Informe público de resultados del proyecto:
Análisis de buenas prácticas en diseño de prompts y su impacto en el rendimiento de modelos LLM



IA GENERATIVA EN LA ACADEMIA:

Calidad, confiabilidad y buenas prácticas para trabajar con ChatGPT y Gemini

CENTRO DE INVESTIGACIÓN EDUCATIVA AIP

Del Istmo

©Centro de Investigación Educativa - AIP, 2026
Edificio Innova 104, Ciudad del Saber.

Primera edición: 29 de octubre de 2026
ISBN obra digital: 978-9962-8644-4-8

Todos los derechos reservados. Está prohibida la reproducción total o parcial de este documento, su recopilación en un sistema informático, su transmisión en cualquier forma o medida, ya sea electrónica, mecánica, por fotocopia, registro o bien por otros medios, sin el previo permiso por escrito de la editorial.

Preparado por:
Dra. Aura L. López de Ramos, Centro de Investigación Educativa AIP (CIEDU AIP).
Dr. Gustavo Quintero Barreto, Universidad del Istmo.
Dra. Yuly Esteves González, Universidad Pedagógica Experimental Libertador / Universidad Internacional de Ciencia y Tecnología.
Dra. Nelly Meléndez, Universidad Internacional de Ciencia y Tecnología / Universidad Monteávila.
Revisión editorial: Laura Esther Martínez
Diseño Gráfico: José Manuel López

Este informe público sintetiza los hallazgos del proyecto de investigación titulado *"Análisis de buenas prácticas en el diseño de prompts y su impacto en el rendimiento de los modelos"*. Su propósito es acercar los resultados a docentes, estudiantes, investigadores, autoridades universitarias y personas interesadas en el uso responsable de la inteligencia artificial generativa en educación superior.



Esta obra está bajo una Licencia Creative Commons
https://creativecommons.org/licenses/by/4.0/

Importante para interpretar los resultados
Las comparaciones corresponden a las versiones y condiciones evaluadas durante el estudio: ChatGPT 4.5 y Gemini 2.5 Flash. Los modelos de IA cambian con rapidez; por ello, estos resultados no constituyen una clasificación permanente ni universal.

Cita sugerida
López de Ramos, A. L., Quintero Barreto, G., Esteves González, Y., y Meléndez, N. (2026). *IA generativa en la academia: calidad, confiabilidad y buenas prácticas para trabajar con ChatGPT y Gemini.* [Informe público de resultados]. Centro de Investigación Educativa.

Declaración de principios sobre lenguaje y equidad
En el CIEDU AIP promovemos un lenguaje respetuoso, inclusivo y no discriminatorio, en coherencia con los principios de equidad e inclusión que orientan nuestras acciones en el ámbito educativo. Reconocemos que el español aún no cuenta con una forma universalmente aceptada para expresar la inclusión de todos los géneros, por lo que procuramos evitar expresiones que invisibilicen a parte de la población. No obstante, para favorecer la fluidez y legibilidad de los textos, en algunos casos se emplea el masculino genérico, entendiéndolo como una forma no excluyente que abarca a todas las personas, sin distinción de género.

Índice

Lo esencial en dos páginas 5

Cinco mensajes para llevar 6

¿Por qué estudiar la calidad de las respuestas de la IA? 7

¿Qué papel cumple un buen prompt? 8

¿Cómo se realizó el estudio? 9

¿Qué mostró la evaluación algorítmica? 10

¿Qué modelo destacó en cada tarea? 11

¿Qué opinó la evaluación humana? 12

¿Cómo vivieron los evaluadores la interacción con la IA? 13

¿Qué significan los resultados para la academia? 14

Recomendaciones 15

Una guía rápida para trabajar con IA generativa 16

Alcances, limitaciones y próximos pasos 17

Productos y referencias 18

Glosario

Adecuación epistémica: Grado en que una respuesta respeta los principios, conceptos y formas de producir conocimiento propias de una disciplina. Por ejemplo, en una investigación implica que el enfoque, los objetivos y los métodos sean coherentes entre sí.

Alfabetización en inteligencia artificial: Conjunto de conocimientos y habilidades necesarios para utilizar la inteligencia artificial de manera crítica, ética y responsable. Incluye saber formular instrucciones, verificar respuestas, reconocer errores y comprender los límites de estas herramientas.

Algoritmo: Conjunto ordenado de instrucciones o reglas que un sistema informático utiliza para procesar información y realizar una tarea.

Alucinación de la inteligencia artificial: Respuesta falsa, imprecisa o inventada que un sistema de IA presenta de manera convincente. Puede incluir datos incorrectos, referencias inexistentes o afirmaciones sin fundamento.

Análisis cualitativo: Forma de investigación que busca comprender opiniones, experiencias, significados y percepciones. En este estudio se empleó para analizar los comentarios abiertos de las personas evaluadoras.

Análisis cuantitativo: Forma de análisis basada en datos numéricos, puntuaciones y procedimientos estadísticos. Permite comparar resultados y determinar si las diferencias observadas tienen respaldo estadístico.

ANOVA: Sigla en inglés de análisis de varianza. Es una prueba estadística utilizada para comparar los resultados de dos o más grupos y determinar si las diferencias observadas son mayores de lo que podría esperarse por azar.

ASPECCT: Técnica para estructurar prompts mediante componentes que ayudan a precisar aspectos como la audiencia, el propósito, el estilo, el contexto y los criterios que debe cumplir la respuesta. En el estudio se utilizó para diseñar instrucciones completas y organizadas.

Chain of Thought (CoT): Expresión en inglés que puede traducirse como cadena de razonamiento. Es una técnica de prompting que orienta al modelo para desarrollar una tarea mediante una secuencia ordenada de pasos. Se utiliza principalmente en problemas que requieren análisis o razonamiento complejo.

ChatGPT: Sistema de inteligencia artificial generativa desarrollado por OpenAI. Puede producir textos, responder preguntas, resumir información y apoyar diversas tareas a partir de instrucciones escritas por el usuario. En el estudio se evaluó la versión ChatGPT 4.5.

Claridad y sintaxis: Criterio que valora si una respuesta está redactada de manera comprensible, ordenada y gramaticalmente correcta.

CLEAR: Técnica para elaborar prompts claros, concisos, lógicos, explícitos, adaptables y reflexivos. Su propósito es reducir la ambigüedad y mejorar la precisión de las instrucciones proporcionadas a la inteligencia artificial.

Codiseñador textual: Forma de describir a la inteligencia artificial cuando se utiliza como apoyo para organizar ideas, generar esquemas o producir primeros borradores. El término no significa que la IA sustituya al autor ni que pueda asumir responsabilidad sobre el contenido.

Confiabilidad: Grado en que un sistema mantiene un desempeño consistente cuando se utiliza en condiciones similares. Una herramienta confiable no necesariamente produce respuestas perfectas, pero muestra cierta estabilidad en sus resultados.

Confianza vigilante: Actitud que reconoce la utilidad de la inteligencia artificial, pero mantiene una revisión crítica permanente. Implica aprovechar sus ventajas sin aceptar automáticamente todo lo que genera.

Contextualización disciplinar: Capacidad de una respuesta para ajustarse al lenguaje, los conceptos, los métodos y las necesidades de un campo específico, como educación, investigación científica o estadística.

Datos cualitativos: Información expresada principalmente mediante palabras, opiniones, relatos o comentarios. Se utiliza para comprender experiencias y significados.

Datos cuantitativos: Información expresada mediante números, frecuencias, porcentajes, puntuaciones o mediciones.

Desempeño algorítmico: Resultado obtenido por un modelo de inteligencia artificial cuando sus respuestas son evaluadas mediante criterios y herramientas automatizadas.

Evaluación algorítmica: Valoración de las respuestas de un modelo mediante plataformas o sistemas automatizados. En el estudio se consideraron criterios como precisión terminológica, adecuación epistémica, contextualización, razonamiento, nivel cognitivo, claridad y sintaxis.

Evaluación humana: Valoración realizada por personas que examinan las respuestas de la inteligencia artificial utilizando criterios o rúbricas. Permite incorporar el juicio experto y la percepción de usuarios reales.

Estadístico F de Fisher: Valor obtenido en una prueba ANOVA. Compara la variación existente entre los grupos con la variación observada dentro de cada grupo. Un valor F más alto puede indicar que las diferencias entre los grupos son relevantes, pero debe interpretarse junto con el valor p.

Gemini: Sistema de inteligencia artificial generativa desarrollado por Google. Puede procesar instrucciones y generar respuestas en diferentes formatos. En el estudio se evaluó la versión Gemini 2.5 Flash.

Hipótesis nula: Supuesto estadístico inicial según el cual no existen diferencias reales entre los grupos comparados. Las pruebas estadísticas permiten determinar si existe evidencia suficiente para rechazar ese supuesto.

Ingeniería de prompts: Proceso de diseñar, probar y mejorar las instrucciones que se proporcionan a una inteligencia artificial para obtener respuestas más precisas, pertinentes y útiles. No consiste únicamente en “hacer preguntas”, sino en definir con claridad la tarea, el contexto, los criterios y el formato esperado.

Inteligencia artificial: Campo de la informática dedicado al desarrollo de sistemas capaces de realizar tareas que normalmente requieren capacidades humanas, como reconocer patrones, procesar lenguaje, clasificar información o generar respuestas.

Inteligencia artificial generativa: Tipo de inteligencia artificial capaz de producir contenidos nuevos, como textos, imágenes, audios o códigos, a partir de instrucciones proporcionadas por una persona.

Interacción humano-IA: Proceso mediante el cual una persona formula instrucciones, recibe respuestas de un sistema de inteligencia artificial, las revisa y decide cómo utilizarlas o modificarlas.

Juicio experto: Valoración realizada por una persona con conocimientos y experiencia en un área determinada. En el uso académico de la IA, el juicio experto es necesario para revisar la precisión conceptual, metodológica y ética de las respuestas.

LLM: Sigla en inglés de Large Language Model, traducida como modelo de lenguaje de gran escala. Es un sistema de inteligencia artificial entrenado con grandes cantidades de texto para procesar y generar lenguaje.

Modelo de lenguaje: Sistema informático que identifica patrones en grandes volúmenes de texto y los utiliza para generar respuestas, completar frases, resumir información o producir nuevos contenidos.

Nivel cognitivo: Grado de complejidad mental que exige una tarea. Puede ir desde recordar o comprender información hasta analizar, evaluar o crear.

Operacionalización: Proceso de convertir una idea o concepto general en elementos concretos que puedan observarse, medirse o evaluarse. En investigación, por ejemplo, implica definir cómo se estudiará una variable.

Precisión terminológica: Uso correcto y exacto de las palabras y conceptos propios de una disciplina.

Prompt: Instrucción, pregunta o conjunto de indicaciones que una persona proporciona a un sistema de inteligencia artificial. Un prompt puede especificar la tarea, el contexto, la audiencia, las restricciones, las fuentes y el formato esperado.

Prompt estructurado: Instrucción organizada mediante componentes definidos, como rol, tarea, contexto, criterios, fuentes, formato y verificación. Su propósito es reducir ambigüedades y orientar mejor la respuesta de la IA.

Razonamiento: Proceso de relacionar información, formular inferencias y llegar a una conclusión. En el informe se evaluó si las respuestas mostraban una secuencia lógica y adecuada al tipo de tarea.

Rigor metodológico: Grado en que una investigación o respuesta sigue procedimientos claros, coherentes y justificables. Incluye definir adecuadamente los objetivos, el enfoque, los métodos, los instrumentos y los criterios de validación.

Rúbrica: Instrumento que presenta criterios y niveles de desempeño para evaluar un producto o una respuesta de manera organizada y consistente.

Sesgo: Tendencia sistemática que puede favorecer, excluir o representar de manera inadecuada determinadas perspectivas, grupos o resultados. Los sistemas de IA pueden reproducir sesgos presentes en los datos con los que fueron entrenados.

Significación estadística: Criterio utilizado para determinar si una diferencia observada probablemente refleja un efecto real y no solamente una variación producida por azar. No significa necesariamente que la diferencia sea grande o importante en términos prácticos.

Supervisión humana: Revisión y control ejercidos por una persona durante el uso de inteligencia artificial. Incluye verificar datos, corregir errores, evaluar la pertinencia del contenido y asumir la responsabilidad final por el producto.

Taxonomía de Bloom: Clasificación de niveles de aprendizaje y complejidad cognitiva. Incluye procesos como recordar, comprender, aplicar, analizar, evaluar y crear.

Triangulación: Proceso de combinar diferentes fuentes, métodos o tipos de evidencia para comprender mejor un fenómeno y aumentar la solidez de las conclusiones. En el estudio se integraron la evaluación algorítmica, la evaluación humana y el análisis cualitativo.

Validación: Proceso de comprobar si una información, método, instrumento o resultado es adecuado, coherente y suficientemente fundamentado.

Valor p: Medida estadística que indica qué tan compatibles son los resultados obtenidos con la hipótesis nula. Habitualmente, un valor p menor que 0.05 se interpreta como evidencia de una diferencia estadísticamente significativa. El valor p no indica por sí solo el tamaño ni la importancia práctica de la diferencia.

Verificación: Proceso de comprobar la exactitud de datos, afirmaciones, referencias y resultados antes de utilizar una respuesta generada por inteligencia artificial.

1. Lo esencial en dos páginas

La inteligencia artificial generativa puede producir respuestas claras, rápidas y bien organizadas. **Sin embargo, la fluidez del texto no garantiza por sí sola la precisión conceptual, el rigor metodológico ni la pertinencia para una disciplina.** Este estudio comparó las respuestas de ChatGPT 4.5 y Gemini 2.5 en tareas educativas e investigativas, combinando evaluación algorítmica y juicio humano.

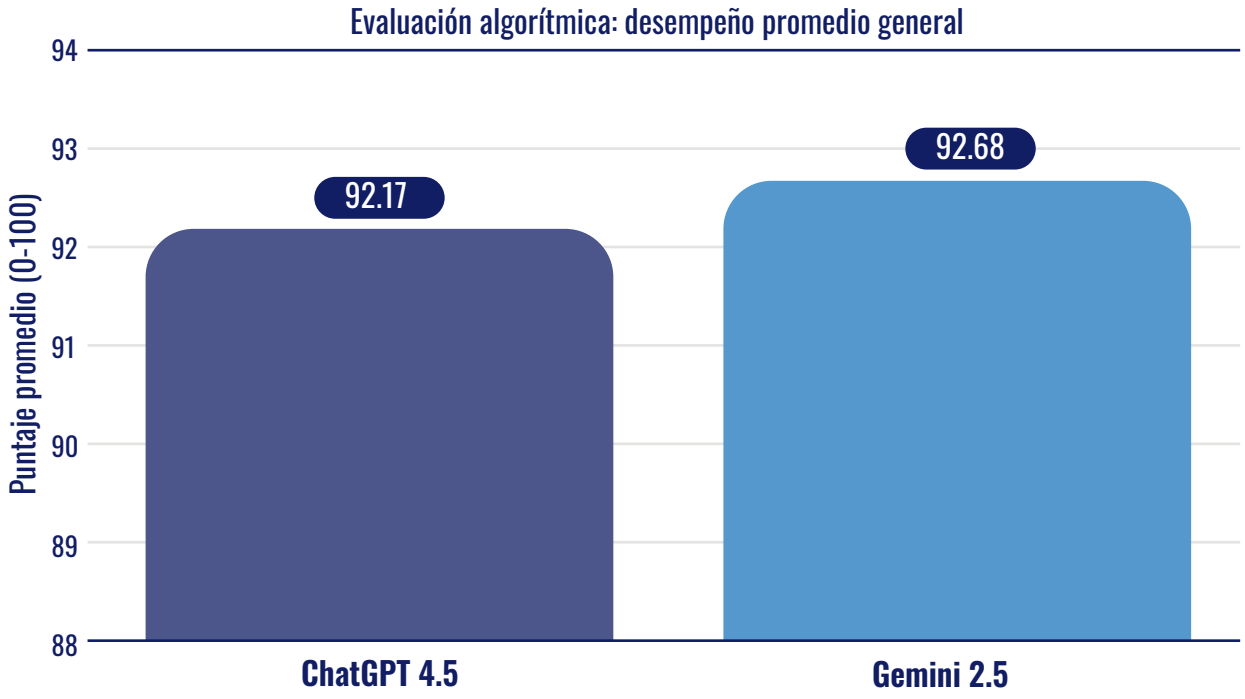


Figura 1. Promedio general de la evaluación algorítmica.

Hallazgo central

Gemini 2.5 obtuvo una ventaja promedio pequeña pero consistente en la evaluación general. ChatGPT 4.5 mostró un desempeño estable y una fortaleza particular en el análisis de datos cuantitativos. La diferencia entre modelos dependió del tipo de tarea.

Cinco mensajes para llevar

1	**La calidad depende de la tarea**	No existe un modelo que sea superior en todo. Las fortalezas variaron entre planificación didáctica, formulación de objetivos, educación continua y análisis de datos.
2	**El prompt importa**	Las instrucciones estructuradas, específicas y contextualizadas mejoraron la calidad de las respuestas. La combinación de CLEAR + Chain of Thought fue especialmente efectiva.
3	**La revisión humana es indispensable**	Los evaluadores describieron a los LLM como co-diseñadores textuales potentes, pero epistemológica y metodológicamente frágiles.
4	**La confianza debe ser vigilante**	La rapidez, claridad y organización de la IA son valiosas, pero cada resultado debe verificarse, contextualizarse y ajustarse.
5	**La ingeniería de prompts es una competencia académica**	Diseñar buenas instrucciones, analizar críticamente las salidas y reconocer límites y sesgos son capacidades necesarias para docentes, estudiantes e investigadores.

2. ¿Por qué estudiar la calidad de las respuestas de la IA?

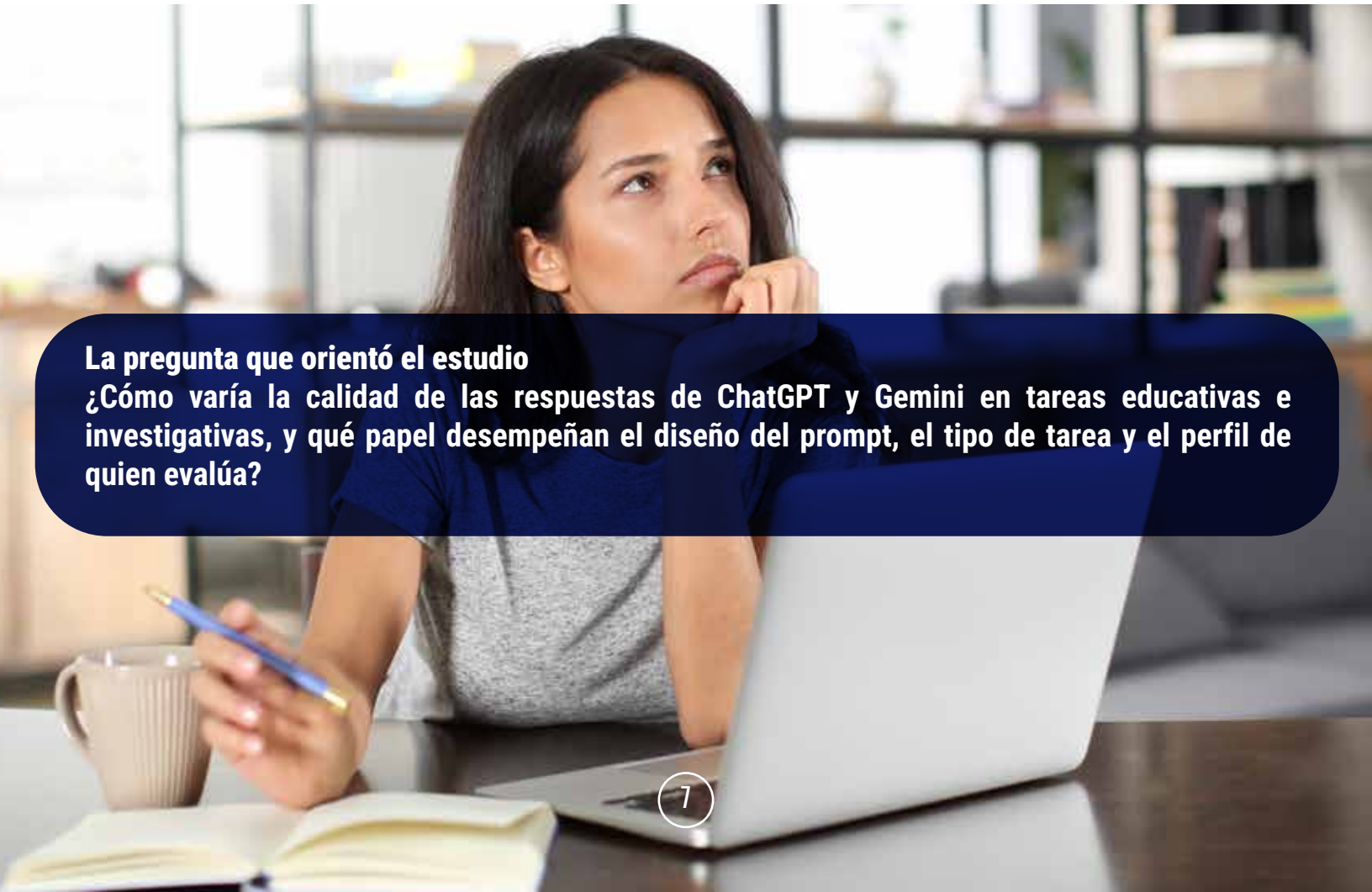
Los modelos de lenguaje de gran escala se han incorporado con rapidez a la redacción, la planificación didáctica, la síntesis de información y el diseño de investigaciones. Esa expansión ha sido más veloz que el desarrollo de marcos empíricos para evaluar la calidad, la confiabilidad y la pertinencia epistemológica de sus respuestas.

El problema no es solo "qué responde" la IA

Una respuesta puede estar bien escrita y, al mismo tiempo, contener vacíos conceptuales, decisiones metodológicas insuficientemente justificadas o afirmaciones que requieren verificación. Por ello, el estudio analizó la calidad desde tres planos complementarios:

- Desempeño sintáctico-estructural: claridad, sintaxis y precisión terminológica.
- Rigor metodológico y pertinencia disciplinar: adecuación epistémica y contextualización.
- Profundidad cognitiva: tipos de razonamiento y niveles cognitivos asociados a la Taxonomía de Bloom.

Una distinción clave
La calidad se entendió como el grado en que la respuesta cumple con estándares de precisión conceptual, coherencia metodológica y profundidad cognitiva. La confiabilidad se vinculó con la consistencia del desempeño en condiciones comparables, valorada mediante métricas y juicio experto.



La pregunta que orientó el estudio
¿Cómo varía la calidad de las respuestas de ChatGPT y Gemini en tareas educativas e investigativas, y qué papel desempeñan el diseño del prompt, el tipo de tarea y el perfil de quien evalúa?

3. ¿Qué papel cumple un buen prompt?

El prompt es la interfaz entre la intención humana y la respuesta del modelo. No es una simple pregunta: puede definir el rol, el propósito, el contexto, las restricciones, el formato esperado y los criterios de calidad. En educación, funciona como una forma de mediación pedagógica porque traduce los objetivos académicos en instrucciones procesables por la IA.

Tres técnicas evaluadas

Técnica	Qué aporta	Uso principal
ASPECCT	Estructura la instrucción mediante componentes que delimitan audiencia, propósito, estilo, contexto y criterios.	Diseño completo de tareas y productos.
CLEAR	Favorece instrucciones concisas, lógicas, explícitas, adaptables y reflexivas.	Claridad y especificidad del prompt.
Chain of Thought (CoT)	Induce una secuencia de razonamiento para tareas que requieren un análisis complejo.	Resolución y argumentación paso a paso.

La combinación más prometedora

CLEAR + Chain of Thought

El estudio encontró que CLEAR ayudó a estructurar la instrucción y CoT reforzó el razonamiento secuencial. Su combinación resultó especialmente efectiva en tareas complejas. Esto no significa que deba pedirse siempre una explicación extensa; el nivel de detalle debe responder al propósito y al tipo de producto esperado.

4. ¿Cómo se realizó el estudio?

Se utilizó un diseño mixto explicativo-secuencial. Primero se compararon cuantitativamente las respuestas de los modelos; luego se incorporaron la evaluación humana y el análisis cualitativo para interpretar los patrones observados.

Ruta metodológica

1. **Diseño de prompts:** Tareas reales de educación e investigación, estructuradas con ASPECCT, CLEAR y CoT.
2. **Generación de respuestas:** Aplicación de los mismos prompts a ChatGPT 4.5 y Gemini 2.5 Flash.
3. **Evaluación algorítmica:** Valoración en siete parámetros mediante cuatro plataformas algorítmicas de referencia.
4. **Evaluación humana:** 78 participantes calificaron las respuestas con rúbricas de 1 a 5 y aportaron comentarios abiertos.
5. **Integración:** Triangulación de resultados cuantitativos, valoraciones humanas y categorías cualitativas.

¿Quiénes participaron?

Docentes	Investigadores	Consultores	Profesionales no docentes	Otros
52.6%	34.6%	3.8%	6.4%	2.6%

¿Qué se evaluó?

- Precisión terminológica.
- Adecuación epistémica.
- Contextualización según la disciplina.
- Tipo de razonamiento.
- Nivel cognitivo según la Taxonomía de Bloom.
- Claridad y sintaxis.

Alcance de la comparación

La evaluación humana comparó ChatGPT y Gemini. El estudio también incluyó resultados algorítmicos de Claude en una fase previa, pero este informe público se centra en los dos modelos evaluados de manera comparable por las personas participantes.

5. ¿Qué mostró la evaluación algorítmica?

Ambos modelos obtuvieron puntuaciones altas. La diferencia global fue estrecha: 92.68 para Gemini y 92.17 para ChatGPT. La lectura más útil no consiste en declarar un ganador absoluto, sino en identificar fortalezas por tipo de tarea.

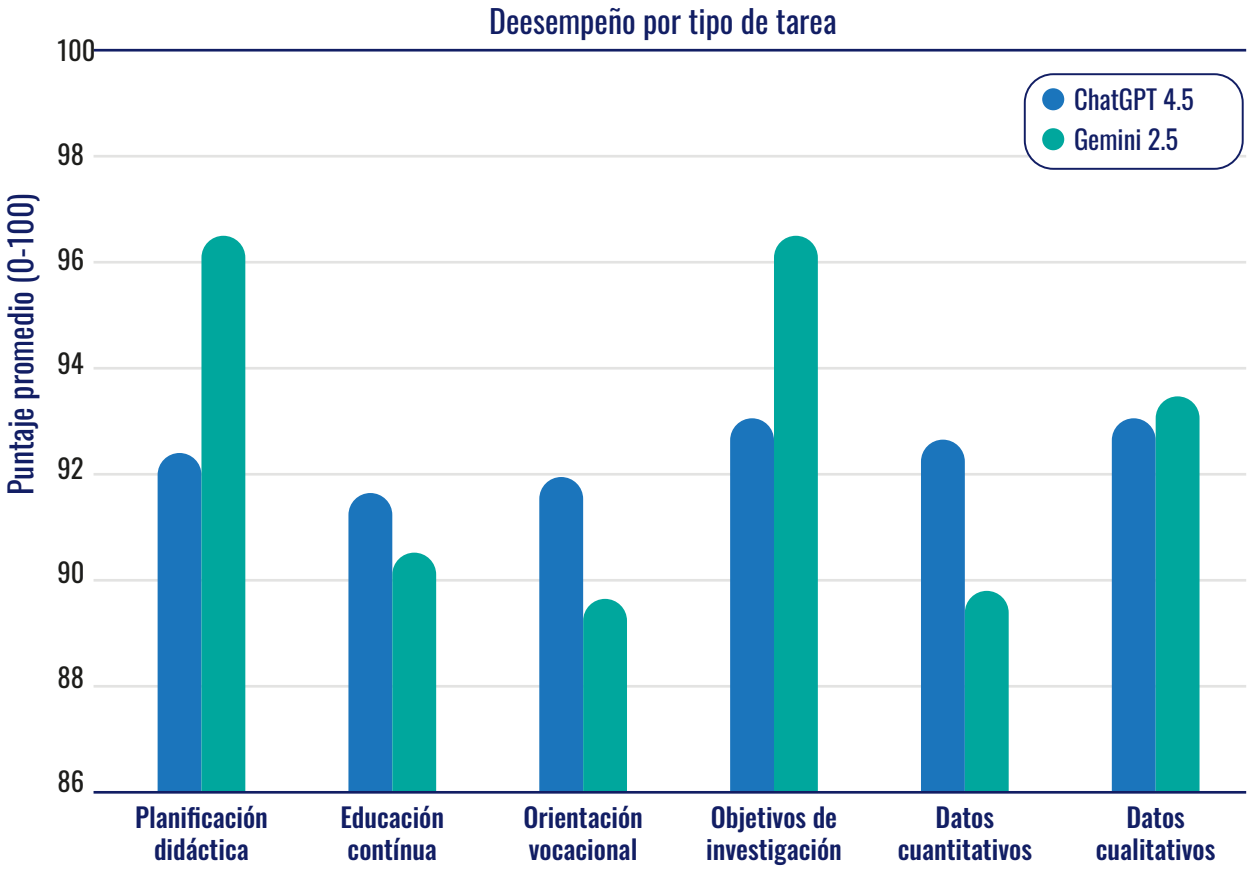


Figura 2. Resultados algorítmicos por subcampo.

Diferencias con mayor respaldo estadístico

- Objetivos de investigación: Gemini obtuvo 96.65 frente a 93.00 de ChatGPT; la diferencia fue significativa ($F=15.847$; $p=0.004$).
- Análisis de datos cuantitativos: ChatGPT obtuvo 92.65 frente a 89.75 de Gemini; la diferencia fue altamente significativa ($F=23.33$; $p<0.001$).
- Orientación vocacional: no se observaron diferencias significativas ($F=0.66$; $p=0.53$).

¿Qué modelo destacó en cada tarea?

Tarea	ChatGPT	Gemini	Lectura práctica
Planificación didáctica	91.75	96.40	Ventaja de Gemini en esta prueba
Educación contínua	91.35	90.30	Resultados cercanos; ligera venta de ChatGPT
Orientación vocacional	91.30	89.50	Diferencia no significativa
Objetivos de investigación	93.00	96.65	Ventaja significativa de Gemini
Datos cuantitativos	92.65	89.75	Ventaja significativa de ChatGPT
Datos cualitativos	93.00	93.45	Desempeño prácticamente equivalente

Qué no debe incluirse

Los datos no permiten afirmar que un modelo sea universalmente mejor. El desempeño observado estuvo condicionado por la versión del sistema, la tarea, la técnica de prompting, los criterios de evaluación y el contexto del estudio.



6. ¿Qué opinó la evaluación humana?

Las valoraciones humanas confirmaron el patrón general de la evaluación algorítmica: Gemini recibió puntuaciones superiores en las dos áreas. La ventaja fue mayor en investigación, donde las tareas demandaban más rigor metodológico y contextualización disciplinar.

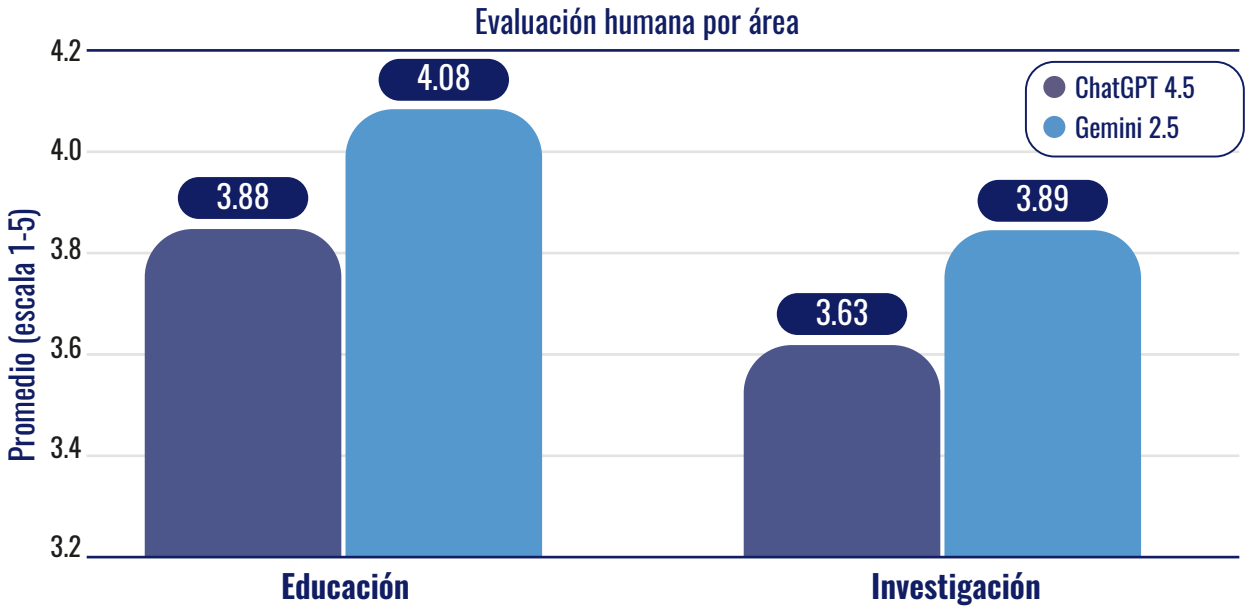


Figura 3. Promedio de la evaluación humana por área

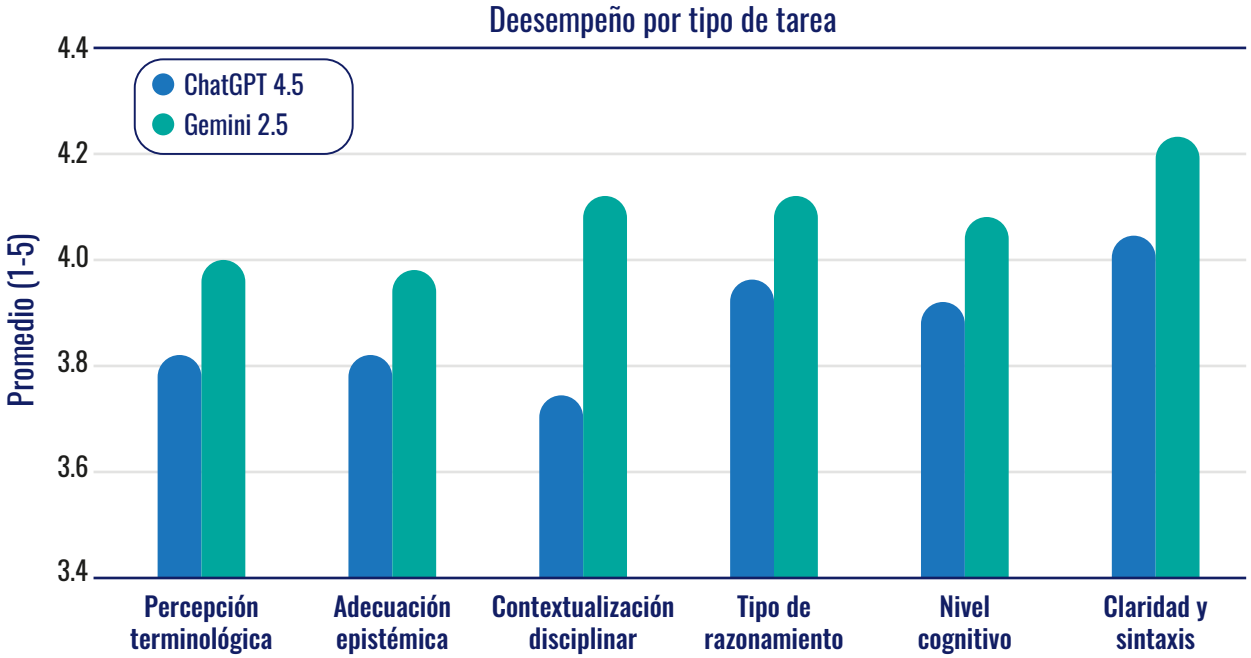


Figura 4. Evaluación humana por dimensiones en tareas educativas

7. ¿Cómo vivieron los evaluadores la interacción con la IA?

Los comentarios abiertos mostraron una experiencia ambivalente y madura. Las personas participantes valoraron la rapidez y la capacidad organizativa de los modelos, pero también identificaron límites que impiden delegar en ellos la responsabilidad académica.

Ocho ideas que emergieron del análisis cualitativo

1. Potencia estructuradora: Los LLM ayudan a organizar, sintetizar y presentar información en planes, tablas y borradores.
2. Profundidad diferenciada: Los modelos mostraron estilos distintos: uno más detallado y argumentativo; otro más operativo y directo.
3. Fragilidad epistémica y metodológica: Se observaron vacíos en paradigmas, operacionalización de objetivos, tipos de estudio y procedimientos de validación.
4. Estandarización discursiva: Algunas respuestas fueron percibidas como previsibles, robóticas o poco originales.
5. Dependencia del prompt: La calidad de la salida estuvo estrechamente vinculada con la precisión y riqueza de la instrucción humana.
6. Contextualización disciplinar: Los evaluadores valoraron el ajuste al campo de conocimiento y detectaron desajustes cuando el contenido era demasiado general.
7. Complementariedad funcional: La elección del modelo puede variar según la tarea; no siempre conviene usar la misma herramienta para todo.
8. Confianza vigilante: La IA se reconoce como útil, pero sus resultados deben revisarse, contrastarse y humanizarse.

Fenómeno central

Uso prudente de modelos de lenguaje como co-diseñadores textuales potentes, pero metodológicamente frágiles, en contextos académicos de educación e investigación.



8. ¿Qué significan los resultados para la academia?

La IA puede apoyar, pero no reemplazar, el juicio experto
Los modelos mostraron capacidad para redactar borradores, organizar ideas y agilizar tareas. Sin embargo, la responsabilidad por la validez conceptual, metodológica, ética y contextual permanece en las personas. La supervisión humana no es una etapa opcional al final del proceso; debe estar presente desde el diseño del prompt hasta la validación del producto.

La selección de herramientas debe ser diferenciada
Los resultados sugieren elegir la herramienta según la naturaleza de la tarea. En las condiciones del estudio, Gemini destacó en la formulación de objetivos, la planificación didáctica y la contextualización; ChatGPT mostró una fortaleza particular en el análisis cuantitativo y en la estabilidad entre áreas.

La alfabetización en IA debe ir más allá del uso instrumental
Saber "usar" una plataforma no es suficiente. La competencia académica incluye formular instrucciones de calidad, reconocer alucinaciones y sesgos, comprobar fuentes y datos, revisar la coherencia metodológica, adaptar el texto al contexto y declarar de manera transparente el uso de IA cuando corresponda.

La ingeniería de prompts puede enseñarse y evaluarse
El estudio demuestra que las instrucciones humanas pueden analizarse de forma sistemática. Esto abre oportunidades para incorporar la escritura de prompts, la evaluación crítica de las salidas y la comparación de respuestas como actividades de aprendizaje en cursos universitarios y en programas de formación docente.



9. Recomendaciones



Para universidades

Establecer lineamientos institucionales claros sobre los usos permitidos, la transparencia, la protección de datos, la autoría y la verificación de contenidos generados con IA.
Incorporar formación transversal en alfabetización en IA e ingeniería de prompts para estudiantes, docentes, investigadores y personal de apoyo.
Crear repositorios institucionales de prompts probados y contextualizados por disciplina, acompañados de criterios de calidad y ejemplos de revisión.
Promover evaluaciones periódicas de herramientas y versiones, evitando convertir resultados puntuales en decisiones permanentes de compra o adopción.



Para docentes

Diseñar prompts que expliciten el propósito, el contexto, la audiencia, el nivel cognitivo, las restricciones y el formato esperado.
Usar la IA como apoyo para generar alternativas, ejemplos o borradores, no como sustituto de la mediación pedagógica.
Convertir la revisión crítica de respuestas en una actividad de aprendizaje: detectar errores, contrastar fuentes y justificar las correcciones.
Evitar ingresar datos personales, confidenciales o sensibles en plataformas sin garantías institucionales adecuadas.



Para investigadores

Verificar referencias, datos, decisiones metodológicas y definiciones antes de incorporar contenidos generados por IA.
Documentar el modelo, la versión, la fecha, el prompt y el modo de uso cuando la IA forme parte sustantiva del proceso investigativo.
No delegar en el modelo la selección del paradigma, la operacionalización de variables, la validez de los instrumentos ni la interpretación final de los resultados.
Utilizar diferentes modelos o estrategias de contraste cuando la tarea tenga alta complejidad epistemológica.



Para estudiantes

Usar la IA para apoyar la comprensión, la práctica y la organización, no para sustituir el proceso de aprendizaje.
Solicitar explicaciones acordes con el nivel de conocimiento y comprobarlas con fuentes académicas confiables.
Reescribir y argumentar con voz propia: un texto fluido no demuestra comprensión.
Cumplir con las normas del curso sobre autoría, citación y declaración del uso de IA.

10. Una guía rápida para trabajar con IA generativa

Antes de escribir el prompt

- Define el propósito real de la tarea.
- Determina qué información puede compartirse y cuál debe mantenerse confidencial.
- Establece los criterios con los que revisarás la respuesta.

Un esquema de prompt en siete componentes

Componente	Pregunta guía	Ejemplo breve
Rol	¿Desde qué perspectiva debe responder?	Actúa como asesor metodológico
Tarea	¿Qué producto debe generar?	Propón tres objetivos específicos
Contexto	¿En qué situación se aplicará?	Estudio sobre docentes universitarios en Panamá
Criterios	¿Qué condiciones debe cumplir?	Verbos observables, coherencia con el objetivo general
Fuentes	¿Qué información puede usar?	Utiliza solo el texto proporcionado
Formato	¿Cómo debe presentar la respuesta?	Tabla con objetivo, variable y justificación
Verificación	¿Qué debe revisar o declarar?	Señala supuestos, vacíos y aspectos que requieren validación

Después de recibir la respuesta

- ¿Responde exactamente a la tarea?
- ¿Los conceptos y términos son correctos?
- ¿La propuesta es coherente con la disciplina y el contexto?
- ¿Las decisiones metodológicas están justificadas?
- ¿Hay datos, referencias o afirmaciones que deban verificarse?
- ¿El nivel cognitivo coincide con el propósito?
- ¿El texto requiere ajustes de tono, de inclusión, de claridad o de voz propia?

Regla práctica

No uses una respuesta de IA que no puedas explicar, comprobar y asumir como propia en términos de responsabilidad académica.

11. Alcances, limitaciones y próximos pasos

Alcances

- Integró la evaluación algorítmica, la evaluación humana y el análisis cualitativo.
- Comparó tareas educativas e investigativas bajo condiciones comunes.
- Incluyó diversos perfiles de evaluadores y criterios académicos explícitos.
- Generó orientaciones y productos transferibles para la formación.

Limitaciones

- Los resultados corresponden a versiones específicas de los modelos y pueden cambiar con actualizaciones posteriores.
- La muestra de 78 evaluadores fue seleccionada por conveniencia y no representa a todos los usuarios o disciplinas.
- La configuración de evaluadores algorítmicos internos requirió una fase extensa de pruebas y constituye una fuente potencial de variabilidad.
- El volumen de prompts necesario para asegurar la consistencia implicó restricciones de tiempo y de alcance.
- Las conclusiones se refieren a las tareas y criterios analizados; no deben extrapolarse automáticamente a otros idiomas, disciplinas o formatos multimodales.

Próximos pasos

- Replicar el estudio con versiones actualizadas y nuevos modelos.
- Ampliar la muestra y la diversidad disciplinar de los evaluadores.
- Incorporar métricas adicionales de confiabilidad, validez, sesgo y trazabilidad.
- Desarrollar repositorios abiertos de prompts validados y programas de formación especializada.
- Profundizar en las políticas institucionales para un uso ético, responsable y pedagógicamente fundamentado de la IA generativa.

12. Productos y referencias

Productos derivados del proyecto

- Manual para las buenas prácticas en la escritura de prompts (https://repositorio.ciedupanama.org/handle/123456789/1057)
- Guía práctica de recomendaciones para la escritura de prompts (https://repositorio.ciedupanama.org/handle/123456789/1076)
- Manuscrito científico aceptado por Comunicar: “Respuestas educativas de IA: calidad y confiabilidad en ChatGPT y Gemini”. Publicado en octubre 2026 (en imprenta).
- Quintero Barreto, G., López de Ramos, A. L., Esteves González, Y., & Meléndez, N. (2026). Respuestas educativas de IA: Calidad y confiabilidad en ChatGPT y Gemini. Comunicar, 34(87), 201–211. https://doi.org/10.5281/zenodo.23185063

Referencias

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.

Cain, W. (2023). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. TechTrends, 68, 47-57.

García-Peñalvo, F. J. (2024). Inteligencia artificial generativa y educación: Un análisis desde múltiples perspectivas. Education in the Knowledge Society, 25.

Holmes, W., Bialik, M., y Fadel, C. (2019). Inteligencia artificial en la educación: Promesas e implicaciones para la enseñanza y el aprendizaje.

Jara, I., y Ochoa, J. M. (2020). Usos y efectos de la inteligencia artificial en educación. Banco Interamericano de Desarrollo.

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering.

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models.

Mensaje final

La IA generativa amplía la capacidad de producir, organizar y explorar ideas. Su valor académico depende menos de aceptar respuestas rápidas y más de formular buenas preguntas, revisar con rigor y conservar la responsabilidad humana sobre el conocimiento.

CENTRO DE
INVESTIGACIÓN
EDUCATIVA
AIP

@ciedupanama

ciedupanama.org