国内外の生成AIの開発状況

Institute of
SCIENCE TOKYO
Supercomputing Research Center
Rio Yokota

第16回 自動チューニング技術の現状と応用に関するシンポジウム

2024/12/26

国内の基盤モデルに関する主要な取り組み

SB Intuitions

・Sarashina-8x70B

ELYZA

・Llama-3-ELYZA-JP-70B

Preferred Elements

・PLaMo-100B

CyberAgent

・CyberAgentLM3-22B

GENIAC 第1期

・ELYZA, Kotoba Tech, 富士通, ABEJA, Sakana AI, NII, Stockmark, Turing, 松尾研, PFE

NII LLM-jp

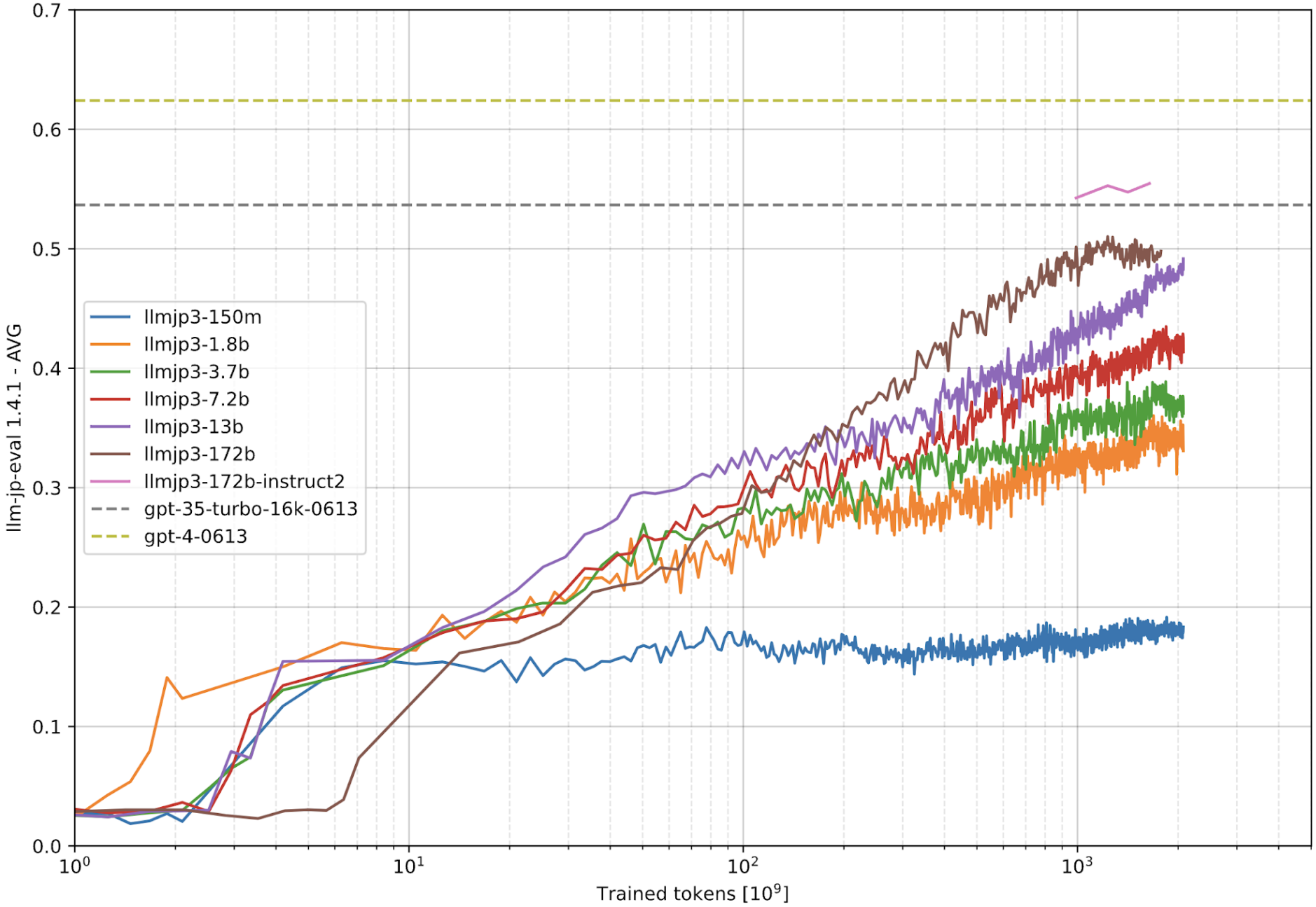
・llm-jp-172B

東京科学大学＋産総研

・Llama-3.1-Swallow-70B

LLM-jp

- Largest effort to train an open LLM from scratch
- Model weights, dataset, checkpoints are open
- Constantly monitoring benchmark performance
- 172B benchmark scores are saturating
- Instruct model beats GPT-3.5



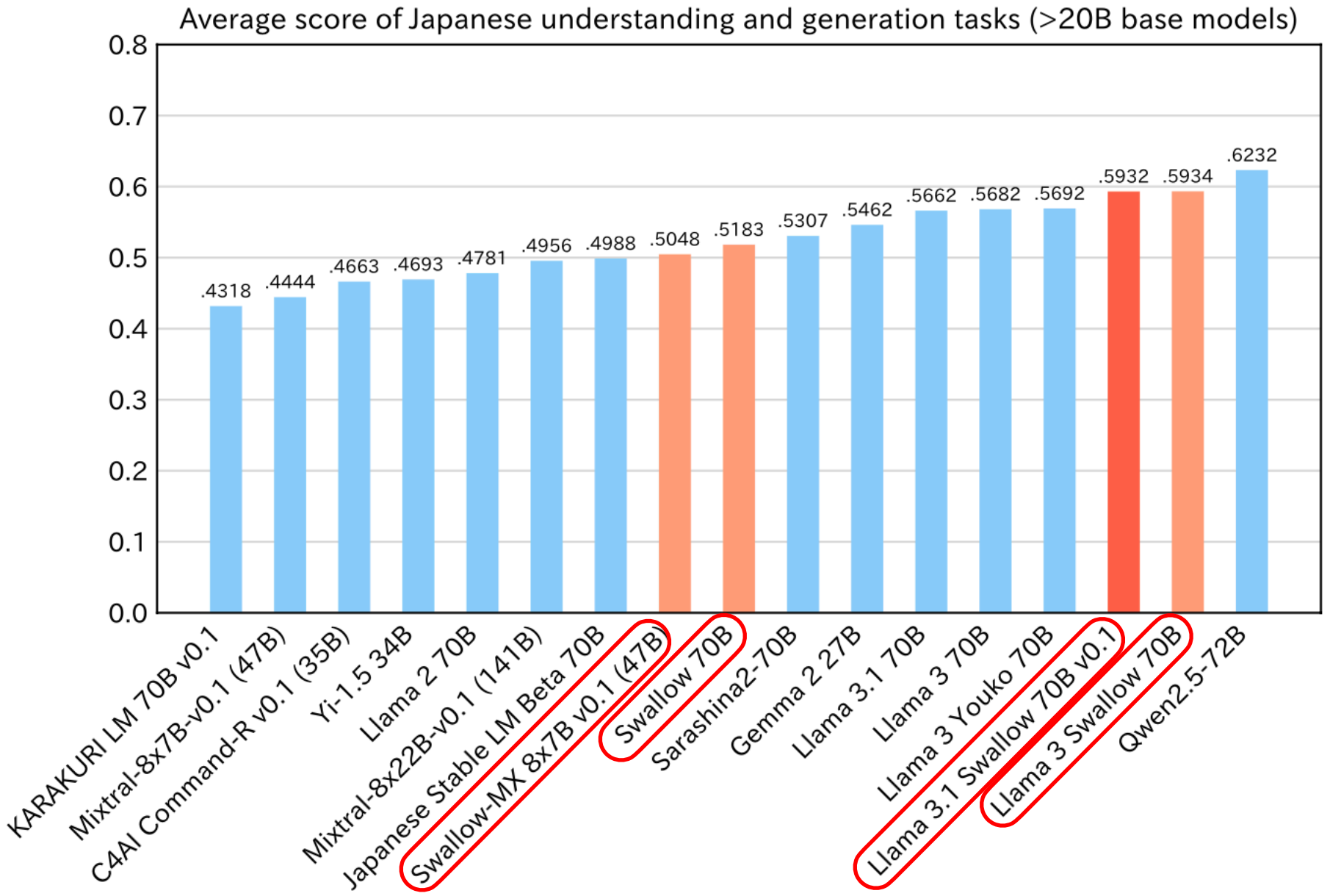
Swallow

・Continual pre-training from open models

・Create original Japanese dataset from CommonCrawl

・Currently #1 among LLMs trained in Japan

・Qwen2.5 is a little better than Swallow



国外の基盤モデルに関する主要な取り組み

OpenAI
- o3

Google
- Gemini 2.0 Flash Thinking

X.ai
- Grok 2

Anthropic
- Claude 3.5 Sonnet

Meta
- Llama 3.3

Mistral AI
- Mistral Large 2

Alibaba
- QwQ

NVIDIA
- Nemotron 4

AWS
- Nova Premier

Microsoft Azure
- Phi 3

TTI
- Falcon 3

Cohere
- Command R+

DeepSeek
- DeepSeek R1

0.1-AI
- Yi Large

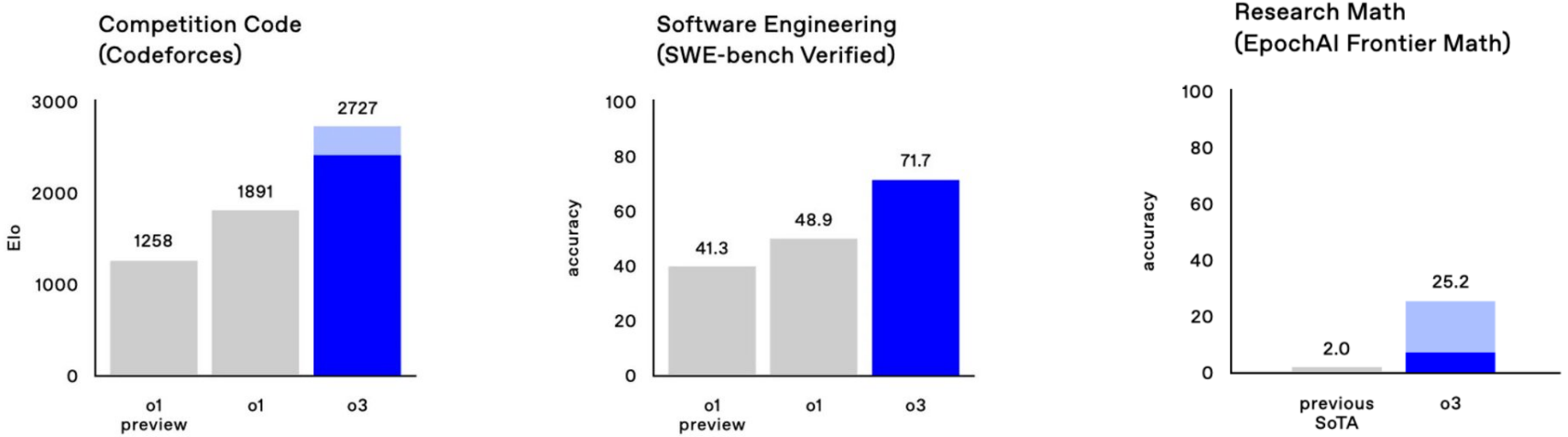
DataBricks
- DBRX

Perplexity
- Sonar 3.1 Large

Reka
- Reka Flash

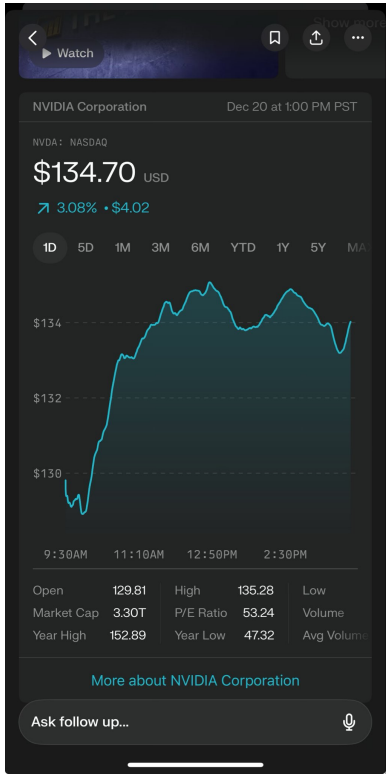
AI21Labs
- Jamba 1.5 Large

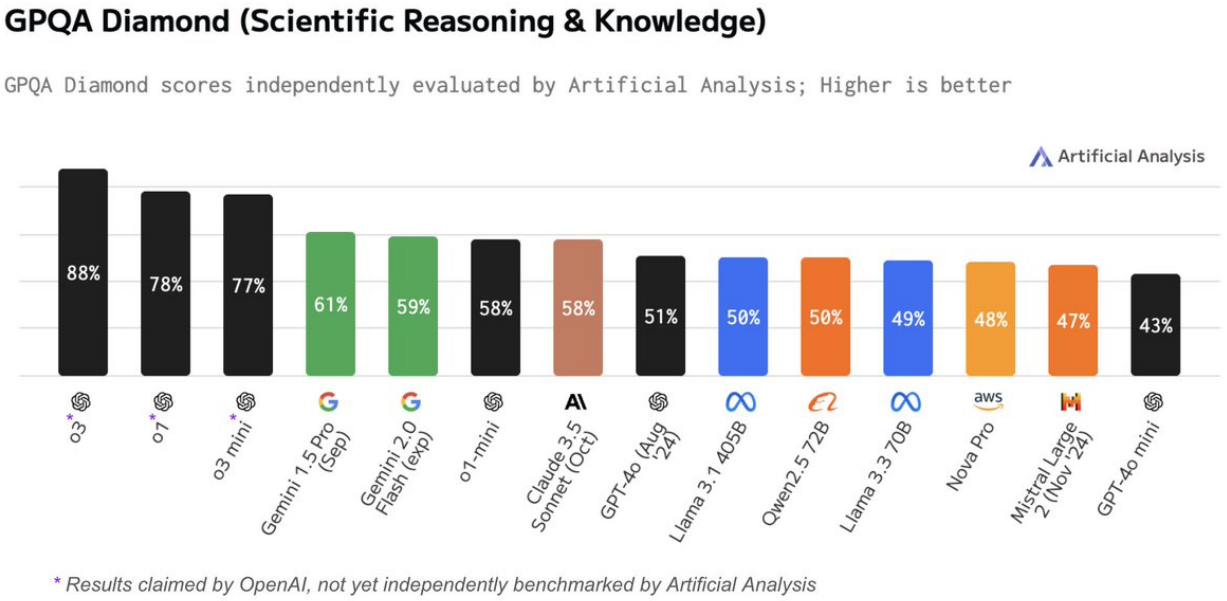
OpenAI o3

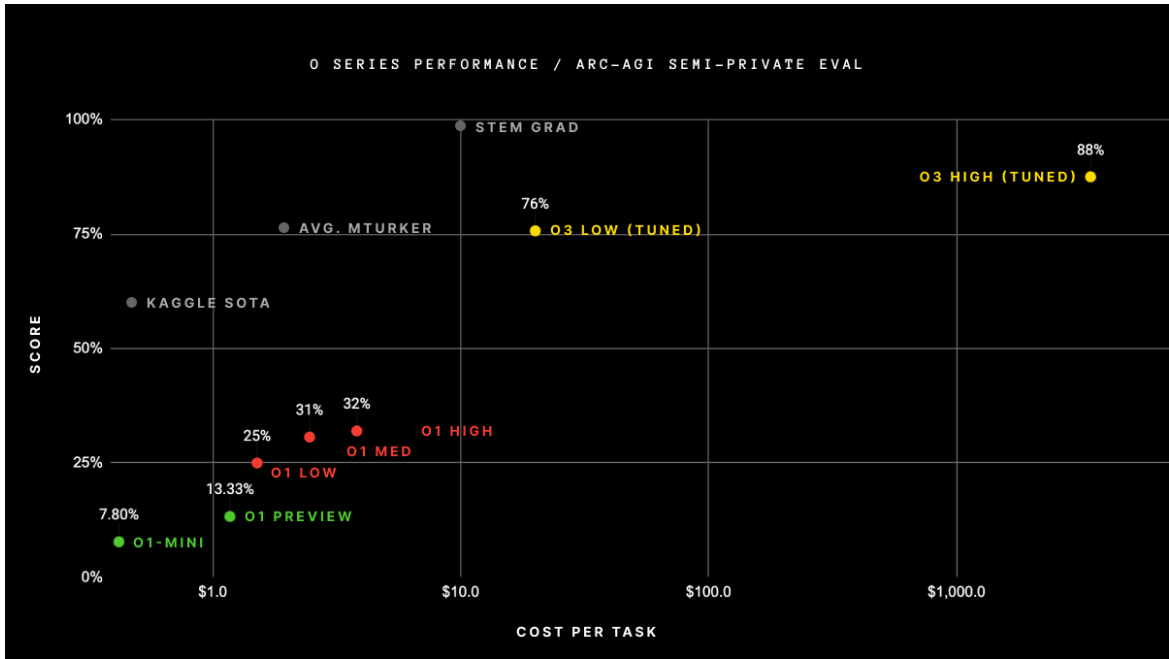

Competition Code
(Codeforces)
Elo
3000
2000
1000
0
1258
1891
2727
o1 preview
o1
o3

Software Engineering
(SWE-bench Verified)
accuracy
100
80
60
40
20
0
41.3
48.9
71.7
o1 preview
o1
o3

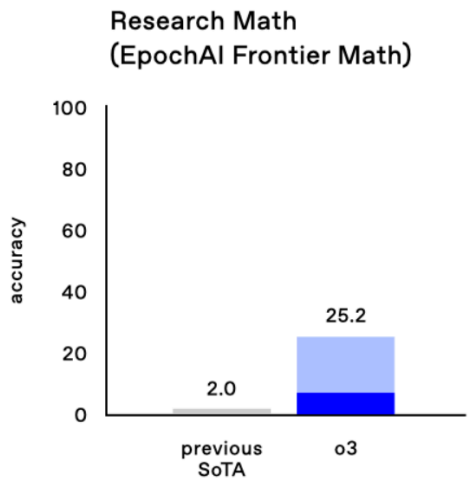
Research Math
(EpochAI Frontier Math)
accuracy
100
80
60
40
20
0
2.0
25.2
previous SoTA
o3


Watch
NVIDIA Corporation
Dec 20 at 1:00 PM PST
NVDA: NASDAQ
$134.70 USD
3.08% • $4.02
1D 5D 1M 3M 6M YTD 1Y 5Y
$134
$132
$130
9:30AM 11:10AM 12:50PM 2:30PM
Open 129.81 High 135.28 Low
Market Cap 3.30T P/E Ratio 53.24 Volume
Year High 152.89 Year Low 47.32 Avg Volume
More about NVIDIA Corporation
Ask follow up...


GPQA Diamond (Scientific Reasoning & Knowledge)
GPQA Diamond scores independently evaluated by Artificial Analysis; Higher is better
Artificial Analysis
88%
78%
77%
61%
59%
58%
58%
51%
50%
50%
49%
48%
47%
43%
o3*
o1*
o3 mini*
Gemini 1.5 Pro (Sep)
Gemini 2.0 Flash (exp)
o1-mini
Claude 3.5 Sonnet (Oct)
GPT-4o (Aug '24)
Llama 3.1 405B
Qwen2.5 72B
Llama 3.3 70B
Nova Pro
Mistral Large 2 (Nov '24)
GPT-4o mini
* Results claimed by OpenAI, not yet independently benchmarked by Artificial Analysis


O SERIES PERFORMANCE / ARC-AGI SEMI-PRIVATE EVAL
SCORE
100%
75%
50%
25%
0%
STEM GRAD
88%
O3 HIGH (TUNED)
76%
O3 LOW (TUNED)
AVG. MTURKER
KAGGLE SOTA
31%
32%
25%
O1 HIGH
O1 MED
O1 LOW
13.33%
7.80%
O1 PREVIEW
O1-MINI
$1.0
$10.0
$100.0
$1,000.0
COST PER TASK

Frontier Math



Prime field continuous extensions

Problem Solution

Let a_n for $n \in \mathbb{Z}$ be the sequence of integers satisfying the recurrence formula

$$a_n = 198130309625a_{n-1} + 354973292077a_{n-2} - 427761277677a_{n-3} + 370639957a_{n-4}$$

with initial conditions $a_i = i$ for $0 \leq i \leq 3$. Find the smallest prime $p \equiv 4 \pmod 7$ for which the function $\mathbb{Z} \to \mathbb{Z}$ given by $n \mapsto a_n$ can be extended to a continuous function on $\mathbb{Z}_p$.

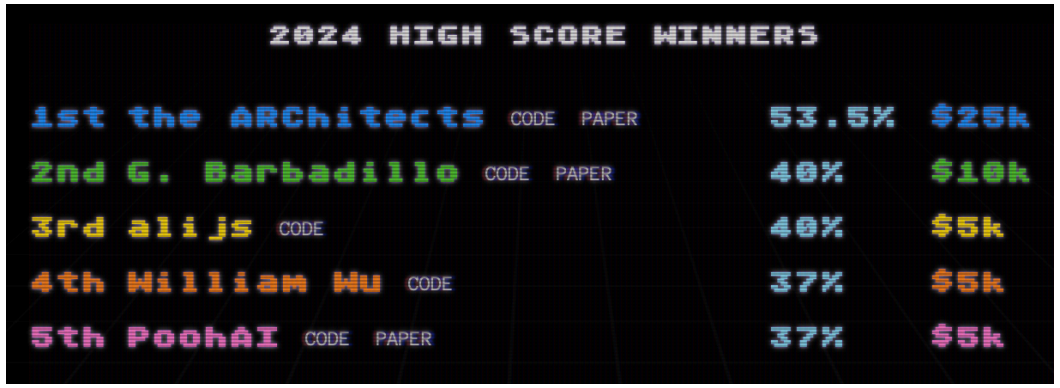
Find the degree 19 polynomial

Problem Solution

Construct a degree 19 polynomial $p(x) \in \mathbb{C}[x]$ such that $X := \{p(x) = p(y)\} \subset \mathbb{P}^1 \times \mathbb{P}^1$ has at least 3 (but not all linear) irreducible components over $\mathbb{C}$. Choose $p(x)$ to be odd, monic, have real coefficients and linear coefficient -19 and calculate $p(19)$.

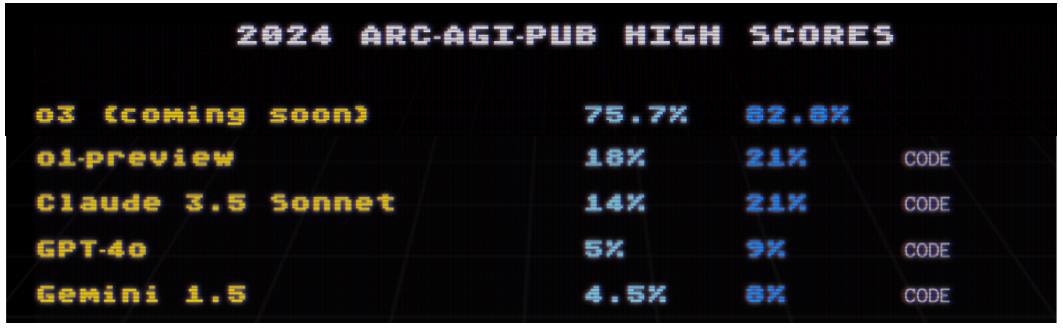
ARC-AGI

ARC-AGI特化モデル



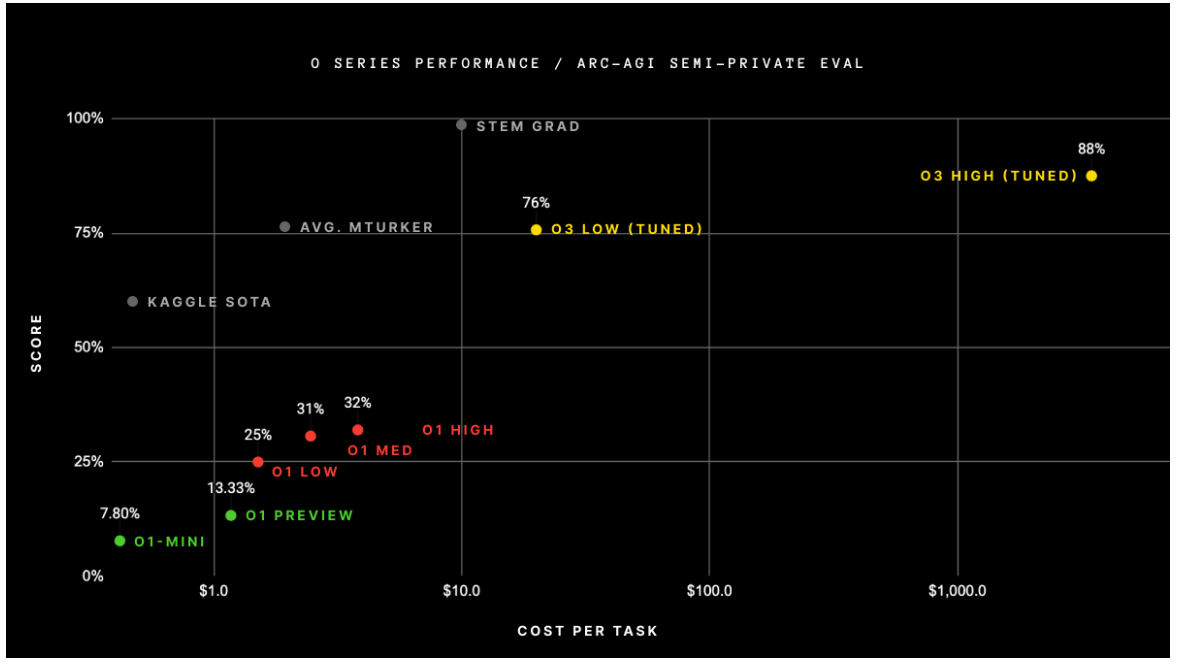
private

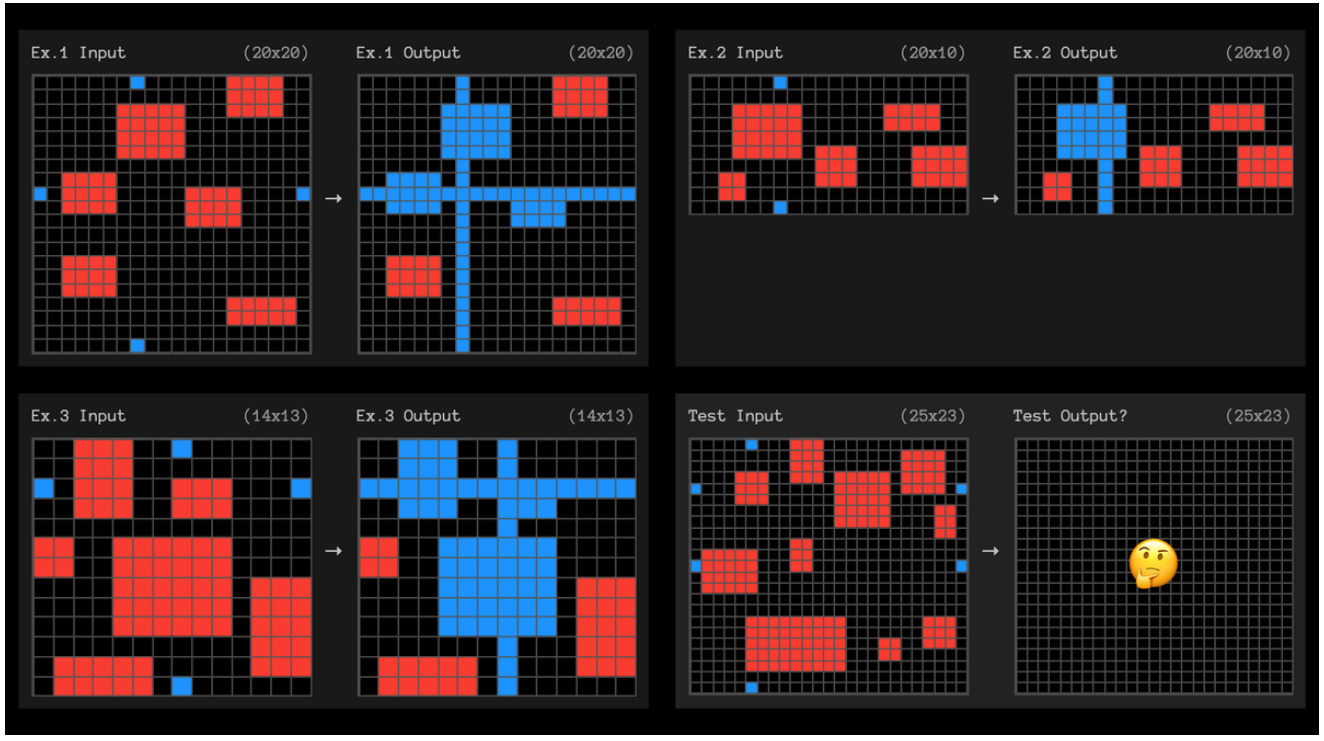
汎用モデル



semi-private

public





o1-like models from China

o1-Coder: an o1 Replication for Coding

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong & Jitao Sang*
School of Computer Science and Technology
Beijing Jiaotong University
Beijing, China
{yuxiangzhang, wushangxi, yqyang, jiangmingshu, jinlinx, 23120361, jtsang}@bjtu.edu.cn

Alibaba International

2024-11-26

Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions

Yu Zhao*, Huifeng Yin*, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, Kaifu Zhang

MarcoPolo Team, Alibaba International Digital Commerce

Generative AI Research

O1 Replication Journey: A Strategic Progress Report – Part 1

Yiwei Qin[1,4*] Xuefeng Li[1,4*] Haoyang Zou[4*] Yixiu Liu[1,4*] Shijie Xia[1,4*]
Zhen Huang[4] Yixin Ye[1,4] Weizhe Yuan[2] Hector Liu[3] Yuanzhi Li[3] Pengfei Liu[1,4†]
[1]Shanghai Jiao Tong University, [2]New York University,
[3]MBZUAI, [4]Generative AI Research Lab (GAIR)

Generative AI Research

O1 Replication Journey – Part 2: Surpassing O1-preview through Simple Distillation Big Progress or Bitter Lesson?

Zhen Huang[4*] Haoyang Zou[4*] Xuefeng Li[1,4*] Yixiu Liu[1,4*] Yuxiang Zheng[1,4*]
Ethan Chern[1,4*] Shijie Xia[1,2,4*] Yiwei Qin[4] Weizhe Yuan[3] Pengfei Liu[1,2,4†]
[1]Shanghai Jiao Tong University, [2]SII, [3]NYU,
[4]Generative AI Research Lab (GAIR)

Benchmark	QwQ 32B-preview	OpenAI o1-preview	OpenAI o1-mini	Claude3.5 Sonnet	Qwen2.5-72B Instruct	GPT-4o
GPQA	65.2	72.3	60.0	65.0	49.0	53.6
AIME	50.0	44.6	56.7	16.0	23.3	9.3
MATH-500	90.6	85.5	90.0	78.3	82.6	76.6
LiveCodeBench	50.0	53.6	58.0	36.3	30.4	33.4

AI for AI を AI for Scienceにつなげる

Trillion Parameter Consortiumの大幅な方針転換

- OpenAIやAnthropicと協議し、AI for Scienceの大幅な方針転換を決めた
- Domain specific foundation model → Domain specific benchmarks
- 方針転換後はOpenAIが協力的でモデルを共有してくれる予定

古典的な深層学習の制約を排除

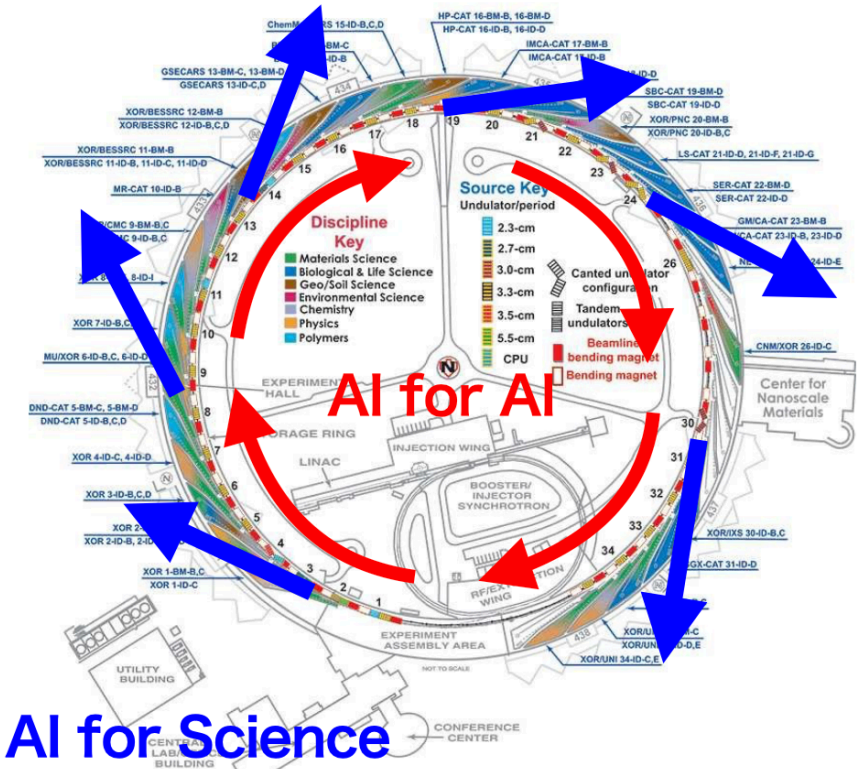
- 分類・要約 → 生成 → インタラクション

モデルにデータを入れるのが最善の方法か？

- 事前学習 → 事後学習 → コンテキスト内学習

データがAIに入っていくのではなく、AIがデータに入っていく

- AIを賢くするためのデータ vs それを使って分析したいデータ
- 前者はBigTechと競争してもリソースの無駄
- 後者を探索するためのAI → 知的な検索・探索

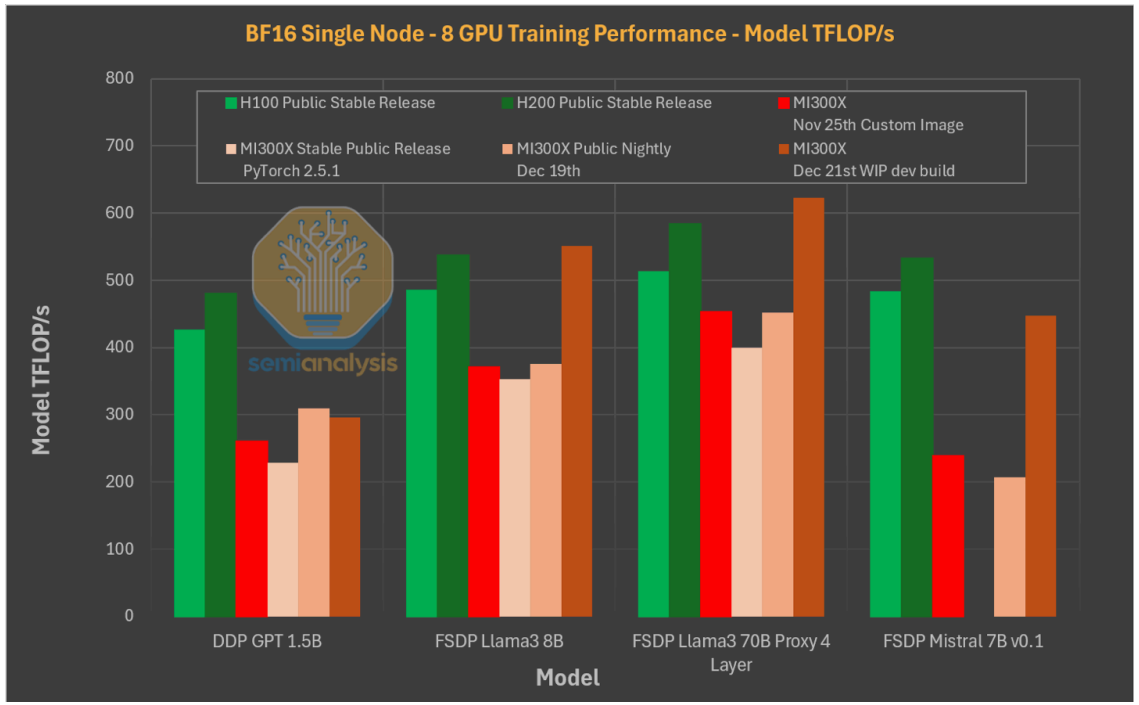


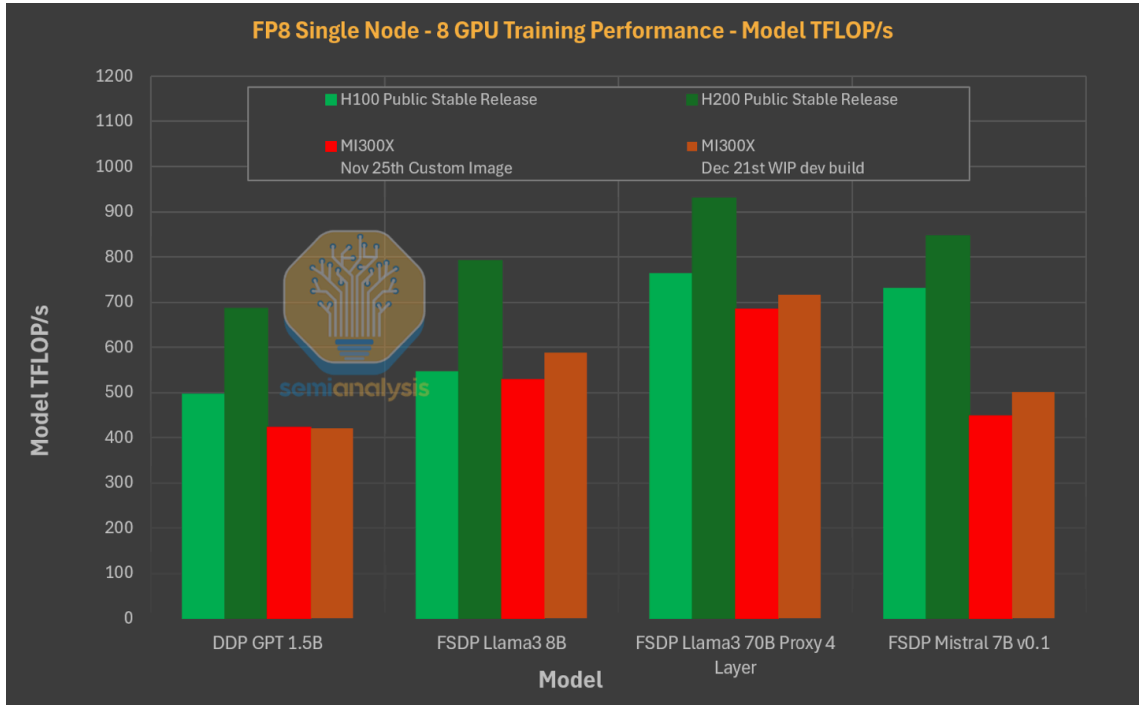
AMD GPUにおけるLLMの性能

- MI300Xは理論性能はH200より高い
- LLMの学習の性能は遥かに低い
- モデルによってはさらに低い
- この性能を出すだけでも大変

H100 vs H200 vs MI300X Basic Specifications			
	H100	H200	MI300X
Watts Per GPU (TDP)	700	700	750
All-in System Watts Per GPU	1,275	1,275	1,275
Memory Capacity (GB)	80GB	141GB	192GB
Memory Bandwidth (GB/s)	3,352	4,800	5,300
FP16/BF16 TFLOPS[1]	989	989	1,307
FP8 / FP6 / Int8 TFLOPS[1]	1,979	1,979	2,615

1. All FLOPS are dense





https://semianalysis.com/2024/12/22/mi300x-vs-h100-vs-h200-benchmark-part-1-training/