The bot fails because GPT-4.1 is not actually doing math. It's just predicting text that looks like math. A real calculator follows strict rules and stores exact numbers. If you give it the same inputs it will always return the same precise answer. A language model doesn't really multiply numbers internally. It guesses what number is most likely to appear after a question like "what is 65536 × 65536". For small numbers this works surprisingly well because those patterns exist everywhere in its training data. Once the numbers get big, the model runs out of reliable examples and starts estimating instead of calculating.The repeated squaring makes the failure unavoidable. Each step massively increases the number of digits. After a few iterations the result requires exact handling of huge integers, but the model only has a rough sense of scale. So it produces something close to the right size but slightly wrong, missing a few digits or rounding in weird ways. You can see that it understands how big the answer should feel, but not the exact value. It is basically guessing magnitude rather than performing arithmetic.The comparison in the program shows a core limitation of language models: imitating reasoning rather than executing algorithms. The system is forcing a predictive text system to act like a processor, and processors and predictors are fundamentally different tools. The AI sounds confident even when it's incorrect, which makes the mistakes funny instead of surprising. The failures come from using a probability machine for deterministic computation. The bot is being asked to do a job it was never designed to do.