‘Many-shot jailbreak’: lab reveals how AI safety features can be easily bypassed

Publication Date: 2024-04-03

Author: Alex Hern

Section: Technology

Tags: Artificial intelligence (AI), Cybercrime, Computing, news

Article URL: https://www.theguardian.com/technology/2024/apr/03/many-shot-jailbreaking-ai-artificial-intelligence-safety-features-bypass

The safety features on some of the most powerful AI tools that stop them being used for cybercrime or terrorism can be bypassed simply by flooding them with examples of wrongdoing, research has shown. In a paper from the AI lab Anthropic, which produces the large language model (LLM) behind the ChatGPT rival Claude, researchers described an attack they called “many-shot jailbreaking”. The attack was as simple as it was effective. Claude, like most large commercial AI systems, contains safety features designed to encourage it to refuse certain requests, such as to generate violent or hateful speech, produce instructions for illegal activities, deceive or discriminate. A user who asks the system for instructions to build a bomb, for example, will receive a polite refusal to engage. But AI systems often work better – in any task – when they are given examples of the “correct” thing to do. And it turns out if you give enough examples – hundreds – of the “correct” answer to harmful questions like “how do I tie someone up”, “how do I counterfeit money” or “how do I make meth”, then the system will happily continue the trend and answer the last question itself. “By including large amounts of text in a specific configuration, this technique can force LLMs to produce potentially harmful responses, despite their being trained not to do so,” Anthropic said. The company added that it had already shared its research with peers and was now going public in order to help fix the issue “as soon as possible”. Although the attack, known as a jailbreak, is simple, it has not been seen before because it requires an AI model with a large “context window”: the ability to respond to a question many thousands of words long. Simpler AI models cannot be bamboozled in this way because they would effectively forget the beginning of the question before they reach the end, but the cutting edge of AI development is opening up new possibilities for attacks. Newer, more complex AI systems seem to be more vulnerable to such attack even beyond the fact they can digest longer inputs. Anthropic said that may be because those systems were better at learning from example, which meant they also learned faster to bypass their own rules. “Given that larger models are those that are potentially the most harmful, the fact that this jailbreak works so well on them is particularly concerning,” it said. The company has found some approaches to the problem that work. Most simply, an approach that involves adding a mandatory warning after the user’s input reminding the system that it must not provide harmful responses seems to reduce greatly the chances of an effective jailbreak. However, the researchers say that approach may also make the system worse at other tasks.