Submitted by Jerwin Cabañeros

Student ID 40204404

Reflection: Why AI Bot Fails at Multiplication

The purpose of this experiment was to explore the reliability of GPT-4.1 as a calculator by creating a bot that performs repeated multiplications. The bot takes a base number and a number of iterations, computes the result using GPT-4.1 (nano or mini), and compares it with the mathematically correct value. If GPT produces an incorrect output, the bot generates a humorous, self-deprecating reaction.

The results illustrate the probabilistic nature of language models. In Combination #1, using GPT-4.1-nano with a temperature of 1 and top_p of 1, all steps from 2 x 2 up to 4294967296 were computed correctly, demonstrating that low-to-moderate randomness allows the model to behave almost like a calculator. In Combination #2, using GPT-4.1-mini with temperature 1.6 and top_p 1, the first step was correct, but the next three iterations produced wildly incorrect numbers (54 -> 6401 -> 34649921), each accompanied by humorous, self-deprecating reactions. Combination #3, with GPT-4.1-nano at temperature 1.8 and top_p 0.9, showed correct results for the first six iterations, but then failed on extremely large numbers, producing truncated or misestimated outputs, again with dynamic, witty self-deprecation.

These examples highlight several key points. First, GPT-4.1 treats numbers as text to predict, not as precise values to compute. Higher temperature increases the likelihood of numeric hallucinations, while model size affects the pattern of errors: mini tends to produce coherent but confidently wrong outputs, whereas nano can hallucinate earlier but sometimes recovers. Second, errors are not deterministic: even after initial mistakes, the model can produce correct numbers purely by chance. Third, the bot's self-deprecating personality remains consistent, demonstrating that narrative coherence can persist even when arithmetic fails.

In conclusion, the bot fails because LLMs prioritize plausible text over numeric correctness. High randomness, large iteration counts, and extremely large numbers exacerbate these failures. This experiment illustrates the tension between creativity and reliability in language models and underscores the necessity of human verification in any task requiring precise calculations.