✨ Member-only story

# 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
26 min read · Mar 30

| Model | | age |
|---|---|---|
| tiiuae/f... | main | |
| timdettmers/guanaco-65b-merged | main | 62.2 |
| CalderaAI/30B-Lazarus | main | 60.7 |
| tiiuae/falcon-40b | main | 60.4 |
| timdettmers/guanaco-33b-merged | main | 60 |
| ausboss/llama-30b-supercot | main | 59.8 |
| huggyllama/llama-65b | main | 58.3 |
| pinkmanlove/llama-65b-hf | main | 58.3 |
| llama-65b | main | 58.3 |
| MetaIX/GPT4-X-Alpasta-30b | main | 57.9 |

Open LLM Leaderboard — a Hugging Face Space by HuggingFaceH4

***Tsinghua University released ChatGLM2 on June 25, 2023, which claimed to beat GPT-4 in Chinese (Leaderboard | C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models (cevalbenchmark.com)).***

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******!"*

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

*BAAI's Aqu* *ler* | *BigScience's* *Polyglot,* *and Pythia* *ternLM* | *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 sub grouped by the list of projects that are fine-tuned LLMs version of those Foundation LLMs, followed by a listing of websites where some organizations and/or individuals who publish fine-tuned LLMs on a frequent basis. These websites are denoted as *Person Name/Organization Name's Hugging Face website*. For example, Tom Jobbins (TheBloke) has over 260+ fine-tuned LLMs on their website.

**Updates:**

- **03/2023:** Added HuggingGPT | Vicuna/FastChat
- **04/2023 Fine-Tuned LLMs:** Baize | Koala | Segment Anything | Galpaca | GPT-J-6B instruction-tuned on Alpaca-GPT4 | GPTQ-for-LLaMA | Dolly 2.0 | StackLLaMA | GPT4All-J | Palmyra Base 5B | Camel 🐫 5B | StableLM | h2oGPT | OpenAssistant Models | StableVicuna | FastChat-T5 | couchpotato888 | GPT4-x-Alpaca | LLaMA Adapter V2 | WizardLM
- **04/2023 Others:** A Survey of Large Language Models | LLMMaps — A Visual Metaphor for Stratified Evaluation of Large Language Models | A brief history of LLaMA models | List of all Foundation Models
- **05/2023 Foundation LLMs:** OpenLLaMA | BigCode StarCoder (Hugging Face + ServiceNow) | Replit-Code (Replit) | Mosaic ML's MPT-7B | Together's RedPajama-INCITE 3B and 7B | TII's Falcon LLM
- **05/2023 Fine-Tuned LLMs:** Pygmalion-7b | Nvidia GPT-2B-001 | crumb's Hugging Face website | Teknium's Hugging Face website | Knut Jägersberg's Hugging Face website | 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 | | PaLM (Concept of Mind) | digitous Huggin 's Hugging Face w ace websit ebsite | Baize V2 | Gorilla (POET?) | QLoRA | ausboss' Hugging Face website | Metal (MetaIX)'s Hugging Face website



- **05/2023 Others:** SemiAnalysis article by Luke Sernau (a senior software engineer at Google) | Chatbot Arena | Ahead of AI #8: The Latest Open Source LLMs and Datasets (Resources section)

- **06/2023 Foundation LLMs:** InternLM | OpenLLaMA 13B | Baichuan Intelligent Technology's baichuan | BAAI's Aquilla | Mosaic ML's MPT-30B | Tsinghua University's ChatGLM2–6B

- **06/2023 Fine-Tuned LLMs:** CalderaAI/30B-Lazarus | elinas' Hugging Face website | Tim Dettmers' Hugging Face website | Tiger Research's Hugging Face website | Eric Hartford's Hugging Face website | pinkmanlove's Hugging Face website | Huggy Llama's Hugging Face website | AllenAI's Tulu 65B | CarperAI's Hugging Face website | Eugene Pentland's Hugging Face website | Concept of Mind's Hugging Face website | ClimateBert's Hugging Face website | LLMs' Hugging Face website | Jon Durbin's Hugging Face website | Benjamin Anderson's Hugging Face website | Georgia Tech Research Institute's Hugging Face website | OptimalScale's Hugging Face website | FlashVenom's Hugging Face website | Michael's Hugging Face website | Pankaj Mathur's Hugging Face website | Manuel Romero's Hugging Face website

- **06/2023 Others:** AlpacaEval Leaderboard

## 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
- Comm



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

*Alpaca.cpp* | *Alpaca-LoRA* | *Baize* | *Cabrita* | CalderaAI/30B-Lazarus | *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* | *AllenAI's Tulu* | *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

**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 ... with ChatGP... Yat-sen Univers...



- Paper: 2304.01196.pdf (arxiv.org)

**Cabrita**

> *A portuguese finetuned instruction LLaMA*

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

**CalderaAI/30B-Lazarus**

> *This model is the result of an experimental use of LoRAs on language models and model merges that are not the base HuggingFace-format LLaMA model they were intended for. The desired outcome is to additively apply desired features without paradoxically watering down a model's effective behavior.*

- Hugging Face: CalderaAI/30B-Lazarus · Hugging Face

**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 [followi](#) [alpaca](#) [n-](#) [结构参考](#)

**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* To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy. *)k GPT-3.5-Turbo Gene*

- 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 o*



- GitHub ++ (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**



> *OpenBuddy* *bal 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-qualit* *odels than the previou* *on both human and* *d 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

**Tulu 65B (AllenAI)**

> *This is the repository for the paper How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources .*
>
> *We explore instruction-tuning popular base models on publicly available datasets. This*

> *repository contains:*
>
> *Training cc*
>
> 
>
> *Evaluation code for the evaluation done in the paper.*
>
> *Script for merging and creating model diffs.*
>
> *As part of this work we introduce Tülu, a suite of LLaMa models fully-finetuned on a strong mix of datasets!*

- GitHub: allenai/open-instruct (github.com)
- Hugging Face: allenai/tulu-65b · 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 🇫🇷 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 odel — YouTuł



## BAAI's Aquila

> *The Aquila language model inherits the architectural design advantages of GPT-3 and LLaMA, replacing a batch of more efficient underlying operator implementations and redesigning the tokenizer for Chinese-English bilingual support. It upgrades the BMTrain parallel training method, achieving nearly 8 times the training efficiency of Magtron+DeepSpeed ZeRO-2 in the training process of Aquila. The Aquila language model is trained from scratch on high-quality Chinese and English corpora. Through data quality control and various training optimization methods, it achieves better performance than other open-source models with smaller datasets and shorter training times. It is also the first large-scale open-source language model that supports Chinese-English-Knowledge, commercial licensing, and complies with domestic data regulations.*

- GitHub: FlagAI/examples/Aquila at master · FlagAI-Open/FlagAI · GitHub
- Hugging Face: BAAI (Beijing Academy of Artificial Intelligence) (huggingface.co)

## Baichuan Intelligent Technology's baichuan

> *baichuan-7B is an open-source large-scale pre-trained model developed by Baichuan Intelligent Technology. Based on the Transformer architecture, it is a model with 7 billion parameters trained on approximately 1.2 trillion tokens. It supports both Chinese and English, with a context window length of 4096. It achieves the best performance of its size on standard Chinese and English authoritative benchmarks (C-EVAL/MMLU).*

- GitHub: baichuan-inc/baichuan-7B: A large-scale 7B pretraining language model developed by BaiChuan-Inc. (github.com)
- Hugging Face: baichuan-inc/baichuan-7B · Hugging Face

## BLOOM (BigScience)

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

- Huggin



- Huggin gface

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 LL



> *Falcon LLM* *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: ...r training large m...

## 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 i 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:

- ChatGI



**ChatGLM2–**

> *ChatGLM2–6B is the second-generation version of the open-source bilingual (Chinese-English) chat model ChatGLM-6B. It retains the smooth conversation flow and low deployment threshold of the first-generation model.*

- GitHub: ChatGLM2–6B/README_EN.md at main · THUDM/ChatGLM2–6B · GitHub

## 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: [illegible] trained on the [illegible]
- 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)

## InternLM



> *We present parameters. InternLM is pre-trained on a large corpora with 1.6T tokens with a multi-phase progressive process, and then fine-tuned to align with human preferences.*
>
> *InternLM is a multilingual large language model jointly developed by Shanghai AI Lab and SenseTime (with equal contribution), in collaboration with the Chinese University of Hong Kong, Fudan University, and Shanghai Jiaotong University.*

- GitHub: InternLM (github.com)

## h2oGPT (h2o.ai)

> *h2oGPT is a large language model (LLM) fine-tuning framework and chatbot UI with document(s) question-answer capabilities. Documents help to ground LLMs against hallucinations by providing them context relevant to the instruction. h2oGPT is fully permissive Apache V2 open-source project for 100% private and secure use of LLMs and document embeddings for document question-answer.*

- 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 (Mosaic ML)

> *MPT 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 is open-source, commercially*

> *usable, and equivalent to LLaMa on evaluation metrics. The MPT architecture contains all the lates* *i for context len* *. The base model and* *!) 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
- Hugging Face: mosaicml/mpt-30b · Hugging Face
- Hugging Face: mosaicml/mpt-30b-chat · Hugging Face

## 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 ... ı AI's LLaMA lar ... of the 7B OpenLLaM ... yTorch and Jax weights of pre-trained OpenLLaMA models, as well as evaluation results and comparison against the original LLaMA models.*
>
> *We are releasing a 7B and 3B model trained on 1T tokens, as well as the preview of a 13B model trained on 600B tokens.*

- GitHub: openlm-research/open_llama (github.com)
- Hugging Face: openlm-research (OpenLM Research) (huggingface.co)

## 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 ogle's PaLM r



## 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 p nificant reason. So uage performanc d 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*



- Websit luding base, ir

- 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



### 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



### 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

#### Benjamin Anderson

- Hugging Face: andersonbcdefg (Benjamin Anderson) (huggingface.co)

**CarperAI**

- Huggin



**chavinlo**

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

**chainyo**

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

**ClimateBert**

- climatebert (ClimateBert) (huggingface.co)

**Concept of Mind**

- conceptofmind (Enrico Shippole) (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)

**elinas**

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

**Eric Hartford**

- Hugging Face: ehartford (Eric Hartford) (huggingface.co)

**Eugene Pentland**

- Hugging Face: eugenepentland (Eugene Pentland) (huggingface.co)

**FlashVenom**

- Hugging Face: flashvenom (FlashVenom) (huggingface.co)

**Georgia Tech Research Institute**

- Huggin [nstitute) (huggi



**Huggy Llam**

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

**Jon Durbin**

- Hugging Face: jondurbin (Jon Durbin) (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

**LLMs**

- LLMs (LLMs) (huggingface.co)

**Manuel Romero**

- Hugging Face: mrm8488 (Manuel Romero) (huggingface.co)

**MetaIX**

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

**Michael**

- Huggin



**OptimalSca**

- Huggin

**Pankaj Mathur**

- Hugging Face: psmathur (Pankaj Mathur) (huggingface.co)

**pinkmanlove**

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

**Teknium**

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

**Tiger Research**

- Hugging Face: TigerResearch (Tiger Research) (huggingface.co)

**Tim Dettmers**

- Hugging Face: timdettmers (Tim Dettmers) (huggingface.co)

**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
- 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

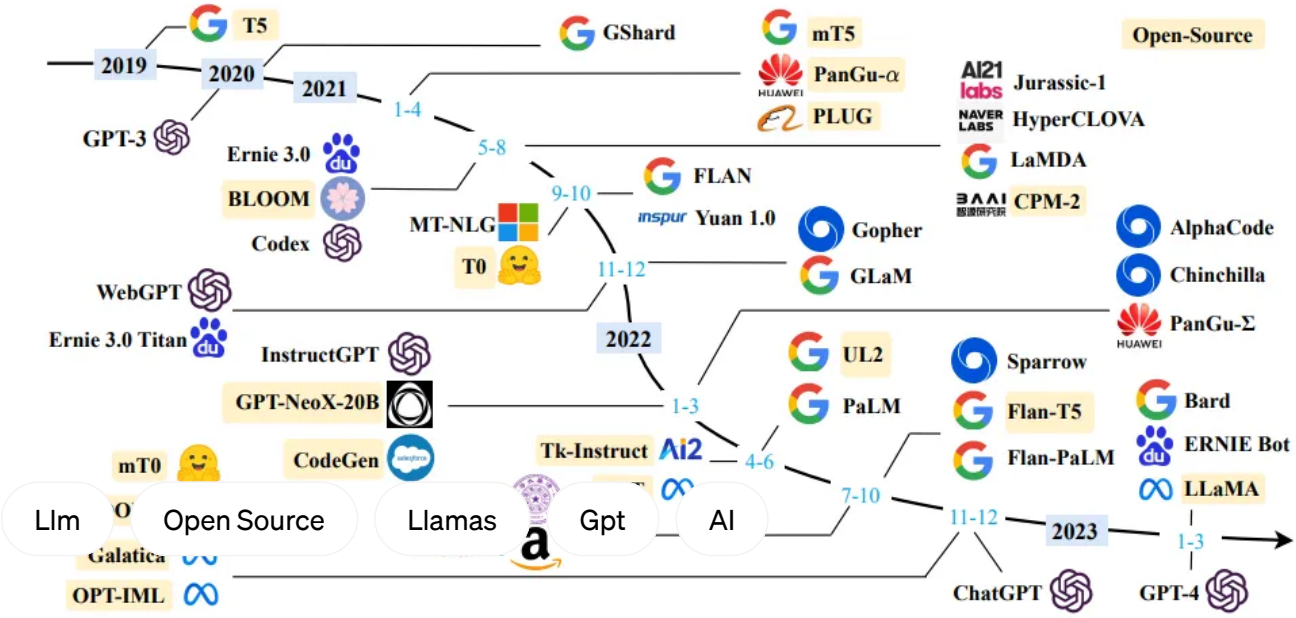


## Resources

- 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






[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

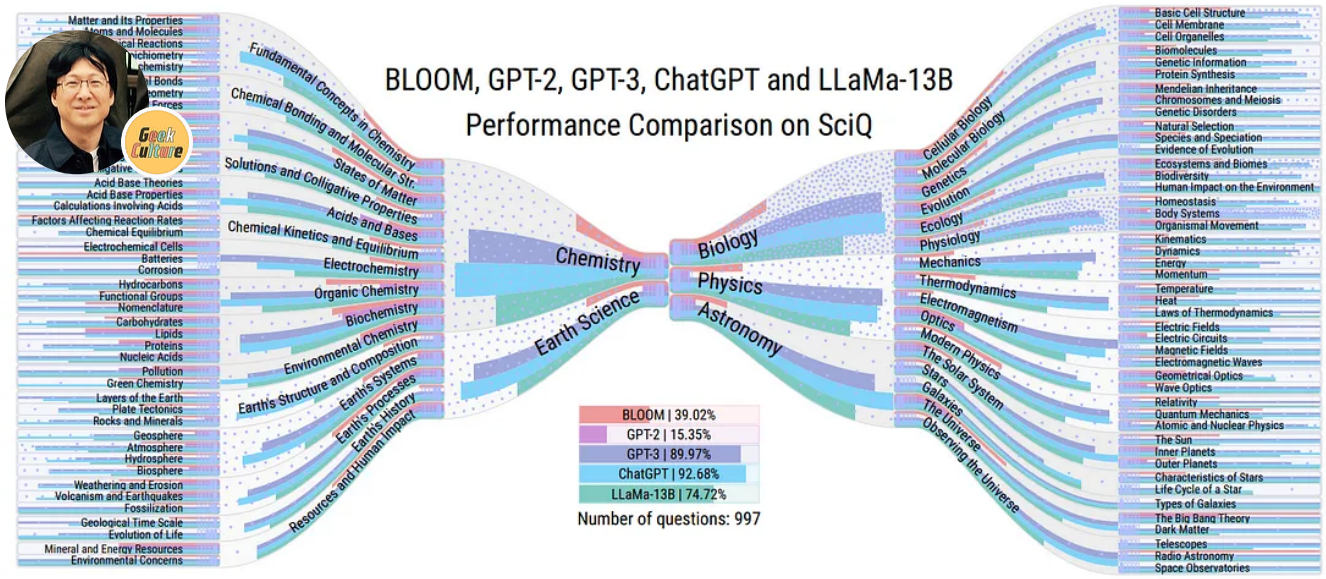
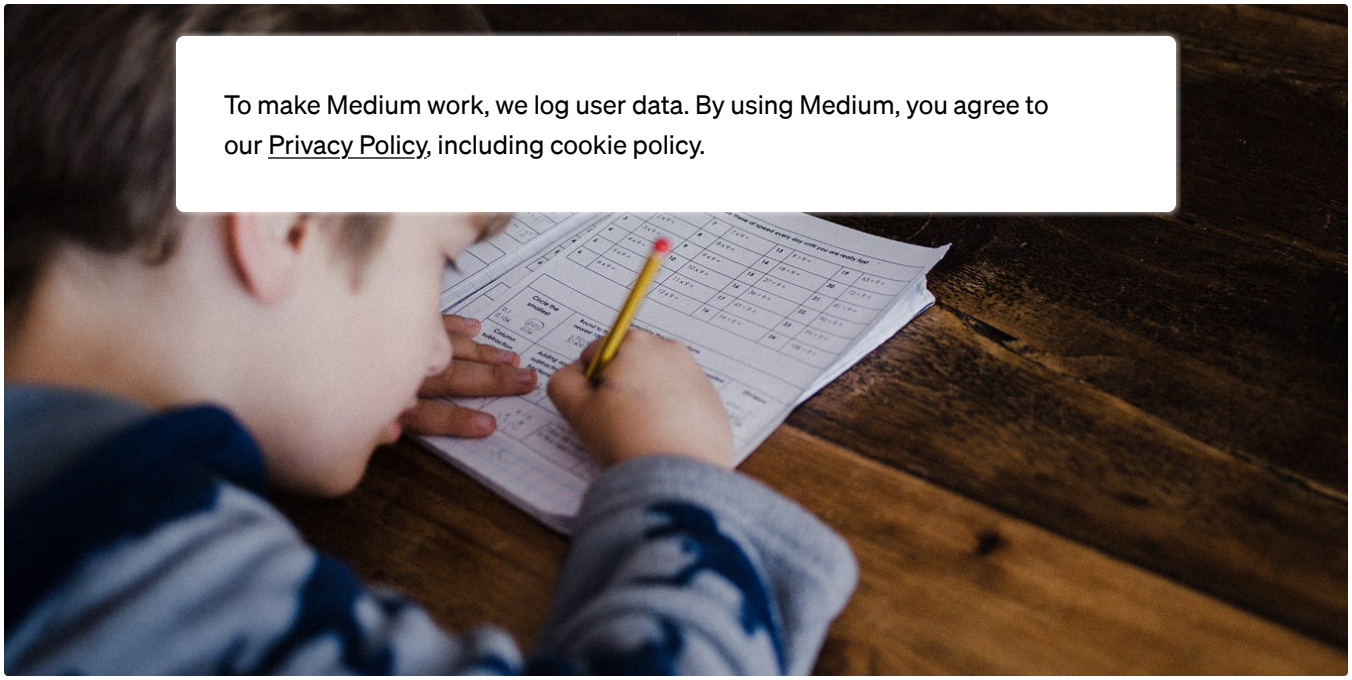



Fig. 4: Comparison of BLOOM, GPT-2, GPT-3,and LLaMa-13B on the stratified SciQ natural sciences Q&A test set. Bars show model accuracy, blue noise number of questions, and discrete progress bar icons model-agnostic difficulty rating - each aggregated per knowledge hierarchy level.

https://arxiv.org/pdf/2304.00457.pdf — Page 7




- AlpacaEval Leaderboard (Alpaca Eval Leaderboard (tatsu-lab.github.io)

Sung Kim in Geek Culture

## How to Detect OpenAI's ChatGPT Output

How to detect if the student used OpenAI's ChatGPT to complete an assignment

· 5 min read · Dec 11, 2022

360 24

parse_expenses.py

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, value, currency).
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2016-01-02 -34.01 USD
        2016-01-03 2.59 DKK
        2016-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
                         float(value),
                         currency))
    return expenses
```

Jacob Bennett in Geek Culture

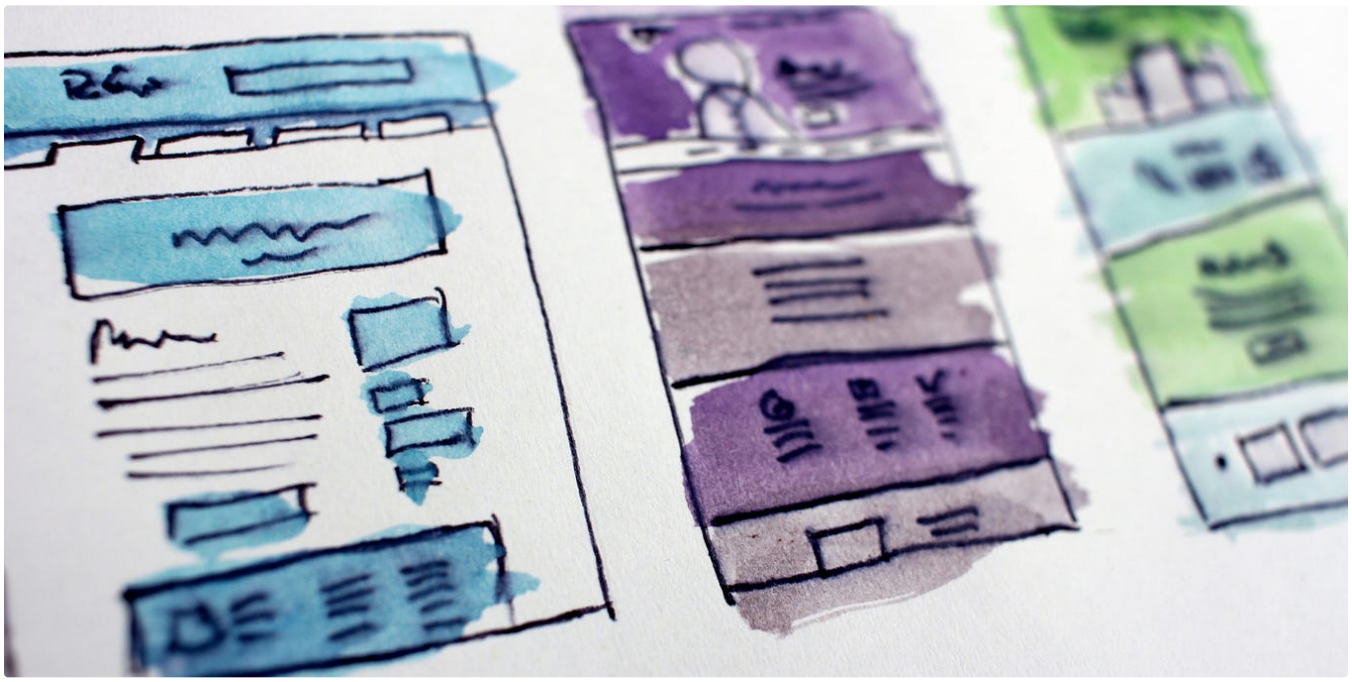
**The 5 pai ineer**

Tools I use tl



✨ · 4 min read · Mar 25

2.6K 25



Rahul Kapoor in Geek Culture

## System Design — Scaling from Zero to Millions Of Users

Note: I have read this great book System Design Interview — An insider's guide by Alex Xu in depth. So most of my definitions and images...

11 min read · Jan 2

578 6

ssification Clustering Pair Classification Retrieval Reranking STS Summarization



k tabs for others

| Embedding Dimensions | Average (56 datasets) | Classification Average (12 datasets) | Clustering Average (11 datasets) | Pair Classification Average (3 datasets) | Reranking Average (4 datasets) | Retrieval Average (15 datasets) | STS Aver (10 data |
|---|---|---|---|---|---|---|---|
| 768 | 61.79 | 73.12 | 44.74 | 86.62 | 57.29 | 49.26 | 83.( |
| 768 | 61.59 | 73.86 | 45.29 | 85.89 | 57.54 | 47.57 | 83.1 |
| 1024 | 61.42 | 73.14 | 43.33 | 85.94 | 56.53 | 49.99 | 82.( |
| 1536 | 60.99 | 70.93 | 45.9 | 84.89 | 56.32 | 49.25 | 80.9 |
| 768 | 60.44 | 72.63 | 42.11 | 85.09 | 55.7 | 48.75 | 80.9 |

Sung Kim in Dev Genius

## So you want to build an AI application powered by LLM: Let's talk about Embedding and Semantic...

Embedding and Semantic Search for an AI application powered by LLM

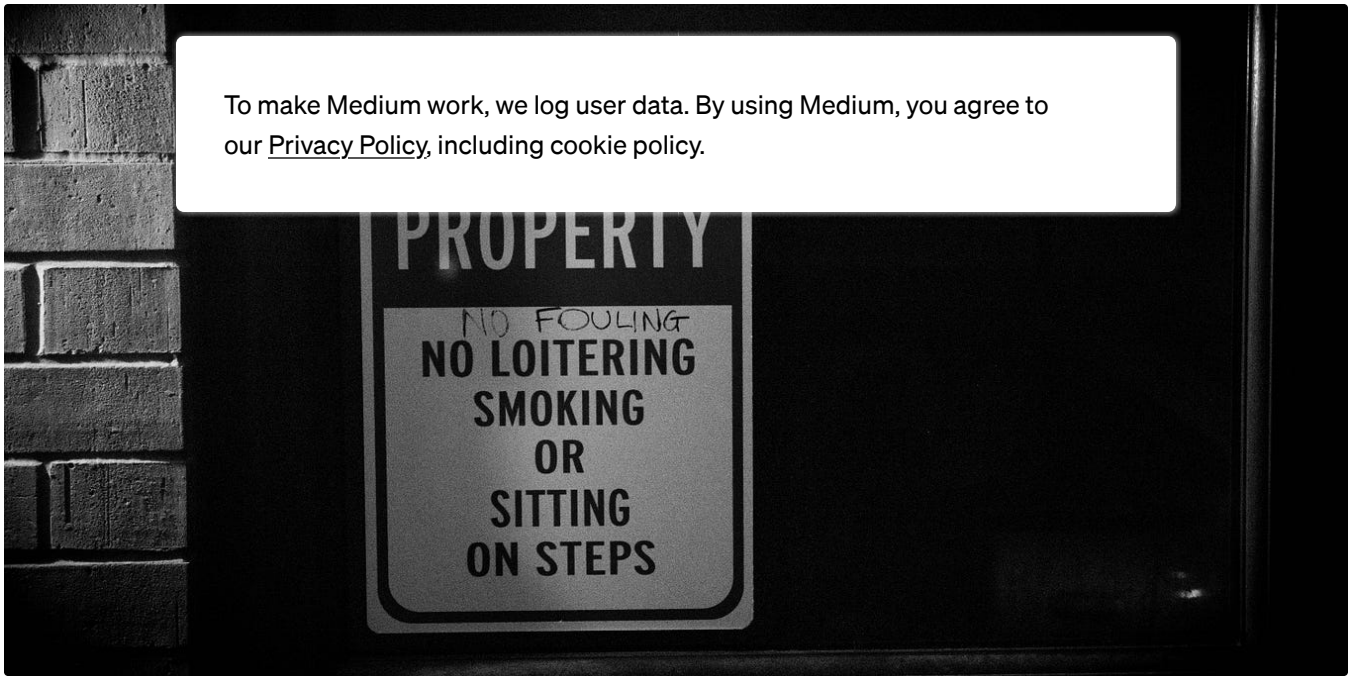
· 9 min read · May 4

275 3

See all from Sung Kim

See all from Geek Culture

## Recommended from Medium





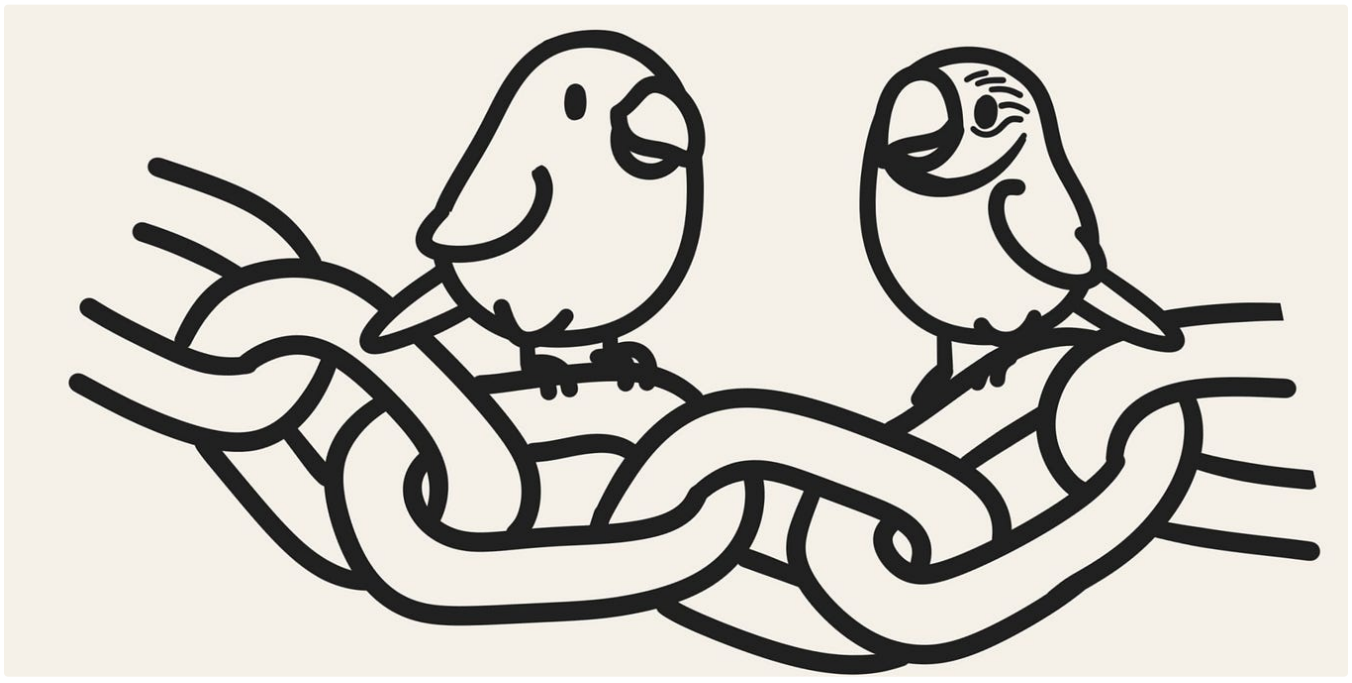

Wei-Meng Lee in Level Up Coding

# Training Your Own LLM using privateGPT

Learn how to train your own language model without exposing your private data to the provider

· 8 min read · May 19





Leonie Monigatti in Towards Data Science

**Getting S LM-Powered**



A LangChain tutorial to build anything with large language models in Python

· 12 min read · Apr 25

Kristen Walters in Adventures In AI

# 5 Ways I'm Using AI to Make Money in 2023

These doubled my income last year

· 9 min read · May 28



Timothy Mugayi in Better Programming

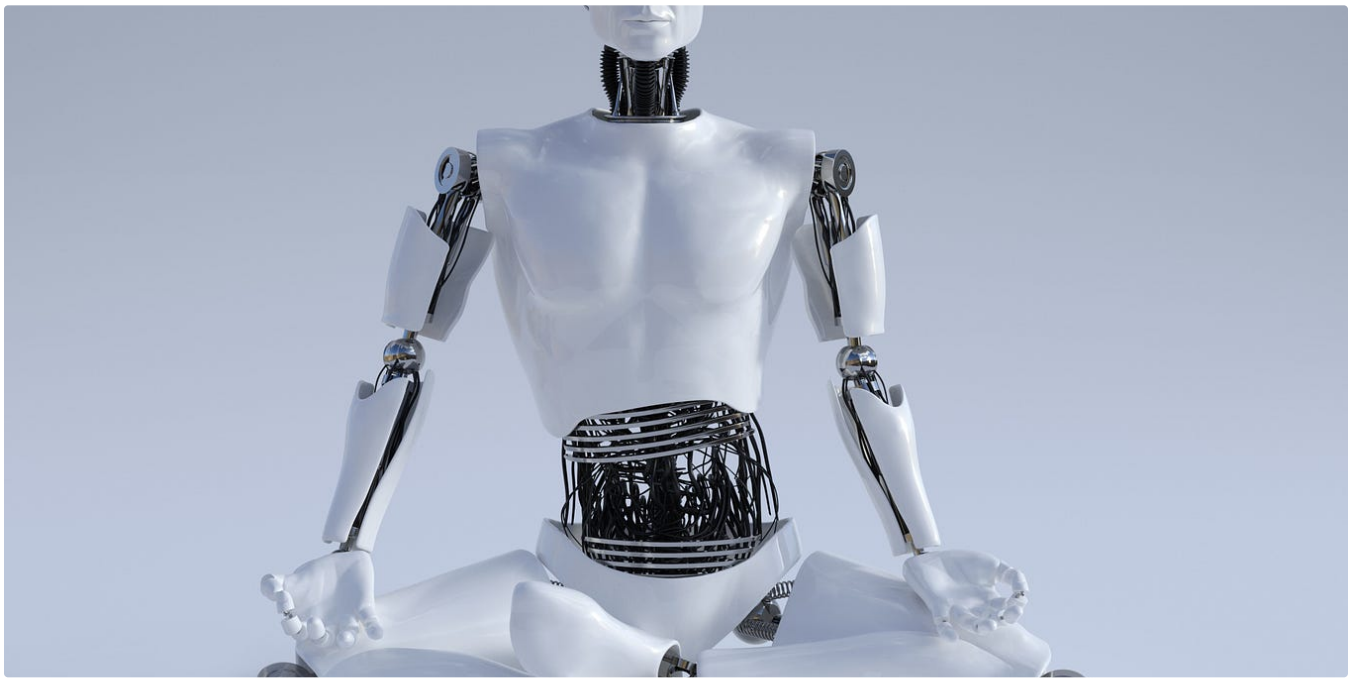
## How To B ge Base

Feed your C



✦ · 11 min read · Apr 7

4.1K 95



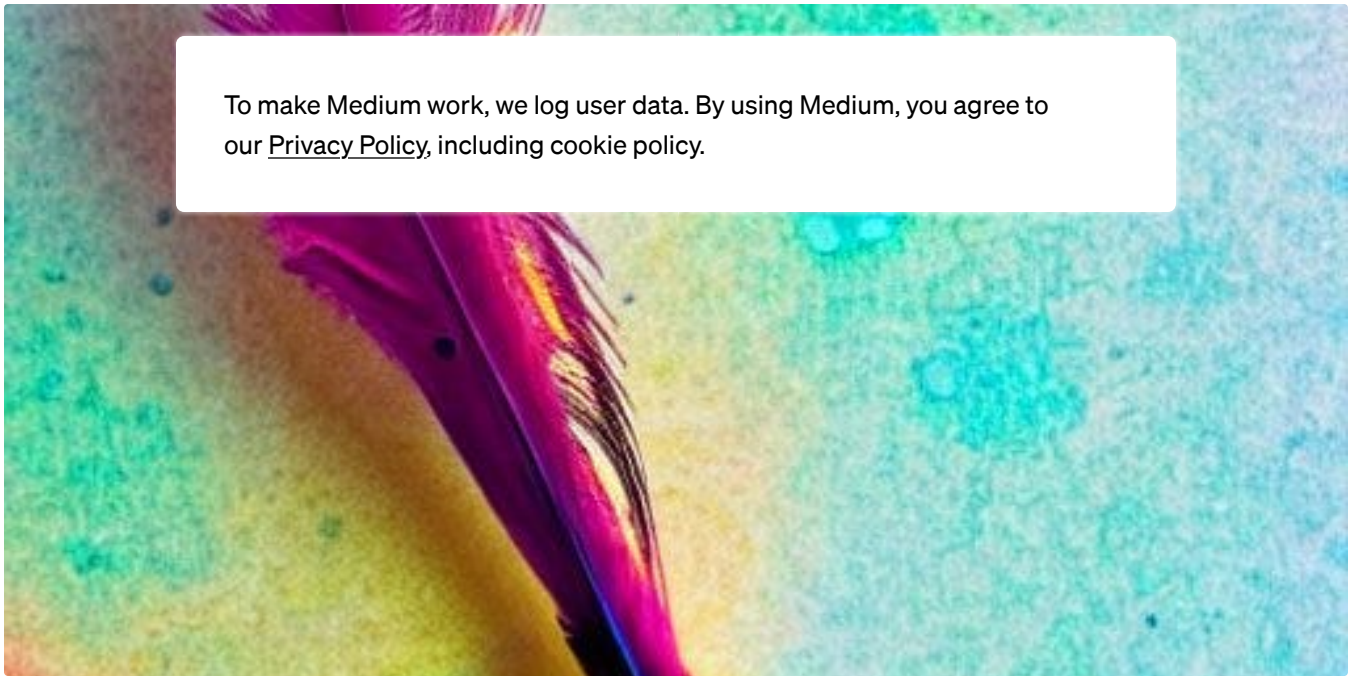
The PyCoach in Artificial Corner

## You're Using ChatGPT Wrong! Here's How to Be Ahead of 99% of ChatGPT Users

Master ChatGPT by learning prompt engineering.

✦ · 7 min read · Mar 17

25K 449



Kory Becker in ITNEXT

# Prompt Engineering: The Magical World of Large Language Models

A data-driven approach to getting the best response

· 11 min read · 5 days ago

124 3