Researchers, scared by their own work, hold back “deepfakes for text” AI

Enlarge / This is fine.

reader comments

129 with 98 posters participating, including story author

Share this story

OpenAI, a non-profit research company investigating “the path to safe artificial intelligence,” has developed a machine learning system called Generative Pre-trained Transformer-2 (GPT-2), capable of generating text based on brief writing prompts. The result comes so close to mimicking human writing that it could potentially be used for “deepfake” content. Built based on 40 gigabytes of text retrieved from sources on the Internet (including “all outbound links from Reddit, a social media platform, which received at least 3 karma”), GPT-2 generates plausible “news” stories and other text that match the style and content of a brief text prompt.

The performance of the system was so disconcerting, now the researchers are only releasing a reduced version of GPT-2 based on a much smaller text corpus. In a blog post on the project and this decision, researchers Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever wrote:

> *Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we*

> *are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas.*

OpenAI is funded by contributions from a group of technology executives and investors connected to what some have referred to as the PayPal "mafia"—Elon Musk, Peter Thiel, Jessica Livingston, and Sam Altman of YCombinator, former PayPal COO and LinkedIn co-founder Reid Hoffman, and former Stripe Chief Technology Officer Greg Brockman. Brockman now serves as OpenAI's CTO. Musk has repeatedly warned of the potential existential dangers posed by AI, and OpenAI is focused on trying to shape the future of artificial intelligence technology—ideally moving it away from potentially harmful applications.

Given present-day concerns about how fake content has been used to both generate money for "fake news" publishers and potentially spread misinformation and undermine public debate, GPT-2's output certainly qualifies as concerning. Unlike other text generation "bot" models, such as those based on Markov chain algorithms, the GPT-2 "bot" did not lose track of what it was writing about as it generated output, keeping everything in context.

For example: given a two-sentence entry, GPT-2 generated a

fake science story on the discovery of unicorns in the Andes, a story about the economic impact of Brexit, a report about a theft of nuclear materials near Cincinnati, a story about Miley Cyrus being caught shoplifting, and a student's report on the causes of the US Civil War.

- Unicorns are real in this AI-generated story.
- John F. Kennedy is re-elected, as a robot. Here's his acceptance speech.
- A fake news story about missing nuclear material gets the name of the US Energy secretary wrong, but how many people know that it's Rick Perry?
- Computer, write me a report about the Civil War.
- What Would Tolkein Write?
- GPT-2, write your own press release.
- This story looks like BuzzFeed content.

Each matched the style of the genre from the writing prompt, including manufacturing quotes from sources. In other samples, GPT-2 generated a rant about why recycling is bad, a speech written by John F. Kennedy's brain transplanted into a robot (complete with footnotes about the feat itself), and a rewrite of a scene from *The Lord of the Rings*.

While the model required multiple tries to get a good sample, GPT-2 generated "good" results based on "how familiar the model is with the context," the researchers wrote. "When prompted with topics that are highly represented in the data (Brexit, Miley Cyrus, *Lord of the Rings*, and so on), it seems to be capable of generating reasonable samples about 50 percent of the time. The opposite is also true: on highly technical or esoteric types of content, the model can perform poorly."

There were some weak spots encountered in GPT-2's word modeling—for example, the researchers noted it sometimes "writes about fires happening under water." But the model

could be fine-tuned to specific tasks and perform much better. “We can fine-tune GPT-2 on the Amazon Reviews dataset and use this to let us write reviews conditioned on things like star rating and category,” the authors explained.

That kind of performance would raise all sorts of concerns about how things like fake reviews—and pretty much any other sort of commenting—could be “deepfaked” for economic or political reasons. And that’s precisely why OpenAI’s researchers are holding off on publishing a more complete version of their model for now.

https://arstechnica.com/?p=1457603