GPT-OSS 20B Discussion

Aug 5th-Aug 12th, 2025, Sebastian Raschka, Xiaoli Alex Chen, et al on Substack and Linkedin

Sebastian Raschka:

OpenAI’s new open-source/open-weight model is finally out.

Will be covering it more in detail. First sneak peak when compared with a Qwen3 model of comparable size: They went for width versus depth.

Frank Hanson:

Interesting design choice. Prioritizing width over depth could mean better parallelization and faster inference

Zach Cohen:

OpenAI's focus on width over depth seems counterintuitive, but it's like choosing a broadsword over a scalpel—both have their merits. The real question is, will this model perform better in diverse contexts without losing precision?

A diagram of a diagram of a diagram

AI-generated content may be incorrect.

Xiaoli Alex Chen:

Reverse-engineered GPT-OSS's architecture from OpenAI's official repo source code[https://lnkd.in/gZaKJw9q] - what a textbook example of a state-of-the-art LLM!

→ 36 Transformer blocks with grouped-query attention and sliding window support in every other layer

→ MoE MLP: 128 experts, top-4 routing per token, two MLP layers per expert with SwiGLU activation

→ Advanced rotary positional encoding (RoPE)

→ 201k vocabulary, 2880 hidden dimension

It's so exciting to see the latest LLM design and implementation details. I'm going to dive deeper into each part when I've got more time. Viva la OSS!

A diagram of a computer system

AI-generated content may be incorrect.

References

[1] <https://github.com/openai/gpt-oss>

[2] [OpenAI’s GPT-5 arrives with PhD-level reasoning and massive context windows, Startupmac, Aug 12th, 2025](https://startupmac.substack.com/p/openais-gpt-5-arrives-with-phd-level)

[3] [The Big LLM Architecture Comparison: From DeepSeek-V3 to Kimi K2: A Look At Modern LLM Architecture Design, Sebastian Raschka, July 19th, 2025](https://sebastianraschka.com/blog/2025/the-big-llm-architecture-comparison.html)