“A really big deal”—Dolly is a free, open source, ChatGPT-style AI model

Databricks

reader comments

97 with

Share this story

On Wednesday, Databricks released Dolly 2.0, reportedly the first open source, instruction-following large language model (LLM) for commercial use that has been fine-tuned on a human-generated data set. It could serve as a compelling starting point for homebrew ChatGPT competitors.

Databricks is an American enterprise software company founded in 2013 by the creators of Apache Spark. They provide a web-based platform for working with Spark for big data and machine learning. By releasing Dolly, Databricks hopes to allow organizations to create and customize LLMs “without paying for API access or sharing data with third parties,” according to the Dolly launch blog post.

Dolly 2.0, its new 12 billion-parameter model, is based on EleutherAI’s pythia model family and exclusively fine-tuned on training data (called “databricks-dolly-15k”) crowdsourced from Databricks employees. That calibration gives it abilities more in line with OpenAI’s ChatGPT, which is better at answering questions and engaging in dialogue as a chatbot than a raw LLM that has not been fine-tuned.

Dolly 1.0, released in March, faced limitations regarding commercial use due to the training data, which contained output from ChatGPT (thanks to Alpaca) and was subject to OpenAI’s terms of service. To address this issue, the team at Databricks sought to create a new data set that would allow commercial use.

To do so, Databricks crowdsourced 13,000 demonstrations of instruction-following behavior from more than 5,000 of its employees between March and April 2023. To incentivize participation, they set up a contest and outlined seven specific tasks for data generation, including open Q&A, closed Q&A, extracting and summarizing information from Wikipedia, brainstorming, classification, and creative writing.

Advertisement

The resulting data set, along with Dolly’s model weights and training code, have been released fully open source under a Creative Commons license, enabling anyone to use, modify, or extend the data set for any purpose, including commercial applications.

By contrast, OpenAI’s ChatGPT is a proprietary model that requires users to pay for API access and adhere to specific terms of service, potentially limiting the flexibility and customization options for businesses and organizations. Meta’s LLaMA, a partially open source model (with restricted weights) that recently spawned a wave of derivatives after its weights leaked on BitTorrent, does not allow commercial use.

On Mastodon, AI researcher Simon Willison called Dolly 2.0 “a really big deal.” Willison often experiments with open source language models, including Dolly. “One of the most exciting things about Dolly 2.0 is the fine-tuning instruction set, which was hand-built by 5,000 Databricks employees and released under a CC license,” Willison wrote in a Mastodon toot.

If the enthusiastic reaction to Meta’s only partially open LLaMA model is any indication, Dolly 2.0 could potentially spark a new wave of open source language models that aren’t hampered by proprietary limitations or restrictions on commercial use. While the word is still out about Dolly’s actual performance ability, further refinements might allow running reasonably powerful LLMs on local consumer-class machines.

“Even if Dolly 2 isn’t good, I expect we’ll see a bunch of new projects using that training data soon,” Willison told Ars. “And some of those might produce something really useful.”

Currently, the Dolly weights are available at Hugging Face, and the databricks-dolly-15k data set can be found on GitHub.

https://arstechnica.com/?p=1931693