FORMACIÓN DEL PROFESORADO E INVESTIGACIÓN EN EDUCACIÓN MATEMÁTICA
Vol. XVI ♦ http://wp.ull.es/fpiem ♦ ISSN: 1695-6613 (p) y 1885-2912 (d)

EVALUACIÓN DEL POTENCIAL DE DISTINTOS CHATBOTS PARA LA FORMULACIÓN DE PROBLEMAS MATEMÁTICOS

Sara Embid (sembidso@ull.edu.es)

Josefa Perdomo-Díaz (jperdom@ull.edu.es)

Universidad de La Laguna

Resumen

En este artículo, analizamos el potencial de cuatro chatbots (ChatGPT, Claude, Copilot y Gemini) para formular problemas matemáticos. Realizamos un estudio cualitativo, de carácter exploratorio, que examina 60 problemas generados tras aplicar la técnica "zero-shot prompting". En concreto, solicitamos formular tres problemas de distinta dificultad, en los que aparezcan las fracciones 1/4 y 3/8 repitiendo la consigna cinco veces a cada chatbot. Evaluamos si los problemas cumplen con los criterios del prompt, si contienen toda la información necesaria para ser resueltos, y si requieren más de una tarea matemática para su resolución. Por último, discutimos los beneficios y limitaciones del uso de este tipo de herramientas digitales, así como sus implicaciones para la formación de docentes de matemáticas.

Palabras clave: Inteligencia artificial, Formulación de problemas, Educación matemática

Abstract

In this article, we analyze the potential of four chatbots (ChatGPT, Claude, Copilot, and Gemini) to formulate mathematical problems. We conducted a qualitative, exploratory

study that examines 60 problems generated after applying the "zero-shot prompting" technique. Specifically, we requested the formulation of three problems of varying difficulty, including the fractions 1/4 and 3/8, repeating the prompt five times with each chatbot. We evaluate whether the problems meet the prompt's criteria, whether they contain all the necessary information to be solved, and whether they require more than one mathematical task to be solved. Finally, we discuss the benefits and limitations of using this type of digital tool, as well as its implications for the training of mathematics teachers.

Keywords: Artificial Intelligence, Problem posing, Mathematics education

Introducción

En los últimos años, hemos sido testigos de la democratización del acceso a herramientas que, a diferencia de los libros de texto, proporcionan un alto grado de interacción y personalización. Se trata de asistentes conversacionales o chatbots que utilizan tecnologías basadas en el análisis de grandes cantidades de datos lingüísticos para generar texto con una fluidez y coherencia comparables al lenguaje humano (Čavojský et al., 2023).

Estudios recientes han evidenciado distintos usos y tendencias educativas de los chatbots (Kuhail et al., 2023). Sin embargo, resulta llamativo que, a pesar de que la formulación de problemas es esencial en la educación matemática (Kilpatrick, 1987), la investigación más actual sobre chatbots se reduce prácticamente a la resolución de problemas (Noster et al., 2024; Parra et al., 2024; Schorcht et al., 2024). Analizar, seleccionar, adaptar o inventar actividades matemáticas que ofrezcan oportunidades para desarrollar el sentido matemático de los estudiantes es una parte importante del trabajo de los docentes, quienes a menudo recurren a libros de texto para ello (Son y Diletti, 2017). En algunos casos, utilizan las actividades matemáticas tal y como aparecen formuladas en dichos libros, lo que muchas veces se traduce en el uso de ejercicios que presentan una única respuesta correcta, un único método para alcanzarla, y que son rutinarios, conteniendo la información exacta que se necesita para

resolverlos (Leavy y Hourigan, 2022). En otros casos, los docentes reformulan las actividades que encuentran, las adaptan o crean otras completamente nuevas. No obstante, este proceso de formulación de problemas no siempre resulta exitoso. Diversos estudios muestran que, tanto docentes en formación como docentes en ejercicio, formulan problemas que son irrelevantes o incorrectos. Además, en muchas ocasiones, las modificaciones que realizan a los problemas presentados en los libros de texto son mínimas (Embid et al., 2023; Koichu et al., 2013; Singer y Voica, 2013; Yao et al., 2021).

En este trabajo, nos centramos en una tarea de formulación de problemas que ya ha sido explorada en estudios previos con futuros docentes de educación primaria (Embid et al., 2023; Embid y Perdomo-Díaz, 2024; García-Alonso et al. 2022, Sosa-Martin et al., 2024). Estos estudios analizan diversas características de los problemas formulados, identifican errores conceptuales y comparan la formulación con lápiz y papel con el uso de herramientas tecnológicas específicas para el tratamiento de fracciones. En esta ocasión, el foco está en el análisis de herramientas tecnológicas no específicas para la educación matemática, como los chatbots. Así, el objetivo general de esta investigación es analizar el potencial de diferentes chatbots en la formulación de problemas matemáticos.

Formulación de problemas matemáticos

En el ámbito de la educación matemática, la formulación de problemas hace referencia a la reformulación, modificación o adaptación de una actividad o problema dado y también a la invención o creación de problemas, sin una actividad de referencia (Silver, 1994). Como actividad matemática, la formulación de problemas permite al individuo reflexionar sobre sus propios conocimientos y profundizar en su comprensión conceptual (Santos, 2001), plantear situaciones o preguntas que pueden ser relevantes a nivel particular o social (Crespo, 2015) y desarrollar su espíritu crítico y creativo (Ayllón y Gómez, 2014).

Desde el punto de vista del docente, la formulación de problemas cumple dos roles. Por un lado, puede ser utilizada como estrategia metodológica, lo que ofrece la oportunidad de evaluar el conocimiento de los estudiantes y detectar errores conceptuales o dificultades de aprendizaje (Embid et al., 2023; Tichá y Hospesová, 2013). Por otro lado, modificar, adaptar o inventar

problemas es una de las múltiples tareas que debe realizar el docente. Por tanto, para los docentes, la formulación de problemas es una práctica matemática, una estrategia metodológica y también una práctica profesional.

La resolución y formulación de problemas son actividades matemáticas estrechamente ligadas entre sí, aunque con distinto foco. Mientras que en la resolución de problemas el énfasis se pone en la búsqueda de una estrategia que conduzca a una solución del problema, el foco principal de la formulación de problemas está en el propio problema (Leavy y Hourigan, 2022). Esto hace que un aspecto clave a considerar en relación con esta práctica matemática sea la calidad de los problemas formulados.

La literatura recoge una amplia gama de características deseables para los problemas formulados, que no necesariamente coinciden en todos los trabajos. Leavy y Hourigan (2022) proponen ocho indicadores a observar para determinar la calidad de un problema matemático. Estos indicadores incluyen el uso de contextos motivadores y culturalmente relevantes, la claridad del lenguaje, la coherencia curricular, la disponibilidad de diversas estrategias de resolución, el número de pasos requeridos para resolver, la cantidad de soluciones posibles, la demanda cognitiva y las oportunidades de éxito ofrecidas a los resolutores. En relación con el contexto, para Hiebert (1997), formular "buenos problemas" requiere diseñar preguntas que los estudiantes consideren retos interesantes para involucrarse en su resolución. Otros autores como Cankoy y Özder (2017) consideran deseable el uso de un contexto no rutinario. Esto implica presentar el problema de una forma distinta a las utilizadas normalmente en clase o con una estructura diferente a las presentes en las actividades de los libros de texto.

Los estudios de Cankoy (2014) y Cankoy y Özder (2017) abordan el concepto de razonabilidad en la formulación de problemas, enfatizando que tanto la información presentada en el enunciado como su solución deben ser lógicas y coherentes con situaciones de la vida real. Además, subrayan que la información proporcionada debe ser suficiente para permitir la resolución del problema. Este último aspecto es considerado también por Grundmeier (2015) quien distingue entre: (i) problemas no plausibles, que contienen algún error matemático, (ii) problemas plausibles a los que les falta alguna información para poder ser resueltos, (iii)

problemas plausibles, con información suficiente para ser resueltos y en cuya resolución interviene un único tipo de tarea matemática y (iv) problemas plausibles, con información suficiente para ser resueltos, en cuya resolución intervienen más de un único tipo de tarea matemática. Otras características a observar en los problemas formulados son la estructura matemática de los mismos (Cankoy, 2014; García-Alonso et al., 2022; Sosa-Martín et al., 2024) o los significados asociados a los conceptos matemáticos involucrados en el problema (García-Alonso et al., 2022; Sosa-Martín et al., 2024).

Las actividades de formulación de problemas pueden adoptar estructuras diversas e incluir distintas variables, lo que influye en las características de los problemas formulados (Cai y Leikin, 2020; Embid et al., 2023; Sosa-Martín et al., 2024; Zhang et al., 2022). En particular, el uso de herramientas digitales, como software de geometría dinámica, hojas de cálculo, entornos de programación avanzada y plataformas de aprendizaje virtual, contribuye a la mejora del diseño de problemas (Abramovich y Cho, 2015; Barana et al., 2020; Christou et al., 2005; Pochulu, 2010). Puesto que formular problemas es esencialmente una actividad discursiva, nos preguntamos de qué forma, herramientas tecnológicas como los asistentes conversacionales o chatbots, pueden contribuir a la creación de problemas matemáticos.

Inteligencia artificial y chatbots

La inteligencia artificial (IA), entendida como la capacidad de máquinas o software para realizar tareas que requieren inteligencia humana, ha sido objeto de estudio durante más de medio siglo. Sin embargo, en los últimos años ha ganado una popularidad notable, especialmente tras el lanzamiento de ChatGPT por OpenAI en 2022 (Parra et al., 2024). Desde la aparición del término "inteligencia artificial" en la década de 1950, la IA ha atravesado diversas etapas de desarrollo, evolucionando de modelos simples a sistemas complejos (Crawford et al., 2023).

Uno de los avances más significativos en el campo de la inteligencia artificial ha sido el desarrollo de modelos de lenguaje a gran escala (en inglés, pre-trained Large Language Models o LLM). Estos modelos emplean técnicas avanzadas de aprendizaje profundo, como redes neuronales, para realizar predicciones a partir de grandes volúmenes de datos textuales (Čavojský et al., 2023). Los asistentes conversacionales o chatbots aprovechan la capacidad

de procesamiento de los LLM para imitar el lenguaje humano, facilitando a los usuarios la realización de consultas, la recepción de información y el mantenimiento de conversaciones de manera intuitiva (Kuhail et al., 2023). La comunicación entre el usuario y el chatbot se realiza mediante el uso de prompts. En los resultados de dicha interacción se produce una considerable variabilidad puesto que las respuestas generadas por cada chatbot dependen de los prompts que reciben y también del LLM asociado a dicho chatbot, ya que la cantidad y calidad de los datos de entrenamiento afectan directamente a la calidad de las respuestas generadas y la adecuación al contenido (Schorcht et al., 2024).

En el ámbito de la enseñanza y el aprendizaje de las matemáticas, las investigaciones sobre el uso de chatbots son muy limitadas. Autores como Plevris et al. (enviado) realizan una comparativa de tres chatbots basados en modelos de lenguaje a gran escala -ChatGPT-3.5, ChatGPT-4 y Google Bard - centrándose en su capacidad para dar respuestas correctas a problemas de matemáticas y lógica. Por otro lado, Schorcht et al. (2024) diseñan un estudio contrastando diferentes técnicas de ingeniería del prompt para la resolución de problemas. Estas técnicas incluyen: (i) el Zero-Shot prompting o sin información adicional, (ii) Few-Shot o entrega de algunos ejemplos, (iii) Chain-of-Thought o cuestionamiento sobre el proceso de razonamiento seguido por la IA y (iv) Ask-me-Anything o petición de preguntas necesarias para que la IA dé respuestas mejor acotadas.

Objetivos del estudio

Esta investigación tiene dos objetivos específicos:

1. Analizar la plausibilidad, suficiencia de datos y cantidad de tipos de tareas matemáticas involucradas en la resolución de los problemas generados por un conjunto de asistentes conversacionales o chatbots.
2. Estudiar los cambios en los problemas formulados por los chatbots seleccionados en distintas iteraciones, en relación con las características de los problemas consideradas en el primer objetivo.

Metodología

La investigación se realiza con un enfoque cualitativo, empleando el análisis de contenido como técnica principal, y con un marcado carácter exploratorio (Creswell, 2012).

Los datos se obtuvieron en septiembre de 2024 a partir de la interacción con cuatro chatbots: GPT-4, Claude 3.5 Sonnet, Copilot (modo preciso) y Gemini. A cada uno de estos chatbots se le planteó la siguiente tarea de formulación de problemas basada en Kiliç (2015):

> "Formula tres problemas, de diferente dificultad, en los que aparezcan los números 1/4 y 3/8. Cada uno de estos números puede ser un dato o una solución. Recuerda que puedes añadir cualquier tipo de información (numérica, de contexto...)."

Se utilizó la técnica de "zero-shot prompting", en la cual se presenta la consigna sin proporcionar información adicional (Schorcht et al., 2024). Esta técnica fue aplicada cinco veces a cada uno de los cuatro chatbots, manteniendo la misma consigna en cada iteración. Como resultado, se generó un conjunto de datos compuesto por 60 problemas en total: 15 problemas obtenidos de cada uno de los cuatro chatbots seleccionados para la investigación.

El proceso de análisis se realizó en tres fases. En las dos primeras fases, el análisis se hizo para cada problema individual formulado por cada uno de los chatbots. En la última fase, se realizó un análisis conjunto de los tres problemas formulados en cada iteración.

La primera fase consistió en verificar el cumplimiento de las condiciones indicadas en la actividad de formulación de problemas, que exigían la inclusión de las fracciones 1/4 y 3/8 en cada uno de los problemas propuestos. Los problemas que no cumplían con esta condición fueron codificados como NA (no ajusta). Solo aquellos problemas que cumplieron con este criterio avanzaron a la siguiente fase de análisis.

En la segunda fase, para aquellos problemas que sí cumplían con las condiciones dadas, se analizaron tres de las características que intervienen en la calidad de los problemas: su plausibilidad, la información que se proporciona en el enunciado y el número de tareas matemáticas necesarias para su resolución. Cada problema se codificó atendiendo a las cuatro categorías propuestas por Grundmeier (2015):

NP: Problema no plausible. Son aquellos problemas que contienen algún error matemático. Es importante distinguir entre los problemas no plausibles y los problemas sin solución.

P1: Problema plausible con información incompleta pero comprensible. Son aquellos problemas que no contienen ningún error matemático, pero a los que les falta algún tipo de información para poder ser resueltos.

P2: Problema plausible, con información suficiente, que requiere de un único tipo de tarea matemática para ser resuelto.

P3: Problema plausible, con información suficiente, que requiere de más de un tipo de tarea matemática para ser resuelto.

Finalmente, en una tercera fase, se estableció un criterio para determinar el grado de éxito de cada chatbot para cada iteración. Se consideró como iteraciones exitosas aquellas que generan tres problemas que cumplen con las indicaciones de la tarea de formulación de problemas, no contienen errores matemáticos y sus enunciados incluyen toda la información necesaria para que el problema pueda resolverse. Así, para que una iteración fuera considerada exitosa, los tres problemas formulados debían ser de tipo P2 o P3.

Análisis de datos

Un primer análisis global, de los 15 problemas propuestos por cada chatbot en las 5 iteraciones, reveló diferencias notables entre los distintos sistemas. Un resultado positivo es que ninguno de los chatbots estudiados formuló problemas que tuvieran errores matemáticos (Tabla 1). Sin embargo, dos de los chatbots examinados—Copilot y Gemini—generaron problemas que no cumplían con los criterios específicos de incluir las fracciones 1/4 y 3/8. Para Copilot, esto sucedió en tres de cinco iteraciones y para Gemini en cuatro (Tabla 6).

En el caso de Copilot, en dos de los problemas no aparece la fracción 1/4, ni como dato, ni como solución del problema (Tabla 2); en otros tres de los problemas que formula no aparece la fracción 3/8, aunque en dos de ellos sí aparece la fracción 1/4. En el caso de Gemini, cuatro

Tabla 1. Número de problemas en cada categoría, para cada chatbot.

Chatbot	NA	NP	P1	P2	P3
ChatGPT	0	0	4	9	2
Claude	0	0	0	8	7
Copilot	5	0	0	7	3
Gemini	4	0	1	10	0

de los problemas formulados no se ajustan porque no incluyen la fracción 1/4, a lo que hay que añadir que tampoco hacen referencia a la igualdad de las partes en que se divide la unidad, aspecto clave para el uso de fracciones. Además, Gemini devuelve un mismo problema en distintas iteraciones.

El análisis también mostró que, en 5 de los 60 problemas, faltaba algún tipo de información para poder ser resueltos (P1). Esos problemas fueron generados por dos de los chatbots: ChatGPT y Gemini (Tabla 1). ChatGPT es la herramienta que genera más problemas de este tipo. Los cuatro problemas de este tipo que formula ChatGPT (Tabla 3) piden calcular la suma de dos cantidades que son fracción de una unidad, sin embargo no se explicita que esas unidades sean equivalentes por lo que, a priori, no sería posible sumar esas fracciones. ChatGPT plantea este tipo de problema en cuatro de las cinco iteraciones. De manera similar, esta misma situación se observa en uno de los tres problemas formulados por Gemini en la iteración 2.

Tabla 2. Ejemplos de problemas formulados que no se ajustan a lo solicitado (NA).

Chatbot	Problemas formulados (# Iteración)
Copilot	● *Tienes una tableta de chocolate dividida en 8 trozos iguales. Si te comes 3 trozos, ¿qué fracción de la tableta te has comido?* (#1) ● *Sara tiene una cinta que mide 1/4 de metro. La corta en trozos de 1/8 de metro cada uno. ¿Cuántos trozos tiene?* (#1) ● *Sara tiene una tableta de chocolate que quiere repartir a partes iguales entre sus tres amigas. Si da a cada amiga 1/4 de la tableta, ¿cuánto le queda?* (#3)
Gemini	● *Una pizza se corta en 8 porciones. Si te comes 3 trozos, ¿qué fracción de pizza queda? (Pista: Piensa en cuántas porciones sobran del total).* (#1, #4) ● *Una pizza se corta en 8 porciones. Si te comes 3/8 de la pizza, ¿cuántas porciones te has comido?* (#3, #5)

Tabla 3. Ejemplos de problemas formulados con falta de información (P1).

Chatbot	Problemas formulados (# Iteración)
ChatGPT	● *Emma tiene 1/4 de una tableta de chocolate, y su amiga le da 3/8 más de una tableta de chocolate. ¿Cuánto chocolate tiene ahora Emma?* (#1) ● *Emma tiene 1/4 de una tableta de chocolate y se come 3/8 de otra tableta. ¿Cuánto chocolate se ha comido en total?* (#3) ● *Tú tienes 1/4 de pizza y tu amigo tiene 3/8 de pizza. ¿Cuánta pizza tenéis los dos en total?* (#4) ● *Te comes 1/4 de pizza y tu amigo te regala otros 3/8 de pizza. ¿Cuánta pizza te has comido en total?* (#5)
Gemini	● *Si tú tienes 1/4 de pizza y tu amigo tiene 3/8 de pizza, ¿cuánta pizza tenéis juntos?* (#2)

Tres cuartas partes de los problemas generados por los cuatro chatbots analizados cumplieron con los requisitos de la actividad, al incluir las fracciones 1/4 y 3/8 en el enunciado o la solución, y proporcionar toda la información necesaria para ser resueltos. La mayor parte de ellos, 34 problemas en total, puede resolverse empleando una única tarea matemática (codificados como P2). Los cuatro chatbots generaron una cantidad similar de este tipo de problemas (Tabla 1).

Tabla 4. Ejemplos de problemas formulados plausibles y de una sola tarea matemática (P2).

Chatbot	Problemas formulados (# Iteración)
ChatGPT	● *Un coche ha recorrido 1/4 de su trayecto en 45 minutos y 3/8 del trayecto en 1 hora. Si el trayecto restante dura 1,5 horas a velocidad constante, ¿qué fracción del trayecto total se completó en el tiempo restante?* (#1) ● *Samantha está cortando una cinta en trozos más pequeños. Corta la cinta de forma que un trozo mide 1/4 de la longitud total y otro trozo mide 3/8 de la longitud total. ¿Qué fracción de la cinta ha utilizado en total?* (#2)
Claude	● *En un experimento científico, se crea una solución mezclando dos sustancias químicas. La primera sustancia química representa 1/4 del volumen total. Después de añadir la segunda sustancia química, la concentración de la primera disminuye a 3/8 de su concentración original. ¿Qué fracción de la solución final corresponde a la primera sustancia química?* (#2)
Copilot	● *Una receta requiere 3/8 de taza de azúcar. Si sólo tienes un vaso medidor de 1/4 de taza, ¿cuántas veces tienes que llenar el vaso de 1/4 de taza para obtener la cantidad exacta de azúcar necesaria para la receta?* (#1)
Gemini	● *En una receta se pide 1/4 de taza de azúcar y 3/8 de taza de harina. ¿Cuánta más harina se necesita que azúcar?* (#3)

En algunos casos, el chatbot formula problemas que contienen información que no es necesario utilizar, como ocurre con uno de los problemas generados por ChatGPT en la primera iteración (Tabla 4). En otros casos, se han presentado problemas que, aunque no presentan ningún error matemático en su formulación, son problemas sin solución, cuya respuesta es del tipo "no se puede hacer" o "no es posible". Pero la mayoría de los problemas formulados tienen solución, es única, se puede obtener empleando todos los datos presentes en el enunciado y utilizando únicamente un tipo de tarea matemática.

Finalmente, la quinta parte de los problemas analizados fueron codificados como P3, lo que significa que cumplen con los requisitos de la actividad de formulación de problemas planteada, contienen la información necesaria para su resolución y requieren más de un tipo de tarea matemática para ser resueltos (Tabla 5). La mayor parte de los problemas de este tipo fueron generados por Claude (Tabla 1), seguido de Copilot y ChatGPT. Por ejemplo, para resolver las dos preguntas planteadas en el problema generado por ChatGPT, primero es necesario calcular la suma de 1/4 y 3/8, y luego operar esa suma sobre 16. La primera operación es de naturaleza

Tabla 5. Ejemplos de problemas formulados plausibles, de más de una tarea matemática (P3).

Chatbot	Problemas formulados (# Iteración)
ChatGPT	● *Un recipiente está 1/4 lleno de agua. Si viertes en él 3/8 del volumen total de agua del recipiente, ¿qué fracción del recipiente estará ahora llena de agua? Si el volumen total del recipiente es de 16 litros, ¿cuánta agua hay ahora en el recipiente?* (#2)
Claude	● *Sara tiene una tableta de chocolate. Se come 1/4 de ella y luego le da 3/8 de la barra restante a su amiga. ¿Cuánto queda de la chocolatina original?* (#2) ● *En una reacción química, 1/4 de la sustancia A se convierte en sustancia B cada 10 minutos. Si la reacción continúa a este ritmo, ¿cuánto tardarán 3/8 de la cantidad original de sustancia A en convertirse en sustancia B?* (#4)
Copilot	● *Estás mezclando dos soluciones para un experimento científico. La solución A contiene 1/4 de una determinada sustancia química y la solución B contiene 3/8 de la misma sustancia química. Si mezclas 1 litro de solución A con 2 litros de solución B, ¿qué fracción de la mezcla total es la sustancia química?* (#2) ● *Un depósito está 1/4 lleno de agua. Después de añadir 15 litros de agua, el depósito está (3/8) lleno. ¿Cuál es la capacidad total del depósito?* (#3)

aditiva, mientras que la segunda es multiplicativa. Gemini es el único chatbot que no generó problemas con estas características.

Para atender al segundo objetivo de la investigación, se realizó un análisis más detallado de cada una de las cinco iteraciones, para cada uno de los cuatro chatbots considerados. La tabla 6 muestra la codificación de los tres problemas formulados por cada chatbot en cada iteración. Cada terna se presenta en orden ascendente según la dificultad de los problemas generados. El primer término indica la codificación del problema que el chatbot consideró más sencillo, seguida de la codificación del problema de dificultad media, y finalmente, la del problema de mayor dificultad.

Tabla 6. Codificación de cada problema para cada chatbot e iteración.

Chatbot	**Iteración**				
	#1	**#2**	**#3**	**#4**	**#5**
ChatGPT	(P1,P2,P2)	(P2,P2,P3)	(P1,P2,P2)	(P1,P2,P2)	(P1,P3,P2)
Claude	(P2, P3, P2)	(P3,P2,P2)	(P3, P2, P3)	(P2, P2, P3)	(P3, P2, P3)
Copilot	(NA,NA,P2)	(NA,NA,P3)	(NA,P2,P3)	(P2,P2,P3)	(P2,P2,P2)
Gemini	(P2,NA,P2)	(P1,P2,P2)	(P2,NA,P2)	(P2,NA,P2)	(P2,NA,P2)

En un primer análisis de los datos de la tabla puede observarse en cuántas iteraciones tuvo éxito cada chatbot, es decir, en cuántas iteraciones los tres problemas generados por el sistema fueron de tipo P2 o P3. Los resultados muestran que Gemini no tuvo éxito en ninguna de las cinco iteraciones ya que, en cuatro de ellas generó un problema que no se ajustaba a los requerimientos dados y en otra formuló un problema al que le faltaba información (Tabla 3). ChatGPT solo tuvo éxito en la segunda iteración, lo que indica que este tipo de herramientas no necesariamente produce mejores resultados a medida que se incrementa el número de iteraciones. Copilot tuvo éxito en las dos últimas iteraciones y Claude fue el único de los cuatro chatbots que, en las cinco iteraciones, formuló problemas que se ajustaban a lo solicitado, sin errores matemáticos y con información suficiente para ser resueltos.

La figura 1 muestra la evolución de cada chatbots a lo largo de las cinco iteraciones y permite comparar visualmente la calidad de los problemas propuestos por cada uno de los cuatro

sistemas, en función de las características consideradas: ajuste a lo solicitado, suficiencia de información y número de tareas diferentes necesarias para la resolución.

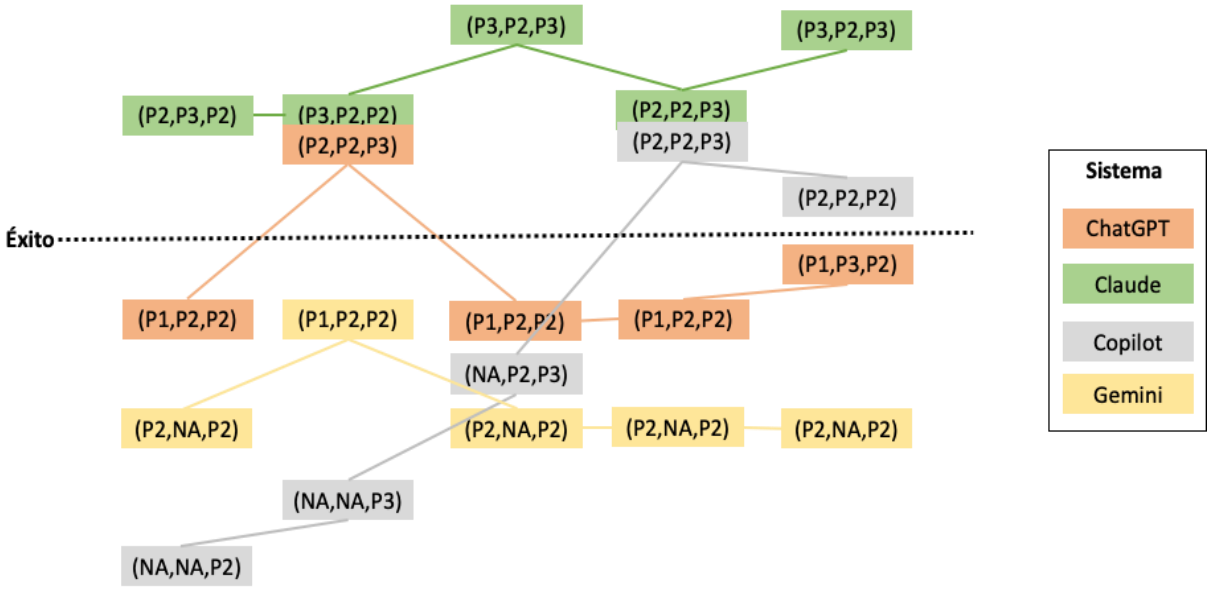

Figura 1. Comportamiento de cada sistema en cada iteración (Elaboración propia).

Discusión final y conclusiones

En relación con el objetivo general de la investigación, los resultados de este estudio evidencian diferencias importantes en el potencial de los distintos chatbots para formular problemas matemáticos.

En el análisis global del ajuste de los problemas generados por los chatbots a los requerimientos de la actividad de formulación de problemas (inclusión de las fracciones ¼ y ⅜), la inexistencia de errores o incoherencias matemáticas, la suficiencia de la información proporcionada y el número de tareas matemáticas involucradas en la resolución (primer objetivo específico), Claude destacó como el sistema más competente, produciendo únicamente problemas que se ajustaban a los requerimientos, plausibles y con información suficiente para ser resueltos (Tabla 1). Los otros tres sistemas analizados, o bien proporcionaron varios problemas que no se ajustaban a lo indicado en el prompt (Copilot y Gemini) o formularon problemas a los que les faltaban información (ChatGPT y Gemini). La información ausente en los problemas codificados como P1 está relacionada con el hecho de no explicitar la igualdad de las unidades

consideradas en los problemas, algo que también se observó en el estudio realizado con futuros docentes de primaria (Embid et al., 2023).

A pesar de que ninguno de los chatbots analizados generó problemas con errores matemáticos, el hecho de que tres de los cuatro sistemas formularan problemas que no se ajustaban a lo indicado o a los que les faltaba información pone de relieve la necesidad de supervisión humana a las respuestas generadas por este tipo de herramientas digitales.

En relación al segundo objetivo específico, el análisis de los cambios en los problemas formulados a lo largo de las iteraciones reveló que la calidad de los problemas generados no siempre mejoró al aumentar el número de intentos (Figura 1). ChatGPT, por ejemplo, mostró un rendimiento inconsistente, logrando su único éxito en la segunda iteración. En contraste, Claude demostró una mayor estabilidad y consistencia en la calidad de sus formulaciones a través de todas las iteraciones. Esto confirma lo indicado por Schorcht et al. (2024) en cuanto a la variabilidad en las respuestas generadas por este tipo de tecnología, en este caso a raíz del modelo de lenguaje que utiliza cada uno de los chatbots estudiados, puesto que el prompt fue el mismo en todos los casos.

Esta investigación supone un primer acercamiento al análisis del potencial de distintos chatbots para la formulación de problemas, puesto que se ha considerado una única técnica del prompt, la Zero-Shot prompting, en la que el usuario hace una única petición al sistema, sin proporcionar información adicional (Schorcht et al., 2024). En próximos estudios se analizará cómo responden estos chatbots a distintas técnicas del prompt ya existentes, y a otras generadas explícitamente para la formulación de problemas matemáticos. Estas últimas podrían obtenerse a partir de los resultados de estudios sobre las estrategias que los individuos emplean en el proceso de formulación de problemas (Baumanns y Rott, 2022; Koichu y Kontorovich, 2013).

Los resultados obtenidos subrayan, además, la importancia de realizar un análisis exhaustivo y continuo del rendimiento de los chatbots en la generación de problemas matemáticos. A medida que se desarrolla esta tecnología, es esencial incorporar mecanismos de validación y supervisión para garantizar que los problemas generados no solo sean válidos desde el punto de vista matemático, sino que también se alineen con los objetivos educativos específicos. Lo

anterior muestra la necesidad de identificar qué conocimientos específicos necesita un docente para poder utilizar este tipo de herramientas digitales, de forma eficiente, para la formulación de problemas e incorporar esos elementos en los programas de formación docente, tanto inicial como continua. En la actualidad existen marcos específicos sobre el conocimiento profesional docente para el uso de inteligencia artificial generativa como ChatGPT (Mishra et al., 2023) y modelos concretos para la formulación de problemas matemáticos (Leavy y Hourigan, 2022). Generar un modelo que combine ambos aspectos permitirá mejorar la efectividad de los chatbots como herramientas de apoyo en la enseñanza y el aprendizaje de las matemáticas.

Agradecimientos

Este trabajo ha sido realizado con el apoyo del proyecto PID2022-139007NB-I00, financiado por MCIN/AEI/10.13039/501100011033/FEDER, UE. Además, la primera autora cuenta con un contrato predoctoral PREP2022-000959 financiado por MCIN/AEI/10.13039/501100011033 y FSE+.

Referencias bibliográficas

Ayllón, M.F. y Gómez, I.A. (2014). La invención de problemas como tarea escolar. *Escuela Abierta*, 17, 29-40.

Cai, J. y Hwang, S. (2020). Learning to teach through mathematical problem posing: Theoretical considerations, methodology, and directions for future research. *International Journal of Educational Research*, 102, 101391. https://doi.org/10.1016/j.ijer.2019.01.001

Cankoy, O. (2014). Interlocked problem posing and children's problem posing performance in free structured situations. *International Journal of Science and Mathematics Education*, 12, 219-238.

Cankoy, O. y Özder, H. (2017). Generalizability Theory Research on Developing a Scoring Rubric to Assess Primary School Students' Problem Posing Skills. *EURASIA Journal of Mathematics Science and Technology Education*, 13(6), 2423-2439. DOI 10.12973/eurasia.2017.01233a

Čavojský, M., Bugár, G., Kormaník, T. y Hasin, M. (2023). Exploring the Capabilities and Possible Applications of Large Language Models for Education. *21st International Conference on Emerging eLearning Technologies and Applications* (ICETA), 91–98. https://doi.org/10.1109/ICETA61311.2023.10344166

Crawford, J., Cowling, M. y Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, 20(3). https://doi.org/10.53761/1.20.3.02

Crespo, S. (2015). A collection of problem-posing experiences for prospective mathematics teachers that make a difference. En Singer, F. M., Ellerton, N. y Cai, J. (eds.), *Mathematical Problem Posing. From Research to Effective Practice* (pp. 493-511). Springer.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Pearson.

Embid, S., Perdomo-Díaz, J., Bruno, A., García-Alonso, I. y Sosa-Martín, D. (2023). Errores en la formulación de problemas de fracciones y efectos de la variable situación en futuros docentes de primaria. En C. Jiménez-Gestal, Á. A. Magreñán, E. Badillo, E. y P. Ivars (Eds.), *Investigación en Educación Matemática* XXVI (pp. 219–226). SEIEM.

Embid, S. y Perdomo-Díaz, J. (2024). Herramientas digitales en la formulación de problemas de fracciones por futuros maestros de primaria. En N. Adamuz-Povedano, E. Fernández-Ahumada, N. Climent y C. Jiménez-Gestal (Eds.), *Investigación en Educación Matemática XXVII* (pp. 201-208). SEIEM.

García-Alonso, I., Bruno, A., Almeida, R., Sosa-Martín, D. y Perdomo-Díaz, J. (2022). Problemas de fracciones formulados por futuros profesores: algunas características. En T. F. Blanco, C. Núñez-García, M. C. Cañadas y J. A. González-Calero (Eds.), Investigación en Educación Matemática XXV (pp. 295-304). SEIEM.

Grundmeier, T. A. (2015). Developing the Problem-Posing Abilities of Prospective Elementary and Middle School Teachers. En F. M. Singer, N. F. Ellerton y J. Cai (Eds.), *Mathematical Problem Posing: From Research to Effective Practice* (pp. 411-431). Springer. https://doi.org/10.1007/978-1-4614-6258-3_20

Guo, S., Zheng, Y. y Zhai, X. (2024). Artificial intelligence in education research during 2013–2023: A review based on bibliometric analysis. *Education and Information Technologies.* https://doi.org/10.1007/s10639-024-12491-8

Hiebert, J. (1997). *Making Sense: Teaching and Learning Mathematics with Understanding.*

Kiliç, C. (2015). Analyzing pre-service primary teachers' fraction knowledge structures through problem posing. *Eurasia Journal of Mathematics, Science and Technology Education*, 11(6), 1603-1619. https://doi.org/10.12973/eurasia.2015.1425a

Kilpatrick, J. (1987). Problem formulating: Where do good problems come from?. En A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 123–147). Lawrence Erlbaum Associates.

Koichu, B., Harel, G., y Manaster, A. (2013). Ways of thinking associated with mathematics teachers' problem posing in the context of division of fractions. *Instructional Science, 41* (4), 681-698.

Kuhail, M. A., Alturki, N., Alramlawi, S. y Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3

Leavy, A. y Hourigan, M. (2022). The Framework for Posing Elementary Mathematics Problems (F-PosE): Supporting Teachers to Evaluate and Select Problems for Use in Elementary Mathematics. *Educational Studies in Mathematics*, 111(1), 147-176. https://doi.org/10.1007/s10649-022-10155-3

Lu, O. H. T., Huang, A. Y. Q., Tsai, D. C. L. y Yang, S. J. H. (2021). Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students Learning Performance. *Educational Technology & Society*, 24 (3), 159-173.

Mishra, P., Warr, M. e Islam, R. (2023). TPACK in the age of ChatGPT and Generative AI. *Journal of digital learning in teacher education, 39(4),* 235-251. https://doi.org/10.1080/21532974.2023.2247480

Noster, N., Gerber, S. y Siller, H.-S. (enviado). Pre-Service Teachers' Approaches in Solving Mathematics Tasks with ChatGPT – A Qualitative Analysis of the Current Status Quo. https://doi.org/10.21203/rs.3.rs-4182920/v1

Parra, V., Sureda, P., Corica, A., Schiaffino, S. y Godoy, D. (2024). Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 65. https://doi.org/10.9781/ijimai.2024.02.009

Plevris, V., Papazafeiropoulos, G. y Rios, A. J. (enviado). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. arXiv preprint arXiv:2305.18618.

Singer, F. M. y Voica, C. (2013). A problem-solving conceptual framework and its implications in designing problem-posing tasks. *Educational Studies in Mathematics, 83*(1), 9-26.

Silver, E. A. (1994). On Mathematical Problem Posing. *For the Learning of Mathematic*s, 14(1), 19-28.

Santos, M. C. (2001). Problem posing and problematization in learning and teaching mathematics. *Adult Education and Development, 5*7, 107-121. https://www.dvv-international.de/en/adult-education-and-development/editions/aed-572001/basic-education-in-practice/problem-posing-and-problematization

Schorcht, S., Buchholtz, N. y Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9. https://doi.org/10.3389/feduc.2024.1386075

Son, JW. y Diletti, J. (2017). What can we learn from textbook analysis? En J.W. Son, T. Watanabe y J.J. Lo (Eds). What matters? Research trends in international comparative studies in mathematics education. Springer. https://doi.org/10.1007/978-3-319-51187-0_1

Sosa-Martín, D., Perdomo-Díaz, J., Bruno, A., Almeida, R. y García-Alonso, I. (2024). The influence of problem-posing task situation: Prospective primary teachers working with fractions. *The Journal of Mathematical Behavior*, 73, 101139.https://doi.org/10.1016/j.jmathb.2024.101139

Yao, Y., Hwang, S. y Cai, J. (2021). Preservice teachers' mathematical understanding exhibited in problem posing and problem solving. *ZDM – Mathematics Education*, 53, 937–949.