

OpenAI releases gpt-oss, its first open-weights language models since GPT-2

Source copy kept by wheresthe.ai for the story "OpenAI releases gpt-oss, its first open-weights language models since GPT-2" (<https://wheresthe.ai/d/openai-gpt-oss-open-weights-2025-08/>).

Source: <https://openai.com/index/introducing-gpt-oss/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: be7783d3552f1fa10a921927b22edcead0b94d305975b338e130a6b1e0f276a1

gpt-oss-120b and gpt-oss-20b push the frontier of open-weight reasoning models

Listen to article 12:24

* Introduction * Pre-training & model architecture * Post-training * Evaluations * Chain-of-thought (CoT)

* Safety & worst-case fine-tuning * Availability * Why open models matter

Introduction

We're releasing gpt-oss-120b and gpt-oss-20b—two state-of-the-art open-weight language models that deliver strong real-world performance at low cost. Available under the flexible Apache 2.0 license, these models outperform similarly sized open models on reasoning tasks, demonstrate strong tool use capabilities, and are optimized for efficient deployment on consumer hardware. They were trained using a mix of reinforcement learning and techniques informed by OpenAI's most advanced internal models, including o3 and other frontier systems.

The gpt-oss-120b model achieves near-parity with OpenAI o4-mini on core reasoning benchmarks, while running efficiently on a single 80 GB GPU. The gpt-oss-20b model delivers similar results to OpenAI o3-mini on common benchmarks and can run on edge devices with just 16 GB of memory, making it ideal for on-device use cases, local inference, or rapid iteration without costly infrastructure. Both models also perform strongly on tool use, few-shot function calling, CoT reasoning (as seen in results on the Tau-Bench agentic evaluation suite) and HealthBench (even outperforming proprietary models like OpenAI o1 and GPT-4o).

These models are compatible with our Responses API([opens in a new window](#)) and are designed to be used within agentic workflows with exceptional instruction following, tool use like web search or Python code execution, and reasoning capabilities—including the ability to adjust the reasoning effort for tasks that don't require complex reasoning and/or target very low latency final outputs. They are entirely customizable, provide full chain-of-thought (CoT), and support Structured Outputs([opens in a new window](#)).

Safety is foundational to our approach to releasing all our models, and is of particular importance for open models. In addition to running the models through comprehensive safety training and evaluations, we also introduced an additional layer of evaluation by testing an adversarially fine-tuned version of gpt-oss-120b under our Preparedness Framework([opens in a new window](#)). gpt-oss models perform comparably to our frontier models on internal safety benchmarks, offering developers the same safety standards as our recent proprietary models. We're sharing the results of that work and more details in a research paper([opens in a new window](#)) and in the model card([opens in a new window](#)). Our methodology was reviewed by external experts and marks a step forward in setting new safety standards for open-weight models.

We've also been working with early partners like AI Sweden([opens in a new window](#)), Orange([opens in a new window](#)), and Snowflake([opens in a new window](#)) to learn about real-world applications of our open models, from hosting these models on-premises for data security to fine-tuning them on specialized datasets. We're excited to provide these best-in-class open models to empower everyone—from individual developers to large enterprises to governments—to run and customize AI on their own infrastructure. Coupled with the models available in our API, developers can choose the performance, cost, and latency they need to power AI workflows.

Pre-training & model architecture

The gpt-oss models were trained using our most advanced pre-training and post-training techniques, with particular focus on reasoning, efficiency, and real-world usability across a wide range of deployment

environments. While we have made other models including Whisper and CLIP available openly, gpt-oss models are our first open-weight language models since GPT-2[1].

Each model is a Transformer which leverages mixture-of-experts (MoE[2]) to reduce the number of active parameters needed to process input. gpt-oss-120b activates 5.1B parameters per token, while gpt-oss-20b activates 3.6B. The models have 117b and 21b total parameters respectively. The models use alternating dense and locally banded sparse attention patterns, similar to GPT-3[3]. For inference and memory efficiency, the models also use grouped multi-query attention, with a group size of 8. We use Rotary Positional Embedding (RoPE[4]) for positional encoding, and natively support context lengths of up to 128k.

Model	Layers	Total Params	Active Params Per Token	Total Experts	Active Experts Per Token	Context Length
gpt-oss-120b	36	117B	5.1B	128	4	128k
gpt-oss-20b	24	21B	3.6B	32	4	128k

We trained the models on a mostly English, text-only dataset, with a focus on STEM, coding, and general knowledge. We tokenized the data using a superset of our tokenizer used for OpenAI o4-mini and GPT-4o: o200k_harmony, which we are also open-sourcing today.

For more on our models' architecture and training, read the model card(opens in a new window).

Post-training

The models were post-trained using a similar process as used for o4-mini, including a supervised fine-tuning stage and a high-compute RL stage. Our objective was to align the models with the OpenAI Model Spec(opens in a new window) and teach it to apply CoT reasoning and tool use before producing its answer. By using the same techniques as our SoTA proprietary reasoning models, the models demonstrate exceptional capabilities after post-training.

Similar to the OpenAI o-series reasoning models in the API, the two open-weight models support three reasoning efforts—low, medium, and high—which trade off latency vs. performance. Developers can easily set the reasoning effort with one sentence in the system message.

Evaluations

We evaluated gpt-oss-120b and gpt-oss-20b across standard academic benchmarks to measure their capabilities in coding, competition math, health, and agentic tool use when compared to other OpenAI reasoning models including o3, o3-mini and o4-mini.

gpt-oss-120b outperforms OpenAI o3-mini and matches or exceeds OpenAI o4-mini on competition coding (Codeforces), general problem solving (MMLU and HLE) and tool calling (TauBench). It furthermore does even better than o4-mini on health-related queries (HealthBench) and competition mathematics (AIME 2024 & 2025). gpt-oss-20b matches or exceeds OpenAI o3-mini on these same evals, despite its small size, even outperforming it on competition mathematics and health.

gpt-oss models do not replace a medical professional and are not intended for the diagnosis or treatment of disease

Example rollouts

gpt-oss-120b is able to quickly aggregate up-to-date information using a browsing tool, including chaining together 10s of subsequent calls.

Chain-of-thought (CoT)

Our recent research has shown that monitoring a reasoning model's CoT can be helpful for detecting misbehavior as long as the model was not trained with direct supervision for aligning the CoT. This perspective is shared(opens in a new window) by others in the industry as well. In line with our principles since launching OpenAI o1-preview, we did not put any direct supervision on the CoT for either gpt-oss model. We believe this is critical to monitor model misbehavior, deception and misuse. Our hope is that releasing an open model with a non-supervised chain of thought gives developers and researchers the opportunity to research and implement their own CoT monitoring systems.

Developers should not directly show CoTs to users in their applications. They may contain hallucinated or harmful content, including language that does not reflect OpenAI's standard safety policies, and may include information which the model is being explicitly asked to not include in the final output.

gpt-oss-120b robustly follows system instructions in its output, but will often explicitly disobey instructions in its CoT.

Safety & worst-case fine-tuning

The gpt-oss models leverage our state-of-art approaches for safety training. During pre-training, we filtered out certain harmful data related to Chemical, Biological, Radiological, and Nuclear (CBRN). During post-training, we used deliberative alignment and the instruction hierarchy([opens in a new window](#)) to teach the model to refuse unsafe prompts and defend against prompt injections.

Once an open-weight model is released, adversaries may be able to fine-tune the model for malicious purposes. We directly assessed these risks by fine-tuning the model on specialized biology and cybersecurity data, creating a domain-specific non-refusing version for each domain the way an attacker might. We then evaluated the capability level of these models through internal and external testing. This testing, as detailed in our accompanying safety paper, indicated that, even with robust fine-tuning that leveraged OpenAI's field-leading training stack, these maliciously fine-tuned models were unable to reach high capability levels according to our Preparedness Framework. This malicious fine-tuning methodology was reviewed by three independent expert groups who made recommendations to improve the training process and evaluations, many of which we adopted. We detail these recommendations in the model card. These processes mark a meaningful advancement for open model safety. These findings informed our decision to release the gpt-oss models. We hope that these models will help accelerate safety training and alignment research across the industry.

To contribute to a safer open source ecosystem, we are hosting a Red Teaming Challenge([opens in a new window](#)) to encourage researchers, developers, and enthusiasts from around the world to help identify novel safety issues. The challenge has a \$500,000 prize fund that will be awarded based on review from a panel of expert judges from OpenAI and other leading labs. At the end of the challenge, we will publish a report and open-source an evaluation data set based on validated findings, so that the wider community can immediately benefit. [Learn more and participate here\(opens in a new window\)](#).

Availability

The weights for both gpt-oss-120b and gpt-oss-20b are freely available for download on Hugging Face and come natively quantized in MXFP4. This allows for the gpt-oss-120B model to run within 80GB of memory, while gpt-oss-20b only requires 16GB.

The models have been post-trained on our harmony prompt format([opens in a new window](#)), and we're open-sourcing a harmony renderer([opens in a new window](#)) in both Python and Rust to make adoption easier. We're also releasing reference implementations for running inference with PyTorch and on Apple's Metal platform, along with a collection of example tools for the model.

We've designed these models to be flexible and easy to run anywhere—locally, on-device, or through third-party inference providers. To support this, we partnered ahead of launch with leading deployment platforms such as Azure, Hugging Face, vLLM, Ollama, llama.cpp, LM Studio, AWS, Fireworks, Together AI, Baseten, Databricks, Vercel, Cloudflare, and OpenRouter to make the models broadly accessible to developers. On the hardware side, we worked with industry leaders including NVIDIA, AMD, Cerebras, and Groq to ensure optimized performance across a range of systems.

As part of today's release, Microsoft is also bringing GPU-optimized versions of the gpt-oss-20b model to Windows devices. Powered by ONNX Runtime, these models support local inference and are available through Foundry Local and the AI Toolkit for VS Code, making it easier for Windows developers to build with open models.

For developers who want fully customizable models they can fine-tune and deploy in their own environments, gpt-oss is a great fit. For those seeking multimodal support, built-in tools, and seamless integration with our platform, models available through our API platform remain the best option. We're continuing to listen closely to developer feedback and may consider API support for gpt-oss in the future.

If you want to try the models, head over to our open model playground([opens in a new window](#)). To learn more about how to use the models using different ecosystem providers or how to fine-tune the models, check out our guides([opens in a new window](#)).

Why open models matter

Releasing gpt-oss-120b and gpt-oss-20b marks a significant step forward for open-weight models. At their size, these models deliver meaningful advancements in both reasoning capabilities and safety. Open models complement our hosted models, giving developers a wider range of tools to accelerate leading edge research, foster innovation and enable safer, more transparent AI development across a wide range of use cases.

These open models also lower barriers for emerging markets, resource-constrained sectors, and smaller organizations that may lack the budget or flexibility to adopt proprietary models. With powerful, accessible tools in their hands, people around the world can build, innovate, and create new opportunities for themselves and others. Broad access to these capable open-weights models created in the US helps expand democratic AI rails.

A healthy open model ecosystem is one dimension to helping make AI widely accessible and beneficial for everyone. We invite developers and researchers to use these models to experiment, collaborate and push the boundaries of what's possible. We look forward to seeing what you build.

* 2025

Author

OpenAI

Contributors

Zoran Martinovic, Zhuohan Li, Zhiqing Sun, Zach Johnson, Yu Yang, Yu Bai, Yang Song, Xin Wang, Wenting Zhan, Volodymyr Kyrylov, Vlad Fomenko, Tyler Bertao, Tong Mu, Timur Garipov, Tarun Gogineni, Suvansh Sanjeev, Steve Mostovoy, Song Mei, Shengjia Zhao, Sebastien Bubeck, Scott McKinney, Scott Lessans, Sandhini Agarwal, Sam Toizer, Sam Altman, Saachi Jain, Romain Huet, Rahul K. Arora, Philippe Tillet, Olivia Watkins, Nivedita Brett, Nikhil Vyas, Miles Wang, Michihiro Yasunaga, Michelle Pokrass, Mia Glaese, Max Schwarzer, Mark Chen, Mario Lezcano-Casado, Marat Dukhan, Lukas Gross, Ludovic Peran, Ludovic Peran, Lindsay McCallum, Lin Yang, Lily (Xiaoxuan) Liu, Leher Pathak, Lama Ahmad, Kristian Georgiev, Kristen Ying, Kimmy Richardson, Kevin Whinnery, Kevin Weil, Kevin Lu, Kevin Fives, Kendal Simon, Katia Gil Guzman, Karan Singhal, Karan Singhal, Kai Chen, Josh McGrath, Jordan Liss, Jongsoo Park, John Hallman, Johannes Heidecke, Jiancheng Liu, Ji Lin, Jason Kwon, Jason Ai, James Park Lennon, Jakub Pachocki, Jacob Huh, Jackie Hehir, Irina Kofman, Huida Qiu, Hongyu Ren, Harshit Sikchi, Hannah Wong, Haitang Hu, Haitang Hu, Haiming Bao, Hadi Salman, Guillaume Leclerc, Greg Brockman, Gideon Myles, Giambattista Parascandolo, Gaby Raila, Foivos Tsimpourlas, Filippo Raso, Eugene Brevdo, Eric Wallace, Enoch Cheung, Elizabeth Proehl, Elaine Ya Le, Edwin Arbus, Eddie Zhang, Dominik Kundel, Dmitry Pimenov, David Robinson, Dane Stuckey, Dana Palmie, Dan Cook, Cyril Zhang, Chris Lu, Chris Koch, Che Chang, Cedric Whitney, Casey Dvorak, Carolina Paz, Brian Zhang, Bowen Baker, Bob Rotsted, Boaz Barak, Ashley Pantuliano, Andy Applebaum, Amy Wendling, Ally Bennett, Alexander Neitz, Alex Paino, Alex Nichol, Alec Helyar, Aidan McLaughlin, Aidan Clark, Adam Goucher