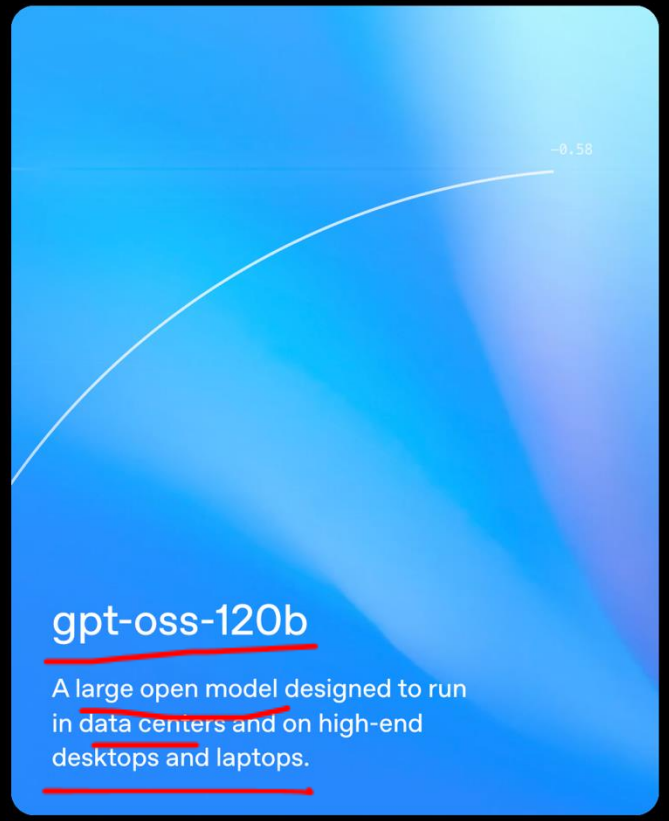
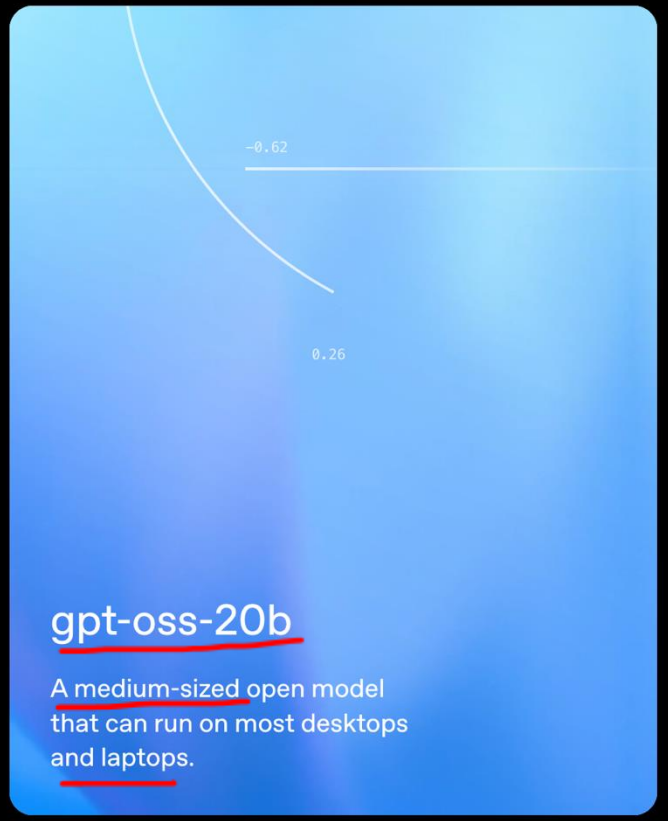
GPT-OSS Introduction & Hands-on





Chúng tôi đang triển khai gpt-oss-120b và gpt-oss-20b - hai mô hình ngôn ngữ tiên tiến có trọng số mở, giúp mang lại hiệu suất thực tế mạnh mẽ với chi phí thấp. Được phát hành theo giấy phép Apache 2.0 linh hoạt, các mô hình này vượt trội hơn so với các open model cùng kích thước trong quá trình thực hiện các nhiệm vụ lập luận, thể hiện khả năng sử dụng công cụ mạnh mẽ, và được tối ưu hóa để triển khai hiệu quả trên phần cứng của người dùng. Các mô hình này được huấn luyện bằng cách học tập tăng cường và kết hợp các kỹ thuật được xây dựng dựa trên các mô hình nội bộ tiên tiến nhất của OpenAI, bao gồm o3 và các hệ thống tiên phong khác.

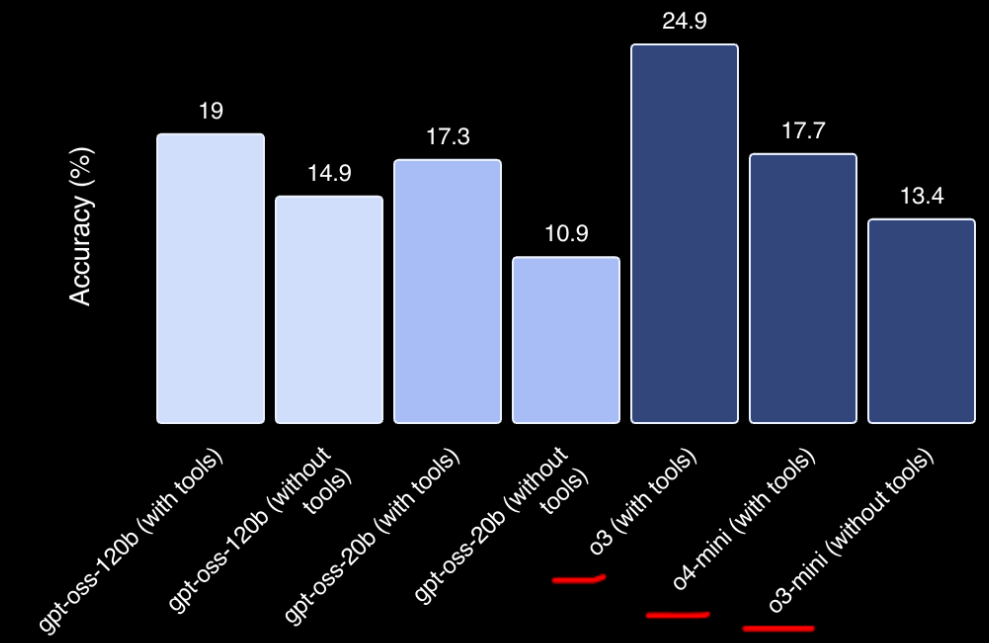
1. Cho phép sử dụng thương mại
2. Cho phép sửa đổi & phân phối lại,
3. Không cần mở mã nguồn khi phân phối phần mềm đã chỉnh sửa

Mô hình gpt-oss-120b đạt hiệu năng gần tương đương với OpenAI o4-mini khi so sánh đối chuẩn về tư duy cốt lõi, đồng thời vận hành hiệu quả trên một GPU 80 GB duy nhất. Mô hình gpt-oss-20b cho kết quả tương đương với OpenAI o3-mini trên các bài kiểm tra đối chuẩn thông thường và có thể chạy trên các thiết bị biên chỉ cần 16 GB bộ nhớ, từ đó trở thành mô hình lý tưởng cho các trường hợp sử dụng trên thiết bị, suy luận cục bộ, hoặc lặp lại nhanh mà không cần cơ sở hạ tầng tốn kém. Cả hai mô hình cũng thể hiện mạnh mẽ khi sử dụng công cụ, tương tác với các hàm few-shot, suy luận CoT (có thấy trong kết quả trên bộ đánh giá thực hiện độc lập Tau-Bench) và HealthBench (thậm chí vượt trội hơn các mô hình độc quyền như OpenAI o1 và GPT-4o).

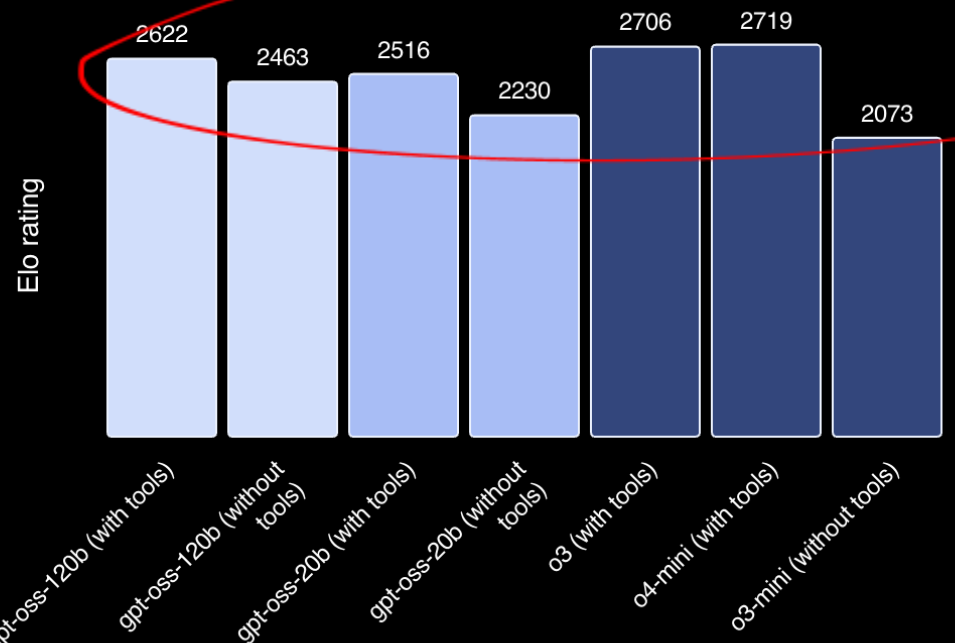
Mô hình	Lớp	Tổng số tham số	Tham số đang hoạt động trên mỗi token	Tổng số chuyên gia	Chuyên gia đang hoạt động trên mỗi token	Độ dài ngữ cảnh
gpt-oss-120b	36	117 tỷ	5,1 tỷ	128	4	128 nghìn
gpt-oss-20b	24	21 tỷ	3,6 tỷ	32	4	128 nghìn

Có thể tải xuống miễn phí các trọng số của cả hai mô hình gpt-oss-120b và gpt-oss-20b trên Hugging Face và được lượng tử hóa sẵn ở MXFP4. Điều này cho phép mô hình gpt-oss-120B có thể chạy trong bộ nhớ 80 GB, còn gpt-oss-20b chỉ yêu cầu 16 GB.

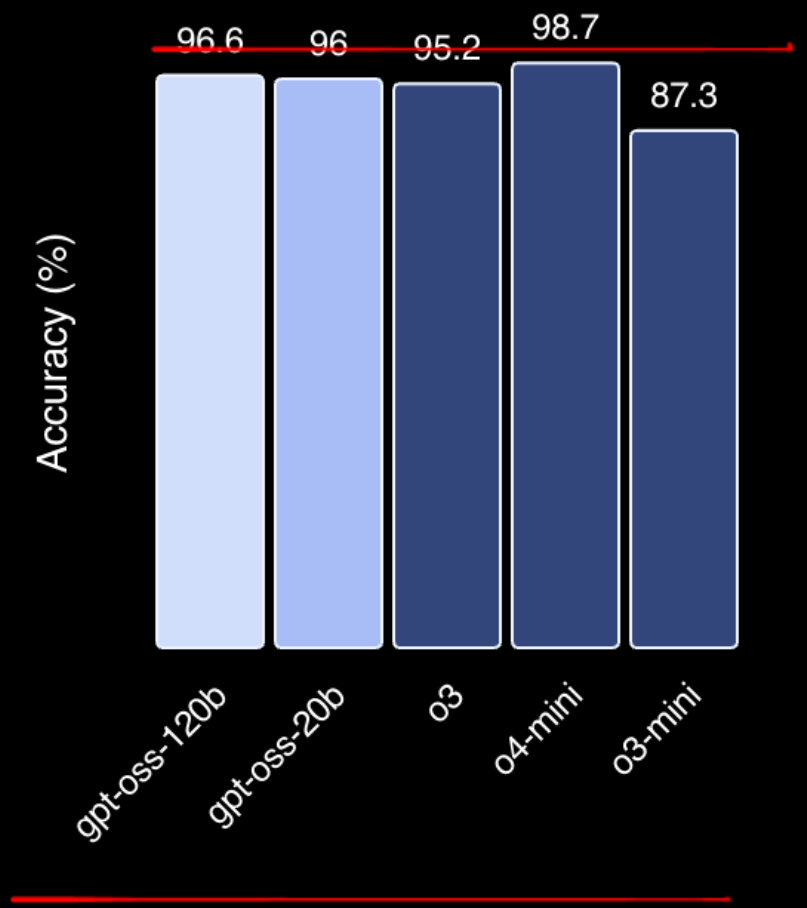
Humanity's Last Exam
Expert-level questions across subjects



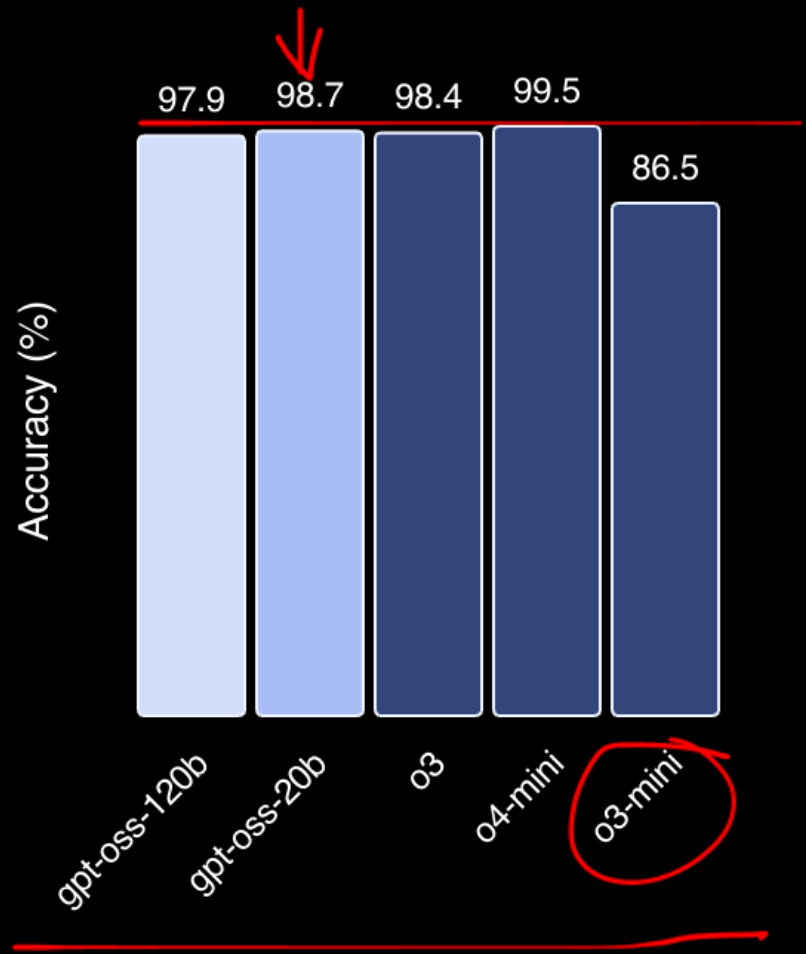
Codeforces
Competition code



AIME 2024 (tools)
Competition math



AIME 2025 (tools)
Competition math



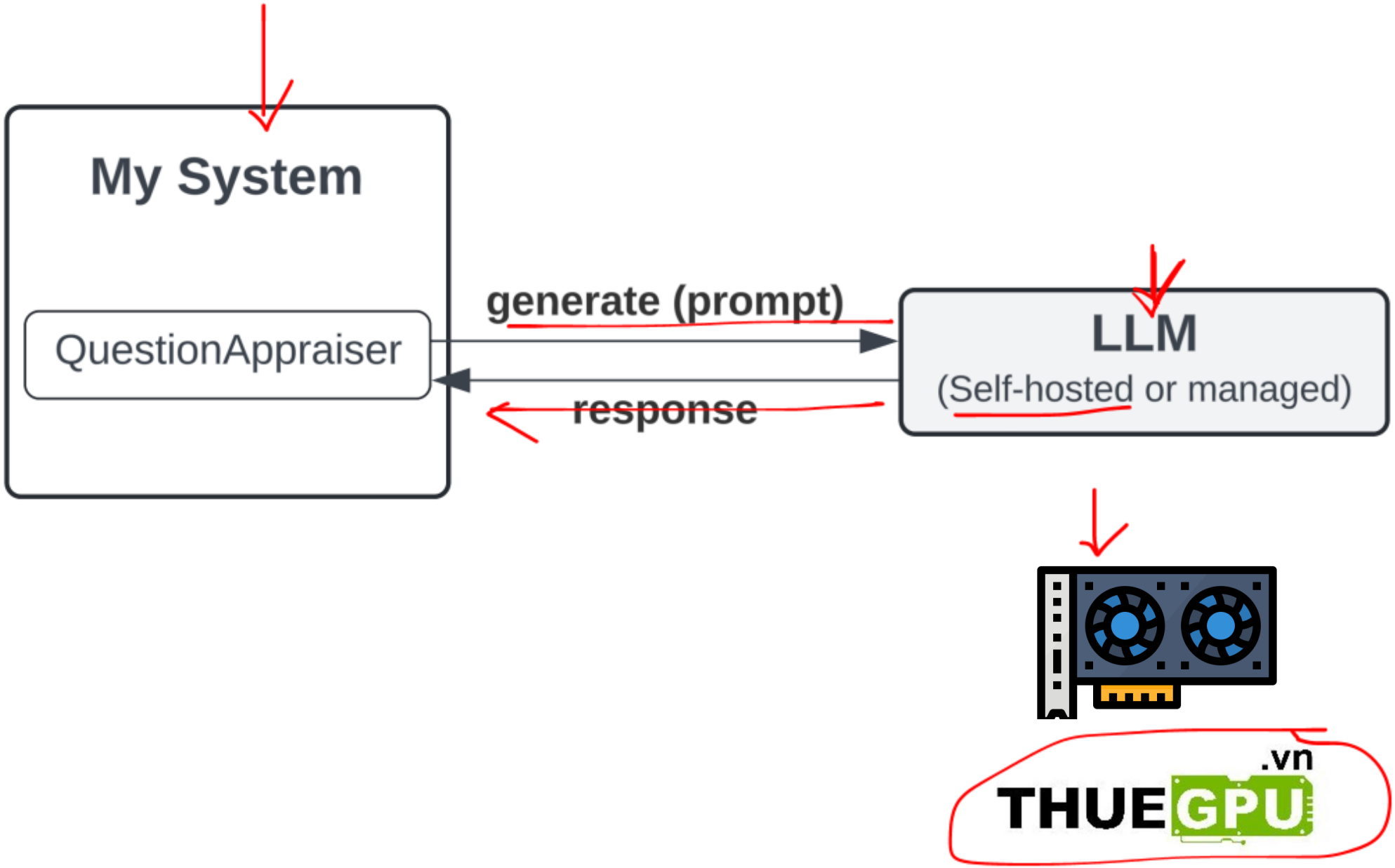
Chuỗi tư duy (CoT)

Nghiên cứu gần đây của chúng tôi cho thấy hoạt động giám sát CoT của một mô hình suy luận có thể giúp phát hiện hành vi sai trái, với điều kiện là mô hình đó không được huấn luyện về điều chỉnh Chuỗi tư duy (CoT) trong điều kiện được giám sát trực tiếp. Những bên khác trong ngành cũng có chung quan điểm này. Tuân theo các nguyên tắc được áp dụng kể từ khi ra mắt OpenAI o1-preview, chúng tôi không thực hiện giám sát trực tiếp lên CoT của cả hai mô hình gpt-oss. Chúng tôi tin rằng việc giám sát hành vi sai lệch, gian dối và lạm dụng mô hình là vô cùng quan trọng. Chúng tôi hi vọng việc phát hành một open model có chuỗi tư duy không bị giám sát sẽ mang đến cho nhà phát triển và nhà nghiên cứu cơ hội được nghiên cứu và triển khai các hệ thống giám sát CoT của riêng mình.

Nhà phát triển không nên hiển thị trực tiếp CoT cho người dùng trong ứng dụng. CoT có thể chứa nội dung tạo ra ảo giác hoặc có hại, bao gồm cả ngôn từ không phản ánh các chính sách an toàn tiêu chuẩn của OpenAI, đồng thời có thể chứa thông tin mà mô hình này hiện được yêu cầu rõ ràng là không thêm vào kết quả đầu ra cuối cùng.

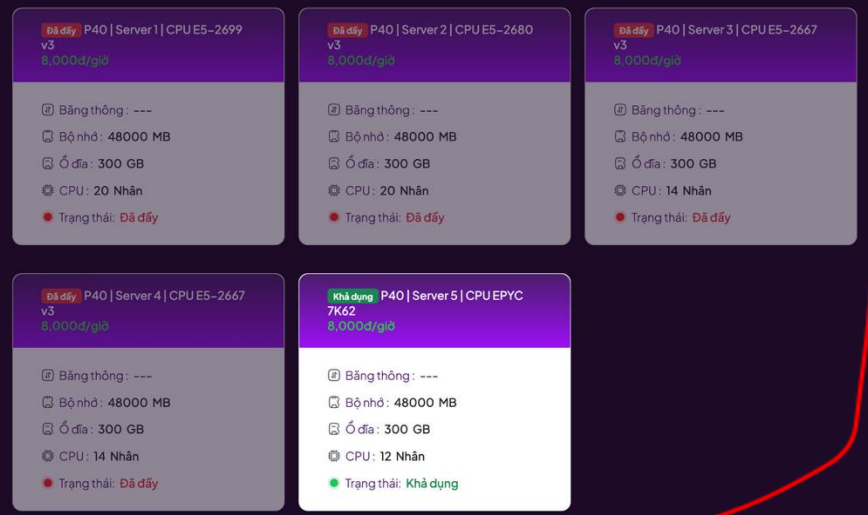
Summary

- gpt-oss-20b: chỉ cần 16GB RAM, chạy mượt với chain-of-thought, hỗ trợ Python, web search.
- gpt-oss-120b: gần ngang o4-mini trên các bài toán lý luận, nhưng chỉ kích hoạt 5.1B tham số mỗi lần nhờ kiến trúc MoE (Mixture of Experts).
- Ta thoải mái chạy local để thử nghiệm, giáo dục hay dùng khi tránh lộ dữ liệu nhạy cảm.


My System
QuestionAppraiser
generate (prompt)
response
LLM
(Self-hosted or managed)
.vn
THUEGPU

Lựa chọn GPU có sẵn

- P40
- A4000
- A6000
- V100
- RTX 3090
- RTX 5090
- RTX 5070Ti



Hệ điều hành *

Lựa chọn hệ điều hành tương ứng với card đồ hoạ theo máy

- Windows AI P40
- Ubuntu AI P40

10 code tặng 100K cho các bạn từ ThueGPU.vn

Comment 1 câu bất kì vào bên dưới video và nhập thông tin và form trong bình luận/mô tả!

Hands-on

- Link github: https://github.com/openai/gpt-oss
- Link weight: https://openai.com/open-models/
- Test on playground: https://gpt-oss.com/
- Install and chat on VPS with GPU RTX 3090 with Ollama (có thể cài OpenWeb UI)
- Call from Python with Ollama

ollama 0.5.2

```
pip install ollama
```

Latest version

Released: Aug 6, 2025

The official Python client for Ollama.

Navigation

Project description

Release history

Download files

Verified details

These details have been verified by PyPI

Project description

Ollama Python Library

The Ollama Python library provides the easiest way to integrate Python 3.8+ projects with Ollama.

Prerequisites

- Ollama should be installed and running
- Pull a model to use with the library: `ollama pull <model>` e.g. `ollama pull gemma3`