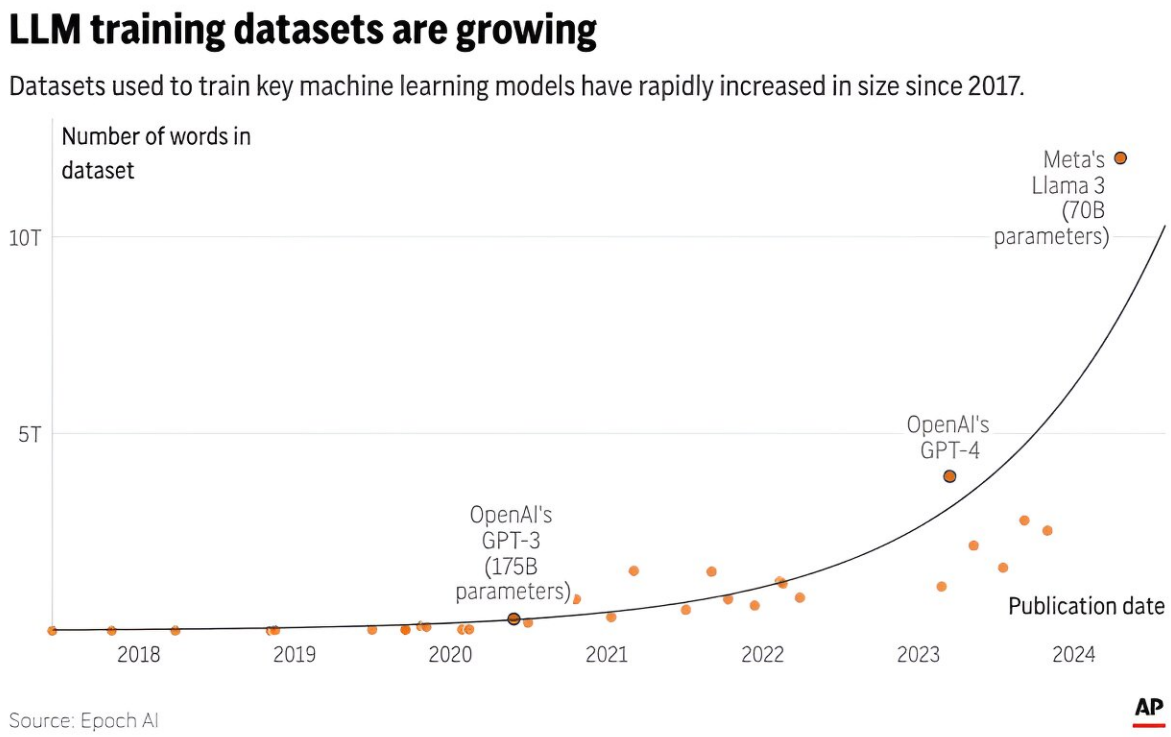
AI 'gold rush' for chatbot training data could run out of human-written text

June 6 2024, by Matt O'brien



Artificial intelligence systems like ChatGPT are gobbling ever-larger collections of human writings they need to get smarter. Credit: AP Digital Embed

Artificial intelligence systems like ChatGPT could soon run out of what keeps making them smarter—the tens of trillions of words people have written and shared online.

A new study released Thursday by research group Epoch AI projects that tech companies will exhaust the supply of publicly available training data for AI language models by roughly the turn of the decade—sometime between 2026 and 2032.

Comparing it to a "literal gold rush" that depletes finite natural resources, Tamay Besiroglu, an author of the study, said the AI field might face challenges in maintaining its current pace of progress once it drains the reserves of human-generated writing.

In the short term, tech companies like ChatGPT-maker OpenAI and Google are racing to secure and sometimes pay for high-quality data sources to train their AI large language models—for instance, by signing deals to tap into the steady flow of sentences coming out of Reddit forums and news media outlets.

In the longer term, there won't be enough new blogs, news articles and social media commentary to sustain the current trajectory of AI development, putting pressure on companies to tap into sensitive data now considered private—such as emails or text messages—or relying on less-reliable "synthetic data" spit out by the chatbots themselves.

"There is a serious bottleneck here," Besiroglu said. "If you start hitting those constraints about how much data you have, then you can't really scale up your models efficiently anymore. And scaling up models has been probably the most important way of expanding their capabilities and improving the quality of their output."

The researchers first made their projections two years ago—shortly before ChatGPT's debut—in a working paper that forecast a more imminent 2026 cutoff of high-quality text data. Much has changed since then, including new techniques that enabled AI researchers to make better use of the data they already have and sometimes "overtrain" on the

same sources multiple times.

But there are limits, and after further research, Epoch now foresees running out of public text data sometime in the next two to eight years.

The team's latest study is peer-reviewed and due to be presented at this summer's International Conference on Machine Learning in Vienna, Austria. Epoch is a nonprofit institute hosted by San Francisco-based Rethink Priorities and funded by proponents of effective altruism—a philanthropic movement that has poured money into mitigating AI's worst-case risks.

Besiroglu said AI researchers realized more than a decade ago that aggressively expanding two key ingredients—computing power and vast stores of internet data—could significantly improve the performance of AI systems.

The amount of text data fed into AI language models has been growing about 2.5 times per year, while computing has grown about 4 times per year, according to the Epoch study. Facebook parent company Meta Platforms recently claimed the largest version of their upcoming Llama 3 model—which has not yet been released—has been trained on up to 15 trillion tokens, each of which can represent a piece of a word.

Traffic on Interstate 35 passes a Microsoft data center on Sept. 5, 2023, in West Des Moines, Iowa. Artificial intelligence systems like ChatGPT could soon run out of what keeps making them smarter — the tens of trillions of words that people have written and shared online. Credit: AP Photo/Charlie Neibergall, File

But how much it's worth worrying about the data bottleneck is debatable.

"I think it's important to keep in mind that we don't necessarily need to train larger and larger models," said Nicolas Papernot, an assistant professor of computer engineering at the University of Toronto and researcher at the nonprofit Vector Institute for Artificial Intelligence.

Papernot, who was not involved in the Epoch study, said building more skilled AI systems can also come from training models that are more

specialized for specific tasks. But he has concerns about training generative AI systems on the same outputs they're producing, leading to degraded performance known as "model collapse."

Training on AI-generated data is "like what happens when you photocopy a piece of paper and then you photocopy the photocopy. You lose some of the information," Papernot said. Not only that, but Papernot's research has also found it can further encode the mistakes, bias and unfairness that's already baked into the information ecosystem.

If real human-crafted sentences remain a critical AI data source, those who are stewards of the most sought-after troves—websites like Reddit and Wikipedia, as well as news and book publishers—have been forced to think hard about how they're being used.

"Maybe you don't lop off the tops of every mountain," jokes Selena Deckelmann, chief product and technology officer at the Wikimedia Foundation, which runs Wikipedia. "It's an interesting problem right now that we're having natural resource conversations about human-created data. I shouldn't laugh about it, but I do find it kind of amazing."

While some have sought to close off their data from AI training—often after it's already been taken without compensation—Wikipedia has placed few restrictions on how AI companies use its volunteer-written entries. Still, Deckelmann said she hopes there continue to be incentives for people to keep contributing, especially as a flood of cheap and automatically generated "garbage content" starts polluting the internet.

AI companies should be "concerned about how human-generated content continues to exist and continues to be accessible," she said.

From the perspective of AI developers, Epoch's study says paying millions of humans to generate the text that AI models will need "is

unlikely to be an economical way" to drive better technical performance.

As OpenAI begins work on training the next generation of its GPT large language models, CEO Sam Altman told the audience at a United Nations event last month that the company has already experimented with "generating lots of synthetic data" for training.

"I think what you need is high-quality data. There is low-quality synthetic data. There's low-quality human data," Altman said. But he also expressed reservations about relying too heavily on synthetic data over other technical methods to improve AI models.

"There'd be something very strange if the best way to train a model was to just generate, like, a quadrillion tokens of synthetic data and feed that back in," Altman said. "Somehow that seems inefficient."

More information: Pablo Villalobos et al, Will we run out of data? Limits of LLM scaling based on human-generated data, *arXiv* (2022). DOI: 10.48550/arxiv.2211.04325