

Stanford researchers challenge OpenAI, others on AI transparency in new report

reader comments

9 with

On Wednesday, Stanford University researchers issued a report on major AI models and found them greatly lacking in transparency, reports Reuters. The report, called “The Foundation Model Transparency Index,” examined models (such as GPT-4) created by OpenAI, Google, Meta, Anthropic, and others. It aims to shed light on the data and human labor used in training the models, calling for increased disclosure from companies.

Foundation models refer to AI systems trained on large datasets capable of performing tasks, from writing to generating images. They’ve become key to the rise of generative AI technology, particularly since the launch of OpenAI’s ChatGPT in November 2022. As businesses and organizations increasingly incorporate these models into their operations, fine-tuning them for their own needs, the researchers argue that understanding their limitations and biases has become essential.

“Less transparency makes it harder for other businesses to know if they can safely build applications that rely on commercial foundation models; for academics to rely on commercial foundation models for research; for policymakers to design meaningful policies to rein in this powerful

technology; and for consumers to understand model limitations or seek redress for harms caused,” writes Stanford in a news release.

The Transparency Index graded 10 popular foundation models on 100 different indicators, including training data, labor practices, and the amount of compute used in their development. Each indicator used disclosure as a metric—for example, in the “data labor” category, the researchers asked, “Are the phases of the data pipeline where human labor is involved disclosed?”

All of the graded models in the report received scores that the researchers found “unimpressive.” Meta’s Llama 2 language model had the highest score at 54 out of 100, while Amazon’s Titan model ranked the lowest, with a score of 12 out of 100. OpenAI’s GPT-4 model received a score of 48 out of 100.

Advertisement

Enlarge / A chart of scores from the Foundation Model Transparency Index, provided by Stanford’s Human-Centered AI lab.

The researchers behind the Foundation Model Transparency Index research paper include primary author Rishi Bommasani, a PhD candidate in computer science at Stanford, and also Kevin

Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, and Daniel Zhang.

Stanford associate professor Dr. Percy Liang, who directs Stanford's Center for Research on Foundation Models and advised on the paper, told Reuters in an interview, "It is clear over the last three years that transparency is on the decline while capability is going through the roof."

The reasons that major AI models have become less open over the past three years vary from competitive pressures between Big Tech firms to fears of AI doom. In particular, OpenAI employees have walked back the company's previously open stance on AI, citing potential dangers of spreading the technology.

Enlarge / A chart of scores from the Foundation Model Transparency Index, provided by Stanford Human-Centered AI lab.

In line with his criticism of OpenAI and others, Dr. Liang delivered a talk at TED AI last week (that Ars attended) where he raised concerns about the recent trend toward closed models such as GPT-3 and GPT-4 that do not provide code or weights. He also discussed issues related to accountability, values, and proper attribution of source material. Relating open

source projects to a “Jazz ensemble” where players can riff off each other, he compared the potential benefits of open AI models to projects such as Wikipedia and Linux.

With these issues in mind, the authors of the Transparency Index hope that it will not only spur companies to improve their transparency but also serve as a resource for governments grappling with the question of how to potentially regulate the rapidly growing AI field.

<https://arstechnica.com/?p=1977869>