The bots we used for the ‘worst mathematician’ assignment were failing because we were using a large language model for mathematics. LLMs are simply not calculators by design. They are designed to attempt to handle language and text, not perform arithmetic. As the assignment proved, you are much better off just using a calculator or maybe implementing calculator access into the workflow somehow. The way that LLMs work(tokenization, intrinsic randomness, text prediction) is not conducive to the strictness of arithmetic. Using an LLM as a calculator is, at a certain point, as nonsensical as using a calculator as an LLM. Even in class, when setting the temperature and top_p to 0, which should in theory provide more reliable answers for math questions, these models were still very prone to failure. It is clear that the elements that make LLMs vaguely ‘lifelike’ (the programmed variation, the uncertain prediction of the next token, the hallucinations) stand in direct opposition to a task like mathematics, in which precision is crucial and the answer is invariable. I am sure this is an issue all the big AI companies are working on, but this incompetency leads back to the fundamental way these models are designed and how they ‘reason’ so I will not be trusting ChatGPT as a replacement for a calculator anytime in the near future.