GOSIM

GOSIM AI Paris 2025

Open-source has caught up, What is next (for AGI)?

Michael Yuan, PhD

The GOSIM origin story

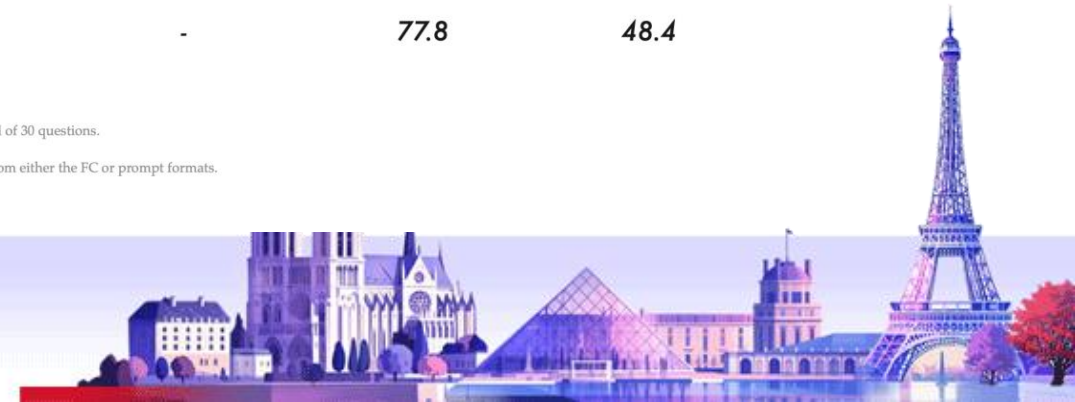
April, 2023

The open-source AI has caught up

GOSIM

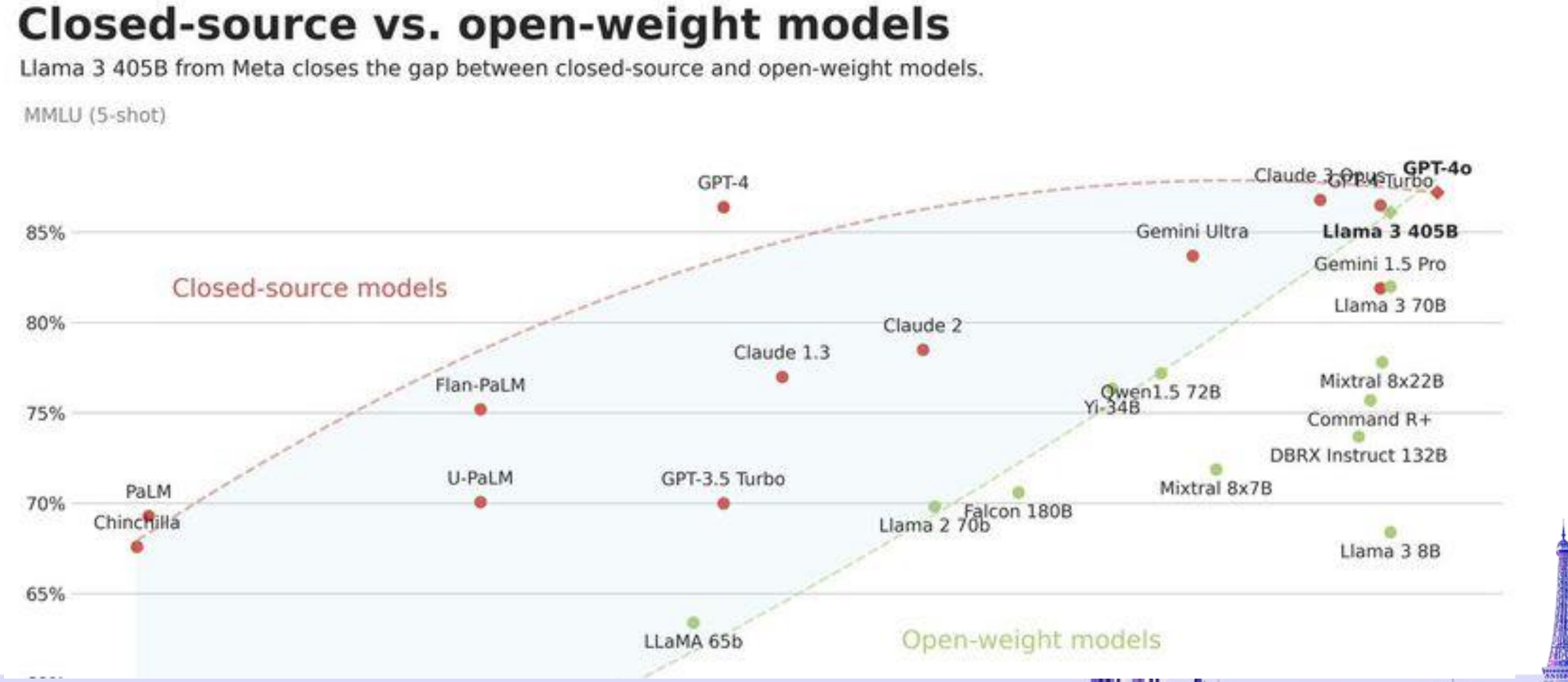
	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.



But what about the dream of AGI?

GOSIM



The open-source path to AGI

GOSIM

- A swarm of agents customized for specific tasks and use cases
  - Fine-tuned LLMs
  - Individualized and proprietary knowledge bases
  - Personalized and authorized tool calls
- Advanced orchestration tools and frameworks
- Deployed on private hardware or robotic devices

The model knowledge and action fit

GOSIM

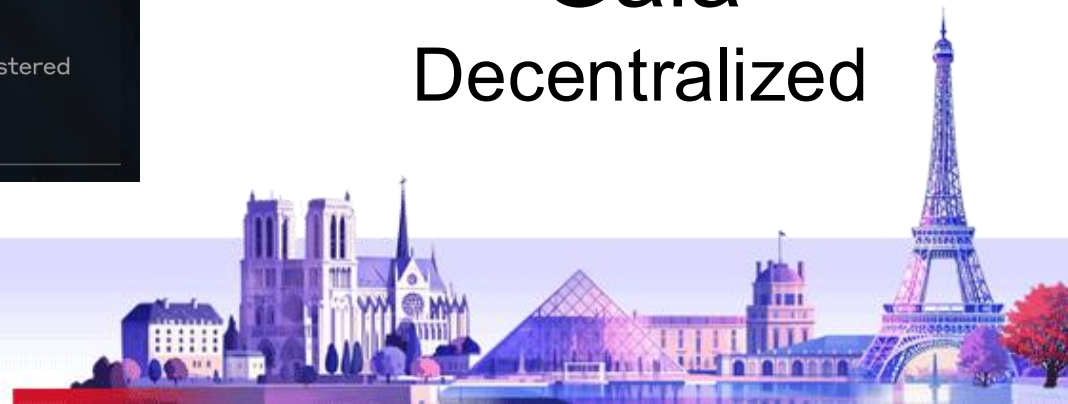
Introducing the Responses API

The Responses API is our new API primitive for leveraging OpenAI's built-in tools to build agents. It combines the simplicity of Chat Completions with the tool-use capabilities of the Assistants API. As model capabilities continue to evolve, we believe the Responses API will provide a more flexible foundation for developers building agentic applications. With a single Responses API call, developers will be able to solve increasingly complex tasks using multiple tools and model turns.

OpenAI
centralized

559,881	2848.4B	3,085
Node Deployments	Total Throughputs	Domains Registered

Gaia
Decentralized



Shameless plug ...

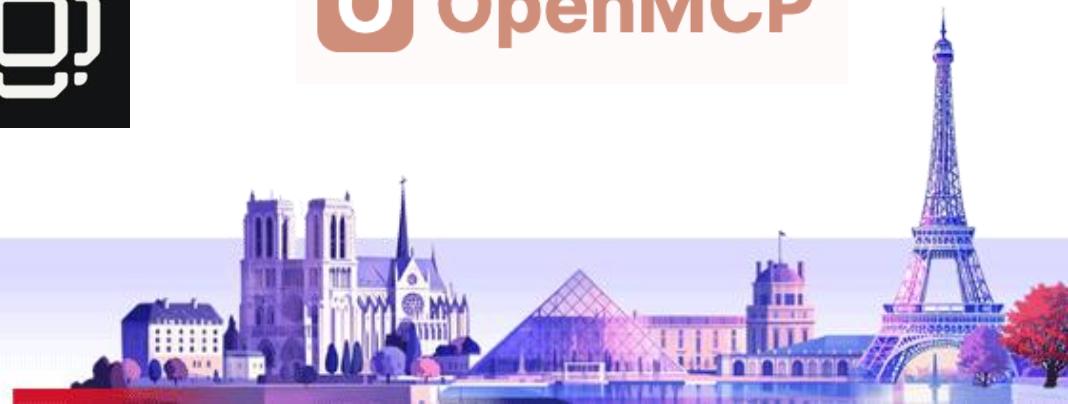
GOSIM

- WasmEdge – a universal AI runtime. Portable across GPU devices, supports many model types. Lightweight and easily deployed on edge servers and devices.
- LlamaEdge — an application and agent framework built on Rust and WasmEdge
- Gaia — a software stack for decentralized and verifiable AI knowledge agent marketplaces
- OpenMCP — a universal adaptor and payment gateway for MCP servers

WasmEdgeRuntime

LlamaEdge

OpenMCP



Plotting the path @ GOSIM Paris

GOSIM

- SOTA open-weight models
- Open-science and fully open-source models
- High performance and cross-platform AI runtimes
- Agent frameworks and MCPs
- Knowledge and vector databases
- Embedded AI and robotics

GOSIM

GOSIM AI Paris 2025

Oh, one more thing ...

Win an all-expense paid trip to the “silicon valley of China”