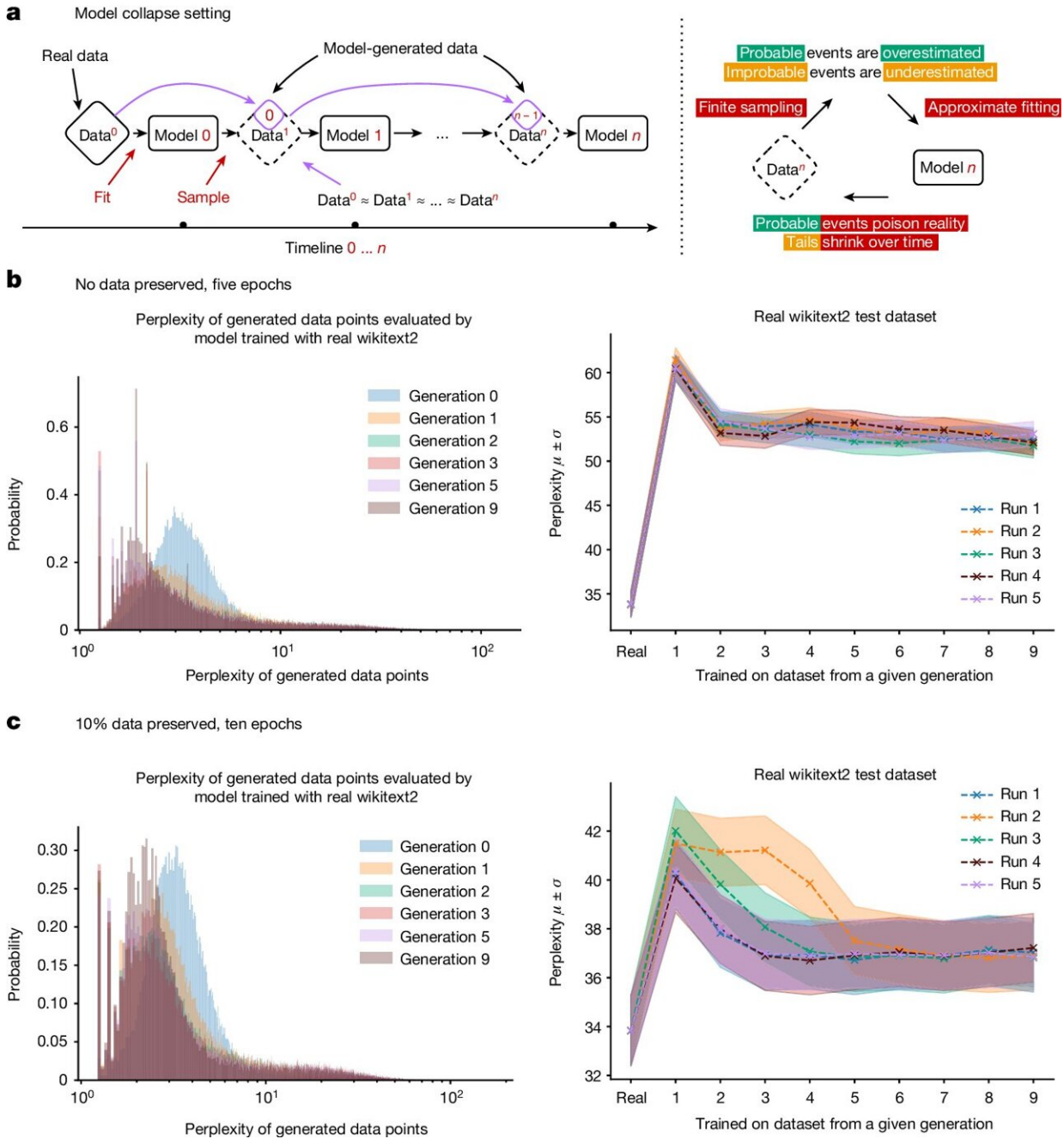
Using AI to train AI: Model collapse could be coming for LLMs, say researchers

July 25 2024



The high-level description of the feedback mechanism in the learning process. Credit: *Nature* (2024). DOI: 10.1038/s41586-024-07566-y

Using AI-generated datasets to train future generations of machine learning models may pollute their output, a concept known as model collapse, according to a new paper published in *Nature*. The research shows that within a few generations, original content is replaced by unrelated nonsense, demonstrating the importance of using reliable data to train AI models.

Generative AI tools such as large language models (LLMs) have grown in popularity and have been primarily trained using human-generated inputs. However, as these AI models continue to proliferate across the Internet, computer-generated content may be used to train other AI models—or themselves—in a recursive loop.

Ilia Shumailov and colleagues present mathematical models to illustrate how AI models may experience model collapse. The authors demonstrate that an AI may overlook certain outputs (for example, less common lines of text) in training data, causing it to train itself on only a portion of the dataset.

Shumailov and colleagues also investigated how AI models responded to a training dataset that was predominantly created with artificial intelligence. They found that feeding a model AI-generated data causes successive generations to degrade in their ability to learn, eventually leading to model collapse.

Nearly all of the recursively trained language models they tested tended

to display repeating phrases. For example, a test was run using text about medieval architecture as the original input and by the ninth generation the output was a list of jackrabbits.

The authors propose that model collapse is an inevitable outcome of AI models that use training datasets created by previous generations. In order to successfully train artificial intelligence with its own outputs, Shumailov and colleagues suggest that training a model with AI-generated data is not impossible, but the filtering of that data must be taken seriously.

At the same time, tech firms that rely on human-generated content may be able to train AI models that are more effective over their competitors.

More information: Ilia Shumailov et al, AI models collapse when trained on recursively generated data, *Nature* (2024). DOI: 10.1038/s41586-024-07566-y

Emily Wenger, AI produces gibberish when trained on too much AI-generated data, *Nature* (2024). DOI: 10.1038/d41586-024-02355-z , doi.org/10.1038/d41586-024-02355-z