



Respuestas educativas de IA: Calidad y confiabilidad en ChatGPT y Gemini

AI Educational Responses: Quality and Reliability in ChatGPT and Gemini

-  Gustavo Quintero Barreto, Universidad del Istmo (Udellstmo), Ciudad de Panamá (Panamá) (gquintero_d2032@udi.edu.pa) (<https://orcid.org/0000-0002-0157-305X>)
-  Aura L. López de Ramos, Centro de Investigación Educativa AIP (CIEDU AIP), Ciudad de Panamá (Panamá) (alopez@ciedupanama.org) (<https://orcid.org/0000-0002-8983-9704>)
-  Yuly Esteves González*, Universidad Pedagógica Experimental Libertador (UPEL), Caracas (Venezuela). Universidad Internacional de Ciencia y Tecnología (UNICyT), Ciudad de Panamá (Panamá) (yuly.esteves@upel.edu.ve) (<https://orcid.org/0000-0003-4153-6103>)
-  Nelly Meléndez, Universidad Monteávila (UMA), Caracas (Venezuela). Universidad Internacional de Ciencia y Tecnología (UNICyT), Ciudad de Panamá (Panamá) (nmelendez21@gmail.com) (<https://orcid.org/0000-0002-2780-2519>)

* Indicates the corresponding author

RESUMEN

La integración de modelos de lenguaje de gran escala en la educación superior ha generado un creciente interés por evaluar la calidad y confiabilidad de sus respuestas en tareas académicas. Sin embargo, persiste una carencia de estudios comparativos que combinen métricas algorítmicas con la percepción de expertos humanos y que consideren el diseño del prompt como variable clave del rendimiento. Este estudio analiza comparativamente el desempeño de ChatGPT 4.5 y Gemini 2.5 en tareas educativas e investigativas mediante un enfoque metodológico mixto explicativo-secuencial. En la fase cuantitativa, las respuestas generadas por cada modelo ante prompts estructurados con técnicas ASPECCT, CLEAR y Chain of Thought fueron evaluadas en siete parámetros de calidad (precisión terminológica, adecuación epistémica, contextualización disciplinar, tipo de razonamiento, nivel cognitivo, claridad y sintaxis). En la fase cualitativa, 78 evaluadores con perfiles académicos diferenciados valoraron las respuestas mediante rúbricas especializadas. Los resultados evidencian que Gemini 2.5 obtuvo el mejor desempeño promedio general (92.68/100), con ventaja significativa en contextualización disciplinar y niveles cognitivos, mientras que ChatGPT 4.5 mostró mayor estabilidad en análisis de datos cuantitativos. El análisis cualitativo reveló una actitud de confianza cautelosa entre los evaluadores, quienes reconocen a los modelos como codiseñadores textuales potentes, pero epistemológicamente frágiles. Se concluye que la eficacia de la IA generativa en contextos académicos depende críticamente del diseño del prompt, la supervisión experta y la adaptación disciplinar, posicionando la ingeniería de prompts como una competencia clave para el uso riguroso de estos sistemas.

ABSTRACT

The integration of large language models into higher education has prompted growing interest in assessing the quality and reliability of their responses in academic tasks. However, there remains a lack of comparative studies combining algorithmic metrics with expert human perception while considering prompt design as a key performance variable. This study comparatively analyzes the performance of ChatGPT 4.5 and Gemini 2.5 across educational and research tasks using an explanatory-sequential mixed methods design. In the quantitative phase, responses generated through prompts structured with ASPECCT, CLEAR, and Chain of Thought techniques were assessed across seven quality parameters. In the qualitative phase, 78 evaluators with differentiated academic profiles rated responses using specialized rubrics. Results show that Gemini 2.5 achieved the highest overall mean score (92.68/100), with significant advantages in disciplinary contextualization and cognitive depth, while ChatGPT 4.5 demonstrated greater stability in quantitative data analysis. Qualitative analysis revealed a stance of cautious trust among evaluators, who acknowledged the models

as powerful textual co-designers yet epistemologically fragile. The study concludes that generative AI effectiveness in academic settings critically depends on prompt quality, expert supervision, and disciplinary adaptation, positioning prompt engineering as an essential academic competency.

PALABRAS CLAVE | KEYWORDS

Inteligencia artificial, educación superior, evaluación, tecnología educativa, competencia digital, modelos de lenguaje. Artificial Intelligence, Higher Education, Assessment, Educational Technology, Digital Competence, Language Models.

1. Introducción y estado de la cuestión

La rápida expansión de la inteligencia artificial generativa, especialmente de los modelos de lenguaje de gran escala (Large Language Models, LLM), está transformando de manera significativa los procesos de enseñanza, aprendizaje e investigación en la educación superior (Wang et al., 2025). Herramientas como ChatGPT y Gemini se han integrado en prácticas académicas como la redacción de textos, la planificación didáctica y el análisis de información, generando nuevas formas de interacción humano-máquina. Sin embargo, este crecimiento ha sido más rápido que la consolidación de marcos de evaluación empírica rigurosos que permitan determinar la calidad, confiabilidad y pertinencia epistemológica de las respuestas generadas en contextos académicos reales. Esta brecha limita la comprensión de su valor educativo y dificulta su integración crítica y fundamentada en la educación superior.

Aunque su uso está ampliamente adoptado, aún es necesario realizar estudios empíricos rigurosos para evaluar la calidad, la confiabilidad y la pertinencia epistemológica de las respuestas que generan los LLM en tareas educativas e investigativas reales. La literatura reciente (Albadarin et al., 2024; Meyer et al., 2023) muestra un predominio de investigaciones exploratorias, estudios de caso aislados o evaluaciones centradas en métricas técnicas internas del modelo, sin incorporar sistemáticamente la percepción de expertos humanos ni analizar el diseño de prompts como variable clave del rendimiento. En esta misma línea, Singh y Strzelecki (2025) señalan que, pese a la rápida difusión de herramientas como ChatGPT en el ámbito académico, aún son escasos los estudios empíricos comparativos que evalúen de manera rigurosa la calidad epistemológica de las respuestas y el papel del diseño del prompt en el desempeño de los modelos.

La revisión sistemática realizada por López Trinidad et al. (2026) evidencia que, si bien la producción científica sobre inteligencia artificial en educación ha crecido de manera sostenida, la mayoría de los estudios se concentran en enfoques técnicos o instrumentales, con una problematización pedagógica limitada y una evaluación empírica escasa de herramientas de IA generativa en contextos educativos reales.

Este contexto revela una laguna importante en la investigación, al menos en tres aspectos. Primero, existe escasez de estudios comparativos sistemáticos entre modelos de IA generativa, como ChatGPT y Gemini, en escenarios educativos e investigativos específicos, evaluados con criterios académicos comparables. Segundo, se observa una carencia de métricas de calidad específicas para prompts educativos reales, elaborados por humanos y alineados con marcos pedagógicos y epistemológicos reconocidos, lo que compromete la validez externa de muchos estudios previos. Tercero, la literatura muestra poca diferenciación entre perfiles de usuarios, al tratar a los evaluadores como un grupo homogéneo y sin considerar cómo varían las percepciones y valoraciones según el rol académico o profesional (docentes, investigadores, consultores, profesionales no docentes, entre otros).

Este estudio busca analizar de manera comparativa cómo funcionan los modelos de IA generativa en tareas educativas e investigativas. Utiliza un enfoque metodológico mixto que combina evaluación cuantitativa y análisis cualitativo, e incluye las perspectivas de distintos tipos de usuarios. A través de rúbricas especializadas se evalúa no solo el rendimiento de los modelos, sino también la experiencia de los evaluadores al interactuar con las respuestas generadas. Esto permite una comprensión más profunda del uso de la IA generativa en contextos académicos.

Este estudio se destaca como uno de los primeros análisis empíricos en el ámbito latinoamericano que combina de manera sistemática la evaluación algorítmica, la percepción de expertos humanos y el análisis cualitativo del uso de LLM en la educación superior. Su aportación al campo tiene un doble enfoque: por un lado, ofrece evidencia científica sobre las fortalezas y limitaciones de diferentes modelos de IA generativa, considerando el tipo de tarea y el perfil del usuario; por otro, brinda orientaciones prácticas para universidades, docentes, investigadores y diseñadores instruccionales interesados en el uso responsable, crítico y pedagógicamente fundamentado de estas tecnologías. En este sentido, los hallazgos contribuyen al debate académico sobre el papel de la IA generativa, que no busca reemplazar el juicio experto, sino funcionar como un co-diseñador textual cuya eficacia depende de la calidad del prompt, la supervisión humana y el contexto en que se aplica.

1.1. Inteligencia artificial generativa

El despliegue de la inteligencia artificial generativa (IAG), y en particular de los modelos de lenguaje de gran escala, está reconfigurando procesos centrales de enseñanza, aprendizaje e investigación en educación

superior. Herramientas como ChatGPT y Gemini han ampliado la disponibilidad de apoyo textual y cognitivo para tareas académicas (por ejemplo, redacción, síntesis, generación de borradores y retroalimentación), abriendo un nuevo espacio de interacción humano-máquina con implicaciones pedagógicas y epistemológicas.

La evidencia empírica sugiere que la retroalimentación generada por IA puede ser percibida por el estudiantado como comparable en utilidad a la retroalimentación docente, aunque con ventajas diferenciadas: la IA aporta rapidez y consistencia estructural, mientras que la mediación humana ofrece matiz interpretativo, juicio pedagógico y apoyo socioemocional (Ali et al., 2025; Escalante et al., 2023). En escritura académica, los estudiantes reportan beneficios en organización textual, gramática y vocabulario, pero también limitaciones en pensamiento crítico, creatividad argumentativa y desempeño oral (Kim et al., 2024; Liu et al., 2024).

En paralelo, una línea emergente integra LLM en sistemas de analítica de aprendizaje para procesar interacciones textuales y rastros digitales de actividad, con el fin de identificar patrones, inferir niveles de comprensión y anticipar riesgos académicos (Mohammadi et al., 2025; Yan et al., 2023). Los estudios de revisión apuntan a que los modelos se emplean para analizar producciones escritas, interacciones en plataformas digitales y otros rastros de actividad del estudiantado, con el objetivo de inferir niveles de comprensión, dificultades recurrentes y patrones de razonamiento, integrando así la lógica de la analítica de aprendizaje con enfoques de evaluación formativa (Ouyang et al., 2023). Desde esta perspectiva, los modelos funcionan como software capaz de transformar grandes volúmenes de datos textuales en indicadores pedagógicamente relevantes, lo que amplía el repertorio de métricas tradicionalmente disponibles para el seguimiento del aprendizaje.

A pesar de los beneficios evidenciados en el uso de modelos generativos en analítica de aprendizaje y en aplicaciones educativas, se plantean desafíos significativos en materia de validez, transparencia y ética del uso de los datos. Estos sistemas suelen basarse en fuentes de datos multimodales (registros de actividad, videos, textos) que frecuentemente están incompletos o sesgados, lo que compromete la validez de las inferencias que puedan realizarse. Resulta necesario desarrollar enfoques sofisticados, interpretables y escalables que permitan superar la opacidad algorítmica que caracteriza a los modelos de IA generativa. En tal sentido, Berghea et al. (2025) indican que la opacidad algorítmica no es una limitación meramente técnica; representa también un obstáculo epistemológico que impide el escrutinio pedagógico de las recomendaciones automatizadas.

1.2. Ingeniería de prompts como mediación pedagógica

Frente a la opacidad y variabilidad de los LLM, la ingeniería de prompts se posiciona como una competencia académica emergente y como una forma de mediación pedagógica: el prompt funciona como interfaz que traduce intencionalidad didáctica y condiciones de tarea en instrucciones procesables por el modelo. Adicionalmente, las estrategias de prompting adaptativo vinculadas a las necesidades del aprendiz (por ejemplo, aproximaciones inspiradas en la zona de desarrollo próximo) han sido asociadas con mejores resultados y reducción de brechas; sin embargo, la evidencia disponible todavía demanda mayor consolidación metodológica y transparencia en los reportes para sostener generalizaciones robustas (Shen et al., 2025).

La evidencia empírica revela que los prompts estructurados, específicos y contextualizados incrementan la exactitud, relevancia y exhaustividad de las respuestas en diversas áreas académicas (Cain, 2023; Lee y Palmer, 2025). En educación superior se ha demostrado que las técnicas como la cadena de pensamiento y los prompts enriquecidos con contexto específico superan a las formulaciones simples en precisión y exhaustividad.

Los prompts que integran el rol del estudiante, su contexto particular y sus objetivos personales favorecen experiencias de aprendizaje personalizadas y participativas, y transforman a docentes y estudiantes en co-creadores del proceso formativo. Asimismo, los prompts adaptativos diseñados según la zona de desarrollo próximo del estudiante han logrado reducir brechas en el rendimiento académico (Liu et al., 2024).

1.3. Estudios comparativos entre modelos de IA generativa

Aunque la producción científica sobre IAG en educación crece, las revisiones sistemáticas identifican un predominio de estudios exploratorios y de evaluaciones centradas en métricas internas del modelo, con menor énfasis en criterios académicos explícitos y en la percepción de expertos humanos. En estudios comparativos se ha observado que distintos modelos pueden mostrar fortalezas diferenciadas según tareas y dominios; por ejemplo, en contextos de educación médica se reportan diferencias de desempeño en pruebas estructuradas entre ChatGPT y Gemini, con ventaja humana en razonamiento complejo (Al-Thani et al., 2025).

1.4. Marco conceptual para la evaluación de respuestas generadas por IA

Para fines de evaluación en contextos académicos, la calidad de la respuesta puede definirse como el grado en que el texto generado cumple estándares de precisión conceptual, coherencia metodológica y profundidad cognitiva; y la confiabilidad, como la consistencia del desempeño en condiciones comparables, valorada mediante juicio experto humano. Con base en ello, la evaluación se organiza en tres dimensiones: (a) desempeño sintáctico-estructural (claridad y precisión terminológica), (b) rigor metodológico y pertinencia disciplinar (adecuación epistémica y contextualización), y (c) profundidad cognitiva (tipo de razonamiento y nivel cognitivo según Bloom).

1.4.1. Desempeño sintáctico-estructural (claridad, sintaxis y precisión terminológica)

Esta dimensión analiza la estructura de superficie textual. Evalúa la solvencia del modelo para actuar como un co-diseñador capaz de orquestar información con rigor gramatical y de desplegar con exactitud el registro léxico de la disciplina. Investigaciones recientes demuestran que los LLM muestran buen desempeño para emular la claridad sintáctica humana y recuperar información fundamental (Guan y Han, 2025). Sin embargo, esta fluidez técnica a menudo oculta una falta de profundidad analítica. Evaluar la precisión terminológica permite discriminar si el modelo domina realmente el corpus de la disciplina o si simplemente ensambla una estructura léxica y gramaticalmente correcta, pero semánticamente vacía.

1.4.2. Rigor metodológico y pertinencia disciplinar (adecuación epistémica y contextualización)

Este bloque constituye el núcleo crítico del instrumento. Mide si la respuesta algorítmica logra anclarse en un paradigma científico explícito y dialogar con las contingencias del entorno local. Niño-Carrasco et al. (2025) documentan cómo los estudiantes tienden a percibir a la IA como un especialista, delegando en la máquina la autoridad epistémica. Este fenómeno desplaza el aprendizaje desde la construcción colaborativa del conocimiento hacia el consumo pasivo de respuestas prefabricadas (Avello-Martínez et al., 2024). Por ello, medir la adecuación epistémica visibiliza la habitual fragilidad metodológica de los LLM, evidente cuando el modelo produce texto impecable, pero fracasa al definir tipos de estudio o procedimientos de validación.

1.4.3. Profundidad cognitiva (tipo de razonamiento y nivel cognitivo según la Taxonomía de Bloom)

Esta última dimensión analiza el límite intelectual que logra simular la plataforma. Discrimina sistemáticamente si la IA se estanca en operaciones mecánicas de base o si exhibe competencia para articular tramas argumentativas lógicas. La literatura confirma que los LLM operan de manera eficiente en los escalones inferiores de la Taxonomía de Bloom, como recordar o resumir, pero requieren un diseño de prompts altamente especializado para ejecutar procesos de evaluación y creación sostenida (Bas Graells et al., 2024). Evaluar este criterio revela el techo cognitivo real del modelo frente a tareas de diseño instruccional complejo.

2. Metodología

2.1. Diseño del estudio

Este estudio adopta un enfoque mixto explicativo-secuencial, en el que la fase cuantitativa se realiza primero y guía el análisis cualitativo. Este diseño permite, en una primera fase, detectar patrones, diferencias y tendencias en la evaluación del rendimiento de los modelos de inteligencia artificial generativa. Luego, en una segunda fase, se profundiza en la interpretación de estos resultados a través de las experiencias y valoraciones discursivas de los participantes. La selección de este diseño está respaldada por la naturaleza del fenómeno analizado: la calidad de las respuestas de los modelos de lenguaje no puede evaluarse únicamente con métricas cuantitativas, sino que también requiere la percepción de expertos humanos, especialmente en aspectos epistemológicos, metodológicos y pedagógicos. El enfoque mixto permite una triangulación sólida entre la evaluación algorítmica, la valoración humana y el análisis cualitativo, lo que refuerza la validez interna y la interpretabilidad de los resultados.

2.2. Participantes

La muestra estuvo compuesta por 78 evaluadores, seleccionados por conveniencia, con experiencia

y vinculación a contextos académicos y profesionales en educación, investigación y tecnologías digitales.

2.3. Perfiles de usuario

Los participantes se clasificaron según su rol principal en los siguientes perfiles: 1) docentes; 2) investigadores(as); 3) consultores(as); 4) profesionales no docentes; 5) otros perfiles (categoría residual, no considerada en los análisis comparativos). La distribución porcentual fue la siguiente: 52.6 % docentes, 34.6 % investigadores, 3.8 % consultores, 6.4 % profesionales no docentes y 2.6 % en la categoría otros.

2.4. Procedimientos de reclutamiento

El reclutamiento se llevó a cabo mediante invitaciones directas y difusión institucional, dirigidas a redes académicas y profesionales relacionadas con el proyecto. La participación fue voluntaria, sin recompensas económicas, y los evaluadores aceptaron participar bajo condiciones de anonimato y con fines académicos.

2.5. Instrumentos

2.5.1. Rúbrica de evaluación de la calidad de la respuesta

Se diseñó una rúbrica ad hoc para evaluar la calidad de las respuestas generadas por los modelos de lenguaje, utilizando una escala Likert de 5 puntos (1 = desempeño muy bajo, 5 = desempeño muy alto). La rúbrica contempló las siguientes dimensiones: 1) precisión terminológica; 2) adecuación epistémica; 3) contextualización según la disciplina; 4) tipo de razonamiento; 5) nivel cognitivo (Taxonomía de Bloom); 6) claridad y sintaxis. Estas dimensiones fueron seleccionadas por su relevancia en la evaluación de productos académicos y educativos generados con apoyo de IA.

2.6. Prompts utilizados

Los prompts evaluados fueron diseñados por el equipo investigador y se organizaron en categorías vinculadas a tareas académicas reales: 1) prompts de carácter técnico; 2) prompts de carácter conceptual; 3) prompts orientados a educación (planificación didáctica, formación docente); 4) prompts orientados a investigación (formulación de objetivos, metodología). Cada prompt se aplicó de manera idéntica a los modelos evaluados, garantizando condiciones comparables para la generación de respuestas.

2.7. Sistemas evaluados

Los sistemas de inteligencia artificial generativa analizados en esta fase del estudio fueron ChatGPT y Gemini. Ambos modelos fueron evaluados exclusivamente a partir de sus respuestas textuales, sin considerar aspectos técnicos internos.

2.8. Procedimiento de recolección de datos

El procedimiento se desarrolló en tres fases secuenciales: a) fase 1: generación de respuestas. Los prompts definidos fueron procesados por ambos modelos de lenguaje, lo que generó respuestas independientes para cada sistema en los ámbitos de la educación y la investigación; b) fase 2: evaluación por usuarios. Las respuestas generadas se presentaron a los participantes. Cada evaluador calificó las respuestas utilizando la rúbrica establecida y, además, registró comentarios abiertos sobre su experiencia de evaluación; y c) fase 3: análisis comparado. Los datos cuantitativos y cualitativos recolectados fueron sistematizados para su análisis comparativo entre modelos, áreas temáticas y perfiles de usuario.

2.9. Selección de modelos y técnicas de prompting

El estudio se centró en dos modelos de lenguaje extenso (LLM) seleccionados con base en evidencia científica, considerando dos criterios fundamentales: a) el balance entre robustez técnica y efectividad práctica en contextos académicos, y b) la calidad del producto tecnológico y la experiencia de usuario, conforme a la Norma Internacional ISO 25010 y la Google Cloud for Educators Guide (Google LLC, 2025). Los modelos seleccionados fueron ChatGPT 4.5 de OpenAI y Gemini 2.5 Flash de Google LLC. La Tabla 1 presenta la comparación técnica de estos modelos.

Tabla 1: Comparación técnica de los LLM seleccionados.

Característica	ChatGPT 4.5 (OpenAI)	Gemini 2.5 Flash (Google)
Multimodalidad	Sí (texto, audio, visión)	Sí (nativo)
Acceso web en tiempo real	Limitado (modo browsing)	Sí (versiones avanzadas)
Desempeño en prompts (ASPECCT, CLEAR, CoT)	Muy alto. Capacidad analítica y razonamiento avanzado	Alto. Razonamiento fuerte con capacidades multimodales
Disponibilidad global	Sí	Sí (excepto algunas regiones)

Fuente: elaboración propia.

Paralelamente, se seleccionaron tres técnicas de prompting fundamentadas en la literatura: ASPECCT y CLEAR, clasificadas como marcos heurísticos simples que funcionan como guías para construir prompts efectivos enfocados en claridad y especificidad (Lo, 2023); y Chain of Thought (CoT), una técnica heurística estratégica de mayor complejidad que induce razonamiento paso a paso, mejorando significativamente la precisión en tareas lógicas y de razonamiento complejo (Wei et al., 2022). La combinación CLEAR + CoT resultó particularmente efectiva al maximizar la precisión mediante la estructura lógica del framework, complementada con el razonamiento secuencial.

3. Análisis y resultados

Los resultados se organizan en tres bloques. Primero se describe la evaluación algorítmica cuantitativa del desempeño de los modelos de lenguaje seleccionados (3.1). A continuación se presentan los resultados del análisis estadístico comparativo (3.2). Finalmente se reportan los hallazgos de la evaluación humana, que comprende tanto la valoración mediante rúbricas como el análisis cualitativo de los comentarios abiertos de los participantes (3.3).

3.1. Evaluación algorítmica del desempeño de los modelos

Los prompts diseñados por el equipo investigador fueron procesados por cada uno de los modelos seleccionados (ChatGPT 4.5 y Gemini 2.5 Flash) en áreas específicas de educación (planificación didáctica, educación continua y orientación vocacional) e investigación (formulación de objetivos, análisis de datos cuantitativos y análisis de datos cualitativos). Las respuestas generadas por cada LLM fueron evaluadas mediante cuatro plataformas algorítmicas de referencia: Stanford HELM, Scale AI, Hugging Face AI y Arthur AI. Para ello se configuraron instrucciones específicas que permitían a cada modelo acceder a su algoritmo evaluador interno y asignar puntajes porcentuales en siete parámetros preestablecidos: precisión terminológica, adecuación epistémica, contextualización disciplinar, tipo de razonamiento, nivel cognitivo según la Taxonomía de Bloom, claridad y sintaxis.

Cada parámetro fue evaluado comparando los puntajes reales asignados por el algoritmo con valores estándar de referencia definidos previamente. Adicionalmente, se solicitó a cada evaluador algorítmico una justificación textual de su puntuación, lo que permitió verificar la coherencia entre la métrica cuantitativa y su fundamentación argumentativa. A modo ilustrativo, en el caso de Gemini 2.5 para planificación didáctica, los puntajes alcanzaron valores cercanos al estándar en todos los parámetros: precisión terminológica 14.25/15, adecuación epistémica 19.2/20, contextualización disciplinar 19.6/20, tipo de razonamiento 14.1/15, nivel cognitivo 19.4/20, claridad 4.9/5 y sintaxis 4.95/5.

Los resultados evidenciaron un patrón consistente al re-escalar los puntajes a una escala porcentual (0-100). Gemini 2.5 alcanzó el mejor desempeño promedio general con 92.68 puntos, especialmente destacado en tareas de investigación como generación de objetivos (96.65) y planificación didáctica (96.40), así como en precisión terminológica, adecuación epistémica y contextualización disciplinar en tareas complejas. ChatGPT 4.5 mantuvo una posición sólida con 92.17 puntos, mostrando fortalezas en análisis de datos cuantitativos (92.65), educación continua (91.35) y una estabilidad consistente entre áreas (rango 91-93).

La validación de coherencia entre justificaciones textuales y métricas cuantitativas confirmó diferencias relevantes entre modelos. Gemini 2.5 presentó justificaciones coherentes en todos los parámetros, superando consistentemente el umbral de 90 en la escala. GPT 4.5 exhibió coherencia en seis de los siete parámetros, con una leve caída en el componente operacional y didáctico (88 puntos).

Respecto a las técnicas de prompting, la combinación CLEAR + Chain of Thought resultó particularmente

efectiva al maximizar la precisión mediante la estructura lógica del framework, complementada con el razonamiento secuencial paso a paso. ASPECCT y CLEAR, en tanto marcos heurísticos simples, demostraron ser más adecuados para la fase de diseño y estructuración del prompt, mientras que CoT potenció el desempeño del modelo durante la generación de respuestas, especialmente en tareas que requerían razonamiento complejo. Gemini 2.5 destacó en combinaciones que requerían acceso a información actualizada, mientras que ChatGPT 4.5 respondió mejor a combinaciones que potencian el razonamiento (CoT) y los ejemplos contextuales.

3.2. Análisis estadístico comparativo

El análisis estadístico mediante ANOVA de un factor y pruebas t para muestras independientes permitió determinar la significancia de las diferencias observadas entre modelos. La Tabla 2 sintetiza los resultados por campo y subcampo.

Campo / Subcampo	Medias por modelo	Resultados ANOVA	Significancia	Ranking
Planificación didáctica	Gemini 96.40; GPT 91.75	—	—	Gemini > GPT
Educación continua	Gemini 90.30; GPT 91.35	—	No significativas	GPT ≈ Gemini
Orientación vocacional	Gemini 89.50; GPT 91.30	F=0.66; p=0.53	No significativas	Modelos similares
Objetivos de investigación	Gemini 96.65; GPT 93.00	F=15.847; p=0.004	Significativas	Gemini > GPT
Datos cuantitativos	Gemini 89.75; GPT 92.65	F=23.33; p<0.001	Altamente significativas	GPT > Gemini
Datos cualitativos	Gemini 93.45; GPT 93.00	—	—	Gemini > GPT

Fuente: elaboración propia. Pruebas ANOVA y t aplicadas según subcampo.

Las pruebas ANOVA confirmaron diferencias significativas entre modelos en investigación cuantitativa ($F=23.33$, $p<0.001$) y en generación de objetivos ($F=15.847$, $p=0.004$). En orientación vocacional, por el contrario, los dos modelos presentaron rendimientos estadísticamente equivalentes ($F=0.66$, $p=0.53$). Estos resultados indican que las diferencias entre modelos se manifiestan con mayor intensidad en tareas de alta complejidad metodológica, mientras que tienden a atenuarse en tareas operativas o de menor demanda epistemológica.

Desde una perspectiva global, Gemini 2.5 alcanzó los desempeños más altos en la mayoría de los subcampos, especialmente en investigación, donde obtuvo ventajas estadísticamente significativas. ChatGPT 4.5 mantuvo una posición intermedia con alta estabilidad y con una fortaleza particular en análisis de datos cuantitativos, único subcampo donde superó a Gemini.

3.3. Resultados de la evaluación humana

3.3.1. Evaluación mediante rúbricas

La evaluación humana se realizó mediante rúbricas con escala Likert de 1 a 5, aplicadas a través de encuestas estructuradas dirigidas a expertos académicos y estudiantes universitarios. Cada participante recibió un prompt con las respuestas generadas por ChatGPT y Gemini mediante la técnica Chain of Thought, las calificó utilizando la rúbrica correspondiente y registró comentarios abiertos sobre su experiencia de evaluación. La elección entre temáticas de educación o investigación fue libre.

	Educación ChatGPT	Educación Gemini	Investigación ChatGPT	Investigación Gemini
Promedio general	3.88	4.08	3.63	3.89
Consultor(a)	3.75	4.08	3.17	3.67
Docente	3.87	4.07	3.93	4.17
Investigador(a)	4.10	4.17	3.53	3.72
Profesional no docente	2.83	3.83	4.04	4.38

Fuente: elaboración propia. Escala Likert 1-5. n=78 participantes.

Los resultados (Tabla 3) muestran que Gemini obtuvo puntuaciones superiores a ChatGPT en todos los perfiles y ambas áreas evaluadas. La diferencia fue especialmente consistente entre docentes e investigadores,

los perfiles con mayor experiencia académica. El área de investigación registró la ventaja más amplia a favor de Gemini (3.89 frente a 3.63), lo que sugiere que la preferencia se intensifica en tareas que demandan mayor rigor metodológico. En ningún criterio individual ChatGPT fue favorecido por encima de Gemini, y para un mismo modelo las respuestas en el área de investigación fueron evaluadas sistemáticamente por debajo de las de educación. Un hallazgo notable fue el comportamiento diferenciado de los profesionales no docentes, quienes otorgaron a ChatGPT la puntuación más baja en educación (2.83) pero la más alta en investigación (4.04), lo que podría reflejar expectativas disciplinares distintas respecto a la calidad del contenido generado.

4. Discusión y conclusiones

Este estudio evaluó el rendimiento de dos modelos de inteligencia artificial generativa en tareas educativas e investigativas, combinando evaluaciones cuantitativas algorítmicas con análisis cualitativos desde distintos perfiles de usuario. Los resultados revelan un patrón consistente: Gemini 2.5 fue preferido de forma sistemática sobre ChatGPT 4.5 en la mayoría de las dimensiones analizadas, con un desempeño promedio general superior (92.68 frente a 92.17 en escala 0-100), especialmente marcado en contextualización disciplinar, claridad y sintaxis, tipo de razonamiento y niveles cognitivos según la Taxonomía de Bloom. Estas diferencias se intensificaron en tareas vinculadas a investigación académica, donde los evaluadores exigen mayor rigor metodológico y profundidad epistemológica. Estos hallazgos coinciden con lo señalado por García-Peñalvo (2024), respecto a que la efectividad de los modelos generativos varía significativamente según el dominio disciplinar y el tipo de tarea requerida.

No obstante, ChatGPT 4.5 mantuvo un rendimiento estable y confiable, con fortalezas específicas en análisis de datos cuantitativos (92.65) y educación continua (91.35), lo que sugiere una complementariedad funcional entre ambos modelos más que una competencia directa. Un hallazgo metodológicamente relevante fue la eficacia de la combinación CLEAR + Chain of Thought como estrategia de prompting, que maximizó la precisión particularmente en Gemini 2.5 y GPT 4.5. Este resultado aporta evidencia empírica a la propuesta de Lo (2023) sobre la complementariedad entre marcos estructurales simples y técnicas estratégicas de razonamiento secuencial, y posiciona la ingeniería de prompts como competencia crítica en entornos académicos.

El análisis cualitativo profundizó en esta interpretación al evidenciar que los participantes consideran a los modelos de lenguaje como codiseñadores potentes de textos, capaces de organizar ideas, generar borradores iniciales y reducir tiempos en tareas académicas. Sin embargo, esta fortaleza viene acompañada de una fragilidad epistemológica y metodológica constante, reflejada en la insuficiente definición de objetivos, la escasa explicitación de paradigmas y la necesidad de validación conceptual continua. En consonancia con las observaciones de Zawacki-Richter et al. (2019) sobre la necesidad de evaluar los sistemas de IA no solo en términos de capacidad generativa, sino también de adecuación pedagógica, los evaluadores manifestaron una actitud de confianza cautelosa: reconocen la utilidad de la IA generativa, pero rechazan su uso acrítico como sustituto del juicio experto humano. Esta postura establece lo que puede denominarse una relación de “confianza vigilante” con los LLM.

La contribución principal de este estudio reside en proporcionar evidencia empírica mediante un enfoque metodológico mixto sobre cómo distintos modelos de IA generativa actúan en contextos académicos reales y son percibidos por usuarios con perfiles variados. A diferencia de investigaciones que se centran exclusivamente en métricas técnicas o evaluaciones algorítmicas (Holmes et al., 2019), este trabajo incorpora la percepción de expertos humanos como elemento clave del análisis, facilitando una comprensión más contextual del uso de la IA generativa en educación superior. Adicionalmente, se demuestra que las instrucciones humanas dirigidas a modelos de lenguaje pueden evaluarse sistemáticamente, tanto cualitativa como cuantitativamente, mediante evaluadores algorítmicos de referencia, desplazando el foco del modelo en sí hacia la interacción humano-IA y consolidando la ingeniería de prompts como método válido de evaluación comparativa.

Desde una perspectiva práctica, los resultados sugieren una estrategia de implementación diferenciada: Gemini 2.5 resulta más adecuado para tareas de alta complejidad metodológica, como la formulación de objetivos, el diseño metodológico y la planificación didáctica, mientras que ChatGPT 4.5 es idóneo para el análisis de datos cuantitativos y tareas operativas. Estos hallazgos enriquecen el debate sobre el uso responsable, ético y pedagógicamente fundamentado de la IA generativa, especialmente en el contexto latinoamericano,

donde los estudios empíricos en esta área son aún escasos (Jara y Ochoa, 2020). Los resultados proporcionan directrices prácticas para universidades, docentes, investigadores y diseñadores instruccionales, evidenciando que la adopción efectiva de estas herramientas no depende exclusivamente del modelo seleccionado, sino también del diseño cuidadoso de los prompts, la supervisión experta y la adaptación al contexto disciplinar. Como productos derivados de la investigación se generaron un manual de buenas prácticas para el diseño de prompts y una guía rápida orientada a docentes y estudiantes universitarios.

Entre las limitaciones del estudio se identifican: a) la complejidad de la identificación y configuración de los evaluadores algorítmicos internos, cuya resolución requirió una extensa fase de pruebas; y b) el volumen de prompts necesarios para asegurar la consistencia y la precisión, que representó un desafío significativo en términos de tiempo y rigor.

Como líneas futuras de investigación, se plantea replicar el diseño con versiones actualizadas de los modelos para evaluar su evolución, ampliar la muestra de evaluadores humanos e incorporar disciplinas adicionales para incrementar la validez externa. Los hallazgos establecen bases sólidas para futuras investigaciones, programas de formación y políticas institucionales orientadas a incorporar la IA generativa como recurso de apoyo para la enseñanza, manteniendo el rigor epistemológico y la responsabilidad científica.

Apoyos

Los autores expresan su agradecimiento a la Universidad del Istmo (Udellstmo) por el financiamiento recibido para llevar a cabo el proyecto “Análisis de buenas prácticas en diseño de prompts y su impacto en el rendimiento de modelos LLM”, correspondiente al PdT N.º PRO-CII-D-03-25. También reconocen el apoyo institucional y académico del Centro de Investigación Educativa AIP (CIEDU AIP), así como la colaboración de la Universidad Monteávila (UMA), la Universidad Pedagógica Experimental Libertador (UPEL) y la Universidad Internacional de Ciencia y Tecnología (UNICyT), cuya participación fue fundamental para el desarrollo, el análisis y la difusión de los resultados de esta investigación.

Declaración de ética y conflicto de interés

Este estudio se llevó a cabo conforme a los principios éticos de la investigación científica. La participación de los evaluadores fue voluntaria, anónima y basada en consentimiento informado, lo que aseguró la confidencialidad de los datos y su utilización exclusiva con fines académicos y de investigación. No se realizó ninguna intervención ni se recopilaron datos sensibles de los participantes. Los autores declaran no tener conflictos de interés financieros, profesionales ni personales que pudieran haber afectado el diseño, la recolección, el análisis, la interpretación de los resultados o la redacción del presente manuscrito.

Declaración de transparencia

Los datos cuantitativos derivados de la evaluación algorítmica y las puntuaciones de las rúbricas aplicadas por los evaluadores humanos se encuentran disponibles para consulta bajo solicitud razonada dirigida al autor de correspondencia. Los comentarios abiertos de los participantes se conservan en su formato original, anonimizados conforme al consentimiento informado otorgado. Los prompts utilizados en el estudio, las rúbricas de evaluación y las matrices de codificación cualitativa están disponibles como material complementario. Los instrumentos de recolección fueron administrados mediante Google Forms y los registros originales se conservan en los archivos del proyecto PRO-CII-D-03-25 de la Universidad del Istmo. El estudio no fue pre-registrado en plataformas de registro previo de investigaciones.

Contribución de autores

Idea, G.Q., A.L.; Revisión de literatura (estado del arte), N.M.; Metodología, A.L.; Análisis de datos, G.Q.; Resultados, G.Q.; Discusión y conclusiones, G.Q., Y.E.; Redacción (borrador original), Y.E.; Revisiones finales, Y.E.; Diseño del proyecto y patrocinios, G.Q.

Referencias

Al-Thani, S. N., Anjum, S., Bhutta, Z. A., Bashir, S., Majeed, M. A., Khan, A. S. y Bashir, K. (2025). Comparative performance of ChatGPT, Gemini, and final-year emergency medicine clerkship students in answering multiple-choice questions: implications for the use of AI in medical education. *International Journal of Emergency Medicine*, 18(1), 146. <https://doi.org/10.1186/s12245-025-00949-6>

- Albadarin, Y., Saqr, M., Pope, N. y Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, 3(1), 60. <https://doi.org/10.1007/s44217-024-00138-2>
- Ali, M., Harbieh, I. y Haider, K. H. (2025). Bytes versus brains: A comparative study of AI-generated feedback and human tutor feedback in medical education. *Medical Teacher*, 48(1), 131–141. <https://doi.org/10.1080/0142159X.2025.2519639>
- Avello-Martínez, R., Gajderowicz, T. y Gómez-Rodríguez, V. G. (2024). ¿ChatGPT es útil para que los estudiantes de posgrado adquieran conocimientos sobre narración digital y reduzcan su carga cognitiva? Un experimento. *Revista de Educación a Distancia (RED)*, 24(78), 8. <https://doi.org/10.6018/red.604621>
- Bas Graells, G., Tinoco Devia, R., Salinas Leyva, L. C. y Sevilla Molina, J. (2024). Systematic review of taxonomies of risks associated with Artificial Intelligence. *Analecta Política*, 14(26), 01–25. <https://doi.org/10.18566/apolit.v14n26.a08>
- Berghea, F., Berghea, E. C., Daia, C. O., Ciuc, D. y Dinca, G. V. (2025). In the search for the perfect prompt in medical AI queries. *Frontiers in Artificial Intelligence*, 8, 1689178. <https://doi.org/10.3389/frai.2025.1689178>
- Cain, W. (2023). Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. *TechTrends*, 68(1), 47–57. <https://doi.org/10.1007/s11528-023-00896-0>
- Escalante, J., Pack, A. y Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57. <https://doi.org/10.1186/s41239-023-00425-2>
- García-Peñalvo, F. J. (2024). Inteligencia artificial generativa y educación: Un análisis desde múltiples perspectivas. *Education in the Knowledge Society (EKS)*, 25, e31942. <https://doi.org/10.14201/eks.31942>
- Google LLC. (2025). Google Cloud for faculty. <https://cloud.google.com/edu/faculty>
- Guan, Q. y Han, Y. (2025). From AI to authorship: Exploring the use of LLM detection tools for calling on “originality” of students in academic environments. *Innovations in Education and Teaching International*, 62(5), 1514–1528. <https://doi.org/10.1080/14703297.2025.2511062>
- Holmes, W., Bialik, M. y Fadel, C. (2019). *Inteligencia artificial en la educación: Promesas e implicaciones para la enseñanza y el aprendizaje*. Centro para el Rediseño Curricular. <https://discovery.ucl.ac.uk/id/eprint/10139722>
- ISO. (2011). ISO/IEC 25010:2011: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. <https://www.iso.org/standard/35733.html>
- Jara, I. y Ochoa, J. M. (2020). *Usos y efectos de la inteligencia artificial en educación*. Banco Interamericano de Desarrollo. <https://doi.org/10.18235/0002380>
- Kim, J., Yu, S., Detrick, R. y Li, N. (2024). Exploring students’ perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*, 30(1), 1265–1300. <https://doi.org/10.1007/s10639-024-12878-7>
- Lee, D. y Palmer, E. (2025). Prompt engineering in higher education: a systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22(1), 7. <https://doi.org/10.1186/s41239-025-00503-7>
- Liu, Y., Park, J. y McMinn, S. (2024). Using generative artificial intelligence/ChatGPT for academic communication: Students’ perspectives. *International Journal of Applied Linguistics*, 34(4), 1437–1461. <https://doi.org/10.1111/ijal.12574>
- Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4), 102720. <https://doi.org/10.1016/j.acalib.2023.102720>
- López Trinidad, N., Lara Inocencio, G., Silverio Hurtado, M. y Pascual Peralta, J. (2026). La inteligencia artificial en la educación: vacíos, tendencias y desafíos éticos. Una revisión de literatura (2019–2024). *Comunicar*, 34(84), 42–56. <https://doi.org/10.5281/zenodo.18113730>
- Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O’Connor, K., Li, R., Peng, P.-C., Bright, T. J., et al. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1), 20. <https://doi.org/10.1186/s13040-023-00339-9>
- Mohammadi, M., Tajik, E., Martínez-Maldonado, R., Sadiq, S., Tomaszewski, W. y Khosravi, H. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. *Computers and Education: Artificial Intelligence*, 8, 100426. <https://doi.org/10.1016/j.caeai.2025.100426>
- Niño-Carrasco, S. A., Castellanos-Ramírez, J. C., Perezchica Vega, J. E. y Sepúlveda Rodríguez, J. A. (2025). Percepciones de estudiantes universitarios sobre los usos de inteligencia artificial en educación. *Revista Fuentes*, 27(1), 94–106. <https://doi.org/10.12795/revistafuentes.2025.26356>
- Ouyang, F., Wu, M., Zheng, L., Zhang, L. y Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education*, 20(1), 4. <https://doi.org/10.1186/s41239-022-00372-4>
- Shen, M., Shen, Y., Liu, F. y Jin, J. (2025). Prompts, privacy, and personalized learning: integrating AI into nursing education—a qualitative study. *BMC Nursing*, 24(1), 470. <https://doi.org/10.1186/s12912-025-03115-8>
- Singh, S. y Strzelecki, A. (2025). Academics as adopters of generative AI: an application of diffusion of innovations theory. *Education and Information Technologies*, 31(2), 621–645. <https://doi.org/10.1007/s10639-025-13835-8>
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., et al. (2025). Large Language Models for Education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6), 51–63. <https://doi.org/10.1109/MSP.2025.3594309>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. En *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 24824–24837). <https://proceedings.com/https://doi.org/10.52202/068431-1800>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martínez-Maldonado, R., Chen, G., Li, X., et al. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>
- Zawacki-Richter, O., Marín, V. I., Bond, M. y Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>