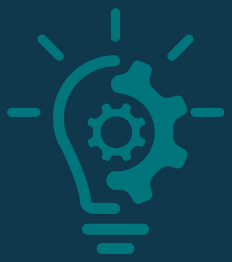
Enthought

Cheat Sheet | Large Language Models+ For Scientific Research

Updated August 2023



LLMs and Tools for R&D

To help scientists and researchers navigate the increasing number of advanced artificial intelligence (AI) options, Enthought experts have put together this summary of Large Language Models (LLMs) and related tools that are most relevant for R&D (updated in early August 2023). We also recommend starting with these additional resources:

> **What Every R&D Leader Needs to Know About ChatGPT and LLMs** on-demand webinar
> **enthought.com/resource/webinar-what-every-rd-leader-needs-to-know-about-chatgpt-and-llms**
>
> **The Practical Guides for Large Language Models**
> **github.com/Mooler0410/LLMsPracticalGuide/tree/main**

The Major Players

Of the major players in AI, only OpenAI is currently offering their LLMs as a commercial service, and then only by invitation (as of this writing). However, many companies have experimental or non-commercial models to explore. Keep IP issues in mind with these.

OpenAI - **openai.com**
OpenAI offers a variety of different LLMs and APIs addressing different use-cases, including fine-tuning models on your own data. Serious commercial use should be via the APIs, which are currently available by invitation.

Meta AI LLaMA 2 - **github.com/facebookresearch/llama/blob/main/MODEL_CARD.md**
A collection of related LLMs released by Meta AI (Facebook). Unlike version 1, version 2 is available for commercial and research purposes.

Google Bard - **bard.google.com**
Google's experimental LLM. No public APIs available yet, and chatbot conversations are used for further training, so not yet ready for commercial use.

Amazon AlexaTM - **github.com/amazon-science/alexa-teacher-models**
Amazon Science's LLM, which can be accessed for non-commercial use via AWS SageMaker.

Anthropic Claude - claude.ai
Unique model because of its large context window (100k+ tokens), allowing it to answer questions about longer documents. API access is only available via inquiries. A chat interface is generally available, but conversations may be used for further training, so not a commercial option.

Hugging Face - **huggingface.co**
Hugging Face provides infrastructure support for LLM and other Machine Learning operations, including hosting, training and deployment of models. They also host some internally developed and open-source models, such as BLOOM.

Open-Source LLMs

If you want to train, fine-tune, or run a LLM on your own, there are a number of models available, ranging from older models from major AI companies to non-commercial research models, to some more recent, permissively licensed models.

Google BERT - **github.com/google-research/bert**
One of the first openly available transformer-based LLMs and available under the permissive Apache 2.0 license. BERT is the foundation for many of the tools for scientific applications of LLMs.

OpenAI GPT-2 - **github.com/openai/gpt-2**
OpenAI's 2nd generation LLM, released under a permissive MIT license. GPT-2 is now 4 years old, so well-behind the state-of-the-art, but ground-breaking at the time.

BLOOM - **bigscience.huggingface.co/blog/bloom**
A multi-lingual LLM by a large consortium of researchers and organizations, including Hugging Face. It is open-sourced under the Responsible AI License (usable commercially with some restrictions, particularly around disclosure and medical use-cases). There is also BLOOMZ which is fine-tuned for following instructions rather than conversation.

Falcon LLM - **huggingface.co/tiiuae**
An LLM released by the Technology Innovation Institute under a permissive Apache 2.0 license. This is used as a basis for a number of other open tools, such as LAION's Open Assistant (https://open-assistant.io/).

MPT-30B - **mosaicml.com/blog/mpt-30b**
A collection of LLMs with different optimizations trained inexpensively on very large input sets. Released by MosaicML under the Apache 2.0 license with the intent that it is commercially usable.

Dolly/Pythia - **huggingface.co/databricks/dolly-v2-12b**
An LLM tuned by Databricks based on the Pythia LLM. It is not cutting edge but is large and released under an MIT license.

Stanford University Alpaca - **crfm.stanford.edu/2023/03/13/alpaca.html**
A model based on Meta's LLaMA v1 produced by the Center for Research on Foundation Models (CRFM) group at Stanford. The model is open-sourced under a non-commercial license and designed to be trained inexpensively on smaller data sets. There are a number of other models derived from this, such as **Vicuna** (**lmsys.org/blog/2023-03-30-vicuna**).

LeRF - **lerf.io**
LeRF combines the ability to reconstruct a 3D scene from a handful of still images using Neural Radiance Fields (NeRF) with LLMs, allowing easy searching of a 3D scene using natural language. The models and code are open source but currently without a license, and so not yet commercially usable.

Toolkits and APIs

To go beyond simple chat applications of LLMs, you will need some tools to connect the models with other services or even libraries to build and train your own models.

Transformers - **huggingface.co/docs/transformers/index**
A toolkit built on top of PyTorch and TensorFlow that provides building blocks for LLMs as well as other state-of-the-art machine learning models. It also integrates with the Hugging Face public API to facilitate building, training and running models in the cloud, as well as accessing many 3rd party models.

LangChain - **python.langchain.com/en/latest/index.html**
LangChain is a toolkit for building LLM-centered applications, particularly agents and assistants. It provides automation for building special-purpose prompts, which work well with LLMs to produce particular types of outputs and integration with other services such as data sources and code execution.

Science-Specific Tools

In the last few years, there have been a number of high-profile papers and toolkits in Material Science and Bioinformatics that use these new ML models. Most of these have source code and model weights freely available, but there are not yet any services built on top of these. They are research-grade software, not production-grade, with many based on LLM techniques that are a generation or two behind the current state-of-the-art. There are likely to be better models in the future.

ChemBERT - **github.com/HyunSeobKim/CHEM-BERT**
Chemical property prediction from SMILES molecular structure representation. There are other models derived from this original work.

ChemCrow - **github.com/ur-whitelab/chemcrow-public**
LangChain-based package for solving reasoning-intensive chemical tasks posed using natural language. This currently needs API access for OpenAI and possibly other APIs depending on the tasks.

ProteinBERT - **github.com/nadavbra/protein_bert**
A framework for building protein property predictors from protein sequence information. The base model is designed to be fine-tuned for arbitrary properties.

TransUNet - **github.com/Beckschen/TransUNet**
Next-generation medical image segmentation using transformer-based models. This has the potential to be cheaper to train and more capable of detecting large-scale structures in an image.

Enformer - **huggingface.co/EleutherAI/enformer-preview**
Transformer-based gene expression and chromatin prediction from DNA sequences. Similar to LLMs, Enformer has the capability of tracking a wider context within a DNA sequence than previous models.