

¿CUÁNTA MATEMÁTICA SABE LA IA? UN ESTUDIO ACERCA DE LAS POTENCIALIDADES Y LIMITACIONES DE LOS AGENTES CONVERSACIONALES PARA LA ENSEÑANZA Y APRENDIZAJE DE LA MATEMÁTICA

Pedro A. Willging, Cristian Scarola

Facultad de Ciencias Exactas y Naturales - Universidad Nacional de La Pampa

Av. Uruguay 151, Santa Rosa, La Pampa

pedro@exactas.unlpam.edu.ar, cristianscarola@exactas.unlpam.edu.ar

Categoría del trabajo: Trabajos de investigación.

Nivel educativo: Secundario, terciario, universitario.

Palabras clave: inteligencia artificial, matemática, recurso didáctico.

Resumen: En este trabajo nos proponemos poner a prueba las capacidades matemáticas de la inteligencia artificial generativa (IAG). Mostramos estándares disponibles para evaluar el desempeño de las herramientas que utilizan IAG para resolver problemas matemáticos. Utilizando algunas aplicaciones de IAG en sus versiones gratuitas (los agentes conversacionales ChatGPT, Gemini, Poe, DeepSeek) evaluamos las respuestas que producen antes problemas que pueden presentarse en un curso de nivel medio de matemática. Encontramos que estas aplicaciones han mejorado su precisión en las respuestas si se las compara con lo que respondían hace poco menos de un año, sin embargo algunas aún fallan en cálculos básicos, fundamentación de procedimientos o interpretación de consignas no explícitas.

Introducción

La inteligencia artificial (IA) ha revolucionado múltiples campos del conocimiento, y la matemática no ha sido la excepción. Debido a sus capacidades para procesar grandes volúmenes de datos, identificar patrones complejos y aprender de manera autónoma, la IA ha abierto nuevas posibilidades tanto en la enseñanza como en la investigación matemática (Panqueban y Huincahue, 2024). Aplicaciones basadas en IA, como tutores virtuales, plataformas de aprendizaje adaptativo y asistentes automatizados de resolución de problemas,

tienen el potencial de ofrecer una experiencia educativa más inclusiva y efectiva. En este momento, la IA ofrece herramientas poderosas para automatizar cálculos, verificar demostraciones e incluso generar nuevas conjeturas. La integración de la IA en la educación matemática abre puertas a la innovación pedagógica. Permite, por ejemplo, la evaluación continua y automática del desempeño, la generación de ejercicios personalizados y el análisis de patrones de error que antes pasaban desapercibidos (Bonilla Marín et. al, 2024; González González, 2023).

En nuestra investigación, exploramos algunas aplicaciones de la inteligencia artificial en el ámbito de las matemáticas, su potencial y sus limitaciones. Esperamos encontrar estrategias que permitan convertir esta tecnología en un recurso didáctico para facilitar el aprendizaje. Nos interesa explorar si la IA, en su estado actual, constituye un aliado para quienes enseñan y aprenden matemáticas. Las preguntas que nos guían son: ¿Existen actualmente aplicaciones de IA que respondan preguntas sobre matemáticas con precisión? ¿Cuáles son los problemas que la IA puede resolver correctamente? ¿Cuáles son sus limitaciones?

La IA: ¿herramienta disruptiva para la matemática?

La inteligencia artificial generativa (IAG) se caracteriza como una tecnología que emplea modelos de aprendizaje profundo para crear contenidos que se asemejan a los que generaría un ser humano, incluyendo imágenes y frases completas que responden a preguntas complejas y diversas (Lim et al., 2023). Si bien se ha experimentado con aplicaciones de IA desde hace un tiempo, su popularidad se magnificó en estos últimos años con la aparición de ChatGPT y otros agentes conversacionales. En los ámbitos educativos interesa lo que la irrupción de este fenómeno puede significar, como herramienta de aprendizaje.

Una manera de evaluar la utilidad de la IA para el aprendizaje de las matemáticas, es abordar un problema del currículo escolar y testear las respuestas que darían los profesores en servicio, comparándolas con las de estas aplicaciones. Este es el enfoque adoptado en los trabajos de Corica, Parra, Sureda, Schiaffino, y Godoy (2024, 2025), donde se exploran las respuestas que ChatGPT 3.5, Bard (Gemini), Bing Chat y un profesor de matemática de un colegio secundario realizan cuando se les propone un problema en el que se involucra un fractal. En esta investigación, se encontraron “diferencias significativas entre los procedimientos, representaciones gráficas y validaciones de las respuestas obtenidas de los chatbots y el profesor”, y agregan: “Alentamos el uso de recursos provenientes de la IAG siempre y cuando se analicen críticamente las respuestas que ofrecen” (p.89).

En otro trabajo de las mismas autoras, se analizan las respuestas cuando la IAG se confronta con un problema de geometría euclidiana dentro del marco conceptual del Aprendizaje Basado en Problemas. Los resultados obtenidos muestran que “la IAG ofrece respuestas inconsistentes, poco variadas, con escasas validaciones matemáticas y cuyo sistema de representación predominante es el coloquial” (p. 128).

Otro enfoque, para evaluar o validar una herramienta de IA en el campo de las matemáticas, es el que adopta el grupo de los creadores de MathVista (Lu et al., 2024), que generan un conjunto de test estandarizados, consistentes en múltiples problemas y ejercicios con los que se evalúa la habilidad de una aplicación para su resolución. MathVista fue diseñado para evaluar la capacidad de los modelos de inteligencia artificial de gran escala (LLMs) en el razonamiento matemático en contextos visuales. Esto incluye tareas que no solo requieren entender texto, sino también interpretar figuras, gráficos, diagramas y escenas para resolver problemas matemáticos complejos. MathVista utiliza como ejemplos 6,141 preguntas extraídas y anotadas de 31 conjuntos de datos existentes, más 3 nuevos conjuntos creados especialmente para este estudio. Para este test: a) GPT-4V se destacó en razonamiento algebraico, geometría, estadística y tareas con gráficos y diagramas. Sin embargo, tuvo dificultades en razonamiento lógico y sentido común numérico; b) Los modelos open-source presentan rendimientos mucho más bajos debido a limitaciones en reconocimiento de texto, comprensión visual y razonamiento matemático; c) Se identificaron errores comunes, como cálculos incorrectos, malos razonamientos, y “alucinaciones” (información inventada sin soporte visual); d) GPT-4V mostró capacidad emergente para auto-verificación, revisando sus propias respuestas y corrigiendo errores, mejorando así su precisión. Los autores concluyen que MathVista es un estándar integral para evaluar el razonamiento matemático en contextos complejos y variados, combinando visión y lenguaje y que aunque los últimos modelos multimodales han avanzado mucho, todavía hay importantes desafíos abiertos para igualar o superar la capacidad humana en esta área (Lu et al., 2024).

Metodología y procedimientos

Para llevar adelante la investigación, seleccionamos algunas aplicaciones de IAG gratuitas, en particular los agentes conversacionales ChatGPT, Gemini, y DeepSeek y el chatbot multiplataforma Poe¹. En el caso de la aplicación Poe, ésta permite crear chatbots personalizados para fines específicos. En nuestras evaluaciones utilizamos dos versiones, una

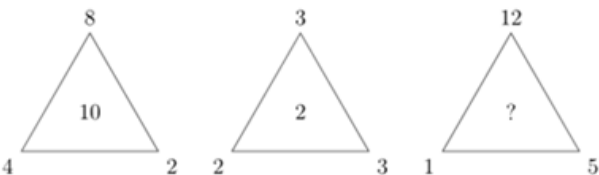
¹ <https://poe.com/>

disponible y ya previamente “entrenada” en resolución de problemas matemáticos y otra que configuramos nosotros, a la que le proporcionamos cuadernillos de precálculo como fuente bibliográfica. A todas estas aplicaciones se les propusieron las mismas situaciones problemáticas, y comparamos sus rendimientos en términos de precisión y riqueza de las respuestas. Para la ingeniería de consultas, utilizamos el método de *Zero-Shot* (Korzynski, Mazurek, Krzypkowska, y Kurasinski, 2023). Exploramos la respuesta de las IAG a preguntas tanto en formato de texto como a través de imágenes, para evaluar también el reconocimiento e interpretación de problemas sin un prompt explícito. Además, esto nos sirvió para saber cómo la IAG reconoce posibles faltas de información, y cómo responde ante dicho faltante. Se hicieron pruebas con problemas tales como: factorización de polinomios, secuencias numéricas según patrones desconocidos, optimización de funciones, reconocimiento de propiedades de funciones a partir de su gráfico, y generación de animaciones con Python. En la próxima sección ilustramos con un caso los resultados obtenidos.

Resultados de nuestras exploraciones

Se muestran los resultados para un problema de secuencia numérica (adaptado de Lu et al, 2024), donde a partir de una imagen con tres triángulos con números en sus vértices, y un número en su centro, para los dos primeros, la consigna es hallar el número faltante en el centro del tercero (ver Figura 1). La resolución es relativamente simple, pues la operación consiste en sumar y restar los vértices correctamente para obtener el número del centro. La consulta fue a través de una imagen y los agentes que lograron interpretarla correctamente fueron Gemini y ChatGPT. Ambos lograron resolver el problema en el primer intento, siendo ChatGPT el que brindó la explicación más completa del mecanismo para reconocer el patrón de la serie. Respecto a los bots de Poe, el bot ya entrenado falló en reconocer alguno de los números en la imagen, y tampoco interpretó que debía utilizar toda la información disponible en la imagen para la consigna, lo que llevó a varios errores de cálculo. A pesar de las múltiples consultas y pistas, no logró una respuesta correcta. El bot personalizado de Poe no interpretó la consigna correctamente en el primer intento, y una vez que lo hizo intentó aplicar propiedades geométricas de los triángulos con información incorrecta. Una vez que explicamos el problema como soporte a la imagen, tuvo numerosos problemas de cálculo básico (sumas y restas), falló al retener información adicional que se le brindó para orientar la resolución y alucinó cometiendo errores para justificar respuestas incorrectas.

El chatbot de DeepSeek tampoco logró interpretar la imagen en la consulta. Ante esto, escribimos en un prompt la consigna del problema, construida por ChatGPT a partir de la imagen utilizada, de manera concisa y clara para consultar nuevamente con DeepSeek. A partir de la consulta en texto, identifica correctamente los datos, analiza distintos patrones posibles y da con el correcto en el primer intento. Además, al solicitarlo, nos brinda un script de Latex para un texto con una explicación paso a paso de la resolución, y un diagrama como soporte visual de la consigna que coincide con la imagen que utilizamos originalmente con los otros modelos.



Pregunta: ¿Cuál es el número que falta en la imagen?

Opciones: (A) 6 (B) 8 (C) 10 (D) 11

Respuesta de DeepSeek: Paso 1: Identificar el patrón Analizamos los dos primeros triángulos para encontrar una relación entre los vértices y el centro. Triángulo 1: Vértices 8, 4, 2 → Centro 10 Probamos combinaciones de suma y resta: $(8 + 4) - 2 = 12 - 2 = 10$ ✓ ¡Coincide! Triángulo 2: Vértices 3, 2, 3 → Centro 2 Aplicamos la misma lógica: $(3 + 2) - 3 = 5 - 3 = 2$ ✓ ¡Coincide! Paso 3: Aplicar al Triángulo 3 Triángulo 3: Vértices 12, 1, 5 → Centro ? La única combinación que produce un valor válido y coincide con las opciones es: $(12 + 1) - 5 = 8$ Respuesta Final El número faltante en el tercer triángulo es 8, correspondiente a la opción (B).	Respuesta de Poe (pre-entrenado): Observa que en cada triángulo, el número interior es la suma de los números inferiores. Triángulo 1: $8 + 4 = 12$ (Hay un pequeño error en la imagen: muestra 10 en lugar de 12. Supongamos que es un error tipográfico y sigamos el patrón). Triángulo 2: $3 + 2 = 5$ (De nuevo, la imagen muestra 2 en lugar de 5. Supongamos que es un error tipográfico y sigamos el patrón). Siguiendo este patrón, para el tercer triángulo: Triángulo 3: $12 + 1 = 13$ Por lo tanto, el número que falta es 13.
---	---

Figura 1: Consigna y respuestas de DeepSeek y Poe (pre-entrenado).

Conclusiones

Además del caso ilustrativo que hemos presentado en este trabajo, realizamos varias pruebas más, incluyendo problemas que abarcan desde conceptos de precálculo, hasta técnicas del análisis en varias variables. En algunas de las consultas, solicitamos que el agente nos brindara soporte visual, logrando resultados satisfactorios en la mayoría de ellos. También exploramos la creación de contenido adaptado a cierto contexto, por ejemplo la explicación de un concepto para un público específico (público general, estudiantes de cierto espacio curricular, etc.), donde los agentes lograron cumplir las expectativas. En base a nuestras

evaluaciones de las aplicaciones, en las situaciones problemáticas a las que las sometimos, podemos obtener algunas conclusiones: a) La diferencia entre los modelos de lenguaje (Gemini, ChatGPT, Deepseek) y las plataformas de metabuscadores de IA (Poe) es considerable en términos de precisión. En estos últimos, a pesar de permitir elegir un modelo base para operar, observamos errores de cálculo básico y recurrencia sobre errores previamente identificados, que llevaron a repetir múltiples consultas, incluso dar una respuesta parcial, sin que ello derive en una respuesta satisfactoria. b) En la “lectura” de problemas a partir de imágenes, dos de los modelos (DeepSeek y Poe) no lograron identificar correctamente números y formas. En concordancia con Lu et al. 2024, comprobamos que unas sólidas capacidades de reconocimiento de texto no garantizan un alto rendimiento. Modelos que identificaron correctamente una consigna y la información necesaria para resolverla en una imagen, no dieron una respuesta satisfactoria de la misma. c) A través de nuestros experimentos y usos de la IAG en matemática a lo largo de estos 3 últimos años, vemos que ha acontecido una gradual mejora en la efectividad y precisión de las respuestas. Hay algunos problemas a los que hoy responde correctamente. Particularmente, cuando se le pide que revise información y explique, genera respuestas adecuadas, pero cuando se requieren cálculos, no es totalmente confiable. d) La IAG es una herramienta que tiene potencial didáctico y puede colaborar con el desarrollo de tareas matemáticas siempre y cuando se la utilice de manera planificada y advirtiendo sus limitaciones.

Referencias

- Bonilla Marín, A., Márquez Díaz, J. E., Benavides Ramírez, L. G., y Gutiérrez Arévalo, F. R. (2024). Inteligencia Artificial Generativa (IAG) en la educación matemática. Encuentro Internacional De Educación En Ingeniería. <https://doi.org/10.26507/paper.3672>
- Corica, A. R., Parra, V., Sureda, P., Schiaffino, S., y Godoy, D. (2024). *Fractal de Koch: análisis de respuestas de IA generativa y un profesor de matemática*, Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, 37, 89-99.
- Corica, A. R., Parra, V., Sureda, P., Schiaffino, S., y Godoy, D. (2025). Análisis y comparación de respuestas aportadas por IA generativa y por profesores de matemática a un problema de geometría euclidiana, Revista Educación Matemática, 37(1), 127-160.
- González González, C. S. (2023). El impacto de la inteligencia artificial en la educación: transformación de la forma de enseñar y de aprender. *Curriculum*, 36, 31–48.
- Korzynski, P., Mazurek, G., Krzypkowska, P. y Kurasinski, A. (2023). La ingeniería de prompts de inteligencia artificial como una nueva competencia digital: Análisis de

tecnologías de IA generativa como ChatGPT. *Entrepreneurial Business and Economics Review*, 11 (3), 25-37. <https://doi.org/10.15678/EBER.2023.110302>

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., y Pechenkina, E. (2023).

Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2), 100790. <https://doi.org/10.1016/j.ijme.2023.100790>

Lu, P. et al. (2023). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*. <https://arxiv.org/pdf/2310.02255>

Panqueban, D., y Huincahue, J. (2024). Inteligencia artificial en educación matemática: Una revisión sistemática. *Revista Uniciencia*, 38(1), 1–22.