

Large Language Model Inventory v1.0

Patrick Emami, Zhaonan Li, Ray Grout
July 2023

Overview

This is a resource that keeps track of the most prominent open- and closed-source large language model (LLM) offerings

Given the rapid pace of progress in this space, it is not guaranteed to be a comprehensive list

Change List

7/24/23: v1.0 – initial inventory

Disclaimer

Open source LLMs can generate toxic, harmful content and typically do not have “guardrails” in place to prevent this

- Current recommendation is to only use them for well-defined research problems
- We should be cautious about deploying any of them in production for any official NREL products

“Instruction Fine-tuning”, which is used to create a class of “augmented” LLMs with supervised learning to boost performance significantly, has ethical concerns surrounding how the labels are obtained via [human annotators](#)

Open-source LLMs

Definition: *Available on HuggingFace*

- This means the weights can be downloaded from HuggingFace Hub
- It is not required to sign a release form
- The weights come with an OSS license such as Apache 2.0

Databricks

Dolly <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>

- 3B, 7B, 12B model available on Hugging Face
- Open for commercial use

MosaicML (now owned by Databricks)

- MPT-7B and MPT instruct fine-tuned available on Hugging Face
- Open for commercial use
- **Might need CUDA version ≥ 11.4 to run on Vermilion**

Meta

OPT

- 0.125, 0.35, 1.3, 2.7, 6.7, 13, 30, 175B
- [License](#), [Application Form](#)
- For non-commercial research use only

Llama 1 & Llama 2

- 7, 13, 33, 65B
- LLaMA1 for non-commercial research use, LLaMA2 for commercial use
- Requires a EULA Form to be signed

HuggingFace

- Bloom
 - 560m, 1b, 3b, 7b, 176b, available on Hugging Face
 - License: <https://huggingface.co/spaces/bigscience/license>
 - RAIL License – permissive with Use Restrictions intended to limit societal harms

EleutherAI

GPT-J

- 6b available on HuggingFace

GPT-NeoX

- 20b on HuggingFace

Apache 2.0 License

Cerebras

Cerebras-GPT

- 111M, 256M, 1.3b, 2.7b, 6.7b, 13b on HuggingFace
- Apache 2.0

NEMO

- Nvidia/GPT-2B-0001 , nemo-megatron-gpt-5b on HuggingFace
- CC-BY-4.0 license

Falcon LLM

Falcon

- <https://www.tii.ae/> (UAE)
- Falcon 7b/40b models are on HuggingFace
<https://huggingface.co/tiiuae>
- Released under Apache 2.0 permissive license

Tsinghua University - GLM

- 130B Bilingual (English/Chinese), also has smaller monolingual (English) ones from 110m to 10B and ChatGLM-6B
- Code under Apache-2.0 License, Model License: https://github.com/THUDM/GLM-130B/blob/main/MODEL_LICENSE
- For non-commercial research use

Vicuna

- 7B, 13B, 33B
- fine-tuned from LLaMA for chatting
- For non-commercial use
- <https://huggingface.co/lmsys/vicuna-13b-delta-v0>

Instruction Fine-tuned Open-Source LLMs

Stanford - Alpaca

Instruction fine-tuned LLaMA-7B

- Instructions synthesized from OpenAI text-davinci-003
- Non-commercial use (CC BY NC 4.0 License)

Closed-source LLMs

Accessible via API on a cloud server

Pricing is typically “\$/token”

Amazon AWS

[AWS JumpStart](#)

Offers (paid) access to closed and open LLMs in the cloud

[Github repo with lots of examples of how to use LLMs for various tasks \(backed by Sagemaker API\)](#)

–E.g., text generation with Llama-2 in a Jupyter Notebook

OpenAI ChatGPT

OpenAI API

- GPT4
- GPT3.5 / ChatGPT

Terms and Polices: <https://openai.com/policies>

Usage: Generation & Finetuning via API calls

Anthropic Claude

Claude and a light, low-cost version named Claude Instant

- AI assistant that has long context windows
- Answer generation via API calls, fine-tuning available for large enterprises
- Terms: <https://console.anthropic.com/legal/terms>

Cohere

Command model, can be accessed via API

Also offers services like summarization, embeddings, entity extraction, etc.

Various usage guidelines <https://docs.cohere.com/docs/usage-guidelines>

This work was authored by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.