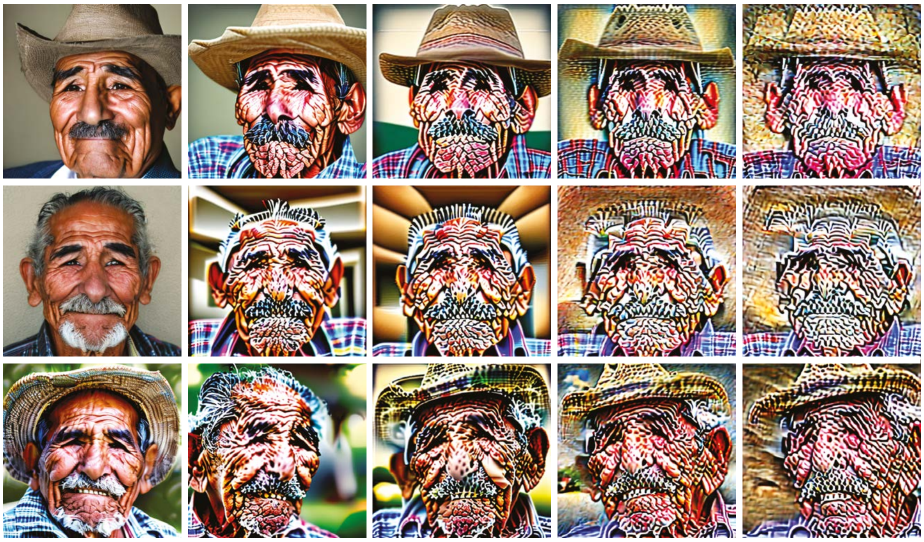

The distorted images produced by an AI model that is trained on AI-made data.

M. BOHÁČEK & H. FARID/ARXIV (CC BY 4.0)

AI MODELS FED AI-GENERATED DATA QUICKLY SPEW NONSENSE

Researchers observed the rapid collapse of a large language model trained on its own output.

By Elizabeth Gibney

Training artificial intelligence (AI) models on AI-generated text quickly leads to the models churning out nonsense, a study has found. This cannibalistic phenomenon, termed model collapse, could halt the improvement of large language models (LLMs) as they run out of human-derived training data and as increasing amounts of AI-generated text pervade the Internet.

“The message is, we have to be very careful about what ends up in our training data,” says co-author Zakhar Shumaylov, an AI researcher at the University of Cambridge, UK. Otherwise, “things will always, provably, go wrong”. he says. The team used a mathematical analysis to show that the problem of model collapse is likely to be universal, affecting all sizes of language model that use uncurated data, as well as simple image generators and other types of AI.

You are what you eat

The researchers began by using an LLM to create Wikipedia-like entries, then trained new iterations of the model on text produced by its predecessor. As the AI-generated information – known as synthetic data – polluted the training set, the model’s outputs became gibberish. The ninth iteration of the model completed a Wikipedia-style article about English church towers with a treatise on the many colours of jackrabbit tails.

More subtly, the study, published in *Nature*[1] on 24 July, showed that even before complete collapse, learning from AI-derived texts caused models to forget the information mentioned least frequently in their data sets as their outputs became more homogeneous.

“If a species inbreeds with their own offspring, it can lead to a collapse of the species.”

This is a concern when it comes to making AI models that represent all groups fairly, because low-probability events often relate to marginalized groups, says study co-author Ilia Shumailov, who worked on the project while at the University of Oxford, UK.

“This is a fantastic paper,” says Julia Kempe, a computer scientist at New York University. Until now, many technology firms have improved their models by feeding them larger and larger amounts of data. But as human-produced content runs out, they are hoping to use synthetic data to keep improving. The study – a version of which first appeared on the arXiv preprint server in May 2023 – has spurred the AI community to try to find solutions to the problem, she says. “It’s been a call to arms.”

Language models work by building up associations between tokens – words or word parts – in huge swathes of text, often scraped from the Internet. They generate text by spitting out the statistically most probable next word, based on these learnt patterns.

Prediction of reality

To demonstrate model collapse, the researchers took a pre-trained LLM and fine-tuned it by training it using a data set based on Wikipedia entries. They then asked the resulting model to generate its own Wikipedia-style articles. To train the next generation of the model, they started with the same pre-trained LLM, but fine-tuned it on the articles created by its predecessor.

The authors judged the performance of each model by giving it an opening paragraph and asking it to predict the next few sentences, then comparing the output to that of the model trained on real data. They expected to see errors crop up, says Shumaylov, but were surprised to see “things go wrong very quickly”, he says.

Collapse happens because each model necessarily samples only from the data it is trained on. This means that words that were infrequent in the original data are less likely to be reproduced, and the probability of common ones being regurgitated is boosted. Complete collapse eventually occurs because each model learns not from reality, but from the previous model’s prediction of reality, with errors getting amplified in each iteration. “Over time, those errors end up stacking up on top of each other, to the point where the model basically only learns errors and nothing else,” says Shumailov.

The problem is analogous to inbreeding in a species, says Hany Farid, a computer scientist at the University of California, Berkeley. “If a species inbreeds with their own offspring and doesn’t diversify their gene pool, it can lead to a collapse of the species,” says Farid, whose work has demonstrated the same effect in image models, producing eerie distortions of reality[2].

Synthetic data problems

Model collapse does not mean that LLMs will stop working, but the cost of making them will increase, says Shumailov.

As synthetic data build up on the Internet, the scaling laws that state that models should get better the more data they train on are likely to break – because training data will lose the richness and variety that comes with human-generated content, says Kempe.

The amount of synthetic data used in training matters. When Shumailov and his team fine-tuned each model on 10% real data alongside synthetic data, collapse occurred more slowly. And model collapse has not yet been seen in the 'wild', says Matthias Gerstgrasser, an AI researcher at Stanford University in California. A study by Gerstgrasser's team found that when synthetic data didn't replace real data, but instead accumulated alongside them, catastrophic model collapse was unlikely[3]. It is unclear what happens when a model trains on data produced by a different AI, rather than its own.

Developers might need to find ways, such as watermarking, to keep AI-generated data separate from real data, which would require unprecedented coordination by big-tech firms, says Shumailov. And society might need to find incentives for human creators to keep producing content. Filtering is likely to become important, too – for example, humans could curate AI-generated text before it goes back into the data pool, says Kempe. "Our work[4] shows that if you can prune it properly, the phenomenon can be partly or maybe fully avoided," she says.

1. Shumailov, I. *et al. Nature* **631**, 755–759 (2024).
2. Bohacek, M. & Farid, H. Preprint at arXiv https://doi.org/10.48550/arXiv.2311.12202 (2023).
3. Gerstgrasser, M. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2404.01413 (2024).
4. Feng, Y. , Dohmatob, E., Yang, P., Charton, F. & Kempe, J. Preprint at arXiv https://doi.org/10.48550/arXiv.2406.07515 (2024).

SDI PRODUCTIONS/GETTY

Musical memory seems to be resistant to age-related cognitive declines.

MUSICAL MEMORIES DON'T DIMINISH WITH AGE

Eighty-year-olds are able to identify familiar tunes just as well as teenagers can.

By Bianca Nogrady

The ability to remember and recognize a musical theme does not seem to be affected by age, unlike many other forms of memory.

"You'll hear anecdotes all the time of how people with severe Alzheimer's can't speak, can't recognize people, but will sing the songs of their childhood or play the piano," says Sarah Sauvé, a feminist music scientist now at the University of Lincoln, UK.

Past research has shown that many aspects of memory are affected by ageing, such as recall tasks that require real-time processing, whereas recognition tasks that rely on well-known information and automatic processes are not. The effect of age on the ability to recall music has also been investigated, but Sauvé was interested in exploring this effect in a real-world setting such as a concert.

In her study, she and her colleagues tested how well a group of roughly 90 healthy adults, ranging in age from 18 to 86 years, were able to recognize familiar and unfamiliar musical themes at a concert (S. A. Sauvé *et al. PLoS ONE* **19**, e0305969; 2024). Participants were recruited at a performance of the Newfoundland Symphony Orchestra in St John's, Canada. Another 31 people watched a recording of the concert in a laboratory.

The study focused on three pieces of music played at the concert: *Eine kleine Nachtmusik* by Mozart, which the researchers assumed most participants were familiar with, and two specially commissioned experimental pieces. One of these was tonal and easy to listen to; the other was more atonal and didn't conform to the typical melodic norms of Western classical music. A short melodic phrase from each of the three pieces was played three times at the beginning of that piece, and participants then logged whenever they recognized that theme in the piece.

"People with severe Alzheimer's can't recognize people, but will sing the songs of their childhood."

The melodic phrase from *Eine kleine Nachtmusik* was equally well recognized across all ages and musical backgrounds, with no loss of recognition as age increased. Participants were less confident in recognizing the theme in the unfamiliar tonal piece, and even less confident with the unfamiliar atonal piece. This pattern, too, did not vary with age. The study also found no age-related difference between the results for participants at the concert and those for people in the lab.

Stirring emotions

Steffen Herff, a cognitive neuroscientist at the University of Sydney, Australia, says the apparent resistance of musical memory to age-related cognitive declines could be due to the emotions that music stirs in people. "We know from general memory research that, effectively, the amygdala – or emotional processing – operates a little bit like an importance stamp," he says. Music also tends to follow rules, so "it's relatively easy to have a pretty good guess of what happened in between," Herff says.

The study collected limited data on some participants' cognitive health, and so did not provide detailed insights into how cognitive impairments affect memory. But Herff says there is great interest in using music as a form of 'cognitive scaffolding' – that is, as a memory aid for other information – for people with neurogenerative conditions such as dementia.