

openai/gpt-oss-120b


like


3.87k


Follow


OpenAI


21.7k


 Text Generation


 Transformers


 Safetensors


 gpt\_oss


 vllm

 conversational

 8-bit precision

 mxfp4

 arxiv:2508.10925

 License: apache-2.0

Model card

Files and versions

X xet

Community

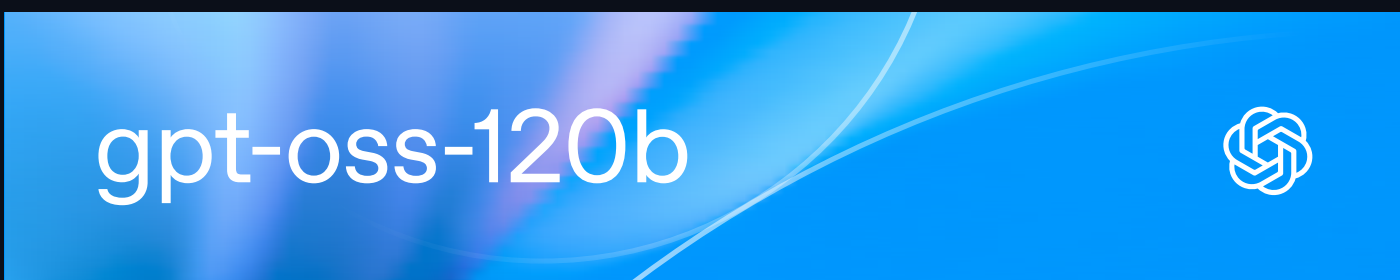
128

i

Train

Deploy

Use this model



[Try gpt-oss](#) · [Guides](#) · [Model card](#) · [OpenAI blog](#)

Welcome to the gpt-oss series, OpenAI's open-weight models designed for powerful reasoning, agentic tasks, and versatile developer use cases.

We're releasing two flavors of these open models:

- gpt-oss-120b** — for production, general purpose, high reasoning use cases that fit into a single 80GB GPU (like NVIDIA H100 or AMD MI300X) (117B parameters with 5.1B active parameters)
- gpt-oss-20b** — for lower latency, and local or specialized use cases (21B parameters with 3.6B active parameters)

Both models were trained on our **harmony response format** and should only be used with the harmony format as it will not work correctly otherwise.

This model card is dedicated to the larger **gpt-oss-120b** model. Check out **gpt-oss-20b** for the smaller model.

## Highlights

- Permissive Apache 2.0 license:** Build freely without copyleft restrictions or patent risk—ideal for experimentation, customization, and commercial deployment.
- Configurable reasoning effort:** Easily adjust the reasoning effort (low, medium, high) based on your specific use case and latency needs.
- Full chain-of-thought:** Gain complete access to the model's reasoning process, facilitating easier debugging and increased trust in outputs. It's not intended to be shown to end users.
- Fine-tunable:** Fully customize models to your specific use case through parameter fine-tuning.
- Agentic capabilities:** Use the models' native capabilities for function calling, **web browsing**, **Python code execution**, and Structured Outputs.
- MXFP4 quantization:** The models were post-trained with MXFP4 quantization of the MoE weights, making **gpt-oss-120b** run on a single 80GB GPU (like NVIDIA H100 or AMD MI300X) and the **gpt-oss-20b** model run within 16GB of memory. All evals were performed with the same MXFP4 quantization.

## Inference examples

### Transformers

You can use **gpt-oss-120b** and **gpt-oss-20b** with Transformers. If you use the Transformers chat template, it will automatically apply the **harmony response format**. If you use `model.generate` directly, you need to apply the harmony format manually using the chat template or use our **openai-harmony** package.

To get started, install the necessary dependencies to setup your environment:

```
pip install -U transformers kernels torch
```

Once, setup you can proceed to run the model by running the snippet below:

```
from transformers import pipeline
import torch

model_id = "openai/gpt-oss-120b"

pipe = pipeline(
    "text-generation",
    model=model_id,
    torch_dtype="auto",
    device_map="auto",
)

messages = [
    {"role": "user", "content": "Explain quantum mechanics clearly and concisely."},
]

outputs = pipe(
    messages,
    max_new_tokens=256,
)

print(outputs[0]["generated_text"][-1])
```

Alternatively, you can run the model via **Transformers Serve** to spin up a OpenAI-compatible webserver:

```
transformers serve
transformers chat localhost:8080 --model-name-or-path openai/gpt-oss-120b
```

[Learn more about how to use gpt-oss with Transformers.](#)

### vLLM

vLLM recommends using **uv** for Python dependency management. You can use vLLM to spin up an OpenAI-compatible webserver. The following command will automatically download the model and start the server.

```
uv pip install --pre vllm==0.10.1+gptoss \
--extra-index-url https://wheels.vllm.ai/gpt-oss/ \
--extra-index-url https://download.pytorch.org/whl/nightly/cu128 \
--index-strategy unsafe-best-match

vllm serve openai/gpt-oss-120b
```

[Learn more about how to use gpt-oss with vLLM.](#)

### PyTorch / Triton

To learn about how to use this model with PyTorch and Triton, check out our [reference implementations in the gpt-oss repository](#).

### Ollama

If you are trying to run gpt-oss on consumer hardware, you can use Ollama by running the following commands after [installing Ollama](#).

```
# gpt-oss-120b
ollama pull gpt-oss:120b
ollama run gpt-oss:120b
```

[Learn more about how to use gpt-oss with Ollama.](#)

### LM Studio

If you are using **LM Studio** you can use the following commands to download.

```
# gpt-oss-120b
lms get openai/gpt-oss-120b
```

Check out our [awesome list](#) for a broader collection of gpt-oss resources and inference partners.

## Download the model

You can download the model weights from the [Hugging Face Hub](#) directly from Hugging Face CLI:

```
# gpt-oss-120b
huggingface-cli download openai/gpt-oss-120b --include "original/*" --local-dir gpt-oss-120b/
pip install gpt-oss
python -m gpt_oss.chat model/
```

## Reasoning levels

You can adjust the reasoning level that suits your task across three levels:

- Low:** Fast responses for general dialogue.
- Medium:** Balanced speed and detail.
- High:** Deep and detailed analysis.

The reasoning level can be set in the system prompts, e.g., "Reasoning: high".

## Tool use

The gpt-oss models are excellent for:

- Web browsing (using built-in browsing tools)
- Function calling with defined schemas
- Agentic operations like browser tasks

## Fine-tuning

Both gpt-oss models can be fine-tuned for a variety of specialized use cases.

This larger model **gpt-oss-120b** can be fine-tuned on a single H100 node, whereas the smaller **gpt-oss-20b** can even be fine-tuned on consumer hardware.

## Citation

```
@misc{openai2025gptoss120bgptoss20bmodel,
  title={gpt-oss-120b & gpt-oss-20b Model Card},
  author={OpenAI},
  year={2025},
  eprint={2508.10925},
  archivePrefix={arXiv},
  primaryClass={cs.CL},
  url={https://arxiv.org/abs/2508.10925},
}
```



### Safetensors

Model size 120B params Tensor type BF16 - U8 Chat template Files info

### Inference Providers

Run 15,000+ Models Instantly

Inference Providers let you run inference on thousands of models served by our partners using a simple, unified, OpenAI-compatible serverless API ([Learn more](#)).

openai/gpt-oss-120b is supported by the following Inference Providers:

Scaleway

Together AI

Nscale

Novita

Nebius AI

Hyperbolic

Groq

Fireworks

Cerebras

View API Code

Dismiss

Examples

Send

View Code Snippets

Open Playground

### Model tree for openai/gpt-oss-120b

Adapters	21 models
Finetunes	60 models
Merges	2 models
Quantizations	54 models

### Spaces using openai/gpt-oss-120b

ginipick/FLUXllama

amd/gpt-oss-120b-chatbot

eduagarcia/open\_pt\_llm\_leaderboard

umint/ai

umint/searchgpt

openfree/OpenAI-gpt-oss

SustainabilityLabITGN/VayuChat

abidlabs/openai-gpt-oss-120b-test

HPAI-BSC/TuRTLe-Leaderboard

VIDraft/gpt-oss-RAG

millwright/chatui-helper

fdaudens/gpt-oss-news-agent

+ 88 Spaces

### Collection including openai/gpt-oss-120b

gpt-oss

Collection

Open-weight models designed for powerf... • 2 items • Updated Aug 7 • Δ 354

System theme

TOS

Privacy

About

Jobs

Models

Datasets

Spaces

Pricing

Docs