Dolly

> Democratizing the magic of ChatGPT with open models

Today we are introducing Dolly, a cheap-to-build LLM that exhibits a surprising degree of the instruction following capabilities exhibited by ChatGPT. Whereas the work from the Alpaca team showed that state-of-the-art models could be coaxed into high quality instruction-following behavior, we find that even years-old open source models with much earlier architectures exhibit striking behaviors when fine tuned on a small corpus of instruction training data. Dolly works by taking an existing open source 6 billion parameter model from EleutherAI and modifying it ever so slightly to elicit instruction following capabilities such as brainstorming and text generation not present in the original model, using data from Alpaca.

The model underlying Dolly only has 6 billion parameters, compared to 175 billion in GPT-3, and is two years old, making it particularly surprising that it works so well. This suggests that much of the qualitative gains in state-of-the-art models like ChatGPT may owe to focused corpuses of instruction-following training data, rather than larger or better-tuned base models. We’ re calling the model Dolly .after Dolly the sheep, the first cloned mammal — because it's an open source clone of an Alpaca, inspired by a LLaMA.(以第一只克隆哺乳动物多莉的名字命名 —— 因为它是羊驼的开源克隆，灵感来自美洲驼。)

Dolly2

the first open source, instruction-following LLM, fine-tuned on a human-generated instruction dataset licensed for **research and commercial use.**

Dolly 2.0 is a 1**2B parameter** language model based on the EleutherAI pythia model family and fine-tuned exclusively on a new, high-quality **human generated instruction following dataset(dolly 15k)**, crowdsourced among Databricks employees.

databricks–dolly–15k **dataset**