Elon Musk Announces Grok, a ‘Rebellious’ AI With Few Guardrails

xAI, Elon Musk’s new company, claims to have built a powerful language model with cutting-edge performance in just two months.

By Will Knight

Nov 06, 2023 11:24 AM · 3 min. read · View original

Large language models have proven stunningly capable over the past year or so, as highlighted most famously by OpenAI’s groundbreaking chatbot, ChatGPT.

These models feed on huge amounts of text taken from books and the web, and then generate text in response to a prompt. They are typically also given further training by humans to make them less likely to produce offensive, rude, or dangerous outputs, and to make them more likely to answer questions in ways that seem coherent and plausibly correct, although they are still prone to producing errors and biases.

The language models developed by OpenAI, Google, and startups like Anthropic, Cohere, and Inflection AI typically refuse to, for example, offer advice on how to commit crimes, and will demure when asked for racy material.

It is unclear from the xAI announcement whether Grok has been trained to be more open to requests deemed inappropriate by other models, or whether it simply has not received the same kind of secondary training.

xAI posted the results of several benchmark tests designed to gauge the capabilities of large language models. Andrei Barbu, a research scientist at MIT, says the results seem similar to other popular models.

xAI says that Grok has so far been tested by a small number of users but will now be made available to a wider group of people who apply for access. Musk said in a post on X that the model would be made available to all X Premium+ subscribers. xAI has not said that it will release any models publicly.

The announcement for xAI says that the company is working on several key challenges involved in advancing AI, including building models that can assess the reliability of their own output and ask for assistance when necessary, and making models that are more robust to "adversarial attacks" designed to make them misbehave. It states: "we will work

towards developing reliable safeguards against catastrophic forms of malicious use.”

Musk was an early investor in generative AI. The billionaire helped OpenAI get its start by investing between $50 and $100 million in the company in 2015. He pulled his support for the company (which at the time was a nonprofit) in 2018 after failing to take control of it.

After OpenAI changed from a nonprofit to a for-profit business and accepted investment from Microsoft—and following ChatGPT’s runaway success—the world’s richest man openly criticized OpenAI and accused its language models of being overly “woke.”

Musk then announced in July 2023 that he had put together a small but well-respected team of AI researchers to develop “less biased" forms of AI.

Some AI researchers have already tried building language models with a more diverse range of political opinions. OpenAI has also said that it will work to remove political biases from its models.

In the year since Musk took control of X, the platform has reinstated a number of controversial users, including those from the far right, in line with its new owner’s stated opposition to moderating content on social media. Multiple studies have found that

disinformation has increased on the platform since Musk took over.