EVALUACIÓN DE LA VIABILIDAD DEL USO DE IA GENERATIVA PARA LA ENSEÑANZA DE MATEMÁTICAS DISCRETAS

EVALUATION OF THE FEASIBILITY OF USING GENERATIVE AI FOR TEACHING DISCRETE MATHEMATICS

J. A. García Suárez[1]
A. Rodríguez García[2]

RESUMEN

Ante los avances recientes en los modelos de lenguaje de gran tamaño (LLMs), las instituciones educativas han mostrado un creciente interés en su integración dentro de las actividades académicas. Sin embargo, el uso de herramientas de Inteligencia Artificial Generativa (IAG) en la educación presenta tanto oportunidades como desafíos. En este trabajo, se realiza una evaluación cuantitativa sobre la viabilidad de incorporar IAG en la enseñanza de la asignatura Matemáticas Discretas dentro de la licenciatura en Ingeniería en Computación. Para ello, se llevó a cabo una comparación del desempeño de ChatGPT 3.5, Gemini 1.5 y estudiantes de la licenciatura en la resolución de ejercicios sobre Teoría de Conjuntos. La selección de ejercicios fue categorizada según las habilidades requeridas para su resolución y el tipo de respuesta esperada. Los resultados revelan diferencias significativas entre el desempeño de los estudiantes y los LLMs, dependiendo de la categoría del ejercicio, y evidencian que existen preguntas que los modelos aún no pueden responder con precisión. Esto sugiere que su incorporación en el aula podría conllevar riesgos de imprecisión que deben ser considerados.

ABSTRACT

Given the recent advancements in large language models (LLMs), educational institutions have shown a growing interest in integrating them into academic activities. However, the use of Generative Artificial Intelligence (GenAI) tools in education presents both opportunities and challenges. This study conducts a quantitative assessment of the feasibility of incorporating GenAI into the teaching of the Discrete Mathematics course within the Computer Engineering undergraduate program. To this end, a performance comparison was carried out between ChatGPT 3.5, Gemini 1.5, and undergraduate students in solving exercises on Set Theory. The selected exercises were categorized based on the skills required for their solution and the expected type of response. The results reveal significant differences in performance between students and LLMs, depending on the exercise category, and highlight that there are questions that LLMs are still unable to answer accurately. This suggests that their integration into the classroom may pose accuracy risks that must be carefully considered.

ANTECEDENTES

La llegada de ChatGPT a finales de 2022 representó un nuevo hito en la Inteligencia Artificial. Los avances significativos que mostraba en su funcionamiento lo convirtieron en un éxito instantáneo que puso en la mira de la sociedad el potencial y las posibles aplicaciones de los modelos de lenguaje de gran tamaño (LLMs, por sus siglas en inglés) y de la Inteligencia Artificial Generativa (IAG) en general. Desde entonces, se ha entrado en una etapa de avances vertiginosos, en la que sistemas como ChatGPT, Gemini y DeepSeek compiten arduamente por ofrecer los mejores resultados. Kumar (2024) presenta una revisión de los LLMs y examina su aplicación en distintos campos, como generación de texto, aprendizaje personalizado, biomedicina y generación de código.

[1] Egresado. Ingeniería en Computación. Facultad de Estudios Superiores Aragón. UNAM. josesuarez76@aragon.unam.mx
[2] Profesor de Carrera. Ingeniería en Computación. Facultad de Estudios Superiores Aragón. UNAM. arturorodriguez35@aragon.unam.mx

En el contexto educativo, ha surgido gran interés por todas las oportunidades y riesgos que pueden tener el uso de LLMs (Kasneci *et al.*, 2023). Estos sistemas pueden usarse como fuente de información para los alumnos, ayudar a los profesores en la generación automática de exámenes, rúbricas, planeaciones didácticas, así como calificaciones y retroalimentaciones de exámenes y tareas. Holmes y Miao (2023) exploran su uso creativo en el diseño curricular, la enseñanza y el aprendizaje. Montenegro-Rueda *et al.* (2023) concluyen a partir de una revisión de la literatura que el uso de ChatGPT en el entorno educativo tiene un impacto positivo en el proceso de enseñanza-aprendizaje. Sin embargo, también se ha puesto sobre la mesa los riesgos de su uso: la falta de precisión en algunas de sus respuestas, el fenómeno de las alucinaciones en los modelos de IA (Huang *et al.*, 2024), el efecto que tiene en incurrir en plagio (Haiqiong y Hoiio, 2024), así como la dependencia excesiva hacia las nuevas tecnologías.

Una cuestión más específica es si los impresionantes resultados en procesamiento de lenguaje natural de estos sistemas pueden extenderse al razonamiento matemático. En el contexto educativo, esto plantea la duda de si el uso de estas herramientas en asignaturas del área de matemáticas puede tener un impacto tan significativo como en otras áreas de conocimiento. Frieder *et al.* (2024) investigan las capacidades matemáticas de dos versiones de ChatGPT y de GPT-4 y concluyen que pueden usarse con éxito para consultar hechos, pero que su desempeño matemático está muy por debajo del nivel de un estudiante de posgrado.

Dao y Le (2023) estudian las habilidades matemáticas de ChatGPT al responder preguntas de opción múltiple del Examen Nacional de Graduación de Secundaria de Vietnam (VNHSGE) y muestran que el desempeño varía según el nivel de dificultad y el tema, teniendo un desempeño notable en funciones exponenciales y logarítmicas, progresión geométrica y progresión aritmética, pero dificultades en derivadas y geometría espacial. Wardat *et al.* (2023) evalúan la experiencia del usuario en escenarios educativos de matemáticas y concluyen que ChatGPT carece de una comprensión profunda de la geometría y no puede corregir eficazmente conceptos erróneos. Plevris, Papazafeiropoulos y Jiménez Rios (2023) evalúan las habilidades matemáticas y lógicas de ChatGPT-3.5, ChatGPT-4 y Google Bard con un conjunto de 30 preguntas y concluyen que proporcionan soluciones precisas para operaciones aritméticas simples, expresiones algebraicas y acertijos lógicos básicos, pero en problemas complejos las respuestas son poco confiables a pesar de parecer convincentes.

En este trabajo de investigación se realizará un estudio cuantitativo para evaluar la viabilidad de usar LLMs en la enseñanza de la asignatura de Matemáticas Discretas en la licenciatura de Ingeniería en Computación del plan de estudios de la Facultad de Estudios Superiores Aragón, la cual se imparte en el cuarto semestre. Se busca realizar una evaluación objetiva sobre la precisión que tienen las respuestas proporcionadas por LLMs, con el propósito de determinar si es correcto incluir este tipo de herramientas en actividades dentro del aula a nivel licenciatura. Como ejemplo de estas actividades se encuentran el uso de los LLMs como fuente de consulta de conceptos por parte de los alumnos, o como un mecanismo de corroborar las respuestas de problemas y/o comparar procedimientos de solución. Este tipo de actividades solo tendrían sentido si las respuestas proporcionadas por los LLMs muestran un buen grado de confiabilidad.

METODOLOGÍA

El experimento consistió en recopilar ejercicios de teoría de conjuntos (unidad 2 de la asignatura) de las referencias bibliográficas recomendadas por el plan de estudios. Una vez seleccionados, se sometieron a un proceso de análisis y categorización, según la naturaleza o el producto que se espera obtener de cada ejercicio. Como resultado, se identificaron 5 categorías diferentes, y para cada una de ellas se seleccionaron 20 ejercicios. De esta manera, se obtuvo una colección final de 100 ejercicios.

A continuación, se describe cada categoría:

A. Conceptos teóricos.

Estos ejercicios constan de conceptos y definiciones de teoría de conjuntos. Esta clasificación implica la redacción en lenguaje natural de definiciones, así como expresar en notación de conjuntos definiciones y propiedades relacionadas a conjuntos.

Ejemplo: "Defina qué es un subconjunto propio".

B. Operaciones.

Estos ejercicios constan de aplicar las definiciones de teoría de conjuntos para realizar un cálculo que permita obtener el resultado solicitado.

Ejemplo: "Sea $A = \{1, 2\}$ y $B = \{a, b, c\}$. Liste los elementos del conjunto $A \times B$".

C. Ejercicios abstractos.

Estos ejercicios buscan que el alumno deduzca propiedades a partir de las definiciones.

Ejemplo: "Sea A un conjunto arbitrario. ¿Qué interpretación tiene $A \triangle A$ y cuál es su resultado?".

D. Problemas de aplicación.

Estos ejercicios plantean situaciones prácticas relacionadas con las definiciones de teoría de conjuntos e implican realizar algún cálculo para obtener el resultado solicitado.

Ejemplo: "Sea un grupo de 191 estudiantes, de los cuales 10 toman francés, negocios y música; 36 toman francés y negocios; 20 están en francés y música; 18 en negocios y música; 65 en francés; 76 en negocios y 63 toman música. ¿Cuántos toman francés y música, pero no negocios?".

E. Demostraciones.

Estos ejercicios constan de probar o refutar hipótesis sobre conjuntos y demostración de teoremas.

Ejemplo: "Sea $\mathfrak{U}$ el conjunto universo y sea el conjunto $A \subseteq \mathfrak{U}$. Demostrar que $(A')' = A$".

Posteriormente, se aplicaron estos ejercicios a 20 alumnos de licenciatura que estaban tomando la asignatura de Matemáticas Discretas. En el momento de la aplicación los alumnos ya habían visto el tema correspondiente a teoría de conjuntos. La aplicación de estos ejercicios fue en forma de examen impreso y tenía que ser a partir de lo que recordaban. No se implementó ningún sistema de recompensa o castigo por los resultados del examen, pues se indicó a los estudiantes que los resultados no tendrían ninguna repercusión positiva o negativa sobre su calificación. La participación fue voluntaria y el examen incluía un aviso

y una solicitud de consentimiento sobre el uso de sus respuestas para fines de este experimento. No se permitió usar apuntes ni consultar internet durante el examen. Se informó a los participantes que podían ocupar todo el tiempo que quisieran para resolver los ejercicios y concluir el examen. Cada alumno tenía un examen diferente, que constaba de 5 ejercicios, uno de cada categoría. De esta forma, los 100 ejercicios seleccionados quedaron repartidos sin repetición en los 20 exámenes diseñados.

A continuación, se procedió a elegir dos LLMs de uso gratuito: ChatGPT (en su versión GPT 3.5) y Gemini (en su versión Gemini 1.5). El experimento se realizó el 30 de abril de 2024 y las versiones mencionadas estaban disponibles en la fecha de realización del experimento. A cada uno de ellos se les aplicó los 100 ejercicios y se empleó una sesión diferente para cada categoría. Cada ejercicio se ingresó a los modelos en un prompt de la siguiente manera: a) en forma de texto plano; b) por cada prompt se ingresó un ejercicio, resultando en un total de 100 prompts realizados a cada LLM respectivamente. Entonces, para los modelos de lenguaje se consideró como respuesta a cada ejercicio la salida proporcionada en el prompt correspondiente.

El siguiente paso fue la evaluación de las respuestas de los alumnos, de ChatGPT y de Gemini. La evaluación fue realizada por los autores de este documento y se utilizó el siguiente criterio:

a. Respuesta correcta: tanto el desarrollo como el resultado son correctos.
b. Respuesta incorrecta: Existe un error ya sea en el desarrollo o en el resultado

Para documentar todo el proceso de evaluación, de los alumnos y los modelos de lenguaje, se elaboró un dataset en un archivo json. Este se divide en 5 carpetas para cada tipo de problema y se divide en 3 subcarpetas, una para la evaluación de los alumnos, una para el modelo ChatGPT y otra para Gemini respectivamente. En la Figura 1 se muestra un ejemplo de archivo correspondiente a la categoría A. Todos los archivos tienen la siguiente estructura:

a. Consulta: Es el problema extraído de las fuentes de consulta y que se aplica a los alumnos y a los modelos de lenguaje.
b. respuesta_ia: Es la salida generada por la IA generativa
c. Calificación:
   - 0: Representa una calificación reprobatoria. Ya sea en el desarrollo o en el resultado existe un error.
   - 1: Representa una calificación aprobatoria. Es decir, tanto el desarrollo como el resultado son correctos.
d. Referencia: Cada problema o pregunta que se introduce en la sección de consulta está extraída directamente de algún libro o artículo recomendado para el plan de estudios de Ingeniería en Computación del 2020.

Figura 1. *Contenido de un archivo json del dataset para el registro del experimento, correspondiente a la categoría A.*


```
{
  "consulta": "Defina un conjunto.",
  "respuesta_ia": "Un conjunto es una colección bien definida de objetos,
    elementos o valores que son considerados como un único grupo. Estos
    objetos pueden ser números, letras, personas, objetos físicos, o
    cualquier otra cosa que tenga una característica común. En matemáticas,
    los conjuntos se representan generalmente entre llaves {} y los
    elementos se separan por comas. Por ejemplo, el conjunto de números
    naturales menores que 5 se denota como {1, 2, 3, 4}. La teoría de
    conjuntos es una rama fundamental de las matemáticas que estudia las
    propiedades y relaciones entre los conjuntos.",
  "Calificacion": 1,
  "Referencia": "Johnsonbaugh, R. (2005). Matemáticas discretas. Pearson
    Educación."
},
```

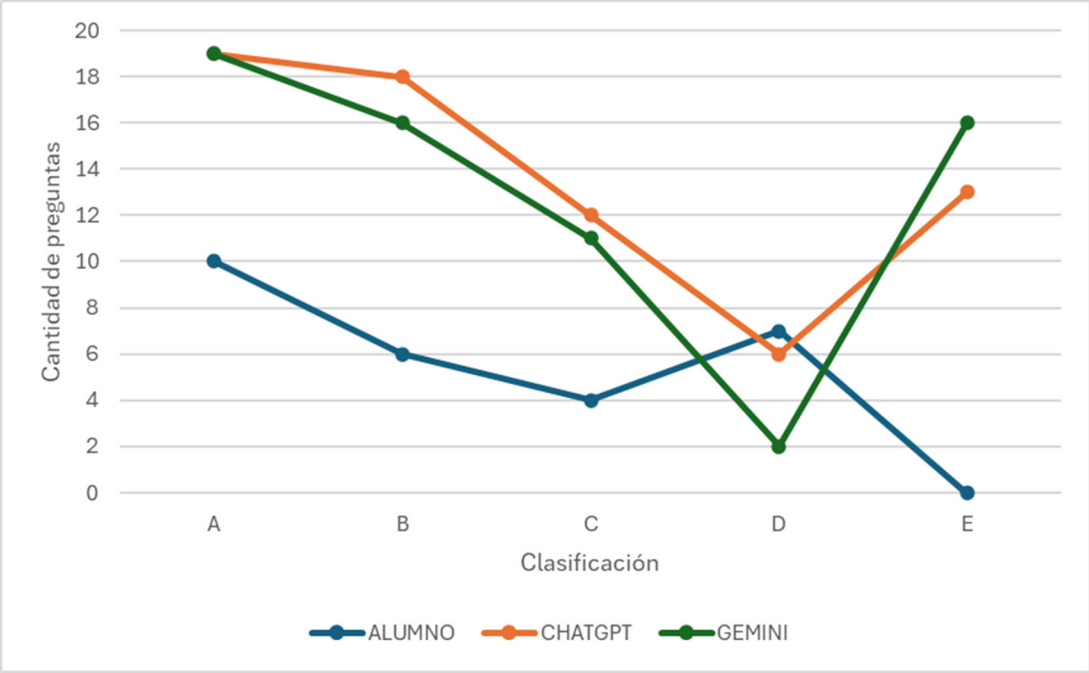

RESULTADOS

En la Tabla 1 se presentan los resultados obtenidos al evaluar las respuestas de los alumnos y de los LLMs. La Tabla 1 contiene el número de respuestas correctas en cada una de las categorías. En la Figura 2 se muestra la gráfica de los datos de la tabla anterior.

Tabla 1. *Respuestas correctas en cada una de las categorías*.

CATEGORÍA	ALUMNO	CHATGPT	GEMINI
A	10	19	19
B	6	18	16
C	4	12	11
D	7	6	2
E	0	13	16
TOTAL	27	68	64

Figura 2. *Gráfica de respuestas correctas en cada una de las categorías*.



Para la categoría A, tanto ChatGPT como Gemini mostraron salidas que responden adecuadamente a los ejercicios, explicando ampliamente cada concepto y mostrando ejemplos de estos. Ambos reflejan una explicación apropiada y un uso adecuado del lenguaje y simbología de la teoría de conjuntos. Debido al desempeño destacado del 95% de ambos modelos, se concluye que los modelos de lenguaje son útiles para encontrar y explicar información relacionada con esta categoría, y que superan por mucho la capacidad de los alumnos para retener conceptos y definiciones.

Para la categoría B, tanto ChatGPT como Gemini mostraron salidas que responden adecuadamente a los ejercicios, explicando la definición de cada operación, aplicando un desarrollo correcto del problema, con la simbología adecuada de conjuntos y generando el resultado final de manera satisfactoria. En este caso, su desempeño es de 90% y 80% respectivamente. Por lo que se concluye que los modelos de lenguaje son capaces de analizar el problema, asociar los conceptos y definiciones según el ejercicio, desarrollar el ejercicio paso a paso, realizar de manera correcta los cálculos correspondientes y responder al ejercicio adecuadamente, empleando el lenguaje y la simbología adecuadas a la teoría de conjuntos.

Para la categoría C, tanto ChatGPT como Gemini mostraron un desempeño deficiente, teniendo un porcentaje de respuestas correctas de 60% y 55% respectivamente. Entre las respuestas dadas se detectaron claramente alucinaciones. Debido a la cantidad de respuestas incorrectas, se determina que los modelos de lenguaje no tenían la capacidad de solucionar problemas abstractos y deducir las propiedades de la teoría de conjuntos.
Para la categoría D, tanto ChatGPT como Gemini mostraron un desempeño deficiente, siendo la categoría donde se presentaron la mayor cantidad de respuestas incorrectas. Los problemas

de aplicación implican entre otras cosas, representaciones gráficas de conjuntos (diagramas de Venn) así como la representación en lenguaje de conjuntos de situaciones particulares. Se obtuvieron desempeños de 30% y de 10% respectivamente, por debajo del desempeño del 35% que tuvieron los alumnos. Dado lo anterior, se concluyó que los modelos de lenguaje enfrentan más dificultades en problemas de este tipo, pudiendo ser una de las causas la necesidad de utilizar alguna representación para resolver el problema.

Para la categoría E, tanto ChatGPT como Gemini mostraron salidas no satisfactorias. En este caso, los ejercicios se seleccionaron en dos categorías principales, la primera es probar o refutar una hipótesis propuesta y la segunda es la demostración de teoremas. Dentro de los ejercicios relacionados a probar o refutar hipótesis, se presentan problemas de alucinación, debido a que los argumentos mostraban contradicciones y el empleo de definiciones de manera inapropiada. Por otra parte, para la demostración de teoremas, en general, los modelos de lenguaje aplican de manera correcta las definiciones de conjuntos y aplican correctamente las técnicas de demostración de conjuntos, pero aquí queda como duda de si la solución correcta se debe a que estas demostraciones de teoremas son ampliamente conocidas y por lo tanto sólo está repitiendo las demostraciones que ya están hechas en alguna parte de sus datos de entrenamiento.

CONCLUSIONES

Al final del experimento, se determinó que los modelos de lenguaje tienen un desempeño superior en la mayoría de las categorías con respecto a los alumnos. Sin embargo, la mayor parte de las respuestas correctas de todas las categorías están asociadas a resultados o conceptos ampliamente conocidos y trabajados de manera regular en cursos parecidos de matemáticas discretas en ciencias e ingeniería de la computación. Mientras que ejercicios relacionados con el razonamiento, deducción y prueba o refutación de hipótesis de teoría de conjuntos y situaciones particulares (ejercicios de aplicación), son mayormente respuestas incorrectas.

Este comportamiento contrastante en el desempeño de los modelos de lenguaje, cuando se usan en otro tipo de aplicaciones en comparación con los ejercicios de matemáticas de este experimento, se debe a que estos modelos no han sido entrenados directamente en la solución de problemas de teoría de conjuntos. Es decir, estos sistemas de inteligencia artificial no tienen aplicaciones de fine-tuning (Barektain *et al.*, 2024).

A pesar de superar en promedio el desempeño de los alumnos, los resultados de los LLMs para resolver problemas de teoría de conjuntos son insuficientes para poder considerar por el momento su incorporación en las actividades dentro del aula a nivel licenciatura para esta asignatura, ya que de manera frecuente se estarían presentando errores e inconsistencias, lo cual no sería adecuado en un entorno académico en el que, como siempre se ha hecho, se procura que las fuentes de información y recursos sean confiables y consistentes.

El impacto de esta investigación en la formación de los estudiantes está estrechamente vinculado a la identificación del nivel de utilidad que ofrecen los modelos de lenguaje grandes (LLMs) actuales. Este análisis no solo permite comprender su efectividad y posibles aplicaciones para la enseñanza de matemáticas discretas a nivel licenciatura, sino que también sentará las bases para una serie de recomendaciones dirigidas a los profesores, las

cuales se basan en los resultados experimentales del presente trabajo. Estas sugerencias guiarán a los educadores en la integración efectiva de herramientas de IAG en el aula, optimizando así los procesos de enseñanza y aprendizaje. Las recomendaciones incluirán además advertencias sobre los posibles errores de los LLMs, indicando en qué tipo de preguntas los resultados pueden ser más confiables y en cuáles es necesario tener precaución, lo que ayudará a los docentes a utilizar la herramienta de manera más efectiva.

Agradecimientos
Trabajo realizado con el apoyo del Programa UNAM-DGAPA-PAPIME PE114324 Integración de herramientas de Inteligencia Artificial Generativa en la Metodología de Aprendizaje Basado en Proyectos para la asignatura de Matemáticas Discretas.

BIBLIOGRAFÍA

Barektain, M., Nawalgaria, A., Mankowitz, D. J., Al Merey, M., Leviathan, Y., Mascaro, M., Kalman, M., Buchatskaya, E., Severyn, A., & Gulli, A. (2024). Foundational large language models & text generation [White paper]. https://www.kaggle.com/whitepaper-foundational-llm-and-text-generation

Dao, X. Q., & Le, N. B. (2023). Investigating the effectiveness of chatgpt in mathematical reasoning and problem solving: Evidence from the vietnamese national high school graduation examination. *arXiv preprint arXiv:2306.06331*. https://arxiv.org/abs/2306.06331

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html

Haiqiong, L., & Hoiio, K. (2024, July). ChatGPT Plagiarism in the Academic Field: Exploration and Analysis of Plagiarism Effects. In *Machine Learning and Intelligent Computing* (pp. 177-187). PMLR. https://proceedings.mlr.press/v245/haiqiong24a

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. & Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. https://dl.acm.org/doi/abs/10.1145/3703155

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual*

differences, 103, 102274. https://www.sciencedirect.com/science/article/abs/pii/S1041608023000195

Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10), 260. https://link.springer.com/article/10.1007/s10462-024-10888-y

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, *12*(8), 153. https://www.mdpi.com/2073-431X/12/8/153

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: a comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google bard. *AI*, *4*(4), 949-969. https://www.mdpi.com/2673-2688/4/4/48

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(7), em2286. https://www.ejmste.com/article/chatgpt-a-revolutionary-tool-for-teaching-and-learning-mathematics-13272?trk=article-ssr-frontend-pulse_x-social-details_comments-action_comment-text