TechXplore

As AI language skills grow, so do scientists' concerns

July 18 2022, by MATT O'BRIEN

Research engineer Teven Le Scao, who helped create the new artificial intelligence language model called BLOOM, poses for a photo, Monday, July 11, 2022, in New York. Credit: AP Photo/Mary Altaffer

The tech industry's latest artificial intelligence constructs can be pretty convincing if you ask them what it feels like to be a sentient computer, or maybe just a dinosaur or squirrel. But they're not so good—and sometimes dangerously bad—at handling other seemingly straightforward tasks.

Take, for instance, GPT-3, a Microsoft-controlled system that can generate paragraphs of human-like text based on what it's learned from a vast database of digital books and online writings. It's considered one of the most advanced of a new generation of AI algorithms that can converse, generate readable text on demand and even produce novel images and video.

Among other things, GPT-3 can write up most any text you ask for—a cover letter for a zookeeping job, say, or a Shakespearean-style sonnet set on Mars. But when Pomona College professor Gary Smith asked it a simple but nonsensical question about walking upstairs, GPT-3 muffed it.

"Yes, it is safe to walk upstairs on your hands if you wash them first," the AI replied.

These powerful and power-chugging AI systems, technically known as "large language models" because they've been trained on a huge body of text and other media, are already getting baked into customer service chatbots, Google searches and "auto-complete" email features that finish your sentences for you. But most of the tech companies that built them have been secretive about their inner workings, making it hard for outsiders to understand the flaws that can make them a source of misinformation, racism and other harms.

"They're very good at writing text with the proficiency of human beings," said Teven Le Scao, a research engineer at the AI startup Hugging Face. "Something they're not very good at is being factual. It looks very coherent. It's almost true. But it's often wrong."

That's one reason a coalition of AI researchers co-led by Le Scao —-with help from the French government—launched a new large language model Tuesday that's supposed to serve as an antidote to closed systems such as GPT-3. The group is called BigScience and their model is BLOOM, for the BigScience Large Open-science Open-access Multilingual Language Model. Its main breakthrough is that it works across 46 languages, including Arabic, Spanish and French—unlike most systems that are focused on English or Chinese.

It's not just Le Scao's group aiming to open up the black box of AI language models. Big Tech company Meta, the parent of Facebook and Instagram, is also calling for a more open approach as it tries to catch up to the systems built by Google and OpenAI, the company that runs GPT-3.

Research engineer Teven Le Scao, who helped create the new artificial intelligence language model called BLOOM, poses for a photo, Monday, July 11, 2022, in New York. Credit: AP Photo/Mary Altaffer

"We've seen announcement after announcement after announcement of people doing this kind of work, but with very little transparency, very little ability for people to really look under the hood and peek into how these models work," said Joelle Pineau, managing director of Meta AI.

Competitive pressure to build the most eloquent or informative system—and profit from its applications—is one of the reasons that most tech companies keep a tight lid on them and don't collaborate on community norms, said Percy Liang, an associate computer science professor at Stanford who directs its Center for Research on Foundation

Models.

"For some companies this is their secret sauce," Liang said. But they are often also worried that losing control could lead to irresponsible uses. As AI systems are increasingly able to write health advice websites, high school term papers or political screeds, misinformation can proliferate and it will get harder to know what's coming from a human or a computer.

Meta recently launched a new language model called OPT-175B that uses publicly available data—from heated commentary on Reddit forums to the archive of U.S. patent records and a trove of emails from the Enron corporate scandal. Meta says its openness about the data, code and research logbooks makes it easier for outside researchers to help identify and mitigate the bias and toxicity that it picks up by ingesting how real people write and communicate.

"It is hard to do this. We are opening ourselves for huge criticism. We know the model will say things we won't be proud of," Pineau said.

While most companies have set their own internal AI safeguards, Liang said what's needed are broader community standards to guide research and decisions such as when to release a new model into the wild.

It doesn't help that these models require so much computing power that only giant corporations and governments can afford them. BigScience, for instance, was able to train its models because it was offered access to France's powerful Jean Zay supercomputer near Paris.

The trend for ever-bigger, ever-smarter AI language models that could be "pre-trained" on a wide body of writings took a big leap in 2018 when Google introduced a system known as BERT that uses a so-called "transformer" technique that compares words across a sentence to predict meaning and context. But what really impressed the AI world was GPT-3, released by San Francisco-based startup OpenAI in 2020 and soon after exclusively licensed by Microsoft.

Research engineer Teven Le Scao, who helped create the new artificial intelligence language model called BLOOM, poses for a photo, Monday, July 11, 2022, in New York. Credit: AP Photo/Mary Altaffer

GPT-3 led to a boom in creative experimentation as AI researchers with paid access used it as a sandbox to gauge its performance—though without important information about the data it was trained on.

OpenAI has broadly described its training sources in a research paper, and has also publicly reported its efforts to grapple with potential abuses of the technology. But BigScience co-leader Thomas Wolf said it doesn't provide details about how it filters that data, or give access to the processed version to outside researchers.

"So we can't actually examine the data that went into the GPT-3 training," said Wolf, who is also a chief science officer at Hugging Face. "The core of this recent wave of AI tech is much more in the dataset than the models. The most important ingredient is data and OpenAI is very, very secretive about the data they use."

Wolf said that opening up the datasets used for language models helps humans better understand their biases. A multilingual model trained in Arabic is far less likely to spit out offensive remarks or misunderstandings about Islam than one that's only trained on English-language text in the U.S., he said.

One of the newest AI experimental models on the scene is Google's LaMDA, which also incorporates speech and is so impressive at responding to conversational questions that one Google engineer argued it was approaching consciousness—a claim that got him suspended from his job last month.

Colorado-based researcher Janelle Shane, author of the AI Weirdness blog, has spent the past few years creatively testing these models, especially GPT-3—often to humorous effect. But to point out the absurdity of thinking these systems are self-aware, she recently instructed it to be an advanced AI but one which is secretly a Tyrannosaurus rex or a squirrel.

"It is very exciting being a squirrel. I get to run and jump and play all day. I also get to eat a lot of food, which is great," GPT-3 said, after Shane asked it for a transcript of an interview and posed some questions.

Shane has learned more about its strengths, such as its ease at summarizing what's been said around the internet about a topic, and its weaknesses, including its lack of reasoning skills, the difficulty of sticking with an idea across multiple sentences and a propensity for being offensive.

"I wouldn't want a text model dispensing medical advice or acting as a companion," she said. "It's good at that surface appearance of meaning if you are not reading closely. It's like listening to a lecture as you're falling asleep."

© 2022 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed without permission.

Citation: As AI language skills grow, so do scientists' concerns (2022, July 18) retrieved 5 May 2024 from https://techxplore.com/news/2022-07-ai-language-skills-scientists.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.