidad de integrar sistemas de IA generativa en la evaluación académica como recurso complementario, particularmente en instituciones con alta carga evaluativa y recursos humanos limitados. Desde una perspectiva aplicada, el estudio aporta lineamientos para la implementación de la IA en la revisión de tesis, subrayando la importancia de calibraciones iterativas y del alineamiento con marcos evaluativos previamente establecidos.

A pesar de su relevancia, el estudio presenta limitaciones que deben considerarse al interpretar sus conclusiones. El análisis se centró exclusivamente en la nota final de cada tesis, sin desagregar los resultados por capítulos o dimensiones evaluativas específicas, lo que impide identificar con precisión en qué aspectos concretos la IA y el juicio humano convergen o divergen. Asimismo, la investigación se llevó a cabo en una sola institución, lo que restringe la generalización de los hallazgos a otros contextos académicos y disciplinares.

En consecuencia, se propone como línea de investigación futura la aplicación del prompt definitivo a un conjunto más amplio y diverso de tesis, incorporando el análisis por capítulos y criterios de evaluación específicos para identificar patrones de concordancia y discrepancia en niveles más detallados. Asimismo, se plantea el desarrollo de un chatbot académico que permita interactuar con el contenido de las tesis y generar retroalimentación personalizada para estudiantes y docentes. Finalmente, se sugiere comparar el desempeño de distintas plataformas de IA generativa en tareas evaluativas, considerando variables como la naturaleza disciplinar de las tesis, la complejidad metodológica y la calidad argumentativa de los trabajos.

En suma, este estudio contribuye a la comprensión del potencial y las limitaciones de la IA en la evaluación académica, reforzando la idea de que su uso debe concebirse como un complemento estratégico al juicio humano, capaz de optimizar procesos sin comprometer la calidad y el rigor académico.

6. Financiación y apoyos

Estudio apoyado y financiado por Universidad de Aconcagua de Chile.

7. Semblanza de los/as autores/as

Dr. Juan Guillermo Muñoz Tapia es Doctor en Educación, Magíster en Educación Superior, Ingeniero en Electricidad y Profesor de Matemáticas. Actualmente es académico en la Universidad de Aconcagua, donde lidera proyectos de investigación sobre inteligencia artificial aplicada a la evaluación académica. Además, se desempeña como director de tesis doctorales en la Universidad Metropolitana de Ciencia y Tecnología (UMECIT, Panamá) y como docente en diversas universidades chilenas, impartiendo asignaturas de matemáticas, estadística y cálculo.

Ha participado como expositor y moderador en congresos internacionales en educación e innovación tecnológica, y cuenta con publicaciones en revistas indexadas en WoS, como Creative Education. Sus principales líneas de investigación son la innovación educativa, la educación matemática, la

ciencia de datos aplicada a la docencia y el uso de inteligencia artificial en procesos de enseñanza-aprendizaje.

Referencias

Adeyele, V. O., & Ramnarain, U. (2024b). Exploring the integration of ChatGPT in inquiry-based learning: Teacher Perspectives. *International Journal of Technology in Education*, *7(2),* 200–217. https://doi.org/10.46328/IJTE.638

Adiwijaya, A., Anggadwita, G., Gozali, A. A., Ramadhana, M. R., & Purbolaksono, M. D. (2025). Digital transformation in higher education: developing a new framework. *Globalisation, Societies and Education*. https://doi.org/10.1080/14767724.2025.2520952

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4337484

Biswas, M. I., Talukder, M. S., & Chen, Y. (2025). Applying the stimulus-organism-behavior-consequence framework to examine the relationship between intention, usage and recommendation of ChatGPT in higher education. *International Journal of Educational Management*, *39(2),* 450–468. https://doi.org/10.1108/IJEM-09-2023-0463

Chiappe, A., Miguel, C. S., & Delgado, F. M. S. (2025). Generative AI vs. Teachers: insights from a literature review. *Pixel-Bit, Revista de Medios y Educacion*, *72,* 119–137. https://doi.org/10.12795/PIXELBIT.107046

Del Mar Sánchez Vera, M. (2024). Artificial Intelligence as a teaching resource: uses and possibilities for teachers. *Educar*, *60(1),* 33–47. https://doi.org/10.5565/REV/EDUCAR.1810

Durmuş Sarıkahya, S., Özbay, Ö., Torpuş, K., Usta, G., & Çınar Özbay, S. (2025). The impact of ChatGPT on nursing education: A qualitative study based on the experiences of faculty members. *Nurse Education Today*, *152*. https://doi.org/10.1016/J.NEDT.2025.106755

Farrelly, T., & Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences*, *13(11)*. https://doi.org/10.3390/EDUCSCI13111109

Jahani, M., Baruah, B., & Ward, A. (2024). Exploring ChatGPT Utilization Among Master's Students in Higher Education. *2024 21st International Conference on Information Technology Based Higher Education and Training, ITHET 2024*. https://doi.org/10.1109/ITHET61869.2024.10837601

Kazimova, D., Tazhigulova, G., Shraimanova, G., Zatyneyko, A., & Sharzadin, A. (2025). Transforming university education with AI: A systematic review of technologies, applications, and implications. *International Journal of Engineering Pedagogy*, *15(1),* 4–24. https://doi.org/10.3991/IJEP.V15I1.50773

Mardiana, H. (2024). Perceived impact of lecturers' digital literacy skills in higher education institutions. *SAGE Open*, *14(3).* https://doi.org/10.1177/21582440241256937

Memon, T. D., & Kwan, P. (2025). A collaborative model for integrating teacher and GenAI into future education. *TechTrends*. https://doi.org/10.1007/S11528-025-01105-W

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13(9).* https://doi.org/10.3390/EDUCSCI13090856

Organisation for Economic Co-operation and Development. (2021). *Education at a glance 2021: OECD indicators.* OECD Publishing. https://www.oecd.org/en/publications/education-at-a-glance-2021_b35a14e5-en.html

Pycka, M., & Zastempowski, M. (2025). Machine learning and artificial intelligence techniques adopted for IT audit. *Management*, *1*, 65–87. https://doi.org/10.58691/MAN/200768

Rahiman, H. U., & Kodikal, R. (2024). Revolutionizing education: Artificial intelligence empowered learning in higher education. *Cogent Education*, *11(1).* https://doi.org/10.1080/2331186X.2023.2293431

Rahmadi, I. F. (2024). Research on digital transformation in higher education: Present concerns and future endeavours. *TechTrends*, *68(4),* 647–660. https://doi.org/10.1007/S11528-024-00971-0

Reyes, J., Batmaz, A. U., & Kersten-Oertel, M. (2025). Trusting AI: does uncertainty visualization affect decision-making? *Frontiers in Computer Science*, *7*. https://doi.org/10.3389/FCOMP.2025.1464348

Robles, L. E. R., & Ek, J. I. (2024). Use of generative artificial intelligence in educational environments: An initial student perspective of the risks and advantages. *IEEE Global Engineering Education Conference, EDUCON*. https://doi.org/10.1109/EDUCON60312.2024.10578776

Romero-Rodríguez, J. M., Ramírez-Montoya, M. S., Buenestado-Fernández, M., & Lara-Lara, F. (2023). Use of ChatGPT at university as a tool for complex thinking: Students' perceived usefulness. *Journal of New Approaches in Educational Research*, *12(2),* 323–339. https://doi.org/10.7821/NAER.2023.7.1458

Sposato, M. (2025). Artificial intelligence in educational leadership: a comprehensive taxonomy and future directions. *International Journal of Educational Technology in Higher Education*, *22(1).* https://doi.org/10.1186/S41239-025-00517-1

Xue, Y., Chen, H., Bai, G. R., Tairas, R., & Huang, Y. (2024). Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in CS1. *Proceedings - International Conference on Software Engineering*, 331–341. https://doi.org/10.1145/3639474.3640076

Zahari, N. A., Nasser, S. S. Q., Mustapa, M., Dahlan, A. R. A., & Ibrahim, J. (2018). A conceptual digital transformation design for international islamic university Malaysia to 'university of the future.' *Proceedings - International Conference on Information and Communication Technology for the Muslim World 2018, ICT4M 2018*, 94–99. https://doi.org/10.1109/ICT4M.2018.00026