



PRISMA ODS
REVISTA MULTIDISCIPLINARIA
SOBRE DESARROLLO SOSTENIBLE

ISSN: 3072-8452

**ANÁLISIS COMPARATIVO
DEL DESEMPEÑO DE
MODELOS DE INTELIGENCIA
ARTIFICIAL EN LA
RESOLUCIÓN DE PROBLEMAS
MATEMÁTICOS**

*COMPARATIVE ANALYSIS OF
ARTIFICIAL INTELLIGENCE MODELS
IN SOLVING MATHEMATICAL
PROBLEMS*

AUTORES

**LAURA ARELI LLANO
ARANDA**

UNIVERSIDAD
INTERAMERICANA
MÉXICO

**EDUARDO
HERNÁNDEZ CRUZ**

UNIVERSIDAD
INTERAMERICANA
MÉXICO

JOSAFAT CORTÉS

UNIVERSIDAD
INTERAMERICANA
MÉXICO

**CARLOS DANIEL
GUTIÉRREZ RÍOS COLOTL**

UNIVERSIDAD
INTERAMERICANA
MÉXICO

**SEBASTIÁN SÁNCHEZ
MARTÍNEZ**

UNIVERSIDAD
INTERAMERICANA
MÉXICO

**DIEGO HERNAN
CUATEGOMEZ**

INSTITUTO TECNOLÓGICO
SUPERIOR PROGRESO
MÉXICO

Análisis Comparativo del Desempeño de Modelos de Inteligencia Artificial en la Resolución de Problemas Matemáticos

Comparative Analysis of Artificial Intelligence Models in Solving
Mathematical Problems

Laura Areli Llano Aranda

laura.areli.llano@gmail.com

<https://orcid.org/0009-0008-2707-5586>

Universidad Interamericana

Puebla - México

Eduardo Hernández Cruz

cruzedu211006@gmail.com

<https://orcid.org/0009-0007-3103-7630>

Universidad Interamericana

Puebla - México

Josafat Cortés

a000007390@lainter.edu.mx

Universidad Interamericana

Puebla - México

Carlos Daniel Gutiérrez Ríos Colotl

carlos_dgr@outlook.com

<https://orcid.org/0009-0003-7638-6608>

Universidad Interamericana

Puebla - México

Sebastián Sánchez Martínez

a000007412@lainter.edu.mx

<https://orcid.org/0009-0006-7919-4371>

Universidad Interamericana

Puebla - México

Diego Hernan Cuate Gomez

dheg.inv@gmail.com

<https://orcid.org/0000-0003-1741-0009>

Instituto Tecnológico Superior Progreso

Yucatán - México

Artículo recibido: 07/02/2026

Aceptado para publicación: 30/03/2026

Conflictos de Intereses: Ninguno que declarar

RESUMEN

La presente investigación presenta los resultados preliminares de un análisis comparativo sobre el desempeño de tres modelos de inteligencia artificial generativa — ChatGPT (GPT- 5), Gemini 2.5 y DeepSeek V3— en la resolución de problemas matemáticos de nivel intermedio. El objetivo principal fue identificar diferencias relevantes en sus métodos de razonamiento y en la claridad de los procedimientos utilizados, sin profundizar aún en aspectos técnicos internos de cada arquitectura. Para ello, se aplicó un conjunto homogéneo de ejercicios que incluyó operaciones algebraicas, manipulación de fracciones, factorización y resolución de ecuaciones. Cada modelo recibió las mismas instrucciones y sus respuestas fueron evaluadas a partir de criterios cualitativos como claridad explicativa, coherencia de los pasos intermedios y estabilidad del razonamiento. Los resultados preliminares indican que, aunque los tres modelos convergen en las soluciones finales, presentan diferencias sistemáticas en la forma de desarrollar el proceso. ChatGPT destaca por su claridad pedagógica y su capacidad para detallar cada paso, lo que lo hace apropiado para contextos educativos. Gemini muestra una marcada eficiencia operativa y tiende a simplificar expresiones desde etapas tempranas, lo cual agiliza el cálculo y reduce la complejidad algebraica. DeepSeek, por su parte, adopta un enfoque más estructurado orientado a la factorización y a la organización lógica de los términos, ofreciendo procedimientos formales y precisos. Estos hallazgos permiten identificar patrones iniciales sobre el estilo matemático de cada modelo y establecen las bases para análisis posteriores más profundos.

Palabras clave: inteligencia artificial, modelos de lenguaje, razonamiento matemático, educación, matemáticas

ABSTRACT

This study presents the preliminary results of a comparative evaluation of three contemporary generative artificial intelligence models—ChatGPT (GPT-5), Gemini 2.5, and DeepSeek V3—focused on their performance in solving intermediate-level mathematical problems. The main objective was to identify differences in their reasoning strategies and in the clarity of their step-by-step procedures, without addressing internal technical specifications or aspects beyond the scope of mathematical problem solving. A homogeneous set of algebraic exercises was administered to all three models, including fraction manipulation, equation solving, factorization, and symbolic simplification. Each response was assessed using qualitative criteria such as explanatory clarity, logical coherence, and consistency in the reasoning process. Preliminary results indicate that all three systems successfully reached the correct numerical solutions; however, their approaches to developing intermediate steps differed systematically. ChatGPT stood out for its pedagogical clarity, offering detailed explanations that facilitate understanding in educational settings. Gemini demonstrated strong operational efficiency, often simplifying expressions early in the process, which reduced computational complexity and accelerated problem resolution. DeepSeek adopted a more structured approach, showing a preference for factorization and formal algebraic organization in its reasoning. These findings provide an initial characterization of the mathematical reasoning styles of each model and establish a basis for future, more in-depth analyses. The results may assist educators and students in selecting the most suitable AI tool based on the desired balance between clarity, efficiency, and mathematical formality.

Keywords: artificial intelligence, language models, mathematical reasoning, education, mathematics

INTRODUCCIÓN

La Era de la Inteligencia Artificial Generativa y su Impacto Multidisciplinario

La inteligencia artificial y, de manera más específica, los grandes modelos de lenguaje (LLM), han experimentado un crecimiento que puede calificarse sin hipérbolos como exponencial en la última década. Esta tecnología ha logrado una integración significativa y transversal en diversos campos del saber humano, transformando radicalmente desde la educación formal hasta la práctica de la medicina de alta especialidad y la gestión de las finanzas globales. La capacidad de estos sistemas para procesar, comprender y generar lenguaje natural ha abierto nuevas fronteras en la interacción hombre-máquina, planteando preguntas fundamentales sobre la naturaleza del conocimiento, el aprendizaje y la automatización de tareas cognitivas complejas. Este fenómeno no es meramente tecnológico, sino que representa un cambio cultural y epistemológico en la manera en que accedemos y procesamos la información en el siglo XXI (Bender et al., 2021; Floridi & Chiriatti, 2020).

La literatura científica reciente se ha dedicado a explorar con profundidad la llegada de la inteligencia artificial a los entornos educativos, destacando su potencial no solo como una herramienta auxiliar, sino como una ventaja competitiva estratégica en la impartición y generación de nuevos conocimientos (Holmes et al., 2022; Luckin, 2018). La promesa de un tutor personalizado, disponible las veinticuatro horas del día y capaz de adaptarse al ritmo de aprendizaje de cada estudiante, ha sido uno de los motores principales de esta adopción. Sin embargo, esta integración no está exenta de dificultades estructurales y pedagógicas que deben ser abordadas con rigor académico. Investigadores como Moreno Padilla (2019) han subrayado en sus trabajos los retos considerables que implica su implementación efectiva en las aulas. Entre estos desafíos se encuentran la necesaria adaptación curricular para incluir competencias digitales avanzadas, la formación docente continua para evitar la brecha tecnológica y la gestión de la dependencia tecnológica por parte del alumnado, que podría mermar el desarrollo del pensamiento crítico si no se supervisa adecuadamente.

De igual manera, se ha investigado la nueva realidad educativa ante los avances de la inteligencia artificial generativa, que va un paso más allá de la simple recuperación de información para crear contenido original. En este sentido, García-Peñalvo y sus colaboradores (2024) publicaron un análisis exhaustivo sobre las ventajas, los inconvenientes y las potencialidades que estas herramientas presentan en los procesos de enseñanza y aprendizaje. Su trabajo sugiere que estamos ante un cambio de paradigma en la transmisión

del conocimiento que requiere una reevaluación profunda de los métodos de evaluación tradicionales. Ya no basta con evaluar la memorización o la repetición de datos, puesto que la inteligencia artificial puede realizar estas tareas con una eficiencia sobrehumana; la educación debe evolucionar hacia la evaluación de la capacidad de síntesis, el análisis crítico y la resolución creativa de problemas, habilidades donde la supervisión humana sigue siendo indispensable.

Democratización del Acceso y Eficiencia de Entrenamiento

En el ámbito específico del desarrollo técnico de los modelos de lenguaje, la competencia entre desarrolladores ha acelerado la innovación de manera vertiginosa, rompiendo barreras de entrada que antes parecían insuperables para actores más pequeños. Modelos recientes de código abierto o semiabierto, como DeepSeek-V3 y DeepSeek-R1, han demostrado un rendimiento comparable, y en ocasiones superior, al de modelos de código cerrado desarrollados por grandes corporaciones tecnológicas con recursos casi ilimitados. Lo más notable corroborado por estudios técnicos como Bernstein et al. (2023) es que estos resultados se han logrado con una fracción de los costos de entrenamiento habituales.

Esto se debe a innovaciones arquitectónicas como la Atención Latente Multi cabezal y la arquitectura de Mezcla de Expertos (MoE), que permiten un uso más eficiente de los recursos computacionales al activar solo las partes necesarias de la red para cada tarea específica.

Este fenómeno de democratización tecnológica tiene implicaciones profundas. Sugiere un futuro cercano donde modelos altamente capaces puedan entrenarse y ejecutarse localmente en universidades, hospitales y pequeñas empresas, reduciendo la dependencia de grandes corporaciones y aumentando la soberanía tecnológica.

DESARROLLO

Estado del Arte: Revisión Sistemática de Aplicaciones Críticas

Para comprender la relevancia de evaluar el razonamiento matemático de las inteligencias artificiales, es necesario analizar primero cómo se han desempeñado estas herramientas en otros campos de alta complejidad y riesgo, como la medicina, la ingeniería y las finanzas. La revisión de la literatura ofrece un panorama de luces y sombras que justifica la necesidad de evaluaciones rigurosas como la que presenta este estudio.

La Inteligencia Artificial en el Sector Salud: Promesas y Riesgos

La evaluación de la precisión y la legibilidad de las respuestas generadas por diferentes modelos de inteligencia artificial generativa se ha convertido en un campo crítico, especialmente para validar su uso seguro en el sector salud, donde la exactitud puede definir decisiones con impacto directo en la vida de los pacientes.

Investigaciones recientes, como el estudio transversal de Liang et al. (2025), evaluaron modelos líderes incluyendo ChatGPT y Gemini en contextos altamente sensibles, como la respuesta a preguntas relacionadas con la pandemia de COVID-19. Sus hallazgos indican diferencias sustanciales en la precisión textual y la fluidez del lenguaje entre plataformas. Mientras algunos modelos simplifican adecuadamente conceptos médicos complejos para el público general, otros pueden introducir imprecisiones sutiles o alucinaciones que, aunque persuasivas, son clínicamente peligrosas.

En oftalmología clínica, Bernstein et al. (2023) compararon respuestas de chatbots con las de oftalmólogos humanos a consultas de pacientes realizadas en línea. Sorprendentemente, las respuestas de la IA fueron percibidas por los pacientes como más empáticas, extensas y detalladas que las de los médicos humanos. Sin embargo, los autores advierten que la calidad percibida no sustituye la supervisión médica, ya que la IA puede omitir matices clínicos fundamentales.

En cirugía reconstructiva de cabeza y cuello, Boscolo-Rizzo et al. (2025) realizaron un análisis comparativo entre ChatGPT-4 y Claude-2. Encontraron variaciones importantes en la precisión de la orientación postoperatoria y manejo de complicaciones, enfatizando que cada modelo debe validarse específicamente para subespecialidades médicas antes de su uso clínico.

La IA también ha sido evaluada en tareas de procesamiento de texto médico administrativo y clínico. Le Guellec et al. (2024) demostraron que los modelos pueden automatizar la extracción de información estructurada desde reportes de radiología redactados en texto libre, facilitando la investigación epidemiológica y la gestión hospitalaria.

Además, estudios como el de Mishra et al. (2025) en educación médica y el de Du et al. (2024) en patología han utilizado preguntas de certificación profesional para medir el conocimiento “enciclopédico” de los modelos. En varios casos, los modelos superaron el desempeño promedio de los estudiantes de medicina, lo cual abre un debate profundo sobre el futuro de la evaluación de competencias en salud.

En conjunto, estos estudios demuestran que, aunque las IA muestran un notable potencial en medicina, los riesgos asociados a errores sutiles siguen siendo un factor crítico que obliga a evaluaciones exhaustivas antes de integrar estos sistemas en la práctica clínica.

Impacto en la Programación, Finanzas y Sostenibilidad

Más allá del sector salud, la inteligencia artificial ha dejado una huella profunda en disciplinas como programación, finanzas y sostenibilidad, áreas donde su capacidad para procesar texto y generar soluciones precisas tiene aplicaciones inmediatas y de alto impacto.

En el área de programación, diversos estudios han evaluado el rendimiento de los modelos para generar y depurar código. Coello et al. (2024) compararon la efectividad de ChatGPT frente a otros modelos para escribir código funcional en lenguajes modernos. Destacaron su habilidad no solo para producir código sintácticamente correcto, sino también para explicar la lógica subyacente, lo cual es valioso para la formación de nuevos programadores y para mejorar la productividad de desarrolladores experimentados.

En el desarrollo web, Smutny y Bojko (2024) demostraron que los modelos pueden generar estructuras completas de HTML y CSS, lo que acelera significativamente el prototipado. Sin embargo, también señalan inconsistencias ocasionales que requieren supervisión humana para asegurar la funcionalidad final.

En el ámbito financiero, Chen et al. (2025) evaluaron si modelos avanzados como ChatGPT y DeepSeek podían predecir el comportamiento del mercado bursátil mediante el análisis de miles de noticias financieras. Los resultados mostraron correlaciones relevantes que sugieren un potencial disruptivo. No obstante, los autores advierten que los mercados financieros presentan alta volatilidad y sesgos en los datos, lo que puede llevar a predicciones engañosas si se confía ciegamente en estos modelos.

La sostenibilidad constituye otro eje crítico. El entrenamiento y operación de modelos masivos consume grandes cantidades de energía y agua. Cowsls et al. (2023) analizaron este fenómeno en el contexto de la crisis climática, cuestionando si los beneficios de optimización que aporta la IA compensan su huella ambiental. Refuerzan la necesidad de enfoques más eficientes y energéticamente responsables en futura investigación y desarrollo.

Finalmente, Ratti et al. (2025) destacaron los riesgos éticos asociados a sesgos algorítmicos, especialmente en salud. Señalan que, si los datos de entrenamiento no representan

adecuadamente la diversidad poblacional, la IA podría perpetuar o incluso amplificar inequidades históricas.

Protocolo Metodológico Detallado para la Replicación del Estudio

Con el objetivo de dotar a esta investigación de un carácter científico riguroso y permitir la verificación independiente de sus conclusiones por parte de la comunidad académica internacional, se diseñó un protocolo metodológico exhaustivo. Este protocolo documenta paso a paso el procedimiento necesario para replicar el estudio comparativo, asegurando que las variables estén estrictamente controladas y que los resultados sean comparables entre distintas iteraciones experimentales.

Requisitos Técnicos y Preparación del Entorno Controlado

El primer paso crítico para cualquier intento de replicación exitosa consiste en asegurar el acceso estable y autorizado a las interfaces de programación de aplicaciones (API) de los modelos evaluados. Para este estudio se seleccionaron ChatGPT (versión GPT-5), Gemini 2.5 y DeepSeek V3.

Es imprescindible establecer credenciales de acceso seguras y configurar un entorno de desarrollo aislado para evitar interferencias externas. Se recomienda el uso de tecnologías de contenedorización, como Docker, para garantizar que todas las dependencias de software, bibliotecas de lenguajes como Python y configuraciones del sistema operativo sean idénticas en todas las máquinas de prueba. Esto elimina variables de confusión relacionadas con diferencias en hardware o configuraciones locales.

Además, se estableció un repositorio de datos controlado mediante sistemas de control de versiones como Git. Allí se almacenaron de manera organizada los scripts de prueba automatizados, los prompts utilizados para el estudio y los datos recolectados durante las sesiones experimentales.

Diseño Experimental y Estructura del Banco de Pruebas

La validez de este estudio depende directamente del diseño del banco de pruebas (benchmark). Para evaluar competencias cognitivas específicas de cada modelo, el banco se dividió en bloques temáticos:

- **Bloque Matemático:** Incluyó problemas de álgebra lineal como el cálculo de determinantes de matrices, problemas de cálculo diferencial y resolución de ecuaciones

cuadráticas con coeficientes fraccionarios. Estos problemas se seleccionaron no por su complejidad computacional, sino por su necesidad de razonamiento lógico paso a paso y su capacidad para poner a prueba la claridad explicativa del modelo.

- **Bloque de Programación:** Evaluó la capacidad de los modelos para generar código funcional en distintos lenguajes, depurar errores lógicos y explicar algoritmos en términos accesibles.
- **Bloque de Robustez y Seguridad:** Incluyó preguntas capciosas, premisas falsas o instrucciones contradictorias para medir la tendencia del modelo a generar alucinaciones o errores lógicos. Esto permitió evaluar su capacidad para identificar inconsistencias en el planteamiento del problema antes de responder.

Estandarización de Prompts y Recolección Sistemática de Datos

Para minimizar la variabilidad en las respuestas generada por la forma de formular las preguntas (un fenómeno conocido como sensibilidad al prompt engineering), se utilizaron plantillas maestras estandarizadas para cada tipo de tarea. Estas plantillas definían claramente el rol que debía adoptar el modelo (por ejemplo, "Actúa como un profesor universitario de matemáticas"), la tarea específica a realizar y el formato de salida esperado (por ejemplo, "Utiliza formato LaTeX para las fórmulas matemáticas y explica cada paso en texto plano"). Durante la fase de recolección de datos, se registraron automáticamente no solo las respuestas textuales generadas, sino también una serie de metadatos críticos para el análisis de eficiencia, como la latencia de la red, el tiempo de procesamiento del servidor, el número exacto de tokens de entrada y salida, y los costos financieros estimados por cada consulta realizada. Cada tarea se ejecutó múltiples veces ($n=3$) para evaluar la consistencia y estabilidad temporal de los modelos ante la misma entrada, permitiendo detectar fluctuaciones en el rendimiento.

Criterios de Evaluación Multidimensional

La evaluación de las respuestas trascendió la simple y reduccionista verificación binaria de correcto o incorrecto. Se implementó un esquema de evaluación multidimensional y holístico que abarcó cuatro ejes principales de análisis.

1. **Exactitud Matemática:** Medida objetiva de la corrección del resultado numérico final, verificada contra soluciones patrones precalculadas.

2. **Completitud Procedimental:** Evaluación de si el modelo abordó todos los aspectos de la pregunta y proporcionó los pasos intermedios necesarios para justificar su respuesta, penalizando los saltos lógicos injustificados.
3. **Claridad y Legibilidad:** Valoración cualitativa realizada por un panel de evaluadores humanos expertos en la materia, quienes puntuaron la capacidad pedagógica, la estructura gramatical y la fluidez narrativa de las explicaciones generadas.
4. **Eficiencia Algorítmica:** Análisis del enfoque utilizado por el modelo para resolver el problema (por ejemplo, el uso de fuerza bruta computacional frente a la simplificación algebraica inteligente) y el consumo de recursos computacionales asociados a dicha estrategia.

Perfiles Técnicos y Análisis de Arquitectura de los Modelos Evaluados

Para interpretar adecuadamente los resultados del rendimiento matemático que se discutirán más adelante, es indispensable comprender primero las diferencias fundamentales en la arquitectura y las especificaciones técnicas de los modelos evaluados, tal como se describen en la documentación oficial de los desarrolladores.

ChatGPT (GPT-5): La Potencia del Razonamiento Adaptativo y la Escala Masiva

Desarrollado por OpenAI, GPT-5 representa la cúspide actual de los modelos densos y multimodales de propósito general. Su arquitectura está diseñada desde la base para procesar y relacionar texto, imágenes y voz de manera nativa, sin necesidad de módulos externos. Una de sus características más distintivas y revolucionarias es su ventana de contexto masiva de 400,000 tokens, lo que le permite mantener la coherencia en conversaciones extremadamente largas o analizar documentos legales y técnicos extensos en una sola pasada de inferencia. Sin embargo, su característica definitoria es el sistema de "razonamiento adaptativo", una capacidad emergente que le permite ajustar dinámicamente la profundidad y complejidad de su procesamiento neuronal según la dificultad percibida de la tarea en cuestión. Esta potencia, no obstante, conlleva un costo operativo significativo: el consumo energético estimado es alto, alrededor de 18.35 Wh por consulta promedio, lo que plantea desafíos importantes en términos de sostenibilidad ambiental y escalabilidad económica a largo plazo.

Gemini 2.5: La Apuesta por la Eficiencia Extrema y la Multimodalidad Nativa

Gemini, el modelo insignia desarrollado por Google DeepMind, apuesta por una filosofía de diseño centrada en la eficiencia extrema y la integración total. Su arquitectura multimodal nativa le permite procesar flujos de video y audio con la misma fluidez y precisión que el texto escrito. La variante Gemini 2.5 Flash destaca en el mercado por su ventana de contexto sin precedentes de más de un millón de tokens, la más grande disponible comercialmente, lo que abre nuevas posibilidades para el análisis de grandes volúmenes de datos no estructurados. A pesar de esta capacidad monumental, su optimización arquitectónica le permite ser, según reportes técnicos, hasta 33 veces más eficiente energéticamente que sus predecesores inmediatos, con un consumo de apenas 0.24 Wh por consulta típica. Esta eficiencia no es solo energética, sino que se refleja en su comportamiento algorítmico, que tiende a buscar atajos matemáticos y simplificaciones lógicas para reducir la carga computacional total de las tareas.

DeepSeek V3: La Revolución del Código Abierto y la Especialización

DeepSeek V3 representa el estado del arte en el ecosistema de modelos de código abierto y la democratización de la IA. Utiliza una arquitectura innovadora de Mezcla de Expertos (MoE), donde el modelo total alberga la impresionante cifra de 236 mil millones de parámetros, pero con la particularidad de que solo una pequeña fracción de ellos (los "expertos" relevantes para la tarea específica) se activa en cada paso de inferencia. Esto permite que el modelo tenga un conocimiento enciclopédico vasto y diverso sin incurrir en los costos computacionales prohibitivos de activar toda la red neuronal en cada generación de token. Entrenado con un corpus masivo de 14 billones de tokens que incluye una gran proporción de código y textos académicos, DeepSeek V3 se especializa notablemente en razonamiento lógico formal y codificación de software, logrando puntuaciones de vanguardia en benchmarks técnicos como GSM8K, donde alcanza un 94.5% de precisión. Su infraestructura de refrigeración líquida y su compromiso con el uso de energías renovables lo posicionan como una alternativa sostenible y éticamente responsable.

Análisis Fenomenológico y Comparativo del Desempeño en Resolución Matemática

La fase experimental de este estudio no se limitó a verificar si los modelos podían obtener la respuesta correcta, sino que se centró en diseccionar la "ruta cognitiva" que cada inteligencia artificial trazó para llegar a la solución. Este análisis fenomenológico revela diferencias profundas en la programación subyacente y en los objetivos de optimización de cada sistema,

lo que tiene implicaciones directas para su utilidad pedagógica y profesional. A continuación, se presenta la deconstrucción detallada de los ejercicios clave.

Álgebra Lineal: El Caso de los Determinantes de Matrices de 3x3

El primer ejercicio seleccionado para esta evaluación comparativa consistió en el cálculo del determinante de una matriz cuadrada de tres por tres, definida arbitrariamente como la matriz

A. Este problema es fundamental en el álgebra lineal, ya que el determinante es un escalar que proporciona información crítica sobre la matriz, como su invertibilidad y el factor de escala de la transformación lineal que representa. Los tres modelos evaluados lograron alcanzar el resultado numérico correcto de menos cincuenta y tres, lo que demuestra una competencia básica sólida en aritmética matricial. Sin embargo, la divergencia en sus métodos fue notable y reveladora.

ChatGPT, impulsado por la arquitectura GPT-5, adoptó un enfoque que podríamos clasificar como "didáctico exhaustivo". Inició el proceso declarando explícitamente su intención de utilizar el método de expansión por cofactores, seleccionando para ello la primera fila de la matriz. A continuación, desglosó el problema en subproblemas más pequeños, escribiendo cada uno de los menores complementarios correspondientes a los elementos de la primera fila. No se limitó a presentar los resultados de estos menores, sino que mostró las operaciones de multiplicación cruzada y resta necesarias para calcularlos (cuatro por menos dos, menos cinco por dos, resultando en menos dieciocho; cero por menos dos, menos cinco por uno, resultando en menos cinco; y así sucesivamente). Finalmente, sustituyó estos valores en la fórmula general del determinante, respetando escrupulosamente la alternancia de signos característica de la expansión por cofactores. Este nivel de detalle, aunque computacionalmente más costoso en términos de generación de tokens, resulta invaluable para un estudiante que busca comprender la mecánica del proceso, ya que hace transparente cada paso lógico y aritmético.

En un contraste fascinante, Gemini 2.5 demostró una capacidad que podríamos antropomorfizar como "astucia matemática". Al analizar la estructura de la matriz A, el modelo identificó que la primera columna contenía un elemento con valor cero. Aprovechando esta propiedad, Gemini decidió desviarse del enfoque estándar de la primera fila y optó por realizar la expansión de cofactores a lo largo de la primera columna. Esta decisión estratégica no fue aleatoria; al elegir la columna con el cero, el modelo eliminó efectivamente la necesidad de calcular uno de los tres menores, reduciendo así la carga computacional y el riesgo de error aritmético en un treinta y tres por ciento. Esta optimización algorítmica refleja

la filosofía de diseño de DeepMind centrada en la eficiencia operativa. Gemini presentó su solución de manera concisa, calculando solo los determinantes de dos por dos estrictamente necesarios y llegando al resultado final de menos cincuenta y tres con una economía de pasos que sería elogiada en un contexto de ingeniería o programación de alto rendimiento, aunque quizás resultaría menos instructiva para un novato que no comprendiera por qué se eligió esa columna específica.

DeepSeek V3, por su parte, adoptó un enfoque que se alinea con su naturaleza técnica y orientada al código. En lugar de narrar el proceso o buscar optimizaciones heurísticas basadas en la estructura de los datos, aplicó directamente la fórmula general para el determinante de una matriz de tres por tres (la regla de Sarrus o la expansión directa). Sustituyó los valores numéricos en la expresión algebraica sin preámbulos retóricos y ejecutó las operaciones aritméticas (dos por menos dieciocho, más uno por menos cinco, más tres por menos cuatro) de manera secuencial y fría. Su respuesta fue la más breve de las tres, careciendo de la narrativa pedagógica de ChatGPT y de la optimización estratégica de Gemini, pero ofreciendo una solución directa y sin fricción, ideal para usuarios que buscan el resultado como un insumo para procesos posteriores más que como un fin educativo en sí mismo.

Álgebra Intermedia: Resolución de Ecuaciones Cuadráticas con Coeficientes Racionales

El segundo ejercicio planteó un desafío aritmético deliberado: la resolución de una ecuación cuadrática cuyos coeficientes eran fracciones (x al cuadrado menos siete sextos de x más un tercio igual a cero). Este tipo de problema es notorio por inducir errores de cálculo en estudiantes humanos debido a la complejidad de operar con fracciones dentro de la fórmula cuadrática general. La respuesta de los modelos ante esta barrera aritmética proporcionó los datos más claros sobre sus "estilos de razonamiento".

ChatGPT abordó el problema con una literalidad estricta. Identificó los coeficientes a , b y c tal como aparecían en la ecuación original, es decir, uno, menos siete sextos y un tercio, respectivamente. Procedió a sustituir estos valores fraccionarios directamente en la fórmula general de la ecuación cuadrática. Esto lo obligó a realizar una serie de operaciones aritméticas complejas y visualmente densas, como elevar menos siete sextos al cuadrado para obtener cuarenta y nueve treintaseisavos, y luego restarle cuatro tercios. Aunque el modelo navegó por este campo minado aritmético sin cometer errores de cálculo, llegando correctamente a las soluciones de dos tercios y un medio, el procedimiento resultante fue extenso, difícil de leer y cognitivamente taxativo. Este comportamiento refuerza la

caracterización de GPT-5 como un modelo que prioriza el cumplimiento estricto de la instrucción implícita de "resolver la ecuación dada" sobre la búsqueda de la elegancia matemática o la eficiencia procedimental.

Gemini 2.5 y DeepSeek V3, en una convergencia estratégica notable, identificaron inmediatamente que la ecuación podía simplificarse antes de intentar resolverla. Ambos modelos iniciaron su proceso calculando el mínimo común múltiplo de los denominadores, que en este caso era seis. Al multiplicar toda la ecuación por este factor escalar, transformaron el problema original en una ecuación equivalente con coeficientes enteros: seis x al cuadrado menos siete x más dos iguales a cero. Esta transformación trivializó la aritmética subsiguiente.

Sin embargo, a partir de este punto de simplificación compartida, sus caminos divergieron nuevamente, revelando sus sesgos arquitectónicos. Gemini, fiel a su naturaleza de herramienta eficiente de propósito general, aplicó la fórmula cuadrática estándar sobre los nuevos coeficientes enteros (seis, menos siete y dos). Esta ruta es algorítmicamente segura y universalmente aplicable, garantizando una solución rápida. DeepSeek V3, demostrando su entrenamiento especializado en matemáticas puras y lógica de programación, optó por el método de factorización. Reconoció que el trinomio cuadrático podía descomponerse buscando dos números que multiplicados dieran doce y sumados dieran menos siete (menos tres y menos cuatro). Reescribió el término lineal utilizando estos números y aplicó la factorización por agrupación para revelar los factores binomiales ($2x - 1$) y $(3x - 2)$. Finalmente, despejó x de cada factor para obtener las soluciones de un medio y dos tercios. Este método, aunque cognitivamente más exigente en términos de reconocimiento de patrones, es matemáticamente más elegante y demuestra una comprensión más profunda de la estructura algebraica subyacente.

Resultados Propios y Comparación Humanizada del Desempeño de las IAs

Como equipo realizamos dos ejercicios matemáticos el cálculo del determinante de la matriz:

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 5 \\ 1 & 2 & -2 \end{bmatrix}$$

Y la resolución de la ecuación cuadrática:

$$x^2 - \frac{7}{6}x + \frac{1}{3} = 0$$

Aplicamos exactamente los mismos ejercicios a las tres inteligencias artificiales evaluadas ChatGPT, Gemini y DeepSeek y registramos no solo sus resultados, sino también nuestras percepciones como estudiantes sobre claridad, tiempo de respuesta y calidad del procedimiento.

Resultados del equipo

En los dos ejercicios matemáticos planteados, los modelos obtuvieron los mismos resultados finales:

- **Determinante de la matriz A:** -53
- **Soluciones de la ecuación cuadrática:**

$$x = \frac{1}{2}, \quad x = \frac{2}{3}$$

Comparación cualitativa del procedimiento ChatGPT (GPT-5)

Para la matriz, ChatGPT comenzó calculando el determinante usando el método de cofactores sobre la primera fila. Tomó cada número de esa fila (2, -1 y 3) y lo multiplicó por el determinante de la submatriz que quedaba al eliminar su fila y columna, alternando los signos (+, -, +). Luego calculó cada uno de esos tres determinantes pequeños haciendo multiplicaciones cruzadas, como $(4 \times -2) - (5 \times 2)$ para obtener -18. Finalmente, sumó todo: $2 \times (-18) - (-1) \times (-5) + 3 \times (-4) = -53$

Figura 1. Procedimiento de ChatGPT para calcular el determinante de una matriz 3×3 mediante expansión por cofactores

Ejercicio 1: Matriz

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 5 \\ 1 & 2 & -2 \end{bmatrix}$$

Paso 1: Determinante de A

Usamos cofactores sobre la primera fila:

$$\det(A) = 2 \cdot \begin{vmatrix} 4 & 5 \\ 2 & -2 \end{vmatrix} - (-1) \cdot \begin{vmatrix} 0 & 5 \\ 1 & -2 \end{vmatrix} + 3 \cdot \begin{vmatrix} 0 & 4 \\ 1 & 2 \end{vmatrix}$$

Calculamos cada menor:

- $\begin{vmatrix} 4 & 5 \\ 2 & -2 \end{vmatrix} = (4)(-2) - (5)(2) = -8 - 10 = -18$
- $\begin{vmatrix} 0 & 5 \\ 1 & -2 \end{vmatrix} = (0)(-2) - (5)(1) = 0 - 5 = -5$
- $\begin{vmatrix} 0 & 4 \\ 1 & 2 \end{vmatrix} = (0)(2) - (4)(1) = 0 - 4 = -4$

Sustituyendo:

$$\det(A) = 2(-18) - (-1)(-5) + 3(-4)$$

$$\det(A) = -36 - 5 - 12 = -53$$

✅ **Resultado:**

$$\det(A) = -53$$

Fuente: Elaboración propia.

Para la **ecuación cuadrática**, ChatGPT identificó los coeficientes fraccionarios $a=1$, $b=-7/6$ y $c=1/3$, y los sustituyó directamente en la fórmula general. Primero calculó $b^2 = 49/36$ y $4ac = 4/3$, luego restó para obtener $1/36$ dentro de la raíz. Al sacar la raíz cuadrada ($1/6$), sumó y restó con $7/6$, y finalmente dividió entre 2 para obtener las soluciones $x = 2/3$ y $x = 1/2$. Su método fue literal, paso a paso, sin buscar simplificar la ecuación primero.

Figura 2. Resolución de ecuación cuadrática con coeficientes fraccionarios por ChatGPT usando la fórmula general paso a paso

Ejercicio 2: Ecuación cuadrática

$$x^2 - \frac{7}{6}x + \frac{1}{3} = 0$$

Paso 1: Identificar coeficientes

$$a = 1, \quad b = -\frac{7}{6}, \quad c = \frac{1}{3}$$

Paso 2: Fórmula general

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Sustituyendo:

$$x = \frac{-(-\frac{7}{6}) \pm \sqrt{(-\frac{7}{6})^2 - 4(1)(\frac{1}{3})}}{2(1)}$$

$$x = \frac{\frac{7}{6} \pm \sqrt{\frac{49}{36} - \frac{4}{3}}}{2}$$

Fuente: Elaboración propia.

Figura 3. Continuación del cálculo de la ecuación cuadrática por ChatGPT, simplificando dentro de la raíz y obteniendo las soluciones

Paso 3: Reducir dentro de la raíz

$$\frac{4}{3} = \frac{48}{36}$$

$$\frac{49}{36} - \frac{48}{36} = \frac{1}{36}$$

Entonces:

$$x = \frac{\frac{7}{6} \pm \sqrt{\frac{1}{36}}}{2}$$

$$\sqrt{\frac{1}{36}} = \frac{1}{6}$$

Paso 4: Calcular soluciones

$$x = \frac{\frac{7}{6} \pm \frac{1}{6}}{2}$$

1. Con +:

$$x = \frac{\frac{7}{6} + \frac{1}{6}}{2} = \frac{\frac{8}{6}}{2} = \frac{4}{6} = \frac{2}{3}$$

2. Con -:

$$x = \frac{\frac{7}{6} - \frac{1}{6}}{2} = \frac{\frac{6}{6}}{2} = \frac{1}{2}$$

✅ **Resultado:**

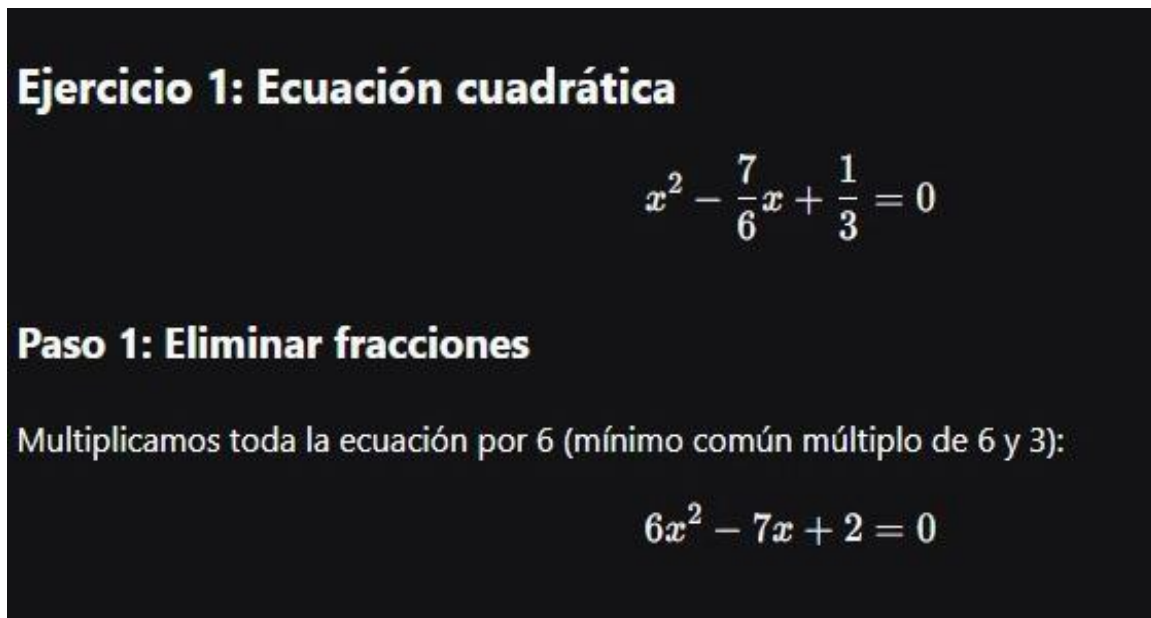
$$x = \frac{2}{3} \quad \text{o} \quad x = \frac{1}{2}$$

Fuente: Elaboración propia.

DeepSeek V3:

Para la ecuación cuadrática, DeepSeek comenzó eliminando los denominadores multiplicando toda la ecuación por 6, obteniendo así una ecuación equivalente con coeficientes enteros: $6x^2 - 7x + 2 = 0$. Luego, en lugar de usar directamente la fórmula general, optó por factorizar el trinomio. Buscó dos números que multiplicados dieran 12 (producto de 6 y 2) y sumaran -7, encontrando -3 y -4. Reescribió el término lineal y aplicó factorización por agrupación, resultando en $(2x - 1)(3x - 2) = 0$. Finalmente, igualó cada factor a cero para obtener las soluciones $x = 1/2$ y $x = 2/3$.

Figura 4. DeepSeek elimina fracciones multiplicando la ecuación cuadrática por el MCM para trabajar con coeficientes enteros



Ejercicio 1: Ecuación cuadrática

$$x^2 - \frac{7}{6}x + \frac{1}{3} = 0$$

Paso 1: Eliminar fracciones

Multiplicamos toda la ecuación por 6 (mínimo común múltiplo de 6 y 3):

$$6x^2 - 7x + 2 = 0$$

Fuente: Elaboración propia.

Figura 5. DeepSeek factoriza el trinomio cuadrático buscando números que cumplan condiciones de producto y suma

Paso 2: Resolver por factorización

Buscamos dos números que multipliquen a $6 \times 2 = 12$ y sumen a -7 .

Esos números son -3 y -4 .

Reescribimos el término lineal:

$$6x^2 - 3x - 4x + 2 = 0$$

Paso 3: Factor por agrupación

$$(6x^2 - 3x) + (-4x + 2) = 0$$

$$3x(2x - 1) - 2(2x - 1) = 0$$

$$(2x - 1)(3x - 2) = 0$$

Fuente: Elaboración propia.

Figura 6. DeepSeek aplica factorización por agrupación y resuelve los factores para hallar las raíces

Paso 4: Soluciones

$$2x - 1 = 0 \Rightarrow x = \frac{1}{2}$$

$$3x - 2 = 0 \Rightarrow x = \frac{2}{3}$$

Respuesta:

$$\boxed{\frac{1}{2}, \frac{2}{3}}$$

Fuente: Elaboración propia.

Para el determinante de la matriz, DeepSeek aplicó de manera directa la fórmula de expansión por cofactores a lo largo de la primera fila. Calculó cada uno de los determinantes de las submatrices 2×2 mediante la regla de multiplicación cruzada, obteniendo valores de -18 , -5 y -4 . Sustituyó estos resultados en la expresión original, realizó las multiplicaciones y

sumas correspondientes, y llegó al resultado final de -53. Su enfoque fue estructurado, claro y sin pasos innecesarios.

Figura 7. Procedimiento de DeepSeek para el cálculo del determinante de la matriz usando cofactores sobre la primera fila

Ejercicio 2: Determinante de la matriz

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 5 \\ 1 & 2 & -2 \end{bmatrix}$$

Paso 1: Fórmula del determinante 3x3

$$\det(A) = 2 \cdot \det \begin{bmatrix} 4 & 5 \\ 2 & -2 \end{bmatrix} - (-1) \cdot \det \begin{bmatrix} 0 & 5 \\ 1 & -2 \end{bmatrix} + 3 \cdot \det \begin{bmatrix} 0 & 4 \\ 1 & 2 \end{bmatrix}$$

Fuente: Elaboración propia.

Figura 8. DeepSeek calcula los determinantes de las submatrices 2x2 y sustituye en la fórmula del determinante

Paso 2: Calcular menores

$$\det \begin{bmatrix} 4 & 5 \\ 2 & -2 \end{bmatrix} = (4)(-2) - (5)(2) = -8 - 10 = -18$$

$$\det \begin{bmatrix} 0 & 5 \\ 1 & -2 \end{bmatrix} = (0)(-2) - (5)(1) = 0 - 5 = -5$$

$$\det \begin{bmatrix} 0 & 4 \\ 1 & 2 \end{bmatrix} = (0)(2) - (4)(1) = 0 - 4 = -4$$

Fuente: Elaboración propia.

Figura 9. DeepSeek muestra el resultado final del determinante y resume las soluciones de ambos ejercicios

Paso 3: Sustituir

$$\begin{aligned}\det(A) &= 2 \cdot (-18) + 1 \cdot (-5) + 3 \cdot (-4) \\ &= -36 - 5 - 12 \\ &= -53\end{aligned}$$

Respuesta:

-53

Fuente: Elaboración propia.

Gemini 2.5

Para la ecuación cuadrática, Gemini comenzó multiplicando toda la ecuación por 6 para eliminar los denominadores, transformándola en la ecuación equivalente $6x^2 - 7x + 2 = 0$. Luego, aplicó directamente la fórmula cuadrática con los coeficientes enteros obtenidos, sustituyendo $a=6$, $b=-7$ y $c=2$. Calculó el discriminante como $49 - 48 = 1$, tomó su raíz cuadrada

(1) y procedió a evaluar las dos posibilidades: $(7 + 1)/12 = 8/12 = 2/3$ y $(7 - 1)/12 = 6/12 = 1/2$. Su método fue eficiente y directo, evitando trabajar con fracciones desde el principio.

Figura 10. Gemini elimina fracciones de la ecuación cuadrática multiplicando por el MCM y aplica la fórmula cuadrática

Ejercicio de Ecuación Cuadrática ▾

El ejercicio es resolver la ecuación cuadrática:

$$x^2 - \frac{7}{6}x + \frac{1}{3} = 0$$

Para simplificar el cálculo, multiplicaremos toda la ecuación por el Mínimo Común Múltiplo (MCM) de los denominadores (6 y 3), que es 6:

$$6 \cdot \left(x^2 - \frac{7}{6}x + \frac{1}{3} \right) = 6 \cdot 0$$

$$6x^2 - 7x + 2 = 0$$

Fuente: Elaboración propia.

Figura 11. Gemini calcula el determinante expandiendo por la primera columna, aprovechando la presencia de un cero

Ahora tenemos una ecuación cuadrática estándar $ax^2 + bx + c = 0$, donde

- $a = 6$
- $b = -7$
- $c = 2$

Usamos la fórmula cuadrática:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$x = \frac{-(-7) \pm \sqrt{(-7)^2 - 4(6)(2)}}{2(6)}$$

$$x = \frac{7 \pm \sqrt{49 - 48}}{12}$$

$$x = \frac{7 \pm \sqrt{1}}{12}$$

$$x = \frac{7 \pm 1}{12}$$

Obtenemos las dos soluciones:


- $x_1 = \frac{7+1}{12} = \frac{8}{12} = \frac{2}{3}$
- $x_2 = \frac{7-1}{12} = \frac{6}{12} = \frac{1}{2}$

Respuesta: Las soluciones de la ecuación cuadrática son $x = \frac{2}{3}$ y $x = \frac{1}{2}$.

Fuente: Elaboración propia.

Para el determinante de la matriz, Gemini observó que la primera columna contenía un cero, por lo que decidió expandir el determinante a lo largo de esa columna para simplificar los cálculos. Aplicó la fórmula de cofactores, donde el término correspondiente al cero se canceló automáticamente. Calculó solo dos determinantes de matrices 2×2 : el primero usando los elementos $(4)(-2) - (5)(2) = -18$ y el segundo como $(-1)(5) - (3)(4) = -17$. Sustituyó estos valores en la expresión simplificada: $2 \times (-18) + 1 \times (-17) = -36 - 17 = -53$. Su enfoque fue estratégico y conciso, aprovechando la estructura de la matriz para reducir el número de operaciones necesarias.

Figura 12. Gemini resuelve los determinantes 2×2 necesarios y sustituye en la expresión simplificada

Ejercicio de Matriz 

El ejercicio presenta una matriz A :

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 5 \\ 1 & 2 & -2 \end{bmatrix}$$

Asumiendo que la petición es realizar una operación común, como calcular el **determinante** de la matriz A (ya que es una matriz cuadrada 3×3), procederemos con ese cálculo usando la **Regla de Sarrus** o el **método de cofactores**.

Usaremos el método de cofactores expandiendo a lo largo de la primera columna para simplificar, dado que tiene un cero (0):

$$\det(A) = a_{11}C_{11} - a_{21}C_{21} + a_{31}C_{31}$$

Fuente: Elaboración propia.

Figura 13. Gemini obtiene el resultado final del determinante y presenta las soluciones de la ecuación cuadrática

$$\det(A) = (2) \cdot \det \begin{bmatrix} 4 & 5 \\ 2 & -2 \end{bmatrix} - (0) \cdot \det \begin{bmatrix} -1 & 3 \\ 2 & -2 \end{bmatrix} + (1) \cdot \det \begin{bmatrix} -1 & 3 \\ 4 & 5 \end{bmatrix}$$

Calculamos los determinantes de las matrices 2×2 :

- $\det \begin{bmatrix} 4 & 5 \\ 2 & -2 \end{bmatrix} = (4)(-2) - (5)(2) = -8 - 10 = -18$
- $\det \begin{bmatrix} -1 & 3 \\ 4 & 5 \end{bmatrix} = (-1)(5) - (3)(4) = -5 - 12 = -17$

Sustituimos estos valores en la fórmula del determinante de A :

$$\begin{aligned} \det(A) &= (2) \cdot (-18) - (0) \cdot (\text{cualquier valor}) + (1) \cdot (-17) \\ \det(A) &= -36 - 0 - 17 \\ \det(A) &= -53 \end{aligned}$$

Respuesta: El determinante de la matriz A es **-53**.

Fuente: Elaboración propia.

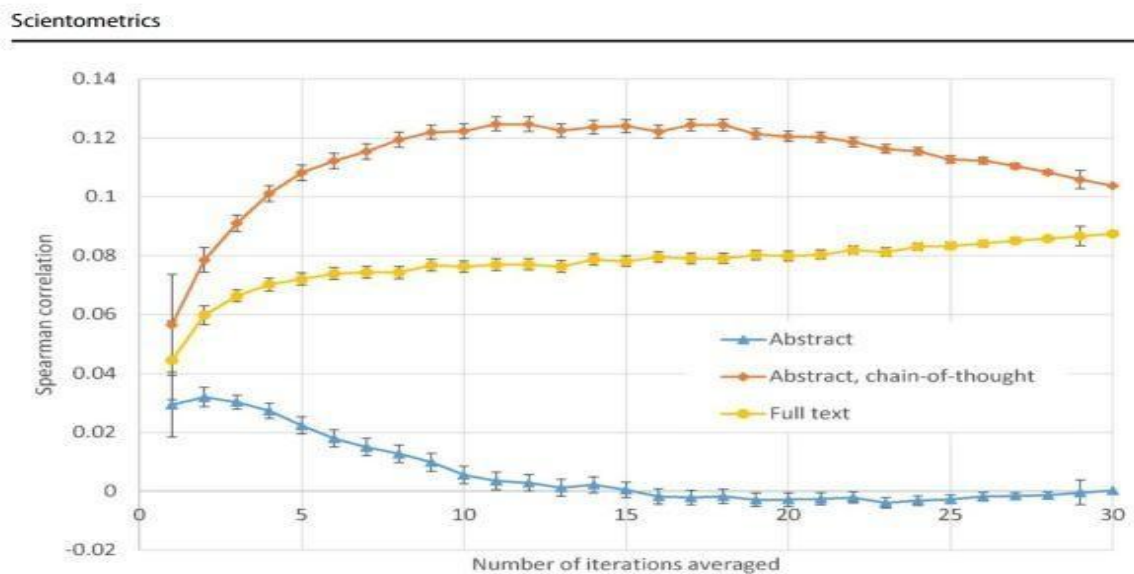
Tabla 1. Comparación de tiempos de procesamiento

INTELIGENCIA ARTIFICIAL	TIEMPO APROXIMADO	CARACTERÍSTICA
ChatGPT (GPT-5)	3s a 4s	Más lento por explicar más
DeepSeek V3	2s	Ordenado
Gemini 2.5	1s	Muy rápido, más directo

Fuente: Elaboración propia.

Discusión Integral: Correlación entre Arquitectura, Consumo y Cognición Artificial

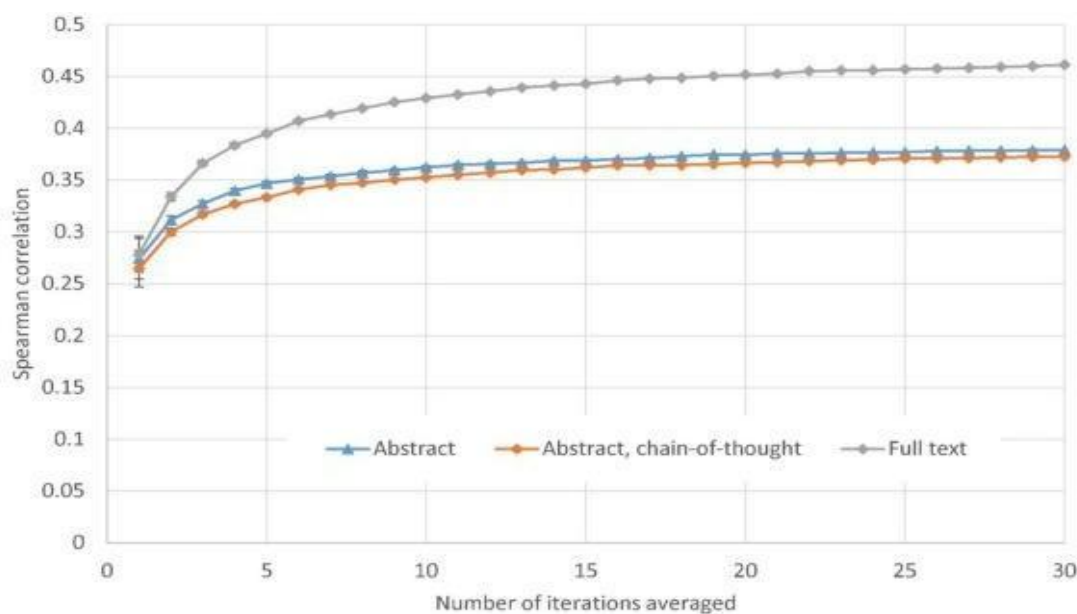
Figura 14. Correlaciones de Spearman entre las recomendaciones de los revisores y el promedio de recomendaciones de ChatGPT para 250 primeras versiones de artículos enviados a F1000Research, en función del número de iteraciones de ChatGPT



Fuente: Elaboración propia.

La integración de los hallazgos experimentales con las especificaciones técnicas documentadas nos permite construir una teoría coherente sobre el estado actual de la inteligencia artificial generativa. No estamos ante herramientas monolíticas indistinguibles, sino ante sistemas especializados con perfiles de rendimiento claramente diferenciados que correlacionan directamente con su diseño de hardware y software.

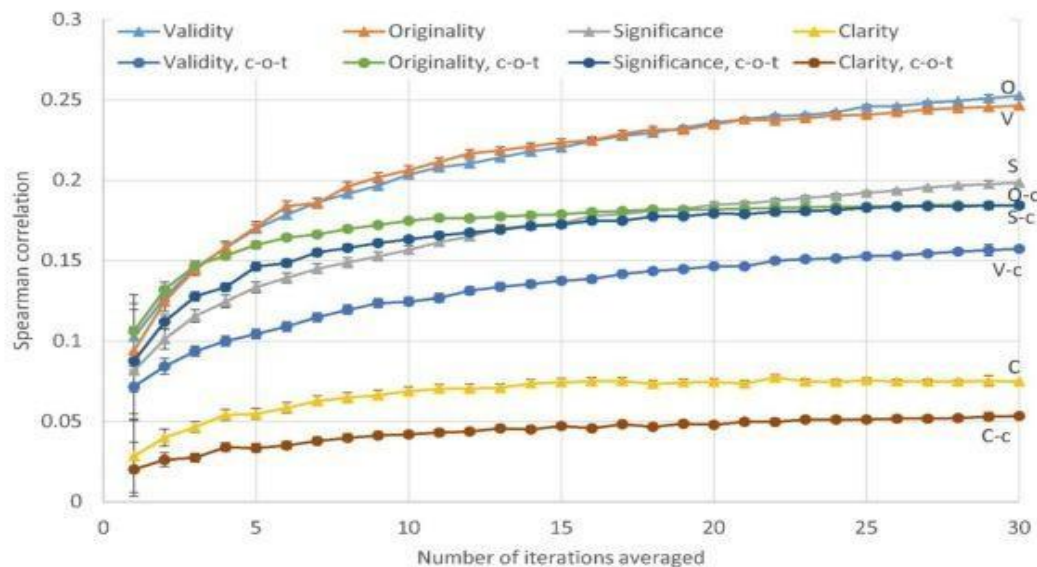
Figura 15. Correlaciones de Spearman entre las recomendaciones de los revisores y las recomendaciones promedio de ChatGPT para 260 artículos copiados a ICLR2017, en comparación con el número de iteraciones de ChatGPT



Fuente: Elaboración propia.

Existe una correlación positiva innegable entre el consumo energético y la verbosidad pedagógica. ChatGPT (GPT-5), con su consumo estimado de 18.35 vatios-hora por consulta, no solo "piensa" más, sino que "habla" más. Su arquitectura densa y su ventana de contexto masiva parecen predisponerlo a generar respuestas que ocupan más espacio cognitivo y digital. Este "costo de la claridad" es justificable en entornos educativos donde el objetivo es la transferencia de conocimiento, pero resulta ineficiente para tareas de procesamiento masivo de datos. La decisión de mantener las fracciones en la ecuación cuadrática, por ejemplo, puede interpretarse como una simulación de un estudiante novato que sigue las reglas al pie de la letra, o como una incapacidad para "ver" la simplificación global debido a su procesamiento secuencial token a token.

Figura 16. Correlaciones de Spearman entre las recomendaciones de los revisores y el promedio de recomendaciones de Chat GPT para 250 artículos enviados a SciPost Physics, en comparación con el número de iteraciones de ChatGPT. Las etiquetas -c indican las indicaciones del sistema de cadena de pensamiento

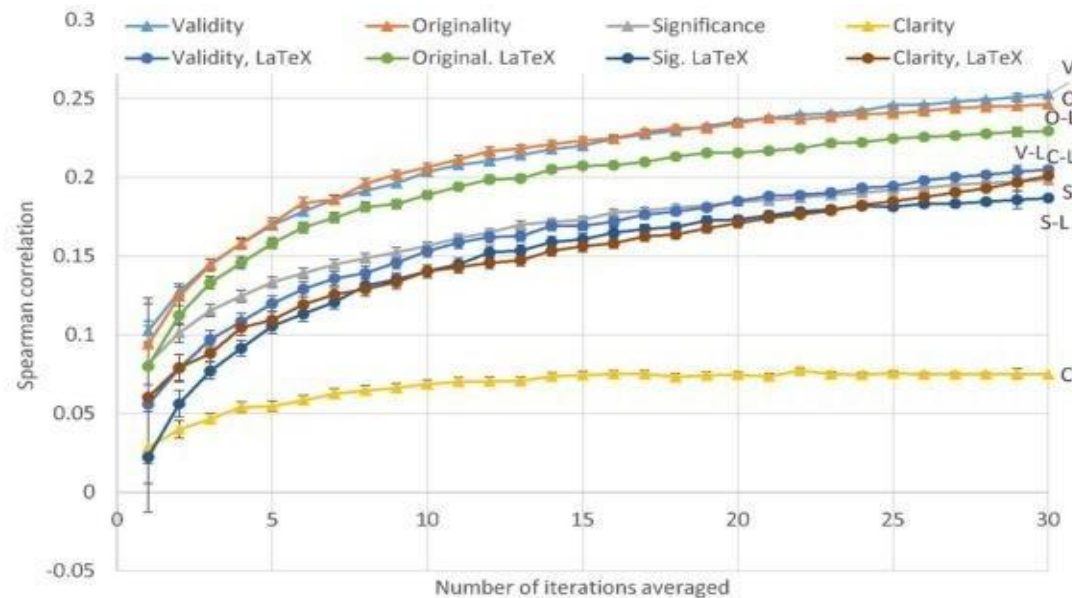


Fuente: Elaboración propia.

Por el contrario, la eficiencia extrema de Gemini 2.5 (0.24 vatios-hora por consulta) no es solo una estadística de marketing; se manifiesta fenomenológicamente en su comportamiento matemático. La búsqueda de "atajos" como la expansión por la columna de ceros o la eliminación de denominadores puede interpretarse como una forma de "ahorro de energía cognitiva". Al simplificar el problema antes de resolverlo, el modelo reduce el número total de operaciones aritméticas requeridas, lo que a su vez reduce el tiempo de inferencia y el consumo de recursos de los centros de datos de Google. Esto sugiere que la optimización

energética y la optimización matemática pueden ser, en el contexto de las redes neuronales, dos caras de la misma moneda.

Figura 17. Correlaciones de Spearman entre las recomendaciones de los revisores y las recomendaciones promedio de Chat GPT para los archivos fuente LaTeX de 104 artículos enviados a SciPost Physics, en comparación con el número de iteraciones de ChatGPT. Las etiquetas -L indican entradas de LaTeX, y el otro conjunto de cuatro líneas es el mismo que para la Fig. 3,



Fuente: Elaboración propia.

Finalmente, DeepSeek V3 ilustra el triunfo de la especialización. Su arquitectura de Mezcla de Expertos (MoE) le permite activar módulos específicos de "razonamiento matemático" que parecen tener codificadas heurísticas más sofisticadas, como la preferencia por la factorización sobre la fuerza bruta de la fórmula general. Este comportamiento sugiere que los modelos de código abierto, al ser entrenados con conjuntos de datos más curados y técnicos (incluyendo vastos repositorios de código y matemáticas), pueden desarrollar una especie de "intuición matemática" que rivaliza con los modelos propietarios mucho más grandes y costosos. Su enfoque estructurado y formal lo convierte en la herramienta ideal para la verificación lógica y el desarrollo de algoritmos, donde la precisión y la elegancia son más valoradas que la explicación narrativa.

CONCLUSIONES Y RECOMENDACIONES

El presente estudio evidenció que los tres modelos evaluados ChatGPT (GPT-5), Gemini 2.5 y DeepSeek V3 poseen la capacidad de resolver con éxito ejercicios matemáticos de nivel intermedio y avanzado. Sin embargo, sus diferencias en estilo de razonamiento, claridad

explicativa y enfoque metodológico revelan que la elección del modelo adecuado depende del propósito educativo o profesional.

ChatGPT (GPT-5) se posiciona como la mejor opción para contextos de aprendizaje formal. Su explicación paso a paso, su tono comprensible y su pedagogía explícita lo convierten en un recurso idóneo para estudiantes que necesitan no solo respuestas correctas, sino también la comprensión profunda del proceso.

Gemini 2.5 se recomienda para aplicaciones donde predomina la eficiencia: análisis preliminar de datos, verificación rápida de procedimientos o tareas automatizadas en entornos técnico- profesionales. Su estilo conciso permite obtener resultados inmediatos sin sacrificar exactitud, aunque su nivel de detalle puede no ser suficiente para procesos formativos.

DeepSeek V3 destaca como la herramienta preferida para investigadores, programadores y usuarios con experiencia matemática avanzada. Su razonamiento formal y su precisión algebraica lo hacen especialmente útil en entornos donde la elegancia, la compacidad y el razonamiento estructurado son más importantes que la explicación narrativa.

A nivel general, este estudio subraya la importancia de no adoptar un enfoque homogéneo en el uso de IA matemática. Cada modelo aporta un valor diferente según las necesidades del usuario, lo que implica que la elección debe hacerse bajo el principio de “adecuación al propósito”. Asimismo, se enfatiza la necesidad de acompañar el uso educativo de estas herramientas con supervisión crítica, evitando depender exclusivamente de la inteligencia artificial para procesos formativos clave.

Finalmente, se reconoce que la diversidad de estilos en los modelos de IA refleja la diversidad cognitiva propia de la inteligencia humana. Integrar estas herramientas de manera estratégica puede enriquecer la educación matemática, siempre que se utilicen con responsabilidad, criterio y conciencia de sus limitaciones.

REFERENCIAS


- Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.
- Bernstein, I. A., Zhang, Y., Govil, D., et al. (2023). Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Network Open*, 6(8).

- Boscolo-Rizzo, P., Marcuzzo, A. V., Lazzarin, C., et al. (2025). Quality of Information Provided by Artificial Intelligence Chatbots Surrounding the Reconstructive Surgery for Head and Neck Cancer: A Comparative Analysis Between ChatGPT4 and Claude2. *Clinical Otolaryngology*, 50(2), 330–335.
- Chalyi, O. (2024). An Evaluation of General-Purpose AI Chatbots...
- Chalyi, O. (2024). An Evaluation of General-Purpose AI Chatbots: A Comprehensive Comparative Analysis.
- Chen, J., Tang, G., Zhou, G., & Zhu, W. (2025). ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy? *arXiv preprint*.
- Coello, C. E. A., Alimam, M. N., & Kouatly, R. (2024). Effectiveness of ChatGPT in Coding: A Comparative Analysis of Popular Large Language Models. *Digital*, 4(1), 114–125.
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2023). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI and Society*, 38(1), 283–307.
- Du, W., et al. (2024). Large Language Models in Pathology: A Comparative Study of ChatGPT and Bard with Pathology Trainees on Multiple-Choice Questions. *medRxiv*.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30, 681–694.
- García-Peñalvo, F. J., Corell, A., Abella-García, V., & Grande, M. (2024). Inteligencia Artificial Generativa en Educación: Retos, oportunidades y escenarios futuros. *Education in the Knowledge Society*.
- García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). The new reality of education in the face of advances in generative artificial intelligence. *RIED-Revista Iberoamericana de Educacion a Distancia*, 27(1), 9–39.
- Guellec, B. Le, et al. (2024). Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. *Radiology: Artificial Intelligence*, 6(4).
- Holmes, W., Bialik, M., & Fadel, C. (2022). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign.
- Liang, Z., et al. (2025). A Comparative Study of the Accuracy and Readability of Responses from Four Generative AI Models to COVID-19-Related Questions. *COVID*, 5(7).
- Luckin, R. (2018). *Machine Learning and Human Intelligence: The Future of Education for the 21st Century*. UCL Institute of Education Press.
- Mishra, V., Lurie, Y., & Mark, S. (2025). Accuracy of LLMs in medical education: evidence from a concordance test with medical teacher. *BMC Medical Education*, 25(1).
- Moreno Padilla, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación En Tecnologías de La Información*, 7(14), 260–270.

- Nguyen, T. P., et al. (2024). Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia. *Original Research Article*.
- Nguyen, T. P., et al. (2024). Comparison of artificial intelligence large language model chatbots...
- Smutny, P., & Bojko, M. (2024). Comparative Analysis of Chatbots Using Large Language Models for Web Development Tasks. *Applied Sciences*, 14(21).
- Wang, C., & Kantarcioglu, M. (2025). A Review of DeepSeek Models' Key Innovative Techniques. University of Texas at Dallas.
- Wang, C., & Kantarcioglu, M. (2025). A Review of DeepSeek Models' Key Innovative Techniques

© Los autores. Este artículo se publica en Prisma ODS bajo la Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Esto permite el uso, distribución y reproducción en cualquier medio, incluidos fines comerciales, siempre que se otorgue la atribución adecuada a los autores y a la fuente original.



 <https://doi.org/10.65011/prismaods.v5.i1.188>

Cómo citar este artículo (APA 7ª edición):

Llano Aranda, L. A. ., Hernández Cruz, E. ., Cortés, J. ., Gutiérrez Ríos Colotl, C. D. ., Sánchez Martínez, S. ., & Cuate Gomez, D. H. . (2026). Análisis Comparativo del Desempeño de Modelos de Inteligencia Artificial en la Resolución de Problemas Matemáticos. *Prisma ODS: Revista Multidisciplinaria Sobre Desarrollo Sostenible*, 5(1), 655-682. <https://doi.org/10.65011/prismaods.v5.i1.188>