What is GPT-OSS?

gpt-oss is a family of open AI models, released under the Apache 2.0 license, allowing for free use.

OpenAI is once again returning to open-source AI models. The developer ChatGPT has released two more large language models (LLMs) under an open license for the first time since GPT-2 in 2019. The naming convention is somewhat confusing, but this time it can be overlooked because gpt-oss-120b and gpt-oss-20b are incredibly interesting models.

What is GPT-OSS?

gpt-oss is a family of open-source AI models, released under the Apache 2.0 license for free use. They are advanced inference models that anyone can download, tweak, and use for almost any purpose – although OpenAI has taken steps to limit how they can be used for malicious purposes or to generate harmful information.

This is a major step forward because, since 2019, all GPT and o-series models have been proprietary. With gpt-oss, OpenAI has lifted the veil of secrecy.

It's also worth noting that gpt-oss-120b and gpt-oss-20b are among the highest-performing open models from North American and European AI labs. For those concerned about how Chinese models are trained and the inherent censorship in their training data, this makes OpenAI's latest models even more important.

gpt-oss-120b and gpt-oss-20b

gpt-oss-120b and gpt-oss-20b are the first two models in this product line, and aside from the fact that they aren't proprietary like other OpenAI models, they look quite similar to the rest:

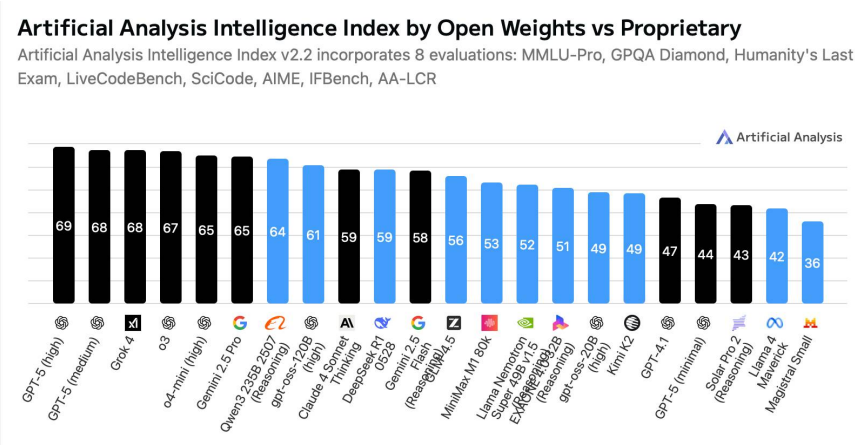
1. Both models utilize a Mixture-of-Experts architecture. The larger model, gpt-oss-120b, has a total of 117 billion parameters across 128 subnets (experts), with the remaining 5.1 billion parameters being activated at some point. The smaller model, gpt-oss-20b, has 21 billion parameters across 32 subnets (experts), with the remaining 3.6 billion parameters being activated at some point.
2. Both models are inference models, so they are capable of using chain reasoning (CoT) to solve complex problems. They have 3 levels of reasoning: Low, Medium, and High.
3. Both models are LLMs (Learning Language Modules), not large multimodal models (LMMs), so they only support text – they don't support audio, images, or any other methods.
4. Both models support a context length of 128k tokens.
5. Both models are tool-enabled, so they can be used for web browsing, coding, and working in agentic systems.

The two models were primarily trained on English text focusing on STEM content, programming, and general knowledge. Data related to chemical, biological, radiological, and nuclear (CBRN) threats were filtered out to ensure maximum safety.

In addition to filtering training data, OpenAI retrained the models using deliberate alignment and instruction hierarchy to prevent them from responding to unsafe prompts and to mitigate the risk of malicious code injection into prompts. These are the same techniques OpenAI uses to secure its proprietary models.

How good is GPT-OSS?

Open-source programming languages ??are experiencing a boom. In particular, Chinese labs like DeepSeek, Qwen, Moonshot, and Z.ai are releasing incredibly competitive open-source models. While the best proprietary models still outperform the best open-source ones, the gap has narrowed considerably. So where do gpt-oss-120b and gpt-oss-20b rank in the standings?



OpenAI claims that gpt-oss-120b delivers similar performance to o4-mini and gpt-oss-20b delivers similar performance to o3-mini in key performance tests, and independent analyses largely support this. In other words, they are indeed very good modern inference models.

But this analysis misses one important detail.

Currently, gpt-oss-120b is the smartest pattern that can run on a single NVIDIA H100 graphics card, and gpt-oss-20b is the smartest pattern that can run on a consumer GPU (or even a laptop with only 16GB of RAM). These patterns are not optimized for raw performance, but instead are designed to be extremely intelligent relative to their number of parameters and subnets.

For example, DeepSeek R1 offers superior performance compared to gpt-oss-120b, but it has a total of 671 billion parameters and 37 billion active parameters (compared to 117 billion parameters and 5.1 billion active parameters), making it consume more than 10 times as much memory. You have to really want every bit of extra performance for the additional cost to be worthwhile.

While the open language modeling space is rapidly evolving, it's safe to say that both gpt-oss-120b and gpt-oss-20b are high-performance, modern models with outstanding efficiency. If OpenAI continues to support them or releases more models in the gpt-oss family, they are likely to remain relevant in the near future.

How to use GPT-OSS-120B and GPT-OSS-20B

Like most open-source models, you can download gpt-oss-120b and gpt-oss-20b from Hugging Face right now. While gpt-oss-120b requires a server-grade GPU to run, you can run gpt-oss-20b on many modern MacBooks.

OpenAI has also partnered with inference providers such as Azure, vLLM, Ollama, LM Studio, AWS, Fireworks, Databricks, Vercel, and OpenRouter to offer gpt-oss-120b and gpt-oss-20b to developers. They provide these two models as APIs with various pricing tiers and a range of features to suit different application needs.

As open-source weighting models, gpt-oss-120b and gpt-oss-20b can be fine-tuned for specific purposes. This can be done by downloading the models yourself or using a third-party inference provider.