



Categories



Hello Dolly: Democratizing the magic of ChatGPT with open models



by Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell and Matei Zaharia

March 24, 2023 in [Company Blog](#)

Share this post



Summary

We show that anyone can take a dated off-the-shelf open source large language model (LLM) and give it magical ChatGPT-like instruction following ability by training it in 30 minutes on one machine, using high-quality training data. Surprisingly, instruction-following does not seem to require the latest or largest models: our model is only 6 billion parameters, compared to 175 billion for GPT-3. We open source the code for our model (Dolly) and show how it can be re-created on Databricks. We believe models like Dolly will help democratize LLMs, transforming them from something very few companies can afford into a commodity every company can own and customize to improve their products.

Background

ChatGPT, a proprietary instruction-following model, was **released** in November 2022 and took the world by storm. The model was trained on trillions of words from the web, requiring massive numbers of GPUs to develop. This quickly led to Google and other companies releasing their own proprietary instruction-following models. In February 2023, Meta released the weights for a set of high-quality (but not instruction-following) language models called **LLaMA** to academic researchers, trained for

over 80,000 GPU-hours each. Then, in March, Stanford built the **Alpaca** model, which was based on LLaMA, but tuned on a small dataset of 50,000 human-like questions and answers that, surprisingly, made it exhibit ChatGPT-like interactivity.

Introducing Dolly

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. We’re in the earliest days of the democratization of AI for the enterprise, and much work remains to be done, but we believe the technology underlying Dolly represents an exciting new opportunity for companies that want to cheaply build their own instruction-following models.

We evaluated Dolly on the instruction-following capabilities described in the **InstructGPT** paper that ChatGPT is based on and found that it exhibits many of the same qualitative capabilities, including text generation, brainstorming and open Q&A. Of particular note in these examples is not the quality of the generated text, but rather the vast improvement in instruction-following capability that results from fine tuning a years-old open source model on a small, high quality dataset.

Generation

Original Model	Dolly
Write a love letter to Edgar Allan Poe.	
Original Model: HISTORY: Edgar Allan Poe rose to fame in the 19th Century. He was often compared to the great Romantic poets and is considered the father of the modern short story and Poe is thought of as	Dolly: Dearest Edgar Allan Poe, I met you first in the dimly lit stairwell of the old house, and I know that the moment I saw you, I never wanted to see you again. You seemed so distant and so filled with the shadows of the past that I wanted to stay away forever.

Original Model	Dolly
<p>the founder of American detective fiction. He has been dead for 130 years.</p> <p>PATRICK CAMPOS: You should never try to write a love letter to a dead person.</p> <p>PATRICK CAMPOS: Poe was more than just the author of The Raven. He was a master of atmosphere.</p> <p>BUDDY FRITZLER: Edgar Allan Poe lived in Baltimore from 1831 until his death in 1849. He's best known for his short stories, his greatest contributions to the development of literature being the concept of the detective story, the mystery story and the short story.</p> <p>BUDDY FRITZLER: Poe's most famous story, "The Gold-Bug", was written in 1843. It's about a treasure hunter who searches for buried treasure. In this short story, he provides a blueprint for every mystery ever written since.</p> <p>BUDDY FRITZLER: The central idea that you have to look for clues is very appealing. It was even used as ...</p>	<p>But I couldn't help but be obsessed with you, and my feelings for you have only grown over the years. Even now, I feel a deep connection with you that I can never explain.</p> <p>I hope that you can find happiness in the life that you now lead. I am sure that you will do great things, and I am forever in your debt.</p> <p>Sincerely,</p> <p>[Your Name]</p>
Write a tweet announcing Dolly, a large language model from Databricks.	
<p>Original Model:</p> <p>In your tweet, provide a short summary of the challenge, how you analyzed it, and how Dolly performed.</p> <p>After your tweet, I will retweet it to Databricks' Twitter account.</p> <p>Clean up your social media page for the next week by removing any mentions of Dolly or Databricks.</p> <p>Data Preparation</p> <p>Our first step is to generate word vector embeddings for each document. We can obtain these vector representations in two different ways:</p> <p>Using a tokenized document set (in this case, everything in our download folder)</p> <p>Loads a word2vec model from the Google Search API (in this case, vector2vec.c)</p>	<p>Dolly:</p> <p>Databricks just announced Dolly – a new language model for natural language processing with large capacity and improved efficiency!</p> <p>#DataScience #NaturalLanguageProcessing</p> <p>[Link to Dolly model release]</p>

Original Model	Dolly
<p>In our case, we want to use the first method, which is easier, but the second may be more robust. We will use the word2vec model above to create a vocabulary file for any new words found in the file. This process gives us a quick way to identify new words and find contextually related terms.</p> <p>We can start by copying word2vec.c to a local directory called /tmp/word2vec.</p> <pre>##/tmp/word2vec/word2vec.c FILE_NAME /tmp/word</pre>	
I'm selling my Nikon D-750, write a short blurb for my ad.	
<p>Original Model:</p> <p>I'm not sure if you can post a location - anything on the website will work.</p> <p>I've only had it a week. It is in mint condition with the charging cable and window case intact. Bought it for \$1800 for in store pickup at a store in San Francisco. Plus tax and the undisclosed Costco price.</p> <p>Some one please let me know that this is a scam and not really an offer.</p> <p>I just received and e-mail that said the camera is at someones house so someone else can pick it up and did I want to message them, well no I'm selling it as I would never ...</p>	<p>Dolly:</p> <p>Looking to get a reliable camera for photo and video work? Check out the Nikon D-750! This sleek, lightweight camera has a 24.2MP sensor and 1080/60p video recording, making it a great choice for capturing those moments in life. The ergonomic grip and nine-point autofocus system make it easy to work with and shoot without hassle. With Wi-Fi, you can upload your pictures easily to your computer or smartphone without having to buy a dedicated external storage. Buy the Nikon D-750 now to experience why it is the best choice for photography and videography!</p>

Open Q&A

Original Model	Dolly
Explain to me the difference between nuclear fission and fusion.	
<p>Original Model:</p> <p>The first thing that strikes one is the similarity in appearance and function between the atom and the sun. The central body in the sun is the source of energy, and the entire structure of the sun is made up of a million tons of the hydrogen gas that's so important to us. It's what all hydrogen fuses into in the sun, and that's why we are made of it: fusion.</p>	<p>Dolly:</p> <p>The difference between nuclear fission and fusion is the method by which the energy of the nucleus is released. In nuclear fission, a single atomic nucleus splits into two smaller nuclei. This process is used to create energy through the release of a large amount of energy in the form of heat. On the other hand, fusion involves two atomic nuclei coming together to form a single nucleus. This process is used to create a much larger amount of energy compared to fission, and the energy is</p>

Original Model	Dolly
<p>The nucleus of the atom, at the center, is made of protons, and the electrons orbit around it. In the center, there is a tiny spherical mass with a lot of protons (and therefore a lot of positive charge) called the nucleus. The electrons aren't so tiny, but their charge is negative, as a positive nucleus and negative electrons make for stability (due to charge balance). Protons are positively charged, so even though they themselves can't get anywhere, they do repel electrons. There is a core of neutrons, that don't have any charge, but are held there by the protons. Protons also repel neutrons.</p> <p>This simple fact is the basic form of our sun, and it is what makes up all the elements in the earth. Hydrogen is the most plentiful element in ...</p>	<p>released through the release of light and heat. Fusion has the potential to be a much more efficient way of creating energy. However, it is much ...</p>

Brainstorming

Original Model	Dolly
Give me a list of 5 science fiction books I should read next.	
<p>Original Model:</p> <p>I know it's early, and I should feel ashamed of myself for even thinking of it, but I have a terrible tendency to procrastinate and read only when I absolutely have to.</p> <p>As a result, I have not read anything this year, in any genre.</p> <p>As soon as I have a set of ideas for an article, I let the ideas and outline do the work of making sure I don't start writing. There is something special about a blank page and an empty mind.</p> <p>That said, I need to change that, and reading the above question made me realize just how badly I want</p>	<p>Dolly:</p> <ol style="list-style-type: none"> 1. 2001: A Space Odyssey by Arthur C. Clarke 2. The Hitchhiker's Guide to the Galaxy by Douglas Adams 3. The Lord of the Rings by J.R.R. Tolkien 4. The Hitchhiker's Guide to the Galaxy by Douglas Adams 5. 1984 by George Orwell.

Why Open Models?

There are many reasons a company would prefer to build their own model rather than sending data to a centralized LLM provider that serves a proprietary model behind an API. For many companies, the problems and datasets most likely to benefit from AI represent their most sensitive and proprietary intellectual property, and handing it over to a third party may be unpalatable. Furthermore, organizations may have different tradeoffs in terms of model quality, cost, and desired behavior. We believe that most ML users are best served long term by directly owning their models.

We are open sourcing a [simple Databricks notebook](#) that you can use to build Dolly yourself on Databricks. Contact us at hello-dolly@databricks.com if you would like to get access to the trained weights.

What's Next?

The release of Dolly is the first in a series of announcements Databricks is making that focus on helping every organization harness the power of large language models. We believe in the incredible power of artificial intelligence to transform the productivity of every organization and individual, and welcome you to join us on this journey. Stay tuned for more in this area in the coming weeks!

Acknowledgments

This work owes much to the efforts and insights of many incredible organizations. This would have been impossible without [EleutherAI](#) open sourcing and training GPT-J. We are inspired by the incredible ideas and data from the [Stanford Center for Research on Foundation Models](#) and specifically the team behind [Alpaca](#). The core idea behind the outsized power of small dataset is thanks to the original paper on [Self-Instruct](#). We are also thankful to [Hugging Face](#) for hosting, open sourcing, and maintaining countless models and libraries; their contribution to the state of the art cannot be overstated.

Disclaimer: Generative AI is an emerging technology and we're in the early stages of research around how to address factual accuracy, bias, offensive responses, general toxicity, and hallucinations in LLMs. Dolly, like other language models, can sometimes exhibit these behaviors and we urge our users to exercise good judgment in designing applications of this technology.

Try Databricks for free

Get Started

Related posts



Announcing General Availability of Databricks Model Serving

March 7, 2023 by [Patrick Wendell](#), [Aaron Davidson](#), [Sue Ann Hong](#), [Kasey Uhlenhuth](#), [Ahmed Bilal](#) and [Josh Hartman](#) in [Platform Blog](#)

ML Virtual Event Enabling Production ML at Scale With Lakehouse
March 14, 9 AM PDT / 4 PM GMT Register Now We are...



Databricks ❤️ IDEs

February 14, 2023 by [Patrick Wendell](#) in [Platform Blog](#)

Happy Valentine's Day! Databricks ❤️ Visual Studio Code. On this lovely day, we are thrilled to announce a new and powerful development experience for...



Finding a Data Platform that Can Do More, With Less

March 8, 2023 by [Isaac Gritz](#), [Andrey Mirskiy](#), [Franco Patano](#), [Pouneh Partowkia](#) and [Katie Cummiskey](#) in [Solutions](#)

In today's economy, the key phrase is "do more with less." Doing more with less is not just about reducing infrastructure cost, but...

[See all Company Blog posts](#)

[Product](#)

[Learn & Support](#)

[Solutions](#)

[Company](#)





Databricks Inc.

160 Spear Street, 13th Floor

San Francisco, CA 94105

1-866-330-0121

© Databricks 2023. All rights reserved. Apache, Apache Spark, Spark and the Spark logo are trademarks of the Apache Software Foundation.

[Privacy Notice \(Updated\)](#) | [Terms of Use](#) | [Your Privacy Choices](#) | [Your California Privacy Rights](#) 