Open Source Models and Hugging Face

Introduction

The world of open-source LLMs is exciting and constantly evolving. This lesson aims to provide an in-depth look at open source models. If you are looking for information on how proprietary models compare to open source models, go to the "Exploring and Comparing Different LLMs" lesson. This lesson will also cover the topic of fine-tuning but a more detailed explanation can be found in the "Fine-Tuning LLMs" lesson.

Learning goals

- Gain an understanding of open source Models
- Understanding the benefits of working with open source Models
- Exploring the open models available on Hugging Face and the Azure AI Studio

What are Open Source Models?

The Open Source Initiative (OSI) has defined 10 criteria for software to be classified as open source. The source code must be openly shared under a license approved by the OSI. For LLM the process is not exactly the same.

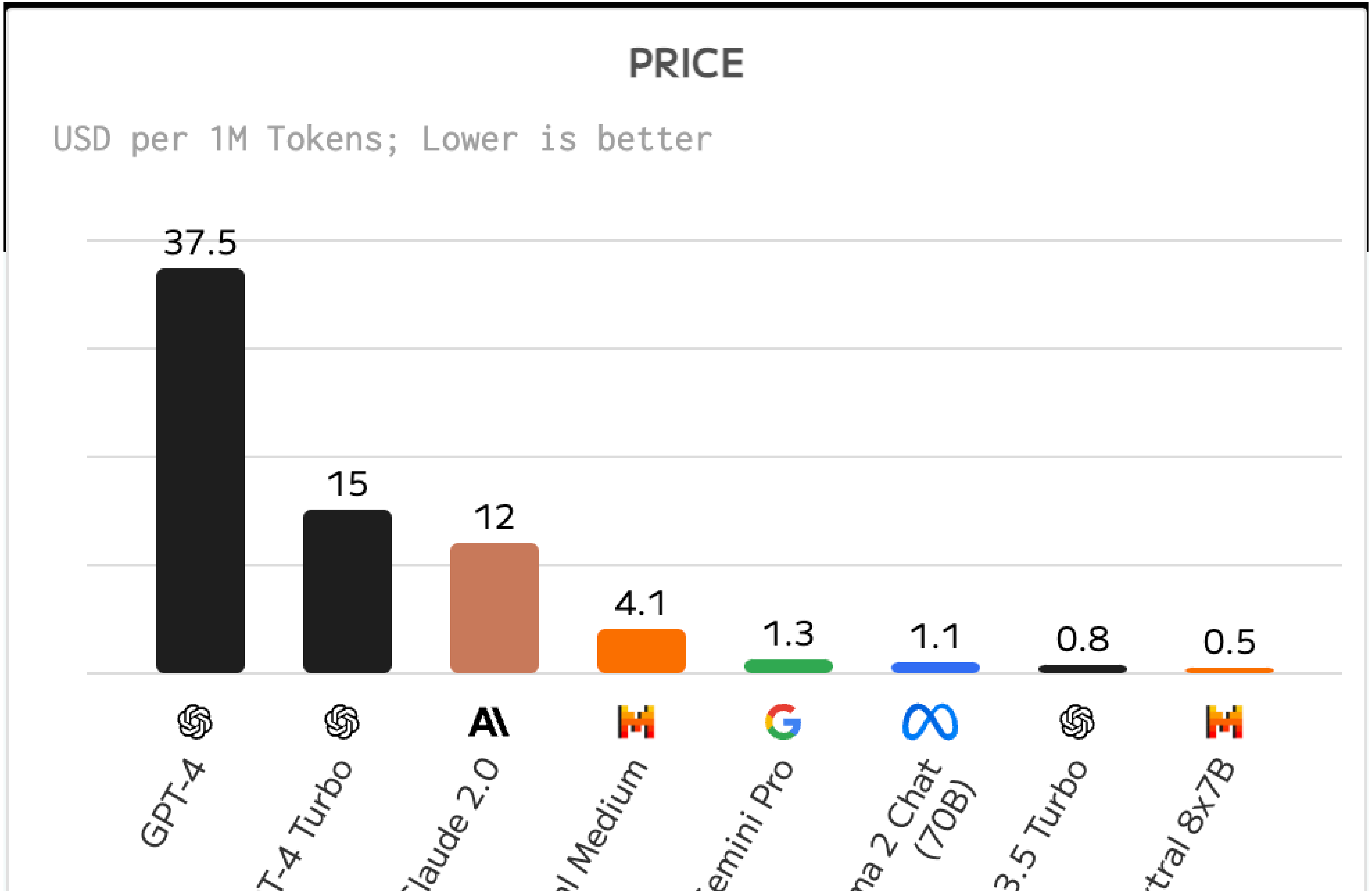
- Datasets used to train the model.
- Full model weights as a part of the training.
- The evaluation code.
- The fine-tuning code.
- Full model weights and training metrics.

There are currently only a few models that match this criteria. The OLMo model created by Allen Institute for Artificial Intelligence (AllenAI) is one that fits this category.

Benefits of Open Models

Highly Customizable - Since open models are released with detailed training information, researchers and developers can modify the model's internals. This enables the creation of highly specialized models that are fine-tuned for a specific task or area of study. Some examples of this are code generation, mathematical operations and biology.

Cost - The cost per token for using and deploying these models is lower than that of proprietary models. When building Generative AI applications, looking at performance vs price when working with these models on your use case should be done.


PRICE
USD per 1M Tokens; Lower is better
37.5
15
12
4.1
1.3
1.1
0.8
0.5
GPT-4
T-4 Turbo
laude 2.0
al Medium
emini Pro
ma 2 Chat
(70B)
3.5 Turbo
tral 8x7B

Exploring Different Open Models

Llama 2

LLama2, developed by Meta is an open model that is optimized for chat based applications. This is due to its fine-tuning method, which included a large amount of dialogue and human feedback. With this method, the model produces more results that are aligned to human expectation which provides a better user experience.

Some examples of fine-tuned versions of Llama include Japanese Llama, which specializes in Japanese and Llama Pro, which is an enhanced version of the base model.

Mistral

Mistral is an open model with a strong focus of high performance and efficiency. It uses the Mixture-of-Experts approach which combines a group of specialized expert models into one system where depending on the input, certain models are selected to be used. This makes the computation more effective as models are only addressing the inputs they are specialized in.

Some examples of fine-tuned versions of Mistral include BioMistral, which is focused on the medical domain and OpenMath Mistral, which performs mathematical computation.

Falcon

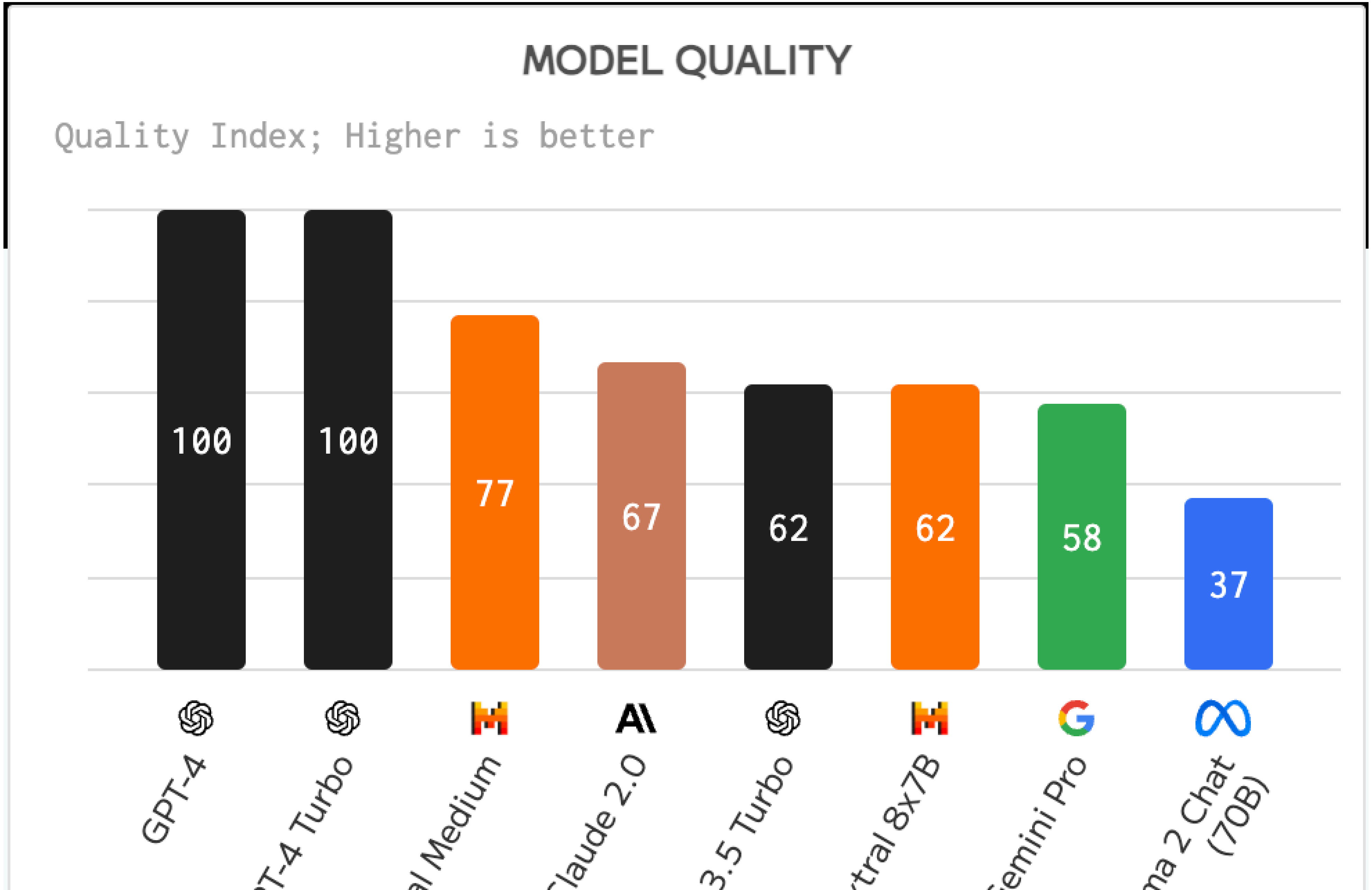
Falcon is an LLM created by the Technology Innovation Institute (**TII**). The Falcon-40B was trained on 40 billion parameters which has been shown to perform better than GPT-3 with less compute budget. This is due to its use of the FlashAttention algorithm and multiquery attention that enables it to cut down on the memory requirements at inference time. With this reduced inference time, the Falcon-40B is suitable for chat applications.

Some examples of fine-tuned versions of Falcon are the OpenAssistant, an assistant built on open models and GPT4ALL, which delivers higher performance than the base model.

How to Choose

There is no one answer for choosing an open model. A good place to start is by using the Azure AI Studio's filter by task feature. This will help you understand what types of tasks the model has been trained for. Hugging Face also maintains an LLM Leaderboard which shows you the best performing models based on certain metrics.

When looking to compare LLMs across the different types, Artificial Analysis is another great resource:


MODEL QUALITY
Quality Index; Higher is better
100
100
77
67
62
62
58
37
GPT-4
T-4 Turbo
al Medium
laude 2.0
3.5 Turbo
tral 8x7B
emini Pro
ma 2 Chat
(70B)