THE AI REVOLUTION IS RUNNING OUT OF DATA. WHAT CAN RESEARCHERS DO?

AI developers are rapidly picking the Internet clean to train large language models such as those behind ChatGPT. Here's how they are trying to get around the problem. **By Nicola Jones**

The Internet is a vast ocean of human knowledge, but it isn't infinite. And artificial intelligence (AI) researchers have nearly sucked it dry.

The past decade of explosive improvement in AI has been driven in large part by making neural networks bigger and training them on ever-more data. This scaling has proved surprisingly effective at making large language models (LLMs) – such as those that power the chatbot ChatGPT – both more capable of replicating conversational language and of developing emergent properties such as reasoning. But some specialists say that we are now approaching the limits of scaling. That's in part because of the ballooning energy requirements for computing. But it's also because LLM developers are running out of the conventional data sets used to train their models.

A prominent study[1] made headlines this year by putting a number on this problem: researchers at Epoch AI, a virtual research institute, projected that, by around 2028, the typical size of data set used to train an AI model will reach the same size as the total estimated stock of public online text. In other words, AI is likely to run out of training data in about four years' time (see 'Running out of data'). At the same time, data owners – such as newspaper publishers – are starting to crack down on how their content can be used, tightening access even more. That's causing a crisis in the size of the 'data commons', says Shayne Longpre, an AI researcher at the Massachusetts Institute of Technology in Cambridge who leads the Data Provenance Initiative, a grass-roots organization that conducts audits of AI data sets.

The imminent bottleneck in training data could be starting to pinch. "I strongly suspect that's already happening," says Longpre.

> **"I DON'T THINK ANYONE IS PANICKING AT THE LARGE AI COMPANIES."**

Although specialists say there's a chance that these restrictions might slow down the rapid improvement in AI systems, developers are finding workarounds. "I don't think anyone is panicking at the large AI companies," says Pablo Villalobos, a Madrid-based researcher at Epoch AI and lead author of the study forecasting a 2028 data crash. "Or at least they don't e-mail me if they are."

For example, prominent AI companies such as OpenAI and Anthropic, both in San Francisco, California, have publicly acknowledged the issue while suggesting that they have plans to work around it, including generating new data and finding unconventional data sources. A spokesperson for OpenAI, told *Nature*: "We use numerous sources, including publicly available data and partnerships for non-public data, synthetic data generation and data from AI trainers."

Even so, the data crunch might force an upheaval in the types of generative AI model that people build, possibly shifting the landscape away from big, all-purpose LLMs to smaller, more specialized models.

Trillions of words

LLM development over the past decade has shown its voracious appetite for data. Although some developers don't publish the specifications of their latest models, Villalobos estimates that the number of 'tokens', or parts of words, used to train LLMs has risen 100-fold since 2020, from hundreds of billions to tens of trillions.

That could be a good chunk of what's on the Internet, although the grand total is so vast that it's hard to pin down – Villalobos estimates the total Internet stock of text data today at 3,100 trillion tokens. Various services use web crawlers to scrape this content, then eliminate duplications and filter out undesirable content (such as pornography) to produce cleaner data sets: a common one called RedPajama contains tens of trillions of words. Some companies or academics do the crawling and cleaning themselves to make bespoke data sets to train LLMs. A small proportion of the Internet is considered to be of high quality, such as human-edited, socially acceptable text that might be found in books or journalism.

The rate at which usable Internet content is increasing is surprisingly slow: Villalobos's paper estimates that it is growing at less than 10% per year, while the size of AI training data sets is more than doubling annually. Projecting these trends shows the lines converging around 2028.

At the same time, content providers are increasingly including software code or refining their terms of use to block web crawlers or AI companies from scraping their data for training. Longpre and his colleagues

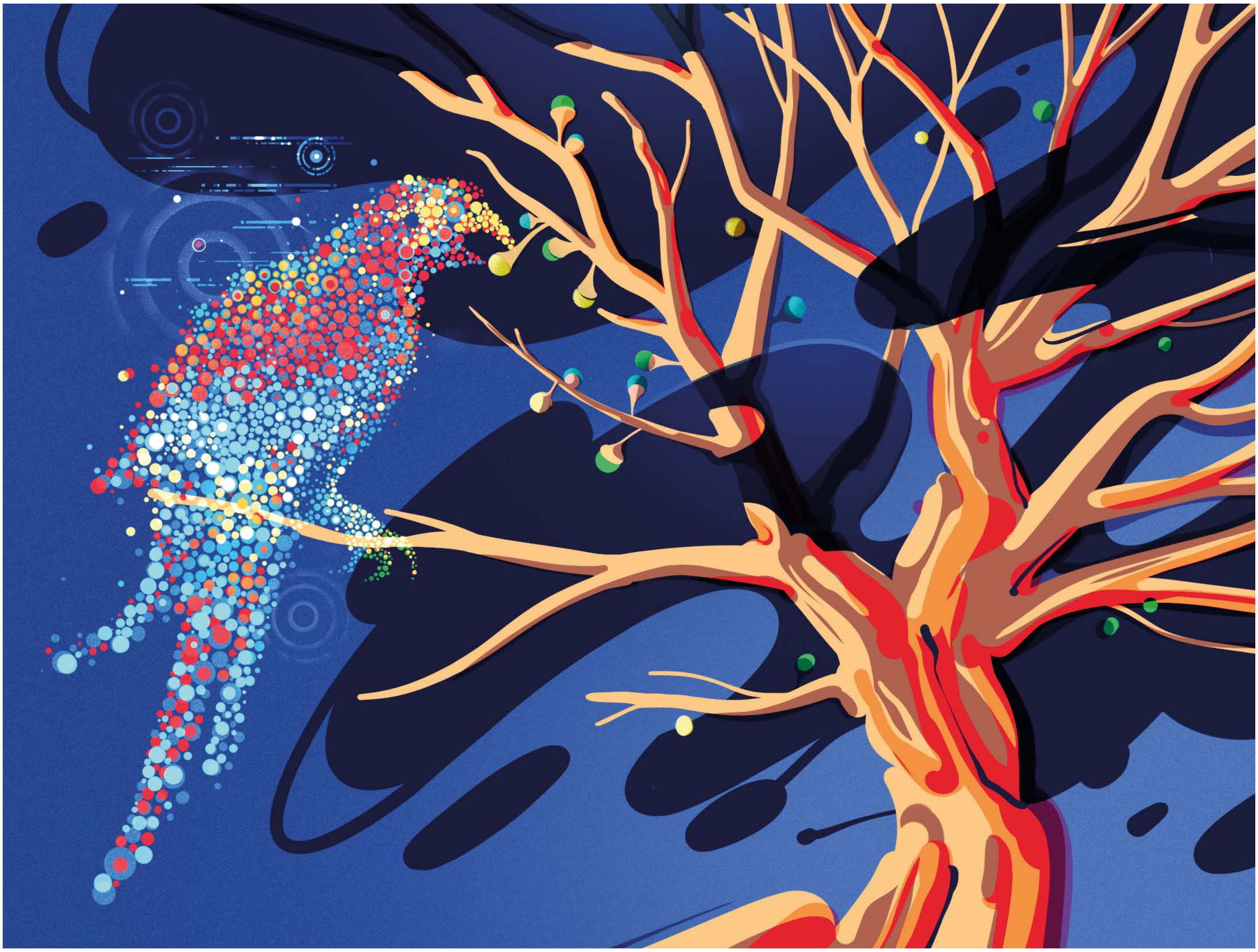


ILLUSTRATION: ADRIÀ VOLTÀ

released a preprint this July showing a sharp increase in how many data providers block specific crawlers from accessing their websites[2]. In the highest-quality, most-often-used web content across three main cleaned data sets, the number of tokens restricted from crawlers rose from less than 3% in 2023 to 20–33% in 2024.

Several lawsuits are now under way attempting to win compensation for the providers of data being used in AI training. In December 2023, *The New York Times* sued OpenAI and its partner Microsoft for copyright infringement; in April this year, eight newspapers owned by Alden Global Capital in New York City jointly filed a similar lawsuit. The counterargument is that an AI should be allowed to read and learn from online content in the same way as a person, and that this constitutes fair use of the material. OpenAI has said publicly that it thinks *The New York Times* lawsuit is "without merit".

If courts uphold the idea that content providers deserve financial compensation, it will make it harder for both AI developers and researchers to get what they need – including academics, who don't have deep pockets. "Academics will be most hit by these deals," says Longpre. "There are many, very pro-social, pro-democratic benefits of having an open web," he adds.

Finding data

The data crunch poses a potentially big problem for the conventional strategy of AI scaling. Although it's possible to scale up a model's computing power or number of parameters without scaling up the training data, that tends to make for slow and expensive AI, says Longpre – something that isn't usually preferred.

If the goal is to find more data, one option might be to harvest non-public data, such as WhatsApp messages or transcripts of YouTube videos. Although the legality of scraping third-party content in this manner is untested, companies do have access to their own data, and several social-media firms say they use their own material to train their AI models. For example, Meta in Menlo Park, California, says that audio and images collected by its virtual-reality headset Meta Quest are used to train its AI. Yet policies vary. The terms of service for the video-conferencing platform Zoom say the firm will not use customer content to train AI systems, whereas OtterAI, a transcription service, says it does use de-identified and encrypted audio and transcripts for training.

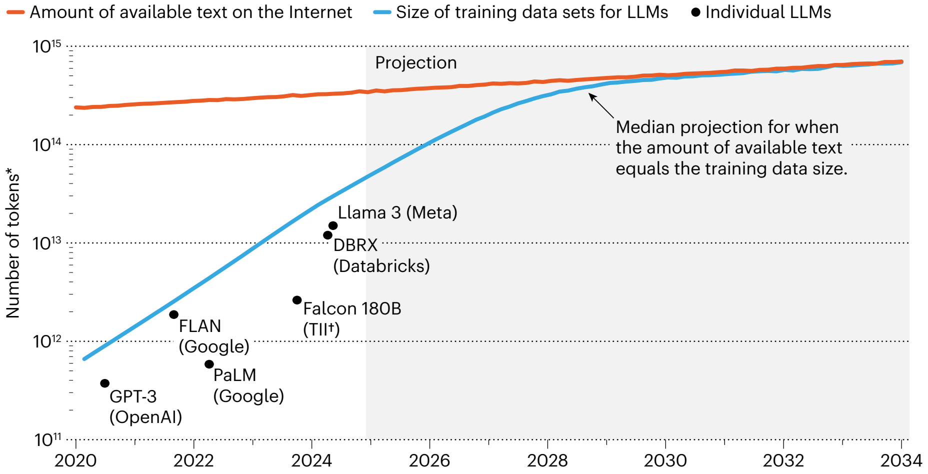
For now, however, such proprietary content probably holds only another quadrillion text tokens in total, estimates Villalobos. Considering that a lot of this is low-quality or duplicated content, he says this is enough to delay the data bottleneck by a year and a half, even assuming that a single AI gets access to all of it without causing copyright infringement or privacy concerns. "Even a ten times increase in the stock of data only buys you around three years of scaling," he says.

Another option might be to focus on specialized data sets such as astronomical or genomic data, which are growing rapidly. Fei-Fei Li, a prominent AI researcher at Stanford University in California, has publicly backed this strategy. She said at a Bloomberg technology summit in May that worries about data running out take too narrow a view of what constitutes data, given the untapped information available across fields such as health care, the environment and education.

But it's unclear, says Villalobos, how available or useful such data sets would be for training LLMs. "There seems to be some degree of transfer learning between many

RUNNING OUT OF DATA

The amount of text data used to train large language models (LLMs) is rapidly approaching a crisis point. An estimate suggests that, by 2028, developers will be using data sets that match the amount of text that is available on the Internet.



*One token is about 0.8 words. †Technology Innovation Institute, Abu Dhabi.

SOURCE: REF. 1

types of data," says Villalobos. "That said, I'm not very hopeful about that approach."

The possibilities are broader if generative AI is trained on other data types, not just text. Some models are already capable of training to some extent on unlabelled videos or images. Expanding and improving such capabilities could open a floodgate to richer data.

Yann LeCun, chief AI scientist at Meta and a computer scientist at New York University who is considered one of the founders of modern AI, highlighted these possibilities in a presentation this February at an AI meeting in Vancouver, Canada. The 10^{13} tokens used to train a modern LLM sounds like a lot: it would take a person 170,000 years to read that much, LeCun calculates. But, he says, a 4-year-old child has absorbed a data volume 50 times greater than this just by looking at objects during his or her waking hours. LeCun presented the data at the annual meeting of the Association for the Advancement of Artificial Intelligence.

Similar data richness might eventually be harnessed by having AI systems in robotic form that learn from their own sensory experiences. "We're never going to get to human-level AI by just training on language, that's just not happening," LeCun said.

If data can't be found, more could be made. Some AI companies pay people to generate content for AI training; others use synthetic data generated by AI for AI. This is a potentially massive source: earlier this year, OpenAI said it generates 100 billion words per day – that's more than 36 trillion words a year, which is about the same size as current AI training data sets. And this output is growing rapidly.

In general, specialists agree, synthetic data seem to work well for regimes in which there are firm, identifiable rules, such as chess, mathematics or computer coding. One AI tool, AlphaGeometry, was successfully trained to solve geometry problems using 100 million synthetic examples and no human demonstrations[3]. Synthetic data are already being used in areas for which real data are limited or problematic. This includes medical data, because synthetic data are free of privacy concerns, and training grounds for self-driving cars, because synthetic car crashes don't harm anyone.

The problem with synthetic data is that recursive loops might entrench falsehoods, magnify misconceptions and generally degrade the quality of learning. A 2023 study coined the phrase Model Autophagy Disorder to describe how an AI model might "go MAD" in this way[4]. A face-generating AI model trained in part on synthetic data, for example, started to draw faces embedded with strange hash markings.

More with less

The alternative strategy is to abandon the 'bigger is better' concept. Although developers continue to build larger models and lean into scaling to improve their LLMs, many are pursuing more-efficient, small models that focus on individual tasks. These require refined, specialized data and better training techniques.

> “EVEN A TEN TIMES INCREASE IN THE STOCK OF DATA ONLY BUYS YOU AROUND THREE YEARS.”

In general, AI efforts are already doing more with less. One 2024 study concluded that because of improvements in algorithms, the computing power needed for an LLM to achieve the same performance has halved every eight months or so[5].

That, along with computer chips specialized for AI and other hardware improvements, opens the door to using computing resources differently: one strategy is to make an AI model re-read its training data set multiple times. Although many people assume that a computer has perfect recall and only needs to 'read' material once, AI systems work in a statistical fashion that means re-reading boosts performance, says Niklas Muennighoff, a PhD student at Stanford University and a member of the Data Provenance Initiative. In a 2023 paper published while he was at the AI firm HuggingFace in New York City, he and his colleagues showed that a model learnt just as much from re-reading a given data set four times as by reading the same amount of unique data – although the benefits of re-reading dropped off quickly after that[6].

Although OpenAI hasn't disclosed information about the size of its model or training data set for its latest LLM, o1, the firm has emphasized that this model leans into a new approach: spending more time on reinforcement learning (the process by which the model gets feedback on its best answers) and more time thinking about each response. Observers say this model shifts the emphasis away from pretraining with massive data sets and relies more on training and inference. This adds a new dimension to scaling approaches, says Longpre, although it's a computationally expensive strategy.

It's possible that LLMs, having read most of the Internet, no longer need more data to get smarter. Andy Zou, a graduate student at Carnegie Mellon University in Pittsburgh, Pennsylvania, who studies AI security, says that advances might soon come through self-reflection by an AI. "Now it's got a foundational knowledge base, that's probably greater than any single person could have," says Zou, meaning it just needs to sit and think. "I think we're probably pretty close to that point."

Villalobos thinks that all of these factors – from synthetic data, to specialized data sets, to re-reading and self-reflection – will help. "The combination of models being able to think by themselves and being able to interact with the real world in various ways – that's probably going to be pushing the frontier."

Nicola Jones is a freelance journalist in Pemberton, Canada.

1. Villalobos, P. *et al. Proc. Mach. Learn. Res.* **235**, 49523–49544 (2024).
2. Longpre, S. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2407.14933 (2024).
3. Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. *Nature* **625**, 476–482 (2024).
4. Alemohammad, S. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2307.01850 (2023).
5. Ho, A. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2403.05812 (2024).
6. Muennighoff, N. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2305.16264 (2023).