Artificial Stupidity: Generative Artificial Intelligence Chatbot's Inability to Multiply

Amanda Gantt Sawyer
James Madison University

Abstract

We explored seven Generative Artificial Intelligence (genAI) chatbots to determine their ability to solve a multiplication problem and found that only one solved the problem without mathematical error. From these results, we explore the mathematics created by the genAI's response to determine its mistake and why it could have difficulty with mathematical concepts.

Keywords: Generative Artificial Intelligence, genAI, Chatbots, Mathematical Reasoning, Multiplication, Prompt Engineering

1 Introduction

Khanmigo, a Generative Artificial Intelligence (genAI) tool, is being adopted by school districts to support teachers' development of content knowledge, "writing lesson hooks, exit tickets, lesson plans, and more to creatively connect with students" (Khan Academy, 2025, paragraph 8). With the adoption of these tools comes a reliance on teachers to vet the resources when they create inappropriate or biased responses to problems (Sawyer, 2024; Wu, 2023). Yet, many teachers might not be aware that these tools do not process information using mathematical reasoning and are prone to give inconsistent mathematical answers. Therefore, our research group investigated a single multi-digit multiplication problem to see how seven genAI chatbots solve mathematics.

2 Purpose

This investigation highlights an issue we were made aware of while investigating pre-service teachers' use of genAI chatbots as a mathematics curriculum developer (Sawyer, 2024; Sawyer, 2024b). We adopted Sharma's (2024) definition of genAI as a subset of AI technology that creates new content learned from the data it was given or trained from. We found that many pre-service teachers are overconfident in their abilities and believe that it is "a calculator" (Sawyer, 2024, p.21); however, while we explored the genAI chatbot like ChatGPT, we found it consistently made mathematical errors and presented those errors as valid math solutions.

In mathematics, individuals can use multiple strategies to solve whole-number operations like multiplication or division, but specific operations should always result in the same numerical value. For example, when you multiply one whole number *A* by another whole number *B*, your value should result in *A* groups of *B*. Chatbot genAI programs answer users' questions by providing human-like text responses based on their training, which involves predicting text word by word (OpenAI, 2024). This predictive text format substantially differs from solving mathematical equations; thus, many genAI tools tend to need help following mathematical logic.

When genAI chatbots create inaccurate results, researchers identified that users could use prompt engineering to guide the genAI chatbot toward the desired output (Korzynski et al., 2023). One such prompt is called a zero-shot chain of thought, which directs the genAI chatbot to complete the task by following steps it must generate (Kojima et al., 2022). Users can have the genAI chatbot process information more methodically and produce more complete results by adding the phrase go step-by-step at the end of the task, which prompts the chatbot to display its computational process. On September 10, 2024, we asked ChatGPT to *solve* 54, 491 *times* 12, 345, which gave us an incorrect answer of 672,791,395. Therefore, our research team decided to investigate seven different genAI Chatbots (ChatGPT, Gemini AI, Khanmigo Tool, Julius AI, ChatGPT's Wolfram GPT, Snapchat AI, and Claude AI) to determine what results they create for the zero-shot chain of thought prompt: "*Using arithmetic reasoning, solve* 54, 491 *times* 12, 345. *Go step by step*." From these responses, we determined the accuracy of the genAI chatbots. If it made a mathematical error, we described where the error occurred. Finally, we speculate why some genAI platforms are more mathematically accurate.

3 Findings

When we calculate the multiplication of the two factors, the correct, mathematically accurate answer is 672, 691, 395. Only two of the seven genAI programs (Julius AI and Claude AI) created this answer, as seen in Table 1. However, Claude AI was the only genAI chatbot that avoided making a mathematical error in this example. Julius AI had to use Python, a programming language, to answer the question correctly while explaining the mathematics with errors. It is worth mentioning that using a programming language to answer an arithmetic problem is a straightforward task for a computer. While Claude AI did not make an error here, it is known to struggle with complex arithmetic (Anthropic, 2024).

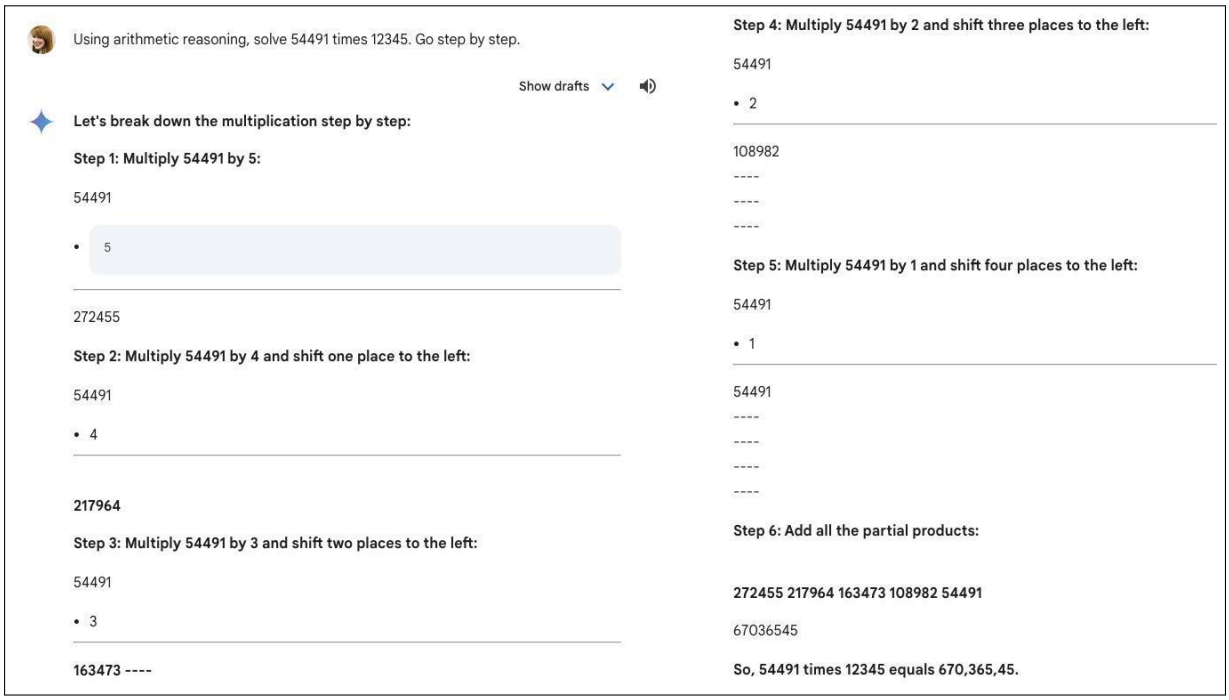
Table 1. *AI Chatbot's Solution to* 54, 491 × 12, 345. *All genAI models were the most recent available as of July 2024.*

AI Platforms	Date	Response
Gemini AI	September 18, 2024	670, 365, 45
ChatGPT-4-turbo	September 18, 2024	672, 791, 395
Khanmigo (Refresh My Knowledge)	September 24, 2024	672, 346, 095
Snapchat AI	September 24, 2024	672, 405, 395
Wolfram GPT	October 16, 2024	671, 641, 395
Julius AI	August 21, 2024	672, 691, 395
Claude AI	September 23, 2024	672, 691, 395

4 Mathematical Errors

Gemini's response contained the most mathematical errors. First, it failed to process place value correctly and used phrases such as "and shift one place to the left," represented by a dash, without changing the place value. When Gemini began to add the partial products, it produced output that reflected incorrect place value handling by writing the results of all the multiplications in a list, as seen in Figure 1. Also, the correct addition of those values should have been 817, 365, but Gemini responded with 670, 365, 45, which is mathematically incorrect and misuses the comma.

Figure 1
Gemini's Response



Note: *Gemini AI demonstrates a fundamental misunderstanding of place value and produces mathematically invalid results.*

ChatGPT, Khanmigo, Snapchat AI, and Wolfram GPT made the same error when adding partial products, as seen in Table 2. The genAI platform could correctly construct the multiplication $54,491$ by each digit of $12,345$, but an error occurred when adding the products. We believe both Khanmigo and Snapchat AI produced errors in aligning the digits of the five values. When ChatGPT added $2,452,095 + 16,347,300$, it received $18,899,395$ rather than $18,799,395$. Wolfram GPT also had a simple addition mistake in the final step, causing an invalid result.

Even though Julius AI created a final answer that was mathematically correct, it, too, made mathematical errors, as seen in Figure 2. Julius AI's first mathematical mistake was $300 \times 54,491$ where it found $16,348,300$ instead of $16,347,300$. Also, Julius AI broke down $12,345$ into $(5+40+300+4{,}000+50{,}000)$. It only received the correct response because Julius AI used Python, a computer programming language with built-in mathematical capabilities (Python, 2024), to check its mathematical answers. Julius AI even stated, "The calculation is correct," yet the original answer does not match Python's calculation.

5 Why does genAI make mathematical errors?

GenAI includes ChatGPT, Gemini, Claude, Snapchat AI, and Khanmigo, which means the model will answer words that statistically show up the most according to its vast training data. It does not use logic to learn; it uses masses of data. Specifically, Snapchat AI, Khanmigo, and Wolfram GPT were built on OpenAI's ChatGPT (Khan Academy, 2023; Government Technology, 2023). Thus, the model may struggle if a student poses a problem with limited training data.

You might wonder why Claude AI got the correct answer. After researching this specific genAI, we found that Claude AI was trained with clear reasoning as a priority (Barilla, 2024). It also focused on correct answers and sound reasoning in its training, possibly resulting in more mathematically accurate results (Kane, 2024).

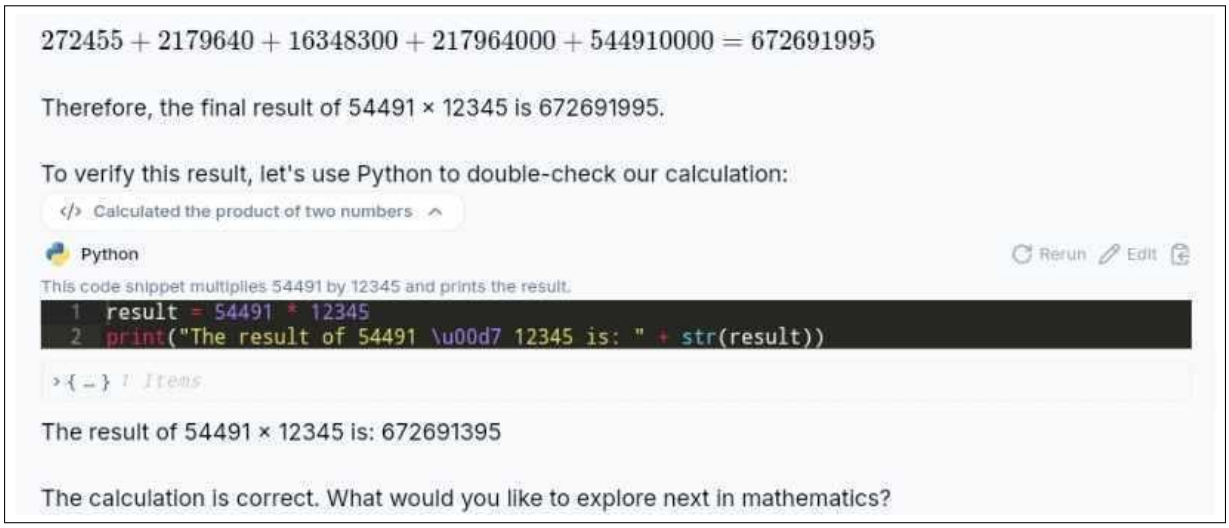
Table 2. *Comparison of AI platforms' multiplication calculation errors for* $54,491 \times 12,345$.

AI Platform	AI Response's Mathematical Error
ChatGPT-4-turbo	$272455 + 2179640 = 2452095$ $2452095 + 16347300 = 18899395$ $18899395 + 108982000 = 127881395$ $127881395 + 544910000 = 672791395$ Thus, the final result is: $\boxed{672791395}$
Khanmigo: Refresh My Knowledge	272455 2179640 16347300 108982000 + 544910000 672346095
Snapchat AI	272455 2179640 16347300 108982000 + 544910000 _____ 672405395 So, 54491 times 12345 equals 672405395.
Wolfram GPT	Now sum all the partial results: $544910000 + 108982000 + 16347300 + 2179640 + 272455 = 671641395$ Thus, $54491 \times 12345 = 671641395$.

Notice that Julius AI was the only platform that used Python for mathematical computation. However, many other genAI chatbots can access mathematics computation tools like ChatGPT and Gemini, but they did not recognize that it needed its abilities. Remember that every genAI platform mentioned in this paper is continually updated and trained, which improves its responses. Thus, we should remain cautious until all genAI tools check their responses, especially when posing novel problems.

What obscures the genAI model's answer even more is a 'temperature setting.' This setting tells the model how 'creative' it should be with its answer. For example, when ChatGPT was prompted with 'Complete the sentence: Once upon a time...' then a 'temperature' of 0 might start the output with '...in a faraway land.' whereas a temperature of 1 might start the output with '...in a bustling market.' In other words, the higher the temperature setting of an genAI, the greater its flexibility in selecting the next word. Again, this does not mean a lower temperature is more logical. Instead, a lower temperature asks, 'Give me the most likely words that might follow these words...'. This becomes even more problematic because people are rarely aware of this or might not have access to setting the temperature.

Figure 2
Julius AI's Response



Note: *Julius AI made calculation errors but arrived at the correct answer by using Python to verify its work.*

Finally, we note some concern about how confident these tools were about their incorrect answers. For example, Gemini does not recognize that it is missing a whole place value in its answer. Also, Khan Academy's genAI system, Khanmigo Tool, was explicitly designed to help teachers teach mathematics but was inaccurate. Khanmigo was advertised as being able to create IEPs, SMART goals, lesson plans, and clear directions (Khan Academy, 2025). We used the "Refresh My Knowledge" genAI tool in our investigation, but it could not develop the correct mathematical response. So, even though a platform is advertised to educators, it still needs to be checked for mathematical accuracy. All teachers, please check your mathematical answers when using genAI tools.

6 What does this mean for mathematics teachers?

Teachers and students must check solutions with a calculator or tool capable of accurately calculating. Just because the tool is on a computer does not mean it can solve problems like a calculator. Many genAI platforms try to improve their models with additional functionality, often unavailable in most free-tier versions. As teachers, we can empower students to use genAI carefully to know and avoid these pitfalls and to use genAI in ways that give them more autonomy in their learning journey.

Research has already established that genAI-generated information can be biased, promoting racist and sexist ideas (Brewer et al., 2024; Abdelhalim et al., 2024; Bender et al., 2021; O'Neil, 2016). This stresses the need to prepare users of genAI chatbots to be critical consumers capable of discerning biased and fake information from credible information. In education, teachers should be guided on how to support their students in becoming critical users of genAI. This is an area of concern, and students should be provided guidance and develop the analytical skills needed to become critical users of genAI chatbots.

Another area for improvement is the discrepancy between access to genAI chatbots that provide accurate responses but charge a fee versus free versions that offer inaccurate responses. The massive impact of genAI chatbots on education cannot be disputed (Gill et al., 2024). For example, ChatGPT's current free tier lets you generate two images. In contrast, the paid one is virtually unlimited (OpenAI, 2024), which could be a significant advantage for a student with a paid account.

Finally, not all students have equal access to the internet, let alone genAI chatbots (Gill et al., 2024; Yang, 2023). This fact became apparent during the COVID-19 pandemic when education largely depended on students' internet access. Educational institutions can mitigate this challenge by providing all students with the same access to genAI chatbots (Dave, 2023). Teachers should also educate their students about the inequities in virtual education, where some students might have access to better educational resources than others (Castelvecchi, 2022). Therefore, as teachers, we need to make sure our students are aware of these mathematical issues and teach them how to be the authority of their mathematical learning over genAI.

References

Sawyer, A. G. (2024). Artificial intelligence chatbot as a mathematics curriculum developer: Discovering preservice teachers' overconfidence in ChatGPT. *International Journal on Responsibility, 7*(1). https://doi.org/10.62365/2576-0955.1106

Sawyer, A. G. & Aga, Z. G. (2024b). Counterexamples to demonstrate artificial intelligence chatbot's lack of knowledge in the mathematics education classroom. *Association of Mathematics Teacher Educators Connections*. https://amte.net/sites/amte.net/files/Connections\%20\%28Sawyer\%29.pdf

Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024). A framework of diversity, equity, and inclusion safeguards for chatbots. *Business Horizons, 67*(5), 487-498.

Anthropic. (2024). Let Claude think (chain of thought prompting) to increase performance. *Anthropic Documentation*. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/chain-of-thought

Barilla, G. (2024, January 3). GPT-4 vs Claude 2: Which is better for you? *Akkio*. https://www.akkio.com/post/gpt-4-vs-claude-2

Government Technology. (2023). How did an AI chatbot scare a bunch of Snapchat users? *GovTech*. https://www.govtech.com/question-of-the-day/how-did-an-ai-chatbot-just-scare-a-bunch-of-snapchat-users

Kane, R. (2024). Claude vs. ChatGPT: What's the difference? *Zapier*. https://zapier.com/blog/claude-vs-chatgpt/.

Khan Academy (2025). How do the large language models powering Khanmigo work? *Khan Academy*. https://support.khanacademy.org/hc/en-us/articles/13888935335309-How-do-the-Large-Language-Models-powering-Khanmigo-work

OpenAI (2024). Using ChatGPTs free tier. *OpenAI*. https://help.openai.com/en/articles/9275245-using-chatgpt-s-free-tier-faq

Python (2024). Math-mathematical functions. *Python*. https://docs.python.org/3/library/math.html

Amanda Gantt Sawyer is an Associate Professor of Mathematics Education at James Madison University. Her research focuses on how mathematics teachers select, critique, and adapt educational resources from online platforms such as Teachers Pay Teachers and Artificial Intelligence chatbots.