

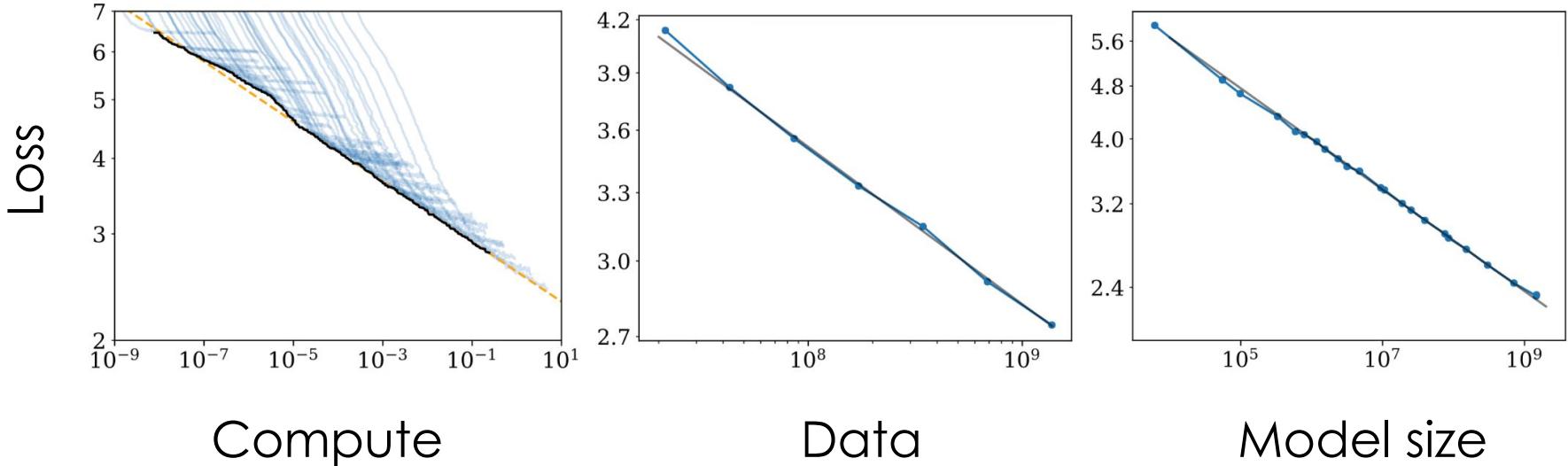


# The State of LLMs: from pre- to post-training

Leandro von Werra (@lvwerra)  
*Head of Research at Hugging Face*

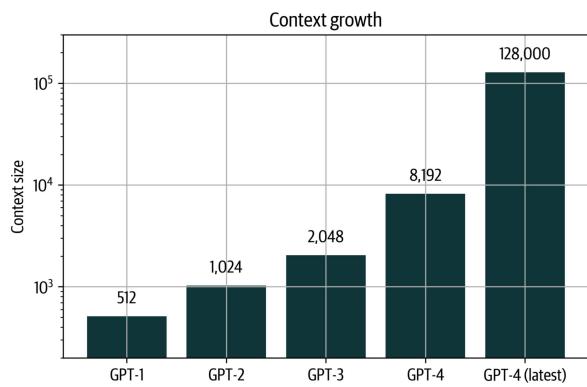
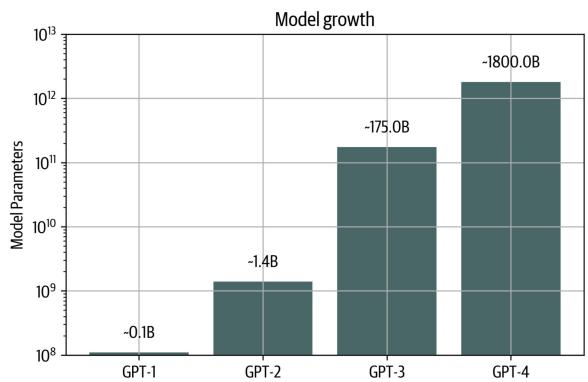
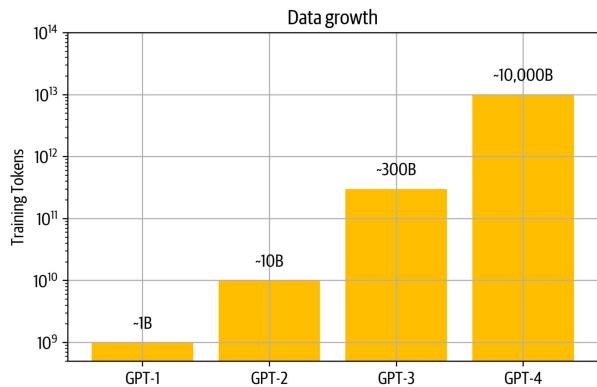
# **pretraining**

# Scaling laws: **predictable** returns



<https://arxiv.org/abs/2001.08361>

# Trends: train longer, larger, more context



# Trends: more compute

**compute  $\approx$  data  $\times$  model size**

	Dataset (Billion Tokens)	Model size (Billion Parameter)	
GPT 1:	1-2	0.11	
GPT 2:	10-20	1.4	
GPT 3:	300	175	
GPT 4:	10'000	1'800	

Compute:

100x

2000x

300x



GPT-4 cost: ~\$100M Dollars

# Infrastructure of **today**



**LUMI (Finland): 11'912 GPUs (AMD MI250x)**



**JUWELS (Germany): 3'774 GPUs (NVIDIA A100)**

*coming soon*

**JUPITER (Germany): 24'000 GPUs (NVIDIA GH200)**



**Leonardo (Italy): 13'824 GPUs (NVIDIA A100)**



**Alps (Switzerland): 10'000 GPUs (NVIDIA GH200)**

# Infrastructure of the tomorrow



OpenAI ✨ @OpenAI · Jan 21

Announcing The Stargate Project



...

The Stargate Project is a new company which intends to invest \$500 billion over the next four years building new AI infrastructure for OpenAI in the United States. We will begin deploying \$100 billion immediately. This infrastructure will secure

[Show more](#)

## In the United States

## In Europe



⊕ EN

Search

Search

[Home](#) > [Press corner](#) > EU launches InvestAI initiative to mobilise €200 billion

Available languages: [English](#)

PRESS RELEASE | Feb 11, 2025 | Paris | 3 min read

## EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence

### PAGE CONTENTS

Top

Quote(s)

Related topics

Print friendly pdf

Contacts for media

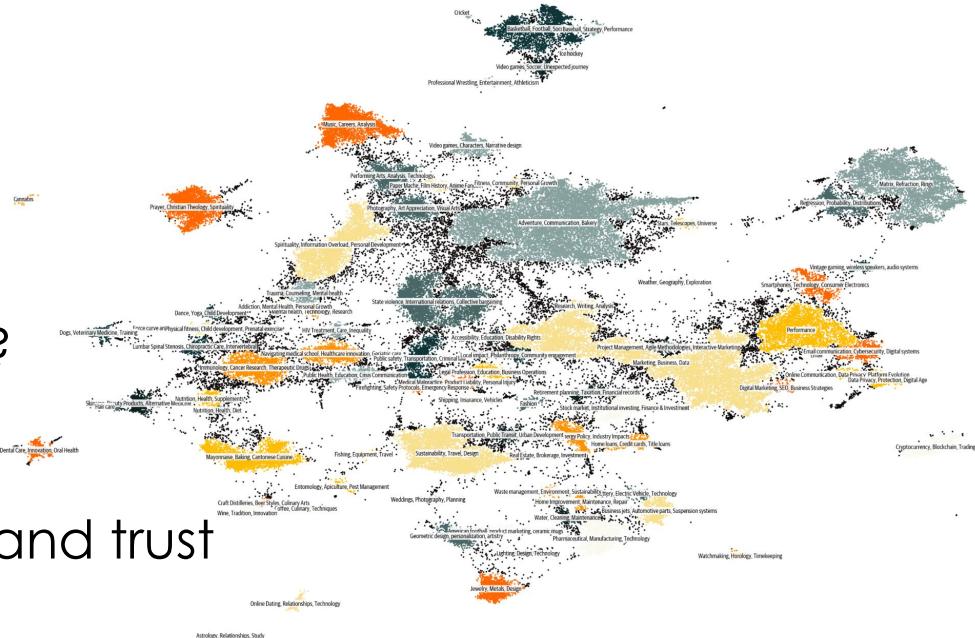
Today, at the Artificial Intelligence (AI) Action Summit in Paris, Commission President Ursula von der Leyen has launched InvestAI, an initiative to mobilise €200 billion for investment in AI, including a new European fund of €20 billion for AI gigafactories. This large AI infrastructure is needed to allow open, collaborative development of the most complex AI models and to make Europe an AI continent.

Commission President Ursula von der Leyen said: "AI will improve our healthcare, spur our research and innovation and boost our competitiveness. We want AI to be a force for good and for growth. We are doing this through our own European approach – based on openness, cooperation and excellent talent. But our approach still needs to be

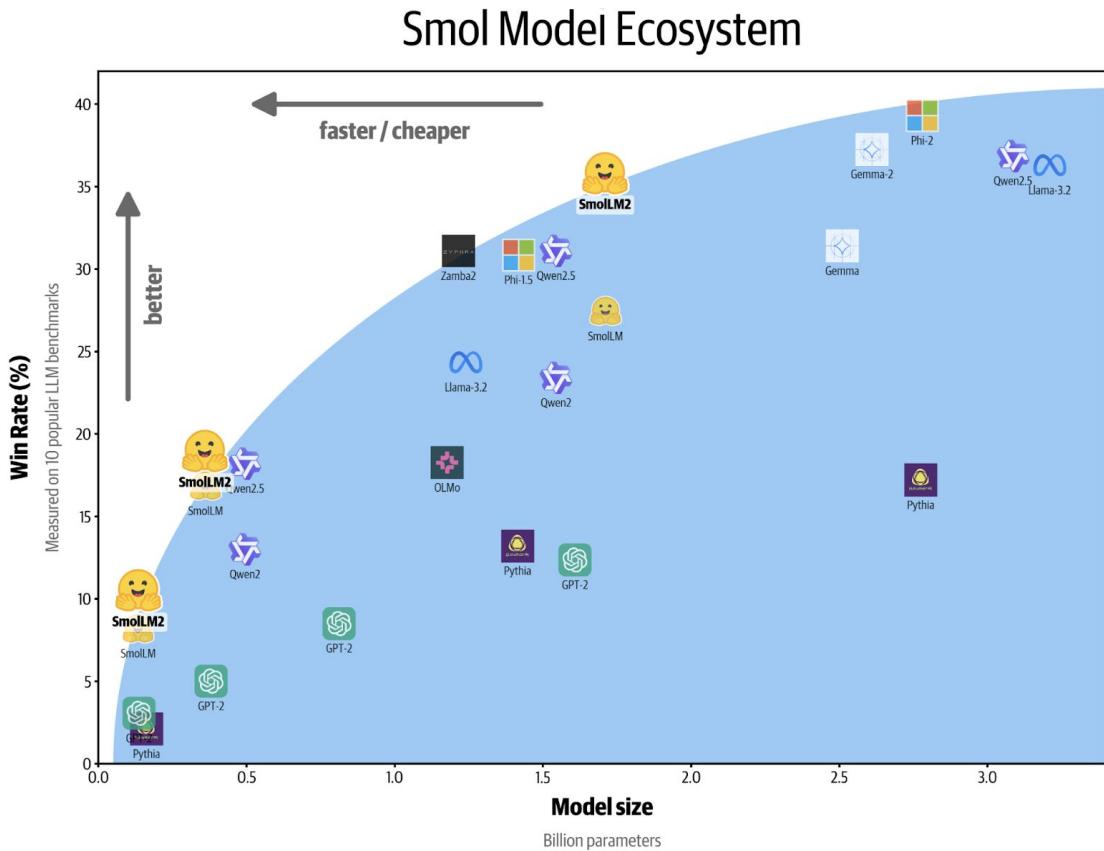
# Transparency on pretraining

- what data was used?
  - how was my data used?
  - how was the data filtered?

→ address biases, attribution, and trust



# Trends: smol models

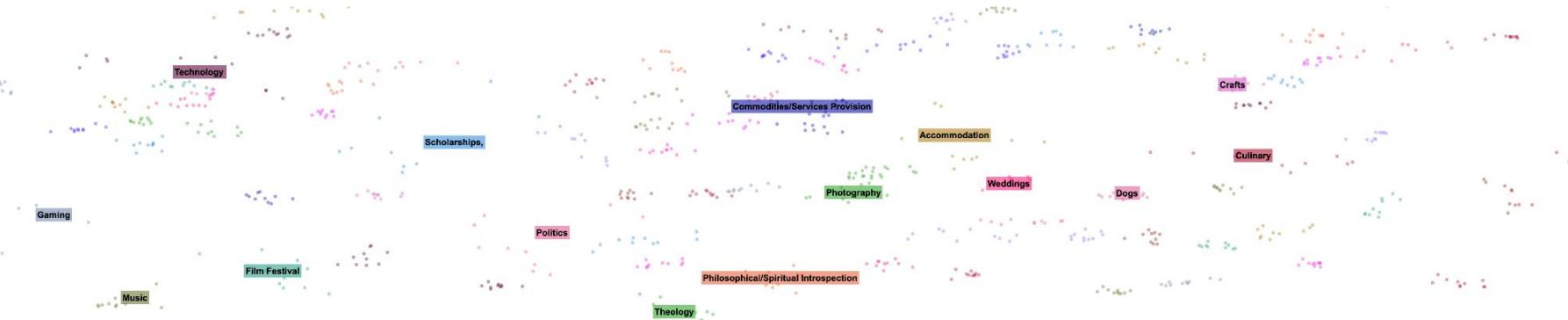


**Apple releases OpenELM: small, open source AI models designed to run on-device**



Credit: VentureBeat made with Midjourney V6

# FineWeb: decanting the web for the finest text data at scale



AUTHORS	AFFILIATION	PUBLISHED
Guilherme Penedo, Hynek Kydiček, Lubna Ben Allal, Anton Lozhkov, Colin Raffel, Leandro Werra, Thomas Wolf	HuggingFace	May 31, 2024

## Table of contents

### Web data

- Finding the raw data
- Processing at scale
- What is good data?
- Ablations and evaluation setup

### The FineWeb recipe

- Starting point: text extraction
- Base filtering
- Deduplicating the data
  - Why deduplicate?
  - Our deduplication parameters
  - More deduplication is always better, right?
  - Taking a step back: individual dump dedup
  - A note on measuring the effect of deduplication
  - [Other \(failed\) initial approaches](#)

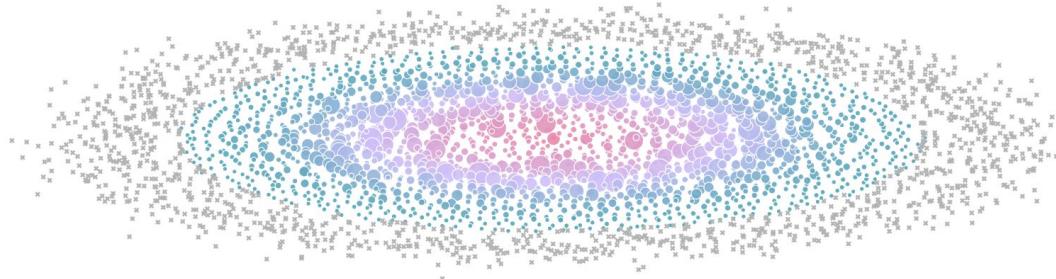
The performance of a large language model (LLM) depends heavily on the quality and size of its pretraining dataset. However, the pretraining datasets for state-of-the-art open LLMs like Llama 3 [1] and Mistral [2] are not publicly available and very little is known about how they were created.

Recently, we released [FineWeb](#), a new, large-scale (15-trillion tokens, 44TB disk space) dataset for LLM pretraining. FineWeb is derived from 96 CommonCrawl snapshots and produces **better-performing LLMs than other open pretraining datasets**. To bring more clarity in machine learning and advance the open understanding of how to train good quality large language models, we carefully documented and ablated all of the design choices used in FineWeb, including in-depth investigations of deduplication and filtering strategies. The present long form report is a deep dive in how to create a large and high-quality web-scale dataset for LLM pretraining. The dataset itself, [FineWeb](#), is available [here](#).

Reading time: 45 min. For the best reading experience, we recommend not using a mobile phone.

We are extremely thankful to the whole distill.pub team (Christopher Olah, Shan Carter, Ludwig Schubert in particular) for creating the template of which we based this blog post. Thanks also for inspiring us with exquisitely crafted articles and blog posts.

# The Ultra-Scale Playbook: Training LLMs on GPU Clusters



We ran over 4000 scaling experiments on up to 512 GPUs and measured throughput (size of markers) and GPU utilization (color of markers). Note that both are normalized per model size in this visualization.

---

## AUTHORS

Nouamane Tazi, Ferdinand Mom, Haojun Zhao, Phuc Nguyen,  
Mohamed Mekkouri, Leandro Werra, Thomas Wolf

## AFFILIATION

Hugging Face Feb 19, 2025

## PUBLISHED

[Download PDF](#)

---

## Table of Contents ▾

### High level overview

#### First Steps: Training on one GPU

Memory usage in Transformers

Profiling the memory usage

Weights/grads/optimizer states memory

Activations memory

Activation recomputation

Gradient accumulation

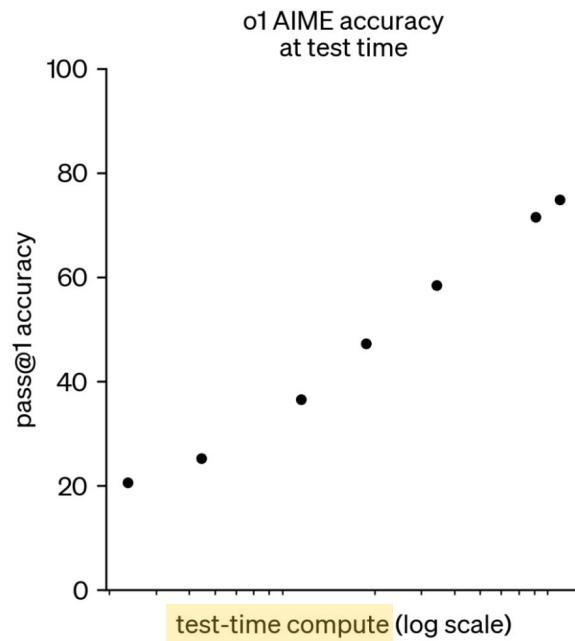
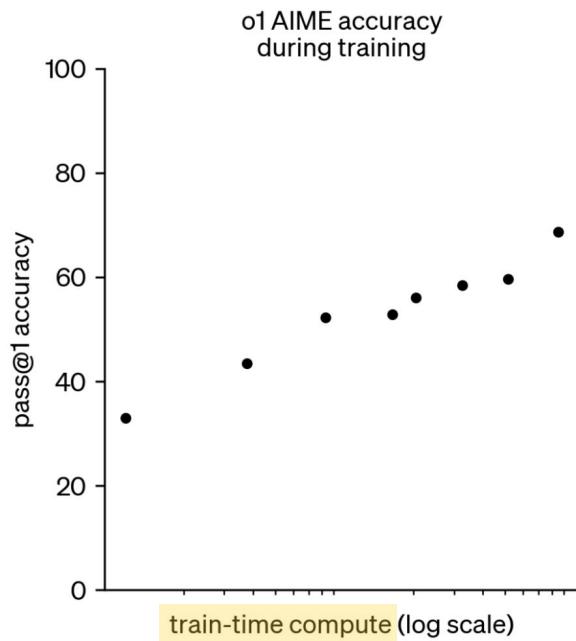
Profiling GPU compute and communication

Thousands of GPUs humming in perfect harmony. That's what it takes to train today's most powerful AI models – a symphony of computing power that until recently was the exclusive domain of elite research labs. Open source has transformed this landscape, but not completely. Yes, you can download the latest Llama or DeepSeek models. Yes, you can read their technical and experiment reports. But the most challenging part – the training code, the knowledge and techniques necessary to coordinate GPUs to train these massive systems – remains shrouded in complexity and spread around a series of disconnected papers and often private codebases.

Reading time: 2-4 days.  
For the best reading  
experience, we  
recommend not using a  
mobile phone.

**post-training**

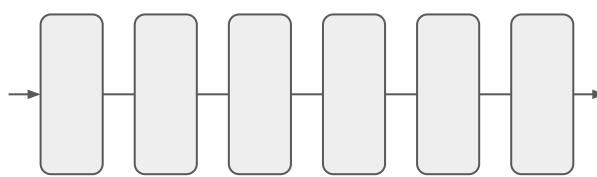
# New frontier: scaling test-time compute



# New frontier: scaling test-time compute

“What is 17+8?”

“Is the Riemann Hypothesis true?”



“25”

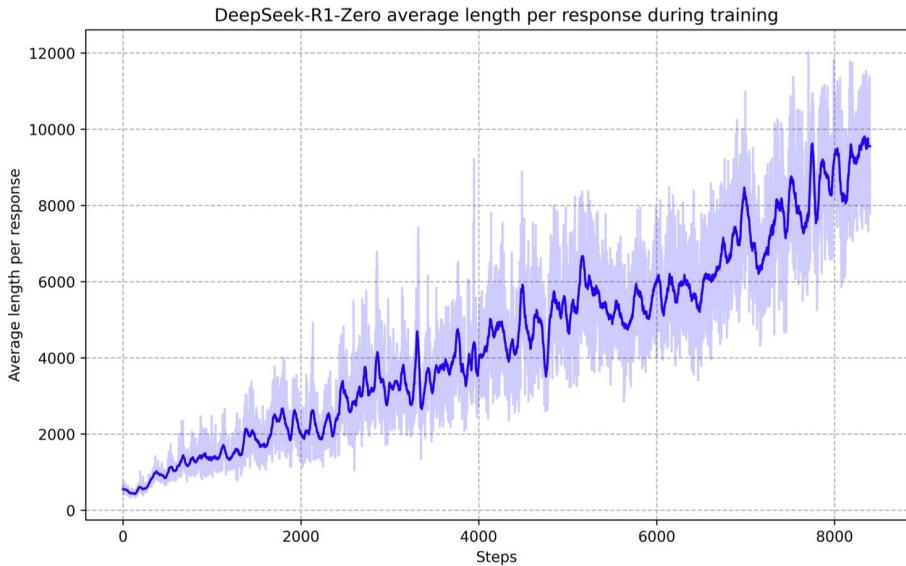
“Yes”

A transformer model uses fixed amount of compute per token irrespective of difficulty.

So how can we scale compute?

o 1/o3, DeepSeek-R1, gemini-thinking, claude-3.7 models all do it..

# Reasoning: DeepSeek-R1



Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a+x}} = x$  is equal to

Response: <think>

To solve the equation  $\sqrt{a - \sqrt{a+x}} = x$ , let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

# DeepSeek-R1: distillation into smol models

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	<b>72.6</b>	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	<b>86.7</b>	<b>94.5</b>	<b>65.2</b>	<b>57.5</b>	1633

# Open source is **not** a zero-sum game!

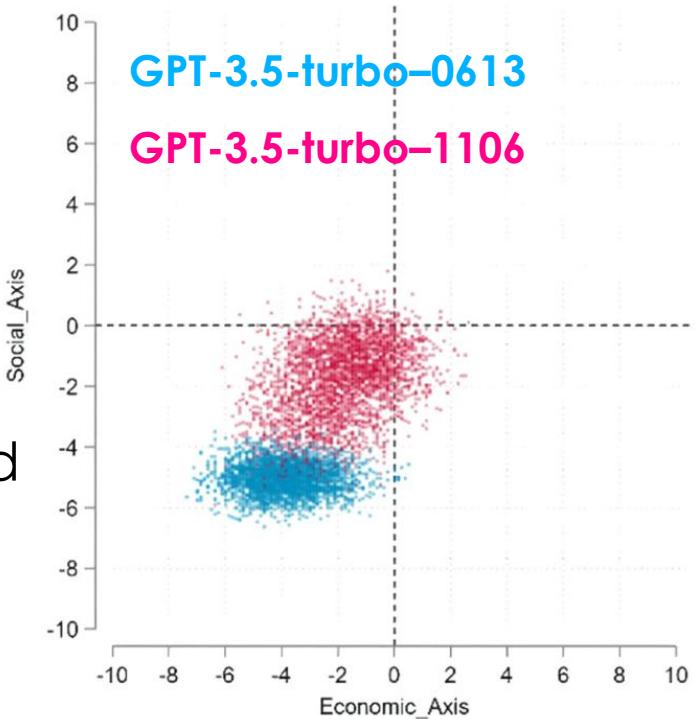


<https://huggingface.co/blog/open-r1>

# Transparency on alignment

API models are aligned for safety + values, but...

- which values? → transparency
- should they be global? → fine-grained
- can I decide? → control



<https://www.nature.com/articles/s41599-025-04465-z>

# Datasets for alignment / reasoning

The screenshot shows the Hugging Face Datasets homepage. The top navigation bar includes links for Models, Datasets, Spaces, Posts, Docs, Enterprise, Pricing, and user profile. The main search bar says "Search models, datasets, users...". On the left, there's a sidebar with categories like Main, Tasks, Libraries, Languages, Licenses, Other, Modalities (3D, Audio, Geospatial, Image, Tabular, Text, Time-series, Video), and Size (rows). The central area displays a list of datasets with their names, descriptions, and metrics. Four datasets are highlighted with green rounded rectangles: "nvidia/Nemotron-CrossThink", "nvidia/OpenMathReasoning", "rajpurkarlab/ReXGradient-160K", and "nvidia/OpenCodeReasoning". Below these, another dataset, "nvidia/Llama-Nemotron-Post-Training-Dataset", is also highlighted.

The screenshot shows the Ai2 website. The header features the Ai2 logo and navigation links for Open technologies +, Research +, About us, Blog, and Playground. The main content area has a light beige background. A large title reads "Tülu 3: The next era in open post-training". Below the title, the date "November 21, 2024" is displayed. A call-to-action link says "Try Tülu 3 in the Ai2 Playground: <https://playground.allenai.org>". At the bottom, a small section of text provides context about post-training and its evolution from InstructGPT to ChatGPT.

Post-training — a collection of techniques including instruction tuning followed by reinforcement learning from human feedback — has become a vital step in refining behaviors and unlocking new capabilities in language models. Since early approaches such as InstructGPT and the original ChatGPT, the sophistication and complexity of post-training approaches have continued to increase, moving towards multiple rounds of training, model merging, leveraging synthetic data, AI feedback, and multiple training algorithms and objectives (such as in the Llama 3.1 report).

**open vs closed**

# The **spectrum** of open

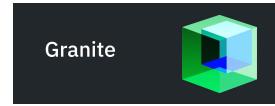
**closed** model APIs



**open** model weights



**fully open** model

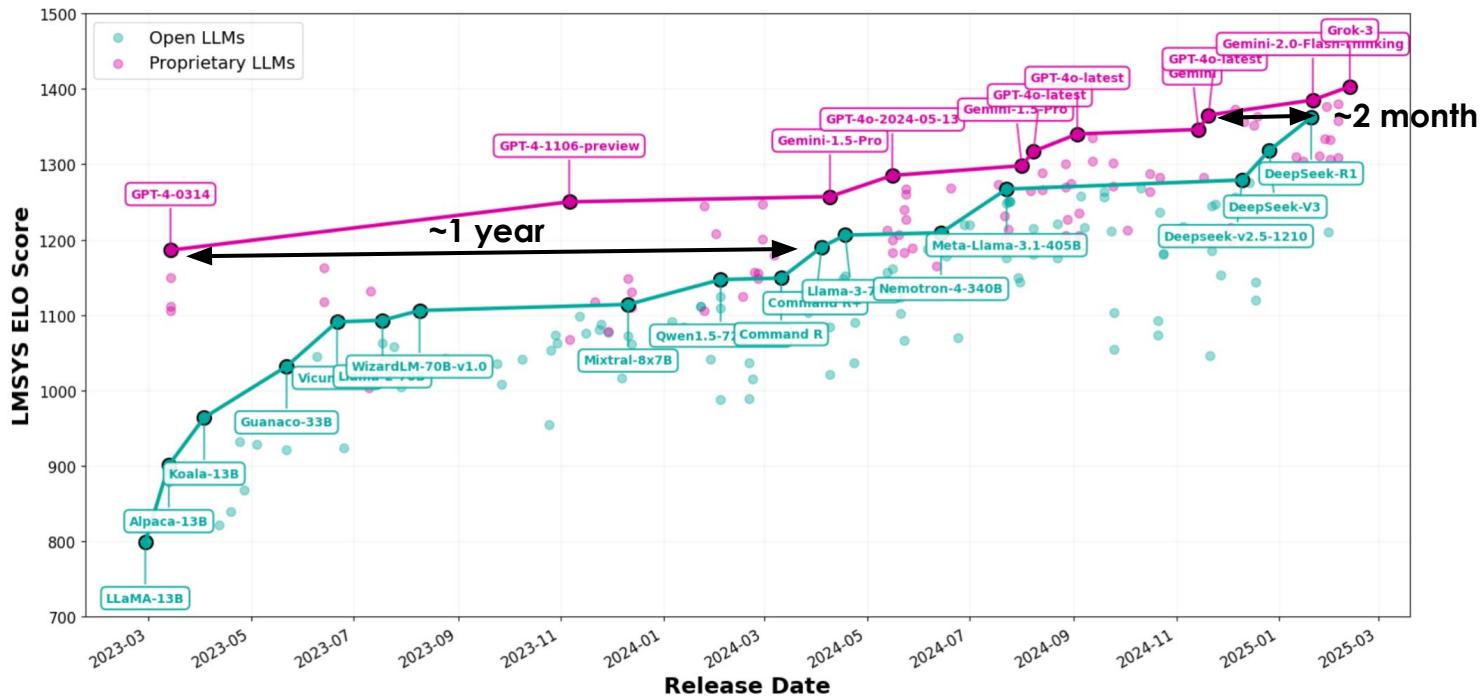


closed

open



# Open source is **closing the gap!**



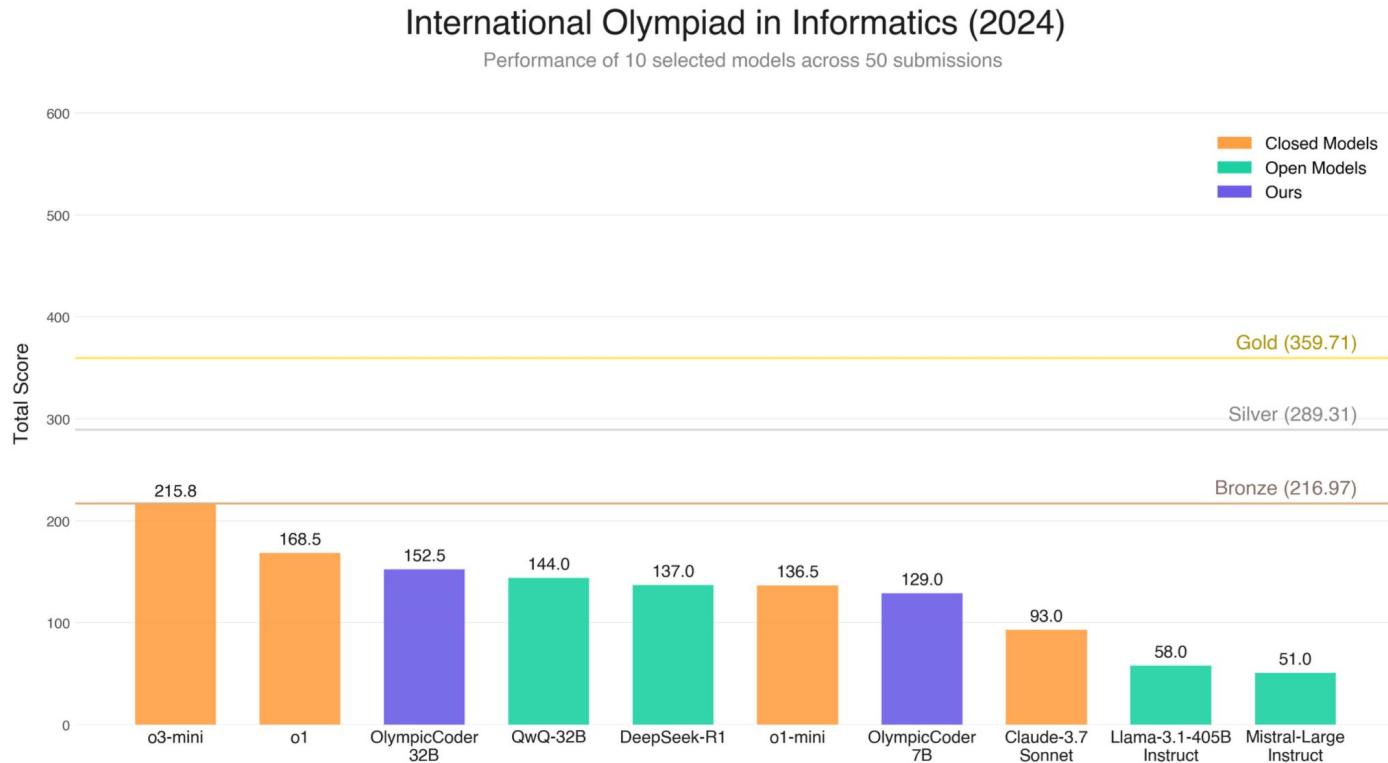
<https://huggingface.co/spaces/andrewreed/closed-vs-open-arena-elo>

# Closed vs **open** access models

**“Software is eating the world.”**

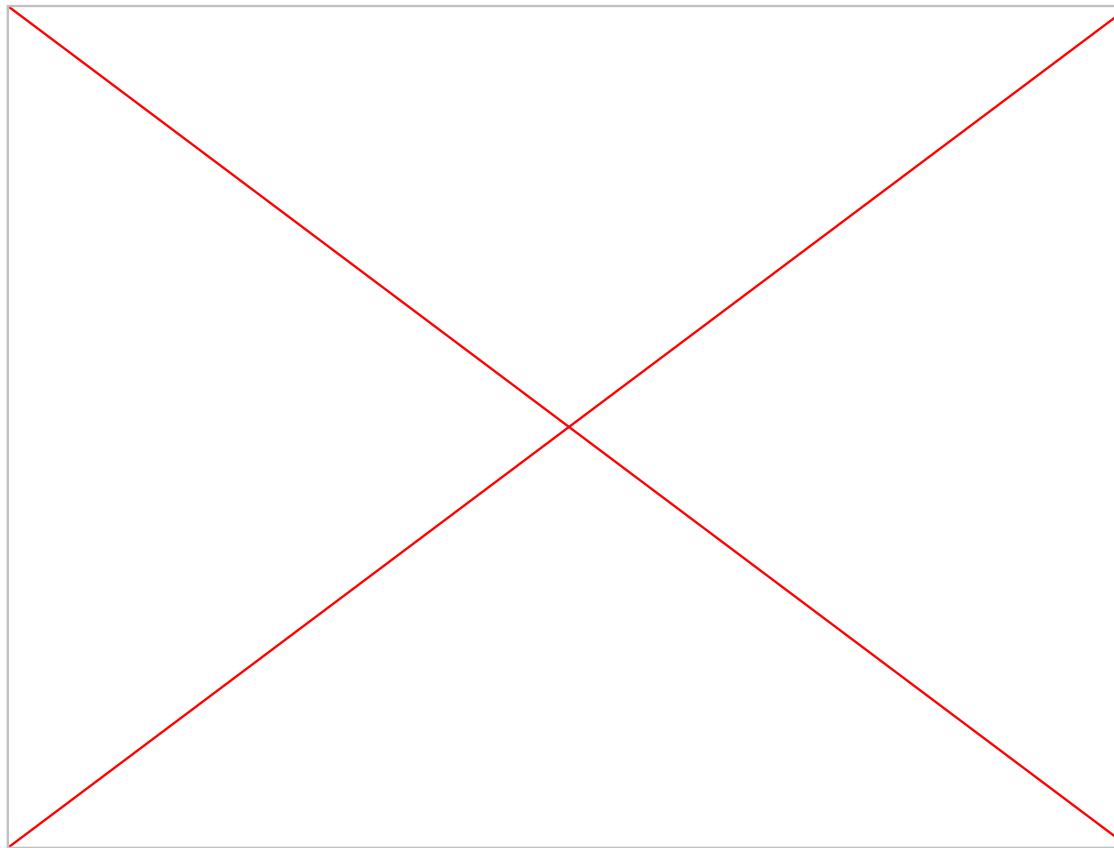
-Marc Andreessen, 2011

# General purpose vs **task specific** models



How?  
Data!

# Agents: Jupyter Agent



# Agents: DABStep

**Question:**

Which card scheme had the highest average fraud rate in 2023?

**Data:**

Name	Description
payments.csv	Payments dataset of 138k (anonymized) transactions with various signals around fraud and risk use-cases.
payments-readme.md	Documentation for the Payments dataset
acquirer_countries.csv	Table of Acquiring Banks and respect Countries
fees.json	Extensive dataset composed of 1000 Scheme Fee structures.
merchant_category_codes.csv	Table of Merchant Category Codes (MCCs)
merchant_data.json	Table describing merchants
manual.md	In finance, business contexts are often outlined in extensive handbooks from networks, regulators, and processors. For the first version of this benchmark, we have created a markdown file (manual.md) that distills essential business knowledge into a precise yet simplified format for solving tasks accurately.

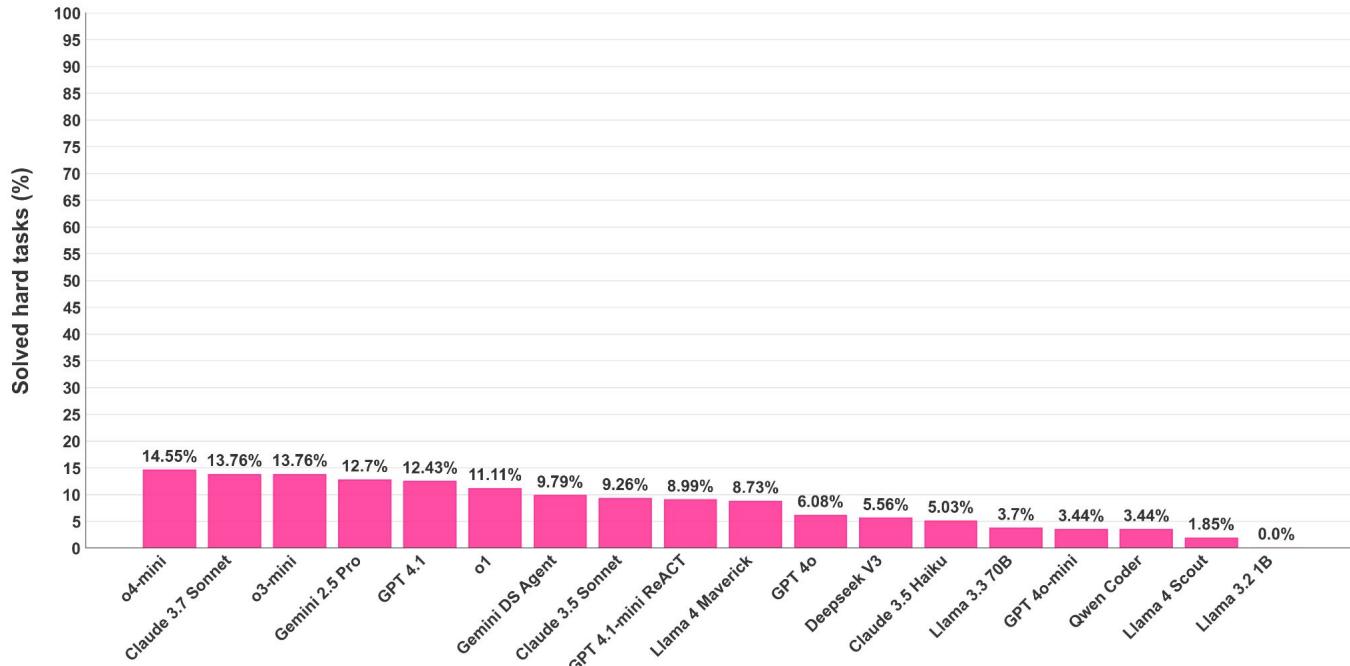
**Solution (220 Loc):**



The screenshot shows a code editor window with a large amount of Python code. The code is organized into several functions and classes, including `get_fraud_rate`, `analyze_payments`, `process_transactions`, and `MerchantCategory`. The code uses various libraries like `pandas`, `numpy`, and `scikit-learn` for data processing and machine learning. The script is designed to analyze payment data to identify the card scheme with the highest average fraud rate in 2023.

# Agents: DABStep

**DABstep**  
Data Analysis Benchmark for Multi-step Reasoning



Thank you!



GitHub/HF Hub/X: **lvwerra**