Self-reflection (Week 3):

AI often seems to perform math autonomously, but it actually predicts likely tokens based on patterns from its training data. As a result, even simple arithmetic like multiplication can fail, as AI is 'remembering results' rather than calculating. Thus, the AI bot keeps failing.

Therefore, early calculations appear correct because small integers are common in the training data, which makes it easier for the AI to 'guess' the right result. However, as the numbers grow larger or more complex, the bot becomes harder to predict the answer due to less exposure to such values in its training data, which starts to make errors.

One of the bot's failures in my testing was computing 4096×4096. It returned "16" instead of the correct "16777216." This could be due to the highest-ranked output for 4096×4096 is "16", not the correct one. Hence, it shows that the bot does not actually calculate 4096×4096 but relies on patterns and probabilities in its training data, which led to an incorrect output.

Another interesting behavior is that even when an earlier step produces an incorrect result, later steps can still be correct. In my testing example, the bot correctly calculated Step 4{8,9} despite a wrong answer in Step 3{8,9} ("16"), because it recognized "16777216" as a known value and predicted the outcome based on patterns rather than recalculating them. This suggests that the bot 'treats' a number as a constant, not a value to calculate.

A further insight is that when the answer was large, the bot often produced outputs where the first digits matched, but the rightmost digits were filled with zeros. While in math, the rightmost (integer) index is typically the easiest to calculate. This suggests that the AI recalls frequent patterns and fills in the rest of the answer based on the statistical likelihood of what the result 'should' look like, without performing actual calculations.

Overall, the GPT-4.1 bot fails not because it is poorly designed, but because it is not a calculator. It chooses the highest-ranked output in its answer database, relying on pattern-based guesses instead of performing actual calculations.