List of Open Sourced Fine-Tuned Large Language Models (LLM)

An incomplete list of open-sourced fine-tuned Large Language Models (LLM) you can run locally on your computer

Sung Kim · Follow
Published in Geek Culture
24 min read · Mar 30

Photo by Liudmila Shuvalova on Unsplash

This is an incomplete list of open-sourced fine-tuned Large Language Models (LLMs) that runs on your local computer, and my attempt to maintain a list since as many as three models are announced on a daily basis.

*I haven't listed them all because you can literally create these models for less than $100. Cabrita, which is one of the models listed here was created for $8 — I find it hard to believe. I am still thinking about whether or not I should create BritneyGPT, but I did create the training dataset for about $20, and it would cost me an additional $50 to use GPU services. I have even thought about the name for the article — "It's BritneyGPT, B******!"*

According to the documentation, you can run these models on a PC with different levels of hardware. For most people, your best bet is **llama.cpp** since it supports seven

models and runs on moderately specced PCs:

LLaMA | *Alpaca* | *GPT4All* | *Chinese LLaMA/Alpaca* | *Vigogne (French)* | *Vicuna* | *Koala* | *OpenBuddy (Multilingual)*

The list is a work in progress where I tried to group them by the Foundation Models where they are:

BigCode's StarCoder | *BigScience's BLOOM* | *Cerebras' Cerebras-GPT* | *EleutherAI's GPT-J, GPT-NeoX, Polyglot, and Pythia* | *GLM* | *Google's Flamingo, FLAN, and PaLM* | *H2O.ai's h2ogpt* | *Meta's GALACTICA, LLaMA, and XGLM* | *Mosaic ML's MPT* | *Nvidia's NeMo* | *OpenLLaMA* | *Replit's Code* | *RWKV* | *StabilityAI's StableLM* | *TII's Falcon LLM* | *Together's RedPajama-INCITE*

They are subgrouped by the list of projects that are reproductions of or based on those Foundation Models.

Updates:

- 03/2023: Added HuggingGPT | Vicuna/FastChat
- 04/2023: Added "A Survey of Large Language Models" | "LLMMaps — A Visual Metaphor for Stratified Evaluation of Large Language Models" | Baize | Koala | Segment Anything | Galpaca | GPT-J-6B instruction-tuned on Alpaca-GPT4 | GPTQ-for-LLaMA | List of all Foundation Models | Dolly 2.0 | StackLLaMA | GPT4All-J | Palmyra Base 5B | Camel 🐫 5B | StableLM | h2oGPT | The Bloke alpaca-lora-65B-GGML | OpenAssistant Models | StableVicuna | FastChat-T5 | couchpotato888 | GPT4-x-Alpaca | LLaMA Adapter V2 | WizardLM, | A brief history of LLaMA models (Resources section)
- 05/2023: Added OpenLLaMA | BigCode StarCoder (Hugging Face + ServiceNow) | Replit-Code (Replit) | Pygmalion-7b | AlpacaGPT4-LoRA-7B-OpenLLaMA | Nvidia GPT-2B-001 | The Bloke's StableVicuna-13B-GPTQ | OpenAlpaca | crumb's Hugging Face website | Teknium's Hugging Face website | Knut Jägersberg's Hugging Face website | SemiAnalysis article by Luke Sernau (a senior software engineer at Google) | Mosaic ML's MPT-7B | gpt4-x-vicuna-13b | LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions | Vigogne | Chinese-LLaMA-Alpaca | OpenBuddy — Open Multilingual Chatbot for Everyone | Chatbot Arena | Together's RedPajama-INCITE 3B and 7B | Ahead of AI #8: The Latest Open Source LLMs and Datasets (Resources section) | PaLM (Concept of Mind) | digitous Hugging Face website | Hugging Face's Open LLM Leaderboard | A'eala's Hugging Face website | chavinlo's Hugging Face website | eachadea's Hugging Face website | chainyo's Hugging Face website | KoboldAI's Hugging Face website | Baize V2 | Gorilla (POET?) | QLoRA | TII's Falcon LLM | ausboss' Hugging Face website | Metal (MetaIX)'s Hugging Face website

LLaMA (Meta)

> *Stanford Alpaca: An Instruction-following LLaMA Model.*

- LLaMA Website: Introducing LLaMA: A foundational, 65-billion-parameter language model (facebook.com)
- Alpaca Website: https://crfm.stanford.edu/2023/03/13/alpaca.html
- Alpaca GitHub: https://github.com/tatsu-lab/stanford_alpaca
- Commercial Use: No

Here is a list of reproductions of or based on Meta's LLaMA or Stanford Alpaca project:

Alpaca.cpp | *Alpaca-LoRA* | *AlpacaGPT4-LoRA-7B-OpenLLaMA* | *Baize* | *Cabrita* | *Chinese-LLaMA-Alpaca* | *Chinese-Vicuna* | *Gorilla (POET?)* | *GPT4-x-Alpaca* | *gpt4-x-vicuna-13b* | *GPT4All* | *GPTQ-for-LLaMA* | *Koala* | *llama.cpp* | *LLaMA-Adapter V2* | *Lit-LLaMA* | *OpenAlpaca* | *OpenBuddy — Open Multilingual Chatbot for Everyone* | *Pygmalion-7b* | *QLoRA* | *StackLLaMA* | *StableVicuna* | *The Bloke alpaca-lora-65B-GGML/StableVicuna-13B-GPTQ/WizardLM-7B-uncensored-GPTQ* | *Vicuna* | *Vigogne* | *WizardLM*

Alpaca.cpp

> *Run a fast ChatGPT-like model locally on your device. The screencast below is not sped up and running on an M2 Macbook Air with 4GB of weights.*

- GitHub: antimatter15/alpaca.cpp: Locally run an Instruction-Tuned Chat-Style LLM (github.com)

Alpaca-LoRA

> *This repository contains code for reproducing the Stanford Alpaca results using low-rank adaptation (LoRA). We provide an Instruct model of similar quality to* `text-davinci-003` *that can run on a Raspberry Pi (for research), and the code is easily extended to the* `13b`*,* `30b`*, and* `65b` *models.*

- GitHub: tloen/alpaca-lora: Instruct-tune LLaMA on consumer hardware (github.com)
- Demo: Alpaca-LoRA — a Hugging Face Space by tloen

AlpacaGPT4-LoRA-7B-OpenLLaMA

- Hugging Face: https://huggingface.co/LLMs
- LLMs Models: https://huggingface.co/LLMs

Baize V2

> *Baize V2 is an open-source chat model fine-tuned with LoRA. It uses 100k dialogs generated by letting ChatGPT chat with itself. We also use Alpaca's data to improve its performance. We have released 7B, and 13B models.*

- GitHub: project-baize/baize: Baize is an open-source chatbot trained with ChatGPT self-chatting data, developed by researchers at UCSD and Sun Yat-sen University. (github.com)
- Paper: 2304.01196.pdf (arxiv.org)

Cabrita

> *A portuguese finetuned instruction LLaMA*

- GitHub: https://github.com/22-hours/cabrita

Chinese-LLaMA-Alpaca

> *In order to promote the open research of large models in the Chinese NLP community, this project open sourced the Chinese LLaMA model and the Alpaca large model with fine-tuned instructions. Based on the original LLaMA, these models expand the Chinese vocabulary and use Chinese data for secondary pre-training, which further improves the basic semantic understanding of Chinese. At the same time, the Chinese Alpaca model further uses Chinese instruction data for fine-tuning, which significantly improves the model's ability to understand and execute instructions. For details, please refer to the technical report (Cui, Yang, and Yao, 2023).*

- GitHub: https://github.com/ymcui/Chinese-LLaMA-Alpaca

Chinese-Vicuna

> *A Chinese Instruction-following LLaMA-based Model*

- GitHub: Facico/Chinese-Vicuna: Chinese-Vicuna: A Chinese Instruction-following LLaMA-based Model — — 一个中文低资源的llama+lora方案，结构参考alpaca (github.com)

Gorilla (POET?)

> `POET` *enables the training of state-of-the-art memory-hungry ML models on smartphones and other edge devices. POET (Private Optimal Energy Training) exploits the twin techniques of integrated tensor rematerialization, and paging-in/out of secondary storage (as detailed in our paper at ICML 2022) to optimize models for training with limited memory. POET's Mixed Integer Linear Formulation (MILP) ensures the solutions are provably optimal!*
>
> *With POET, we are the first to demonstrate how to train memory-hungry SOTA ML models such as BERT and ResNets on smartphones and tiny ARM Cortex-M devices*

> 💪

- Website: Gorilla (berkeley.edu)
- GitHub: ShishirPatil/poet: ML model training for edge devices (github.com)

GPT4-x-Alpaca

> *GPT4-x-Alpaca is a LLaMA 13B model fine-tuned with a collection of GPT4 conversations, GPTeacher. There's not a lot of information on its training and performance.*

- Hugging Face: chavinlo/gpt4-x-alpaca · Hugging Face

gpt4-x-vicuna-13b

> *As a base model used https://huggingface.co/eachadea/vicuna-13b-1.1. Finetuned on Teknium's GPTeacher dataset, unreleased Roleplay v2 dataset, GPT-4-LLM dataset, and Nous Research Instruct Dataset. Approx 180k instructions, all from GPT-4, all cleaned of any OpenAI censorship/"As an AI Language Model" etc.*

- Hugging Face: NousResearch/gpt4-x-vicuna-13b · Hugging Face

GPT4All

> *Demo, data and code to train an assistant-style large language model with ~800k GPT-3.5-Turbo Generations based on LLaMa.*

- GitHub: nomic-ai/gpt4all: gpt4all: a chatbot trained on a massive collection of clean assistant data including code, stories and dialogue (github.com)
- GitHub: nomic-ai/pyllamacpp: Official supported Python bindings for llama.cpp + gpt4all (github.com)
- Review: Is GPT4All your new personal ChatGPT? — YouTube

GPTQ-for-LLaMA

> *4 bits quantization of LLaMA using GPTQ. GPTQ is SOTA one-shot weight quantization method.*

- GitHub: qwopqwop200/GPTQ-for-LLaMa: 4 bits quantization of LLaMA using GPTQ (github.com)

Koala

> *Koala is a language model fine-tuned on top of LLaMA. Check out the blogpost! This documentation will describe the process of downloading, recovering the Koala model weights, and running the Koala chatbot locally.*

- Blog: Koala: A Dialogue Model for Academic Research — The Berkeley Artificial Intelligence Research Blog
- GitHub: EasyLM/koala.md at main · young-geng/EasyLM (github.com)
- Demo: FastChat (lmsys.org)
- Review: Investigating Koala a ChatGPT style Dialogue Model — YouTube
- Review: Running Koala for free in Colab. Your own personal ChatGPT? — YouTube

llama.cpp

> *Inference of LLaMA model in pure C/C++*

- GitHub: ggerganov/llama.cpp: Port of Facebook's LLaMA model in C/C++ (github.com)
- Supports three models: LLaMA, Alpaca, and GPT4All

LLaMA-Adapter V2

> *Official implementation of 'LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention' and 'LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model'.*

- GitHub: ZrrSkywalker/LLaMA-Adapter: Fine-tuning LLaMA to follow Instructions within 1 Hour and 1.2M Parameters (github.com)

Lit-LLaMA

> *Independent implementation of LLaMA that is fully open source under the Apache 2.0 license. This implementation builds on nanoGPT.*

- GitHub: Lightning-AI/lit-llama: Implementation of the LLaMA language model based on nanoGPT. Supports quantization, LoRA fine-tuning, pre-training. Apache 2.0-licensed. (github.com)

OpenAlpaca

> *This is the repo for the OpenAlpaca project, which aims to build and share an instruction-following model based on OpenLLaMA. We note that, following OpenLLaMA, OpenAlpaca is permissively licensed under the Apache 2.0 license. This repo contains*
>
> *- The data used for fine-tuning the model.*
> *- The code for fine-tuning the model.*
> *- The weights for the fine-tuned model.*
> *- The example usage of OpenAlpaca.*

- GitHub: yxuansu/OpenAlpaca: OpenAlpaca: A Fully Open-Source Instruction-Following Model Based On OpenLLaMA (github.com)

OpenBuddy — Open Multilingual Chatbot for Everyone

> *OpenBuddy is a powerful open-source multilingual chatbot model aimed at global users, emphasizing conversational AI and seamless multilingual support for English, Chinese, and other languages. Built upon Facebook's LLAMA model, OpenBuddy is fine-tuned to include an extended vocabulary, additional common characters, and enhanced token embeddings. By leveraging these improvements and multi-turn dialogue datasets, OpenBuddy offers a robust model capable of answering questions and performing translation tasks across various languages.*

- GitHub: https://github.com/OpenBuddy/OpenBuddy

Pygmalion-7b

> *Pygmalion 7B is a dialogue model based on Meta's LLaMA-7B. This is version 1. It has been fine-tuned using a subset of the data from Pygmalion-6B-v8-pt4, for those of you familiar with the project.*

- Hugging Face: https://huggingface.co/PygmalionAI/pygmalion-7b

QLoRA

> *We present QLoRA, an efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA). Our best model family, which we name Guanaco, outperforms all previous openly released models on the Vicuna benchmark, reaching 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU. QLoRA introduces a number of innovations to save memory without sacrificing performance: (a) 4-bit NormalFloat (NF4), a new data type that is information theoretically optimal for normally distributed weights (b) Double Quantization to reduce the average memory footprint by quantizing the quantization constants, and © Paged Optimizers to manage memory spikes. We use QLoRA to finetune more than 1,000 models, providing a detailed analysis of instruction following and chatbot performance across 8 instruction datasets, multiple model types (LLaMA, T5), and model scales that would be infeasible to run with regular finetuning (e.g. 33B and 65B parameter models). Our results show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA. We provide a detailed analysis of chatbot performance based on both human and GPT-4 evaluations showing that GPT-4 evaluations are a cheap and reasonable alternative to human evaluation. Furthermore, we find that current chatbot benchmarks are not trustworthy to accurately evaluate the performance*

> *levels of chatbots. We release all of our models and code, including CUDA kernels for 4-bit training.*

GitHub: artidoro/qlora: QLoRA: Efficient Finetuning of Quantized LLMs (github.com)

StableVicuna

> *We are proud to present StableVicuna, the first large-scale open source chatbot trained via reinforced learning from human feedback (RHLF). StableVicuna is a further instruction fine tuned and RLHF trained version of Vicuna v0 13b, which is an instruction fine tuned LLaMA 13b model. For the interested reader, you can find more about Vicuna here.*

- Website: Stability AI releases StableVicuna, the AI World's First Open Source RLHF LLM Chatbot — Stability AI
- Hugging Face: StableVicuna — a Hugging Face Space by CarperAI
- Review: StableVicuna: The New King of Open ChatGPTs? — YouTube

StackLLaMA

> *A LlaMa model trained on answers and questions on Stack Exchange with RLHF through a combination of: Supervised Fine-tuning (SFT), Reward / preference modeling (RM), and Reinforcement Learning from Human Feedback (RLHF)*

Website: https://huggingface.co/blog/stackllama

The Bloke alpaca-lora-65B-GGML

> *Quantised 4bit and 2bit GGMLs of changsung's alpaca-lora-65B for CPU inference with llama.cpp.*

- Hugging Face: TheBloke/alpaca-lora-65B-GGML · Hugging Face

The Bloke's StableVicuna-13B-GPTQ

> *This repo contains 4bit GPTQ format quantised models of CarterAI's StableVicuna 13B. It is the result of first merging the deltas from the above repository with the original Llama 13B weights, then quantising to 4bit using GPTQ-for-LLaMa.*

- Hugging Face: TheBloke/stable-vicuna-13B-GPTQ · Hugging Face

The Bloke's WizardLM-7B-uncensored-GPTQ

> *These files are GPTQ 4bit model files for Eric Hartford's 'uncensored' version of WizardLM. It is the result of quantising to 4bit using GPTQ-for-LLaMa. Eric did a fresh 7B training using the WizardLM method, on a dataset edited to remove all the "I'm sorry.." type ChatGPT responses.*

- Hugging Face: TheBloke/WizardLM-7B-uncensored-GPTQ · Hugging Face

Vicuna (FastChat)

> *An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality.*

- GitHub: lm-sys/FastChat: The release repo for "Vicuna: An Open Chatbot Impressing GPT-4" (github.com)
- Review: Vicuna — 90% of ChatGPT quality by using a new dataset? — YouTube

Vigogne

> *This repository contains code for reproducing the Stanford Alpaca in French FR using low-rank adaptation (LoRA) provided by 🤗 Hugging Face's PEFT library. In addition to the LoRA technique, we also use LLM.int8() provided by bitsandbytes to quantize pretrained language models (PLMs) to int8. Combining these two techniques allows us to fine-tune PLMs on a single consumer GPU such as RTX 4090.*

GitHub: https://github.com/bofenghuang/vigogne

WizardLM

> *An Instruction-following LLM Using Evol-Instruct. Empowering Large Pre-Trained Language Models to Follow Complex Instructions*

- GitHub: nlpxucan/WizardLM: WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions (github.com)
- Review: WizardLM: Evolving Instruction Datasets to Create a Better Model — YouTube

BLOOM (BigScience)

> *BigScience Large Open-science Open-access Multilingual Language Model.*

- Hugging Face: bigscience/bloom · Hugging Face
- Hugging Face Demo: Bloom Demo — a Hugging Face Space by huggingface

Here is a list of reproductions of or based on the BLOOM project:

- BLOOM-LoRA | Petals

BLOOM-LoRA

> *Low-Rank adaptation for various Instruct-Tuning datasets.*

- GitHub: linhduongtuan/BLOOM-LORA: Due to restriction of LLaMA, we try to reimplement BLOOM-LoRA (much less restricted BLOOM license here https://huggingface.co/spaces/bigscience/license) using Alpaca-LoRA and Alpaca_data_cleaned.json (github.com)

Petals

> *Generate text using distributed 176B-parameter BLOOM or BLOOMZ and fine-tune them for your own tasks.*

- GitHub: bigscience-workshop/petals: 🌸 Run 100B+ language models at home, BitTorrent-style. Fine-tuning and inference up to 10x faster than offloading (github.com)

Cerebras-GPT (Cerebras)

> *A Family of Open, Compute-efficient, Large Language Models. Cerebras open sources seven GPT-3 models from 111 million to 13 billion parameters. Trained using the Chinchilla formula, these models set new benchmarks for accuracy and compute efficiency.*

- Website: Cerebras-GPT: A Family of Open, Compute-efficient, Large Language Models — Cerebras
- Hugging Face: cerebras (Cerebras) (huggingface.co)
- Review: Checking out the Cerebras-GPT family of models — YouTube

Falcon LLM

> *Falcon LLM is TII's flagship series of large language models, built from scratch using a custom data pipeline and distributed training library.*

- Website: tiiuae (Technology Innovation Institute) (huggingface.co)
- Hugging Face: tiiuae/falcon-40b-instruct · Hugging Face
- Hugging Face: tiiuae/falcon-7b-instruct · Hugging Face
- Review: Falcon Soars to the Top — The NEW 40B LLM Rises above the rest. — YouTube

Flamingo (Google/Deepmind)

> *Tackling multiple tasks with a single visual language model*

- Website: Tackling multiple tasks with a single visual language model

Here is a list of reproductions of or based on the Flamingo project:

- Flamingo — Pytorch | OpenFlamingo

Flamingo — Pytorch

> *Implementation of Flamingo, state-of-the-art few-shot visual question answering attention net, in Pytorch. It will include the perceiver resampler (including the scheme where the learned queries contributes keys / values to be attended to, in addition to*

> *media embeddings), the specialized masked cross attention blocks, and finally the tanh gating at the ends of the cross attention + corresponding feedforward blocks.*

- GitHub: https://github.com/lucidrains/flamingo-pytorch

OpenFlamingo

> *Welcome to our open source version of DeepMind's Flamingo model! In this repository, we provide a PyTorch implementation for training and evaluating OpenFlamingo models. We also provide an initial OpenFlamingo 9B model trained on a new Multimodal C4 dataset (coming soon). Please refer to our blog post for more details.*

- GitHub: mlfoundations/open_flamingo: An open-source framework for training large multimodal models (github.com)

FLAN (Google)

> *This repository contains code to generate instruction tuning dataset collections. The first is the original Flan 2021, documented in Finetuned Language Models are Zero-Shot Learners, and the second is the expanded version, called the Flan Collection, described in The Flan Collection: Designing Data and Methods for Effective Instruction Tuning and used to produce Flan-T5 and Flan-PaLM.*

- GitHub: google-research/FLAN (github.com)

Here is a list of reproductions of or based on the FLAN project:

- FastChat-T5 | Flan-Alpaca | Flan-UL2

FastChat-T5

> *We are excited to release FastChat-T5: our compact and commercial-friendly chatbot! that is Fine-tuned from Flan-T5, ready for commercial usage! and Outperforms Dolly-V2 with 4x fewer parameters.*

- GitHub: lm-sys/FastChat: The release repo for "Vicuna: An Open Chatbot Impressing GPT-4" (github.com)
- Hugging Face: https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/huggingface_api.py

Flan-Alpaca

> *Instruction Tuning from Humans and Machines. This repository contains code for extending the Stanford Alpaca synthetic instruction tuning to existing instruction-tuned models such as Flan-T5. The pretrained models and demos are available on HuggingFace*

- GitHub: declare-lab/flan-alpaca: This repository contains code for extending the Stanford Alpaca synthetic instruction tuning to existing instruction-tuned models such as Flan-T5. (github.com)

Flan-UL2

> *Flan-UL2 is an encoder decoder model based on the `T5` architecture. It uses the same configuration as the `UL2 model` released earlier last year. It was fine tuned using the "Flan" prompt tuning and dataset collection.*

- Hugging Face: google/flan-ul2 · Hugging Face
- Review: Trying Out Flan 20B with UL2 — Working in Colab with 8Bit Inference — YouTube

GALACTICA (Meta)

> *Following Mitchell et al. (2018), this model card provides information about the GALACTICA model, how it was trained, and the intended use cases. Full details about how the model was trained and evaluated can be found in the release paper.*

- GitHub: galai/model_card.md at main · paperswithcode/galai (github.com)

Here is a list of reproductions of or based on the GALACTICA project:

- Galpaca

Galpaca

> *GALACTICA 30B fine-tuned on the Alpaca dataset.*

- Hugging Face: GeorgiaTechResearchInstitute/galpaca-30b · Hugging Face
- Hugging Face: TheBloke/galpaca-30B-GPTQ-4bit-128g · Hugging Face

GLM (General Language Model)

> *GLM is a General Language Model pretrained with an autoregressive blank-filling objective and can be finetuned on various natural language understanding and generation tasks.*

Here is a list of reproductions of or based on the GLM project:

- ChatGLM-6B

ChatGLM-6B

> *ChatGLM-6B is an open bilingual language model based on General Language Model (GLM) framework, with 6.2 billion parameters. With the quantization technique,*

> *users can deploy locally on consumer-grade graphics cards (only 6GB of GPU memory is required at the INT4 quantization level).*
>
> *ChatGLM-6B uses technology similar to ChatGPT, optimized for Chinese QA and dialogue. The model is trained for about 1T tokens of Chinese and English corpus, supplemented by supervised fine-tuning, feedback bootstrap, and reinforcement learning wit human feedback. With only about 6.2 billion parameters, the model is able to generate answers that are in line with human preference.*

- GitHub: THUDM/ChatGLM-6B: ChatGLM-6B：开源双语对话语言模型 | An Open Bilingual Dialogue Language Model (github.com)

GPT-J (EleutherAI)

> ***GPT-J*** *is an open source artificial intelligence language model developed by EleutherAI.[1] GPT-J performs very similarly to OpenAI's GPT-3 on various zero-shot down-streaming tasks and can even outperform it on code generation tasks.[2] The newest version, GPT-J-6B is a language model based on a data set called The Pile.[3] The Pile is an open-source 825 gibibyte language modelling data set that is split into 22 smaller datasets.[4] GPT-J is similar to ChatGPT in ability, although it does not function as a chat bot, only as a text predictor.[5]*

- GitHub: https://github.com/kingoflolz/mesh-transformer-jax/#gpt-j-6b
- Demo: https://6b.eleuther.ai/

Here is a list of reproductions of or based on the GPT-J project:

- Dolly | GPT-J-6B instruction-tuned on Alpaca-GPT4

Dolly (Databricks)

> *Databricks' Dolly, a large language model trained on the Databricks Machine Learning Platform, demonstrates that a two-years-old open source model (GPT-J) can, when subjected to just 30 minutes of fine tuning on a focused corpus of 50k records (Stanford Alpaca), exhibit surprisingly high quality instruction following behavior not characteristic of the foundation model on which it is based. We believe this finding is important because it demonstrates that the ability to create powerful artificial intelligence technologies is vastly more accessible than previously realized.*

- GitHub: databrickslabs/dolly: Databricks' Dolly, a large language model trained on the Databricks Machine Learning Platform (github.com)
- Review: Meet Dolly the new Alpaca model — YouTube

GPT-J-6B instruction-tuned on Alpaca-GPT4

> *This model was finetuned on GPT-4 generations of the Alpaca prompts, using LoRA for 30.000 steps (batch size of 128), taking over 7 hours in four V100S.*

- Hugging Face: vicgalle/gpt-j-6B-alpaca-gpt4 · Hugging Face

GPT4All-J

> *Demo, data, and code to train open-source assistant-style large language model based on GPT-J*

- GitHub: nomic-ai/gpt4all: gpt4all: an ecosystem of open-source chatbots trained on a massive collections of clean assistant data including code, stories and dialogue (github.com)
- Review: GPT4ALLv2: The Improvements and Drawbacks You Need to Know! — YouTube

GPT-NeoX (EleutherAI)

> *This repository records EleutherAI's library for training large-scale language models on GPUs. Our current framework is based on NVIDIA's Megatron Language Model and has been augmented with techniques from DeepSpeed as well as some novel optimizations. We aim to make this repo a centralized and accessible place to gather techniques for training large-scale autoregressive language models, and accelerate research into large-scale training.*

- GitHub: EleutherAI/gpt-neox: An implementation of model parallel autoregressive transformers on GPUs, based on the DeepSpeed library. (github.com)

h2oGPT (h2o.ai)

> *Our goal is to make the world's best open source GPT!*

- GitHub: h2oai/h2ogpt: Come join the movement to make the world's best open source GPT led by H2O.ai (github.com)
- Hugging Face: H2ogpt Oasst1 256 6.9b App — a Hugging Face Space by h2oai

HuggingGPT (Microsoft)

> *HuggingGPT is a collaborative system that consists of an LLM as the controller and numerous expert models as collaborative executors (from HuggingFace Hub).*

- GitHub: microsoft/JARVIS: JARVIS, a system to connect LLMs with ML community (github.com)

MPT-7B (Mosaic ML)

> *MPT-7B is a GPT-style model, and the first in the MosaicML Foundation Series of models. Trained on 1T tokens of a MosaicML-curated dataset, MPT-7B is open-source, commercially usable, and equivalent to LLaMa 7B on evaluation metrics. The MPT architecture contains all the latest techniques on LLM modeling — Flash Attention for efficiency, Alibi for context length extrapolation, and stability improvements to mitigate loss spikes. The base model and several variants, including a 64K context length fine-tuned model (!!) are all available.*

- Website: Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs (mosaicml.com)
- GitHub: mosaicml/llm-foundry (github.com)
- Review: MPT-7B — The First Commercially Usable Fully Trained LLaMa Model — YouTube

NeMo — GPT-2B-001 (Nvidia)

> *GPT-2B-001 is a transformer-based language model. GPT refers to a class of transformer decoder-only models similar to GPT-2 and 3 while 2B refers to the total trainable parameter count (2 Billion) [1, 2]. This model was trained on 1.1T tokens with NeMo.*

- Hugging Face: https://huggingface.co/nvidia/GPT-2B-001

OpenAssistant Models

> *Conversational AI for everyone.*

- Website: Open Assistant (open-assistant.io)
- GitHub: LAION-AI/Open-Assistant: OpenAssistant is a chat-based assistant that understands tasks, can interact with third-party systems, and retrieve information dynamically to do so. (github.com)
- Hugging Face: OpenAssistant (OpenAssistant) (huggingface.co)

OpenLLaMA

> *In this repo, we release a permissively licensed open source reproduction of Meta AI's LLaMA large language model. In this release, we're releasing a public preview of the 7B OpenLLaMA model that has been trained with 200 billion tokens. We provide PyTorch and Jax weights of pre-trained OpenLLaMA models, as well as evaluation results and comparison against the original LLaMA models. Stay tuned for our updates.*

- GitHub: openlm-research/open_llama (github.com)

PaLM (Google)

> *PaLM demonstrates the first large-scale use of the Pathways system to scale training to 6144 chips, the largest TPU-based system configuration used for training to date. The training is scaled using data parallelism at the Pod level across two Cloud TPU v4 Pods, while using standard data and model parallelism within each Pod. This is a significant increase in scale compared to most previous LLMs, which were either trained on a single TPU v3 Pod (e.g., GLaM, LaMDA), used pipeline parallelism to scale to 2240 A100 GPUs across GPU clusters (Megatron-Turing NLG) or used multiple TPU v3 Pods (Gopher) with a maximum scale of 4096 TPU v3 chips.*

- Website: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance — Google AI Blog (googleblog.com)

Here is a list of reproductions of or based on the PaLM project:

- PaLM (Concept of Mind)

PaLM (Concept of Mind)

Introducing three new open-source PaLM models trained at a context length of 8k on C4. Open-sourcing LLMs is a necessity for the fair and equitable democratization of AI. The models of sizes 150m, 410m, and 1b are available to download and use here.

- GitHub: conceptofmind/PaLM: An open-source implementation of Google's PaLM models (github.com)

Palmyra Base 5B (Writer)

> *Palmyra Base was primarily pre-trained with English text. Note that there is still a trace amount of non-English data present within the training corpus that was accessed through CommonCrawl. A causal language modeling (CLM) objective was utilized during the process of the model's pretraining. Similar to GPT-3, Palmyra Base is a member of the same family of models that only contain a decoder. As a result, it was pre-trained utilizing the objective of self-supervised causal language modeling. Palmyra Base uses the prompts and general experimental setup from GPT-3 in order to conduct its evaluation per GPT-3.*

- Hugging Face: Writer/palmyra-base · Hugging Face

Here is a list of reproductions of or based on the Palmyra Base project:

- Camel 5B

Camel 🐫 5B

> *Introducing Camel-5b, a state-of-the-art instruction-following large language model designed to deliver exceptional performance and versatility. Derived from the foundational architecture of Palmyra-Base, Camel-5b is specifically tailored to*

> *address the growing demand for advanced natural language processing and comprehension capabilities.*

- Hugging Face: Writer/camel-5b-hf · Hugging Face

Polyglot (EleutherAI)

> *Large Language Models of Well-balanced Competence in Multi-languages. Various multilingual models such as mBERT, BLOOM, and XGLM have been released. Therefore, someone might ask, "why do we need to make multilingual models again?" Before answering the question, we would like to ask, "Why do people around the world make monolingual models in their language even though there are already many multilingual models?" We would like to point out there is a dissatisfaction with the non-English language performance of the current multilingual models as one of the most significant reason. So we want to make multilingual models with higher non-English language performance. This is the reason we need to make multilingual models again and why we name them 'Polyglot'.*

- GitHub: EleutherAI/polyglot: Polyglot: Large Language Models of Well-balanced Competence in Multi-languages (github.com)

Pythia (EleutherAI)

> *Interpreting Autoregressive Transformers Across Time and Scale*

- GitHub: EleutherAI/pythia (github.com)

Here is a list of reproductions of or based on the Pythia project:

- Dolly 2.0

Dolly 2.0 (Databricks)

> *Dolly 2.0 is a 12B parameter language model based on the EleutherAI pythia model family and fine-tuned exclusively on a new, high-quality human generated instruction following dataset, crowdsourced among Databricks employees.*

- Website: Free Dolly: Introducing the World's First Open and Commercially Viable Instruction-Tuned LLM — The Databricks Blog
- Hugging Face: databricks (Databricks) (huggingface.co)
- GutHub: dolly/data at master · databrickslabs/dolly (github.com)
- Review: Dolly 2.0 by Databricks: Open for Business but is it Ready to Impress! — YouTube

RedPajama-INCITE 3B and 7B (Together)

> *The first models trained on the RedPajama base dataset: a 3 billion and a 7B parameter base model that aims to replicate the LLaMA recipe as closely as possible. In addition, we are releasing fully open-source instruction-tuned and chat models.*

- Website: Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models — TOGETHER
- Hugging Face: togethercomputer/RedPajama-INCITE-Base-3B-v1 · Hugging Face, togethercomputer/RedPajama-INCITE-Chat-3B-v1 · Hugging Face, and togethercomputer/RedPajama-INCITE-Instruct-3B-v1 · Hugging Face
- Hugging Face: togethercomputer/RedPajama-INCITE-Base-7B-v0.1 · Hugging Face, togethercomputer/RedPajama-INCITE-Chat-7B-v0.1 · Hugging Face, and togethercomputer/RedPajama-INCITE-Instruct-7B-v0.1 · Hugging Face

Replit-Code (Replit)

> `replit-code-v1-3b` *is a 2.7B Causal Language Model focused on Code Completion. The model has been trained on a subset of the Stack Dedup v1.2 dataset. The training mixture includes 20 different languages, listed here in descending order of number of tokens:*
> `Markdown`, `Java`, `JavaScript`, `Python`, `TypeScript`, `PHP`, `SQL`, `JSX`, `reStructuredText`, `Rust`, `C`, `CSS`, `Go`, `C++`, `HTML`, `Vue`, `Ruby`, `Jupyter Notebook`, `R`, `Shell`
> *In total, the training dataset contains 175B tokens, which were repeated over 3 epochs -- in total,* `replit-code-v1-3b` *has been trained on 525B tokens (~195 tokens per parameter).*

- Hugging Face: https://huggingface.co/replit/replit-code-v1-3b

The RWKV Language Model

> *RWKV: Parallelizable RNN with Transformer-level LLM Performance (pronounced as "RwaKuv", from 4 major params: R W K V)*

- GitHub: BlinkDL/RWKV-LM
- ChatRWKV: with "stream" and "split" strategies and INT8. 3G VRAM is enough to run RWKV 14B :) https://github.com/BlinkDL/ChatRWKV
- Hugging Face Demo: HuggingFace Gradio demo (14B ctx8192)
- Hugging Face Demo: Raven (7B finetuned on Alpaca) Demo
- RWKV pip package: https://pypi.org/project/rwkv/
- Review: Raven — RWKV-7B RNN's LLM Strikes Back — YouTube

Segment Anything (Meta)

> *The Segment Anything Model (SAM) produces high quality object masks from input prompts such as points or boxes, and it can be used to generate masks for all objects in an image. It has been trained on a dataset of 11 million images and 1.1 billion masks, and has strong zero-shot performance on a variety of segmentation tasks.*

- Website: Introducing Segment Anything: Working toward the first foundation model for image segmentation (facebook.com)
- GitHub: facebookresearch/segment-anything: The repository provides code for running inference with the SegmentAnything Model (SAM), links for downloading the trained model checkpoints, and example notebooks that show how to use the model. (github.com)

StableLM (StabilityAI)

> *A new open-source language model, StableLM. The Alpha version of the model is available in 3 billion and 7 billion parameters, with 15 billion to 65 billion parameter models to follow. Developers can freely inspect, use, and adapt our StableLM base models for commercial or research purposes, subject to the terms of the CC BY-SA-4.0 license. StableLM is trained on a new experimental dataset built on The Pile, but three times larger with 1.5 trillion tokens of content. We will release details on the dataset in due course. The richness of this dataset gives StableLM surprisingly high performance in conversational and coding tasks, despite its small size of 3 to 7 billion parameters (by comparison, GPT-3 has 175 billion parameters)*

- Website: Stability AI Launches the First of its StableLM Suite of Language Models — Stability AI
- GitHub: Stability-AI/StableLM: StableLM: Stability AI Language Models (github.com)
- Hugging Face: Stablelm Tuned Alpha Chat — a Hugging Face Space by stabilityai
- Review: Stable LM 3B — The new tiny kid on the block. — YouTube

StartCoder (BigCode)

> *BigCode is an open scientific collaboration working on responsible training of large language models for coding applications. You can find more information on the main website or follow Big Code on Twitter. In this organization you can find the artefacts of this collaboration: StarCoder, a state-of-the-art language model for code, The Stack, the largest available pretraining dataset with perimssive code, and SantaCoder, a 1.1B parameter model for code.*

- Website: https://huggingface.co/bigcode

- Hugging Face: https://huggingface.co/spaces/bigcode/bigcode-editor and https://huggingface.co/spaces/bigcode/bigcode-playground
- Review: Testing Starcoder for Reasoning with PAL — YouTube

XGLM (Meta)

> *The XGLM model was proposed in Few-shot Learning with Multilingual Language Models.*

- GitHub: https://github.com/facebookresearch/fairseq/tree/main/examples/xglm
- Hugging Face: https://huggingface.co/docs/transformers/model_doc/xglm

Other Repositories

A'eala

- Hugging Face: Aeala (A'eala) (huggingface.co)

ausboss

- Hugging Face: https://huggingface.co/ausboss

chavinlo

- Hugging Face: chavinlo (Chavez) (huggingface.co)

chainyo

- Hugging Face: chainyo (Thomas Chaigneau) (huggingface.co)

couchpotato888

- Hugging Face: couchpotato888 (Phil Wee) (huggingface.co)

crumb

- Hugging Face: https://huggingface.co/crumb

digitous

- Hugging Face: digitous (Erik) (huggingface.co)

eachadea

- Hugging Face: eachadea (eachadea) (huggingface.co)

Knut Jägersberg

- Hugging Face: https://huggingface.co/KnutJaegersberg

KoboldAI

- Hugging Face: KoboldAI (KoboldAI) (huggingface.co)

LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions

> *LaMini-LM is a collection of small-sized, efficient language models distilled from ChatGPT and trained on a large-scale dataset of 2.58M instructions. We explore different model architectures, sizes, and checkpoints, and extensively evaluate their performance across various NLP benchmarks and through human evaluation.*

- Paper: [2304.14402] LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions (arxiv.org)
- GitHub: mbzuai-nlp/LaMini-LM: LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions (github.com)
- Review: LaMini-LM — Mini Models Maxi Data! — YouTube

MetaIX

- Hugging Face: https://huggingface.co/MetaIX

Teknium

- Hugging Face: https://huggingface.co/teknium

I hope you have enjoyed this article. If you have any questions or comments, please provide them here.

List of all Foundation Models

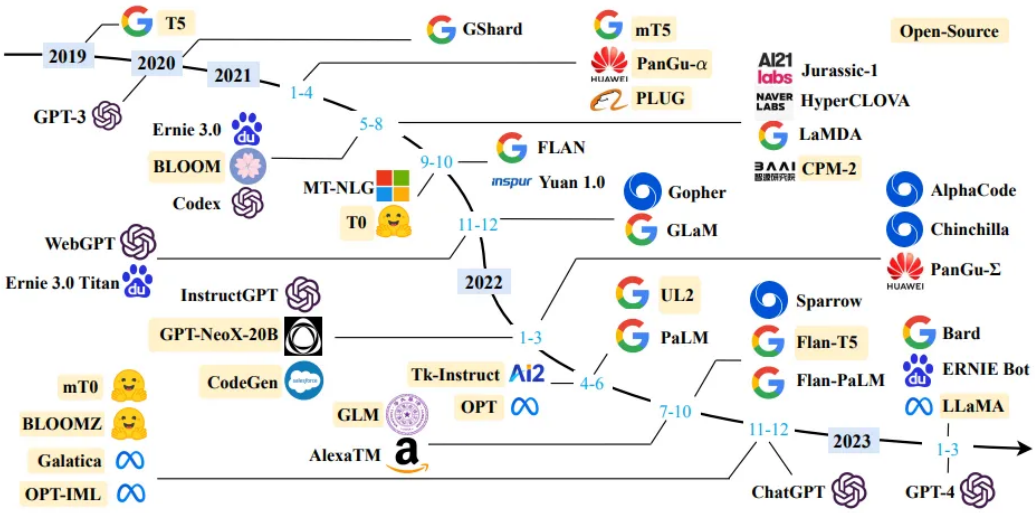
Sourced from: A List of 1 Billion+ Parameter LLMs (matt-rickard.com)

- GPT-J (6B) (EleutherAI)
- GPT-Neo (1.3B, 2.7B, 20B) (EleutherAI)
- Pythia (1B, 1.4B, 2.8B, 6.9B, 12B) (EleutherAI)
- Polyglot (1.3B, 3.8B, 5.8B) (EleutherAI)
- J1/Jurassic-1 (7.5B, 17B, 178B) (AI21)
- J2/Jurassic-2 (Large, Grande, and Jumbo) (AI21)
- LLaMa (7B, 13B, 33B, 65B) (Meta)
- OPT (1.3B, 2.7B, 13B, 30B, 66B, 175B) (Meta)
- Fairseq (1.3B, 2.7B, 6.7B, 13B) (Meta)
- GLM-130B YaLM (100B) (Yandex)
- YaLM (100B) (Yandex)
- UL2 20B (Google)
- PanGu-α (200B) (Huawei)
- Cohere (Medium, XLarge)
- Claude (instant-v1.0, v1.2) (Anthropic)
- CodeGen (2B, 6B, 16B) (Salesforce)

- NeMo (1.3B, 5B, 20B) (NVIDIA)
- RWKV (14B)
- BLOOM (1B, 3B, 7B)
- GPT-4 (OpenAI)
- GPT-3.5 (OpenAI)
- GPT-3 (ada, babbage, curie, davinci) (OpenAI)
- Codex (cushman, davinci) (OpenAI)
- T5 (11B) (Google)
- CPM-Bee (10B)
- Cerebras-GPT

Resources

- PRIMO.ai Large Language Model (LLM): https://primo.ai/index.php?title=Large_Language_Model_(LLM)
- A Survey of Large Language Models: [2303.18223] A Survey of Large Language Models (arxiv.org)



[2303.18223] A Survey of Large Language Models (arxiv.org) — Page 5

- LLMMaps — A Visual Metaphor for Stratified Evaluation of Large Language Models: https://arxiv.org/abs/2304.00457

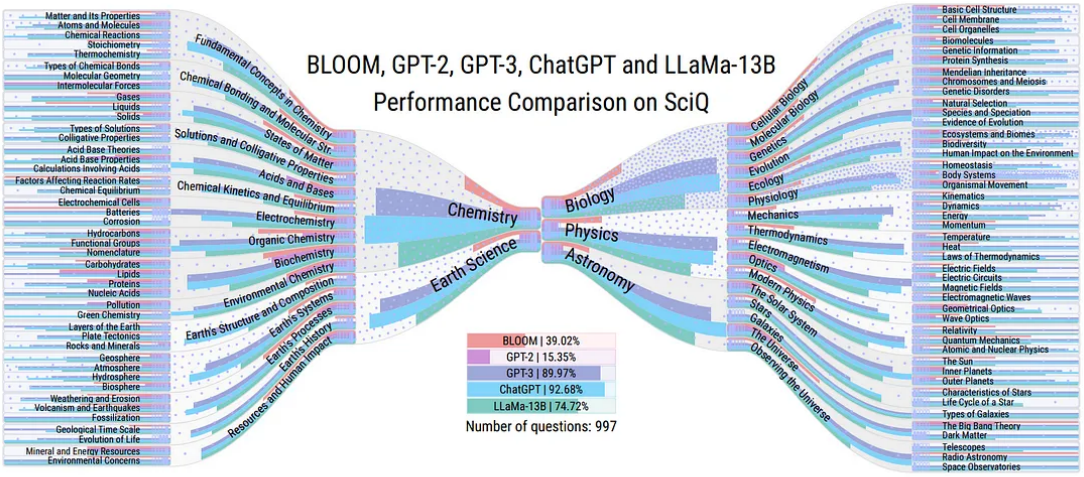


Fig. 4: Comparison of BLOOM, GPT-2, GPT-3,and LLaMa-13B on the stratified SciQ natural sciences Q&A test set. Bars show model accuracy, blue noise number of questions, and discrete progress bar icons model-agnostic difficulty rating - each aggregated per knowledge hierarchy level.

- A brief history of LLaMA models (A brief history of LLaMA models — AGI Sphere (agi-sphere.com))
- Google “We Have No Moat, And Neither Does OpenAI” (https://www.semianalysis.com/p/google-we-have-no-moat-and-neither)
- Chatbot Arena (Chat with Open Large Language Models (lmsys.org))
- Ahead of AI #8: The Latest Open Source LLMs and Datasets
- Open LLM Leaderboard (Open LLM Leaderboard — a Hugging Face Space by HuggingFaceH4)

Llm | Open Source | Llamas | Gpt | AI

Follow

Written by Sung Kim

2K Followers · Writer for Geek Culture

A business analyst at heart who dabbles in machine learning, data science, data engineering, and project management.