Here’s why ChaptGPT and other AI models may not always improve over time

When OpenAI released its latest text-generating artificial intelligence, the large language model GPT-4, in March, it was very good at identifying prime numbers. When the AI was given a series of 500 such numbers and asked whether they were primes, it correctly labeled them 97.6 percent of the time. But a few months later, in June, the same test yielded very different results. GPT-4 only correctly labeled 2.4 percent of the prime numbers AI researchers prompted it with—a complete reversal in apparent accuracy. The finding underscores the complexity of large artificial intelligence models: instead of AI uniformly improving at every task on a straight trajectory, the reality is much more like a winding road full of speed bumps and detours.

Even OpenAI has acknowledged that, when it comes to GPT-4, “while the majority of metrics have improved, there may be some tasks where the performance gets worse,” as employees of the company wrote in a July 20 update to a post on OpenAi’s blog. Past studies of other models have also shown this sort of behavioral shift, or “model drift,” over time. That alone could be a big problem for developers and researchers who’ve come to rely on this AI in their own work.

This is an excerpt. Read the full article here