At first my bot was failing on purpose, it was only outputting a result of “-1”, forcing the bot to respond with "Damn, chat I did not cook. Send me V-bucks and your parents' credit card info to improve my capabilities.". After tweaking the model to just calculate the inputs, and no nothing more than output the result and if wrong react. That's when I started to see where in the script/model where things went wrong in terms of calculations. The model may be failing due to the complexity of the question, the token limit and or temperature setting. If I increase the numbers, the model simply fails and gives an output of 1, while python remains consistent. With the token limit, I set it to 20. Which isn't too high (~80 characters). This would give the model more room to compute, rather than cutting off the calculation or outputting not enough numbers. With temperature, the model will try to get creative and give more variety in its responses. In this case, it's a very simple task of multiplication of results and there's no need to be extra. Which is why python was able to perform the task so well. It does not try anything else other than computing the calculations. There could also be an issue with how the AI script generation interpreted the mathematical question. To give an understanding of where the GPT-4o model fails, let's look at a couple examples:
“
(2,1) -> (2,6)
GPT-4 Final Answer: 18446744073709551616
Python Final Answer: 18446744073709551616
GPT-4o got it right!

But…

(2,7)
GPT-4 Final Answer: 340282366920938463426481119284
Python Final Answer: 340282366920938463463374607431768211456
Damn chat, I did not cook. Send me V-bucks and your parents' credit card info to improve my capabilities.