data makes the generated data more closely resemble original data, he said. "It's like you have a distribution of the synthetic data, you have a distribution of the real data, and you want to close the gap between them as much as possible," he said.

Improving the quality of synthetic data could also help with another challenge LLMs are facing as they try to improve: a dearth of new data on which to train. Scientists from Epoch AI, a research institute that focuses on trends in AI, have predicted the world will run out of new text to train on sometime between 2026 and 2032. With no new data on which to train future generations of LLMs, progress could stagnate. "The interesting question is, can synthetic data lead to not just stagnation, but actual improvement in the model?" asked Pablo Villalobos, a staff researcher at Epoch.

With curation of high-quality synthetic data, he said, the question becomes "whether this can be done iteratively so that each model generates better data that is used to train another model in basically the opposite of model collapse, in some virtuous circle." He is not yet sure whether such improvement is possible, but sees some signs it could be.

Other issues arise from training new models on generated data that do not quite reach the level of model collapse. For instance, Koyejo said, synthetic data could increase the likelihood that LLMs will discriminate against people in minority groups. Because any minority is by definition a smaller part of the data distribution, losing the tails of the distribution could make minorities disappear entirely. "Data tends to anchor on majority subgroups," he said. "It tends to be good at capturing the most popular themes and less good at capturing tails, so less-represented demographics can get erased in various ways."

While such erasure is something that could happen, he added, the issue has not been well studied. His colleague Diyi Yang, an assistant professor in the natural language processing group at Stanford, said there has been very little research into the question of how model collapse affects diversity issues. "Part of the reason is that, if you think about any existing big models, a lot of the training dynamics or checkpoints of those models actually are not really transparent or publicly available," she said.

"The interesting question is, can synthetic data lead to not just stagnation, but actual improvement of the model?"

In the end, Gal argued, model collapse is an important consideration, but not the matter of imminent disaster that some news coverage has made it out to be. "It's a matter for the tech companies who build these models to be aware of how the models are being used and how the models are being trained, in order to avoid training on synthetic data that they themselves generated."

Further Reading

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y.
AI models collapse when trained on recursively generated data, *Nature*, 2024, DOI: 10.1038/s41586-024-07566-y https://www.nature.com/articles/s41586-024-07566-y

Feng, Y., Dohmatob, E., Yang, P., Charton, F., and Kempe, J.
Beyond model collapse: Scaling up with synthesized data requires reinforcement, arXiv, 2024, https://doi.org/10.48550/arXiv.2406.07515

Gerstgrasser, M. et al.
Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data, arXiv, 2024, https://doi.org/10.48550/arXiv.2404.01413

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M.
Will we run out of data? Limits of LLM scaling based on human-generated data, arXiv, 2022, https://doi.org/10.48550/arXiv.2211.04325

What is generative AI model collapse? How can we stop it? Super Data Science podcast; https://www.youtube.com/watch?v=knVIecsI-OM

Neil Savage is a science and technology writer based in Lowell, MA, USA.

ACM Member News

MAINSTREAMING FORMAL VERIFICATION METHODS

Santosh Nagarakatte is a professor of computer science at Rutgers University, where he also serves as Undergraduate Program Director of Computer Science.

Nagarakatte earned his undergraduate degree in computer engineering from the National Institute of Technology Karnataka in India, and his master's degree in computer science from the Indian Institute of Science. After receiving a Ph.D. in computer science from the University of Pennsylvania, he joined the staff at Rutgers in January 2013, where he has remained.

"My research centers on building abstraction tools and techniques for safe and secure computing systems," Nagarakatte said, adding that his focus is on using formal methods, mathematically rigorous techniques that ensure system correctness, for the verification of mainstream software.

Nagarakatte said one of the foundations of computing systems is that they should behave as intended, providing the system is correct and everything works as expected.

"When you have compiled your code, if the program is behaving weirdly you think, 'oh, I made a mistake in my code.' You wouldn't expect your compiler is wrong," he said.

Nagarakatte has been building lightweight tools that automatically check for correctness. For instance, LLVM is a compiler for which Nagarakatte co-developed Alive, a domain-specific language that verifies LLVM optimizations. His challenge is getting the average computer programmer to embrace these tools.

"My research aims to make these verification tools mainstream so that every programmer can use these formal verification methods in their day-to-day work."
—John Delaney