Published in Geek Culture

Sung Kim Follow

Mar 30 · 14 min read · ✦ · Listen

Save

# 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

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.

***Right now, I am most excited about StabilityAI's StableLM, released on April 19, 2023. I hope it is good.***

*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 three models and runs on moderately specced PCs:

- LLaMA
- Alpaca
- GPT4All

The list is a work in progress where I tried to group them by the Foundation Models where they are: BigScience's BLOOM; Cerebras' Cerebras-GPT, EleutherAI's GPT-J, GPT-NeoX, Polyglot, and Pythia; GLM; Google's Flamingo, and FLAN; H2O.ai's h2ogpt; Meta's GALACTICA, LLaMA, and XGLM; RWKV; and StabilityAI's StableLM. They are subgrouped by the list of projects that are reproductions of or based on those Foundation Models.

**Updates:**

- 03/31/2023: Added HuggingGPT and Vicuna/FastChat (I have not tried GPT4All, but Vicuna/FastChat is pretty good)

- 04/02/2023: Added “A Survey of Large Language Models” and “LLMMaps — A Visual Metaphor for Stratified Evaluation of Large Language Models” to the Resources section.
- 04/04/2023: Added Baize and Koala.
- 04/05/2023: Added Segment Anything (Not really LLM, but it will be really helpful in CV pipeline.)
- 04/09/2023: Added Galpaca, GPT-J-6B instruction-tuned on Alpaca-GPT4, GPTQ-for-LLaMA, and List of all Foundation Models
- 04/11/2023: Added Dolly 2.0, StackLLaMA, and GPT4All-J
- 04/17/2023: Added Palmyra Base 5B and Camel 🐫 5B
- 04/19/2023: Added StableLM (I really hope this is good), h2oGPT, and The Bloke alpaca-lora-65B-GGML

## Alpaca / LLaMA (Meta/Stanford)

> *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
- Baize
- Cabrita

- Chinese-Vicuna
- GPT4All
- GPTQ-for-LLaMA
- Koala
- llama.cpp
- Lit-LLaMA
- StackLLaMA
- The Bloke alpaca-lora-65B-GGML
- Vicuna

### 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

> *Baize 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, 13B, and 30B 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-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)

### 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



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

### 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)

### 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)*
>
> *- 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



### 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

## 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

### 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:

- Flan-Alpaca
- Flan-UL2

### 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

> *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

> ***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

> *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

> *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

> *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)

## 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

> *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

> *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

## 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

> *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

> *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

## XGLM

> *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

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

## List of all Foundation Models

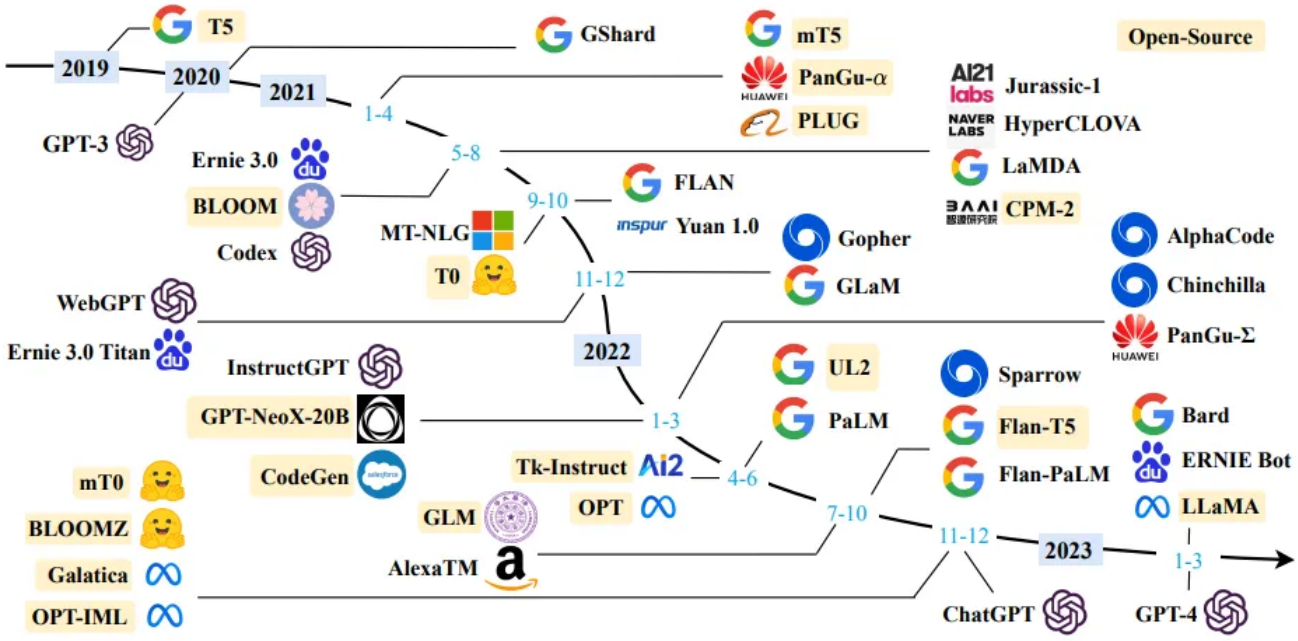
Sourced from: A List of 1 Billion+ Parameter LLMs (matt-rickard.com)

- 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)

- UL2 20B (Google)
- PanGu-α (200B) (Huawei)
- Cohere (Medium, XLarge)
- Claude (instant-v1.0, v1.2) (Anthropic)
- CodeGen (2B, 6B, 16B) (Salesforce)
- 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

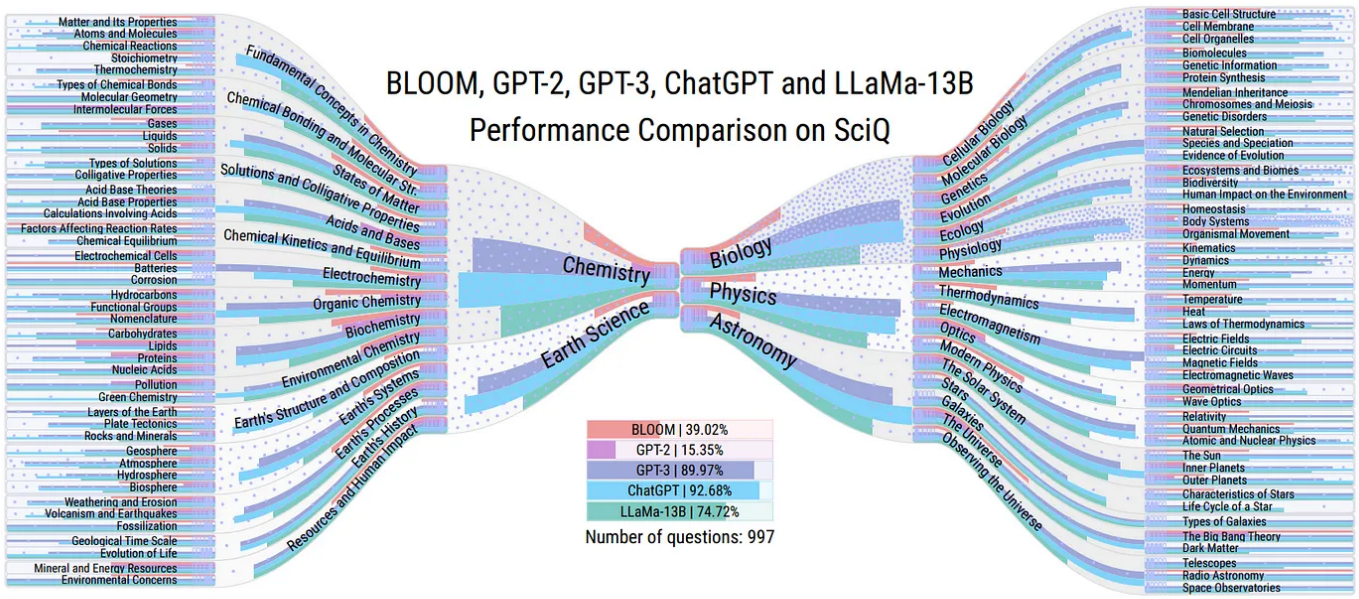



Fig. 4: Comparison of BLOOM, GPT-2, GPT-3,and LLaMa-13B on the stratified SciQ natural sciences Q&A test set. Bars show model accuracy, blue noise number of questions, and discrete progress bar icons model-agnostic difficulty rating - each aggregated per knowledge hierarchy level.

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

Llm Open Source Llamas Gpt AI

---