

 adaption

Shaping the next era of intelligence

My work has centered around innovation at the edge of what is possible



Sara



100+
published
research
papers



7 awards,
including one
Best Paper
award for our
Aya Model

Led the **launch** of
foundation models:

- Command A
- Aya Vision 8b
- Aya Vision 32b
- Aya Expanse 32b
- Aya Expanse 8b
- Aya 101
- R+
- R
- Command R7B



Current AI is Monolithic



We serve the same
model to billions of
people.

Everyone is expected to
become expert prompt
engineers.

It is also very inefficient.
We spend the same
amount of compute on
all problems.

We are Ripe for a
Revolution

1

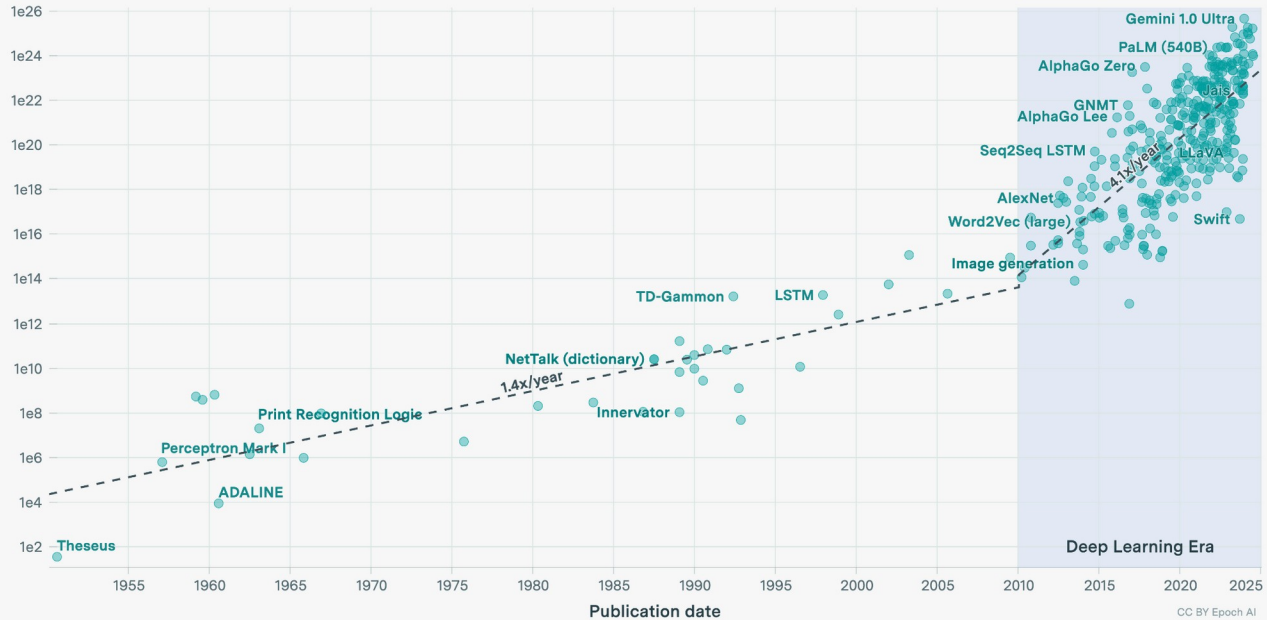
How Did We Get Here? The Limits of Scaling

A Decade of “Bigger is Better”

Notable AI Models

EPOCH AI

Training compute (FLOP)



This is captured by Rich Sutton as the “bitter lesson”

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

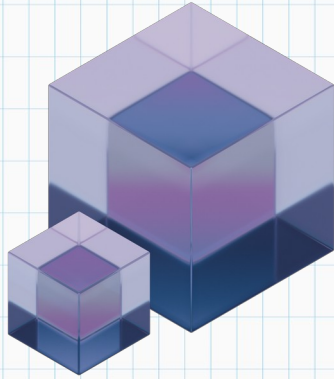
In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that “brute force” search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

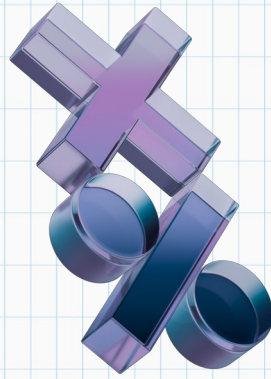
“... the only thing that matters in the long run is the leveraging of compute.”

In a punch to the ego of every computer scientist out there, what is being implied is that nothing in computer science history has worked as well as letting a model learn patterns for itself coupled with scaling the algorithm.

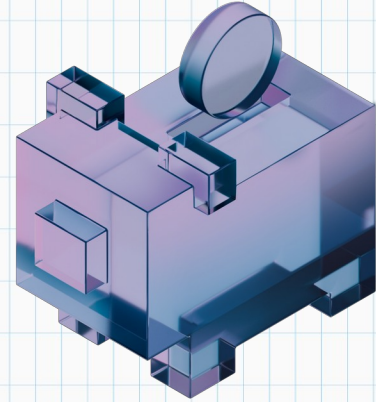
Compute has been the default bet



Lower risk
Inventing new
optimization techniques
is hard.



Predictable
Predictable gains fit
quarterly cycles and are
hard to argue against.



Fundable
Capital raises are built on
the belief that compute
matters most.

The need for compute has resulted in jokes about being GPU rich or GPU poor.



“Today we can’t think without holding a piece of metal.”

Michael Jordan

It has shifted contributions from academic to industry research.

Number of Parameters of Significant Machine Learning Systems by Sector, 1950–2022

Source: Epoch, 2022 | Chart: 2023 AI Index Report

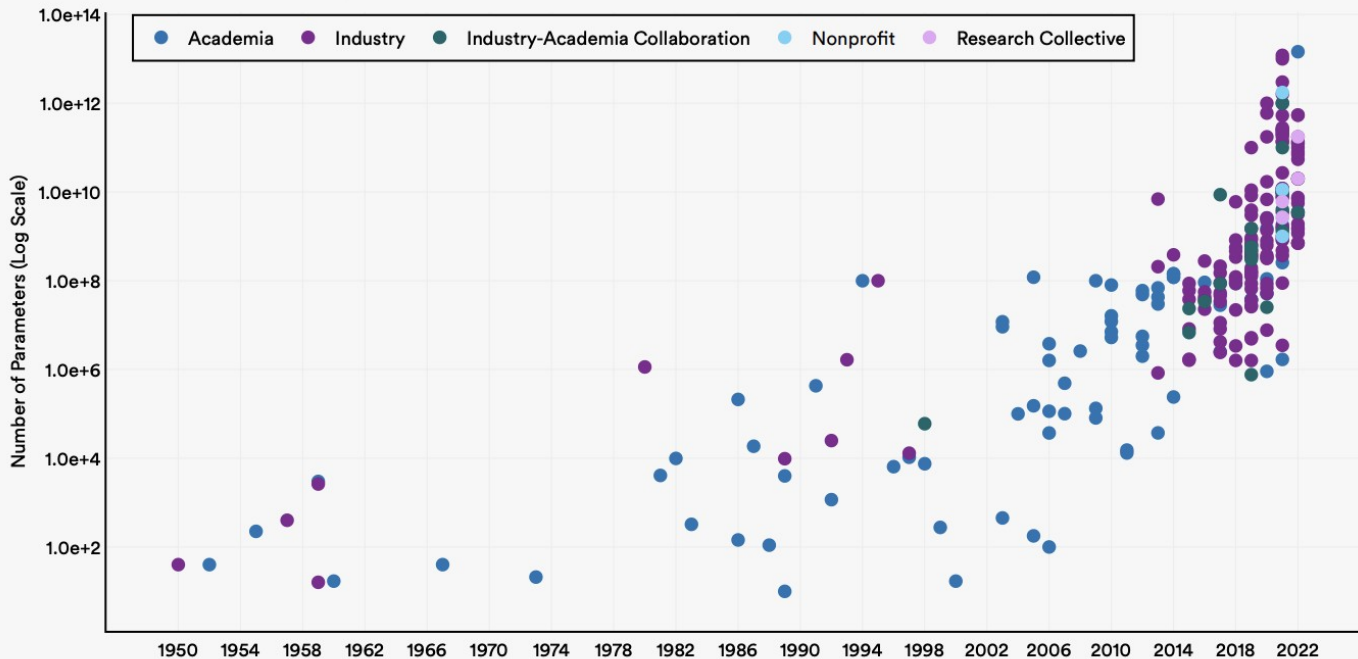


Figure 1.2.9

And has determined who gets to participate in breakthroughs and who doesn't.

Number of Significant Machine Learning Systems by Country, 2002–22 (Sum)

Source: AI Index, 2022 | Chart: 2023 AI Index Report

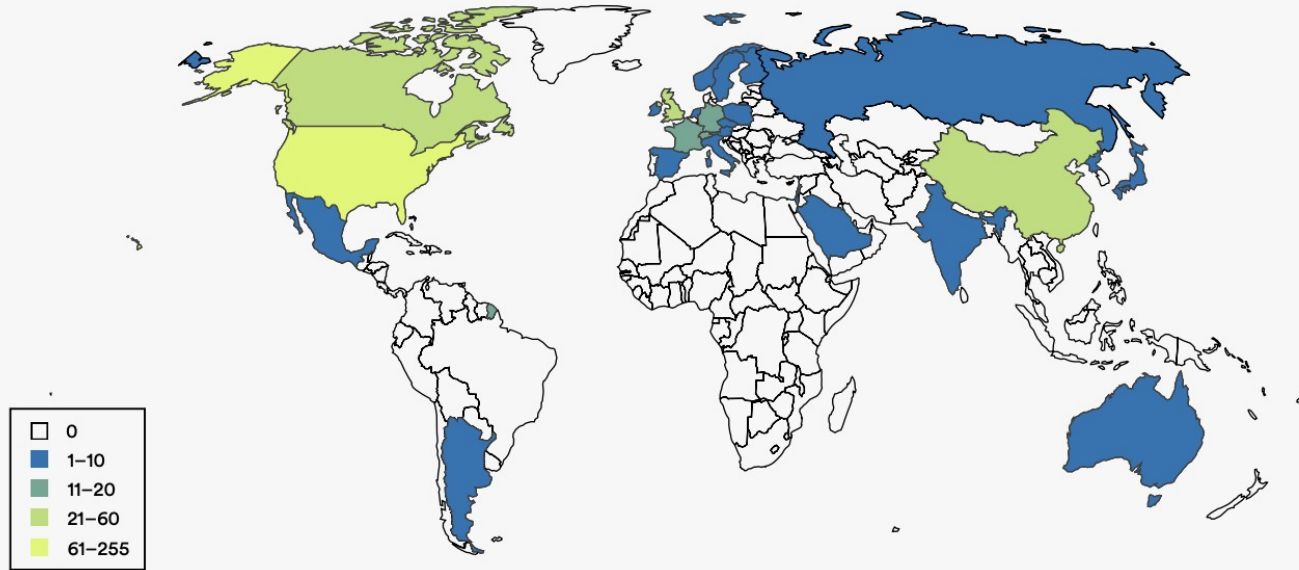


Figure 1.2.5

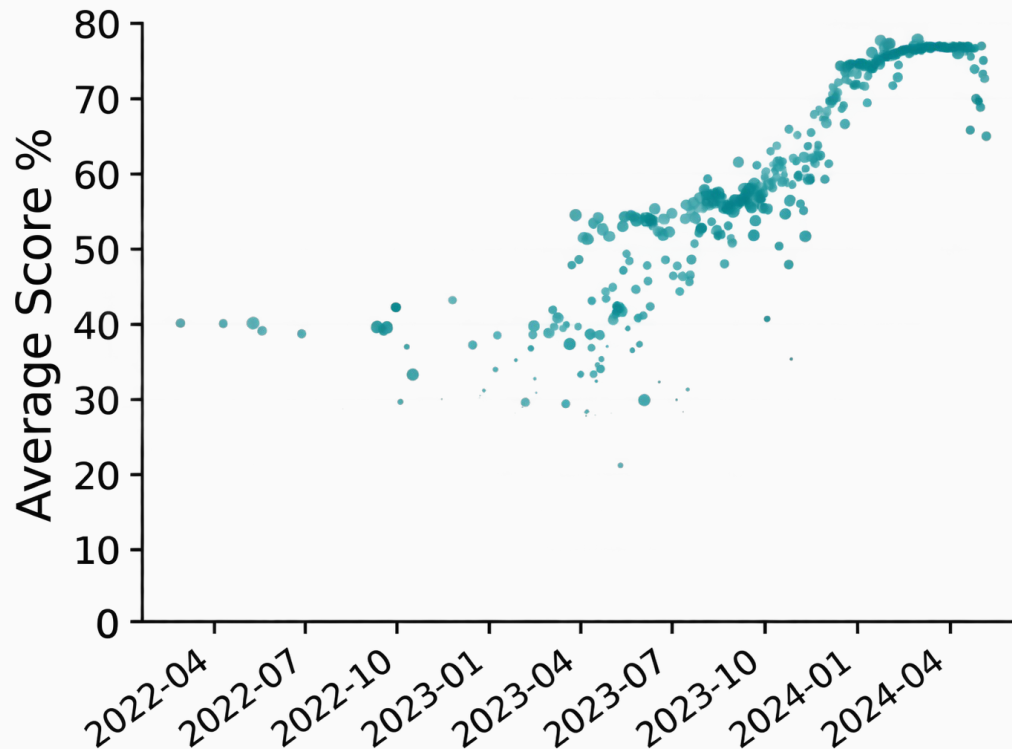
Is Sutton right?

It is controversial to suggest that scaling is over.

But recent evidence contradicts the view that monolithic large models are the future.

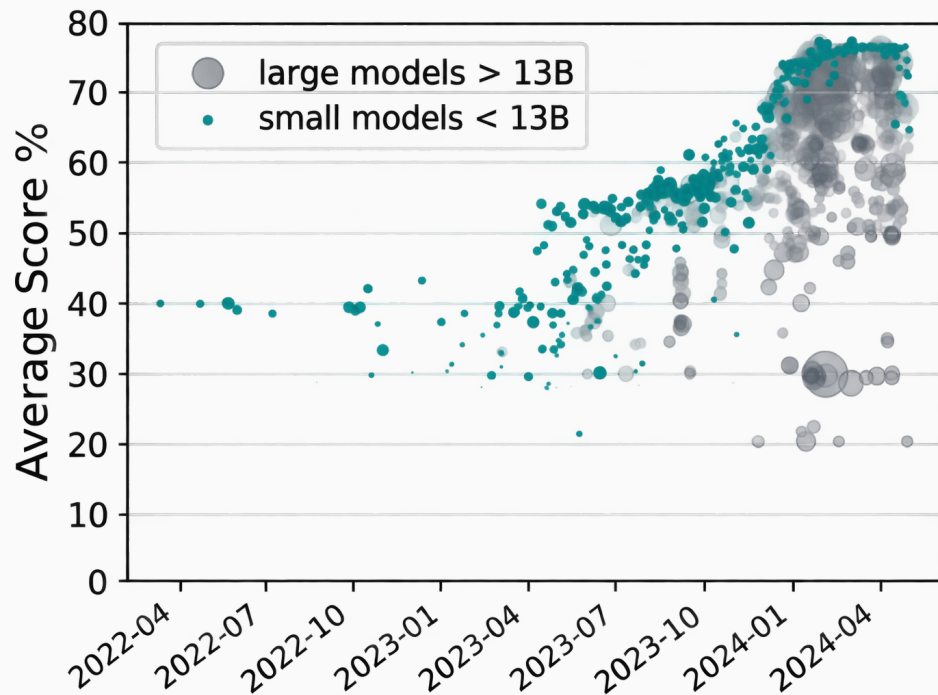
Same size, far more capable

Models under 13B on the LLM open leaderboard over time.



Small Models Frequently Outperform Larger Ones

All models over 13B (grey) that underperform the best daily model under 13B submitted to the LLM open leaderboard (green).



Recent models have required far more compute but have generally not been considered a major step forward.

 Interconnects | Nathan Lambert
<https://www.interconnects.ai>

GPT-4.5: "Not a frontier model"? - by Nathan Lambert

Feb 28, 2025 — GPT-4.5 is a point on the graph that scaling is still coming, but trying to make sense of it in a day-by-day transition is hard.

Analysis and Technology

Is OpenAI hitting a wall with huge and expensive GPT-4.5 model?

Some researchers think OpenAI's giant and expensive latest model is a sign that tech companies cannot keep making progress by continually scaling up

By **Matthew Sparkes**

📅 28 February 2025



POST

 **Yuchen Jin**  
@Yuchenj_UW

Everyone says Opus 5.5 is crazy. Best model. Massive jump. Blah blah.

I still don't think it's better than Astra.

Frontier LLM coding capability has plateaued after Opus 4.8.

9:00 PM · Sep 26, 2026 · **145.2K** Views

 322  56  1K  112 

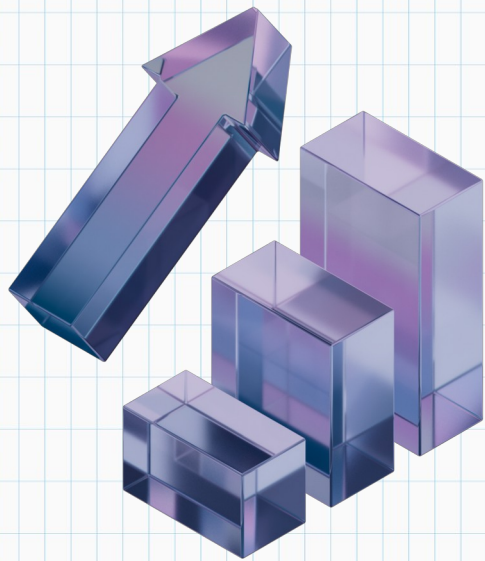
Relevant ▾ View quotes >

Meta's Llama 4: Revolution or Disappointment?

A Deep Dive into AI's Newest Contender

 **Dinmay kumar Brahma**  6 min read · Apr 14, 2025

In the rapidly evolving landscape of artificial intelligence, Meta has just unleashed its latest creation: Llama 4. Released on April 5, 2025, this new suite of AI models represents Meta's most ambitious foray yet into the competitive world of foundation models. But does it live up to the hype, or is it just another incremental step in an increasingly crowded field? Let's dissect what makes these models unique, where they excel, where they fall short, and what it all means for the future of AI.

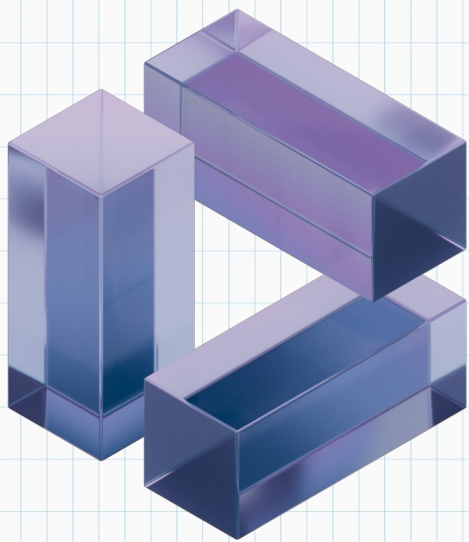


A new era of intelligence is here.

Scaling pre-training is dead (or slowly dying).

Rate of return for compute is much higher in post-training, test time scaling, adaptive compute, co-design with hardware.

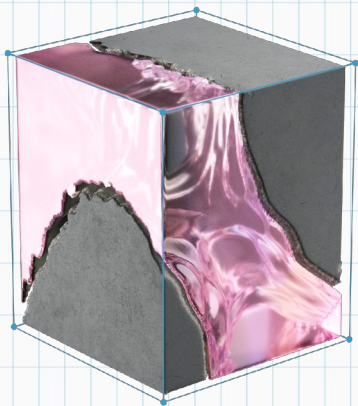
2



When scale is not the most important ingredient.

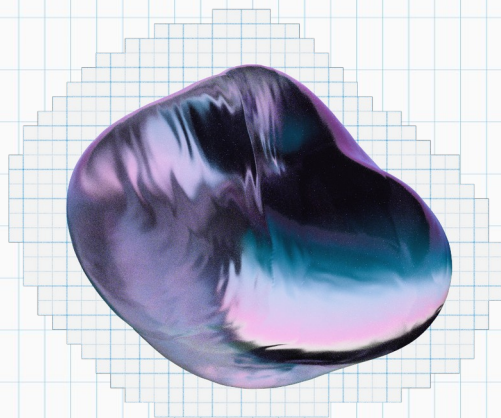
Speed and ease of building intelligence becomes more important.

What if we could allow anyone to build the same intelligence found in frontier labs.



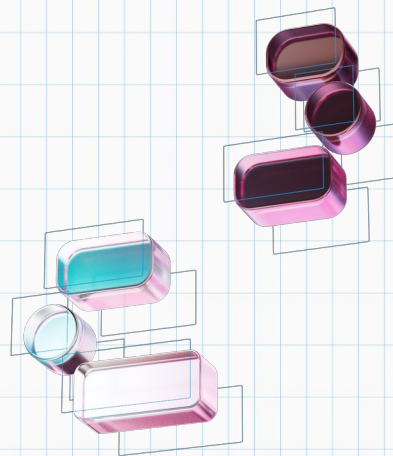
Adaptive Data

Data is the foundation of intelligence. We dynamically shape data at scale to target new objectives and rapidly scale quality.



Adaptive Intelligence

We are building the next generation of adaptive intelligence. Evolved for any industry, language or specialization.



Adaptive Interfaces

An innovation hub focused on re-imagining the interface of the human AI interaction.

Less than

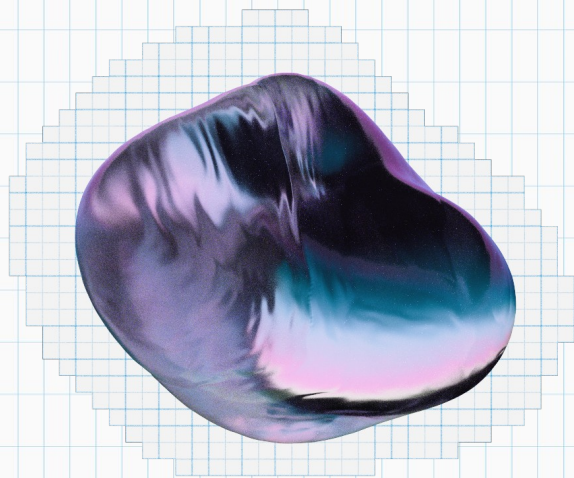
0.00001%

of the world know how to
train a frontier model

What if we made that

0.000010%

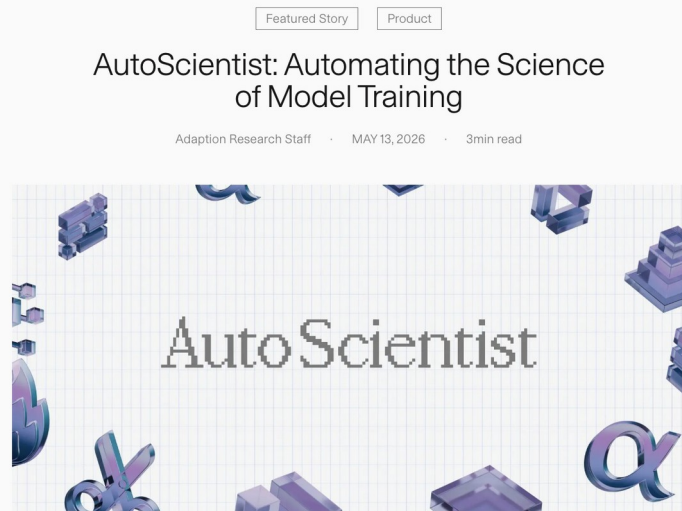
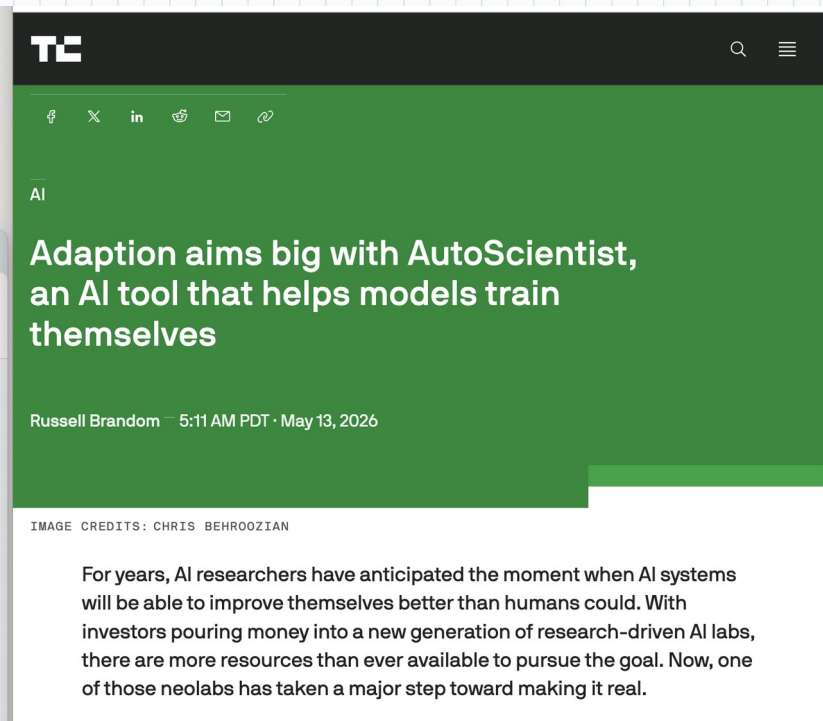
We are building continuously self-improving models



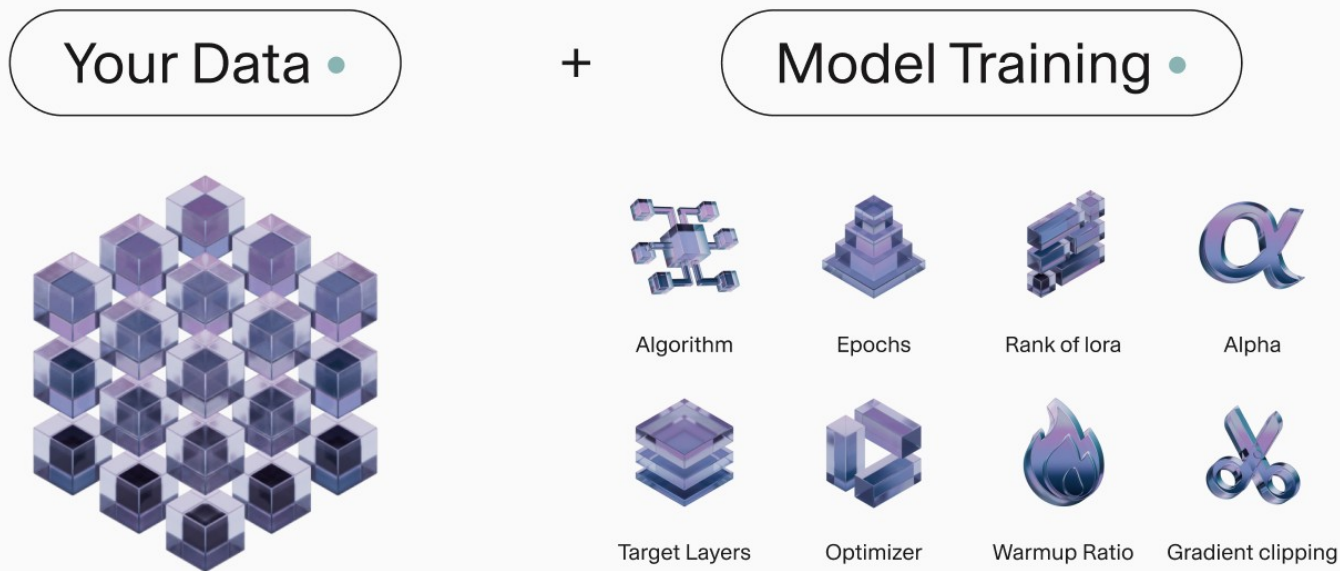
Adaptive Intelligence

We are building the next generation of adaptive intelligence. Evolved for any industry, language or specialization.

We recently released AutoScientist.



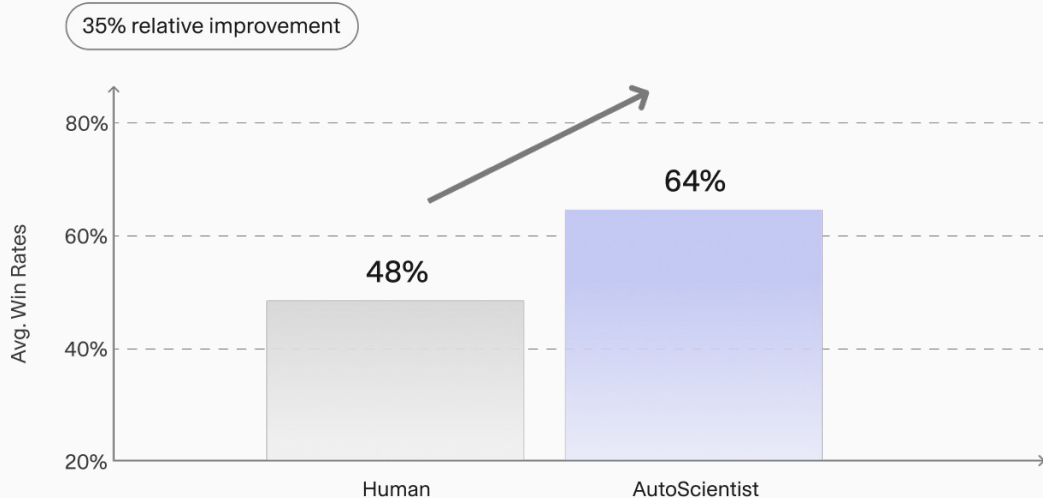
AutoScientist self-improves and automates the full research loop behind model training and alignment.



AutoScientist

outperforms our research staff at making decisions about
how to train frontier models

AutoScientist Outperforms Human-Configured Training



AutoScientist for Harness Design

Harness Optimization

Improved Qwen3.6-27B pass rate from 67% to 85% on the Harvey Legal Agent Benchmark.

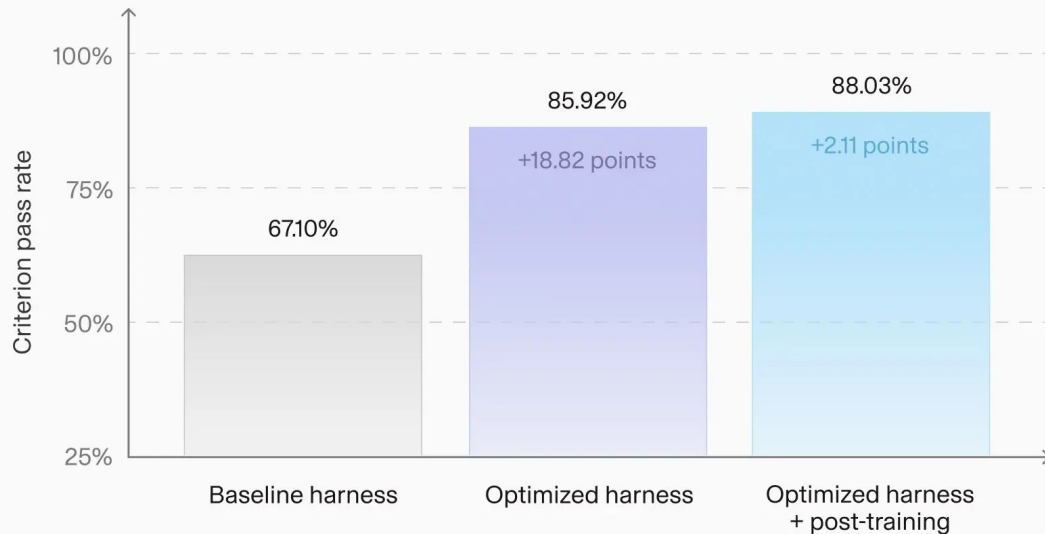
Token Efficiency

Achieved performance within 0.79 percentage points of Prime Agent while reducing input token volume by 54%.

Post-Training

Further increased the pass rate from 85% to 88%, while increasing the complete task pass rate from 5% to 9%.

Qwen performance improved in two stages



AutoScientist for Agent Memory

Token Reduction

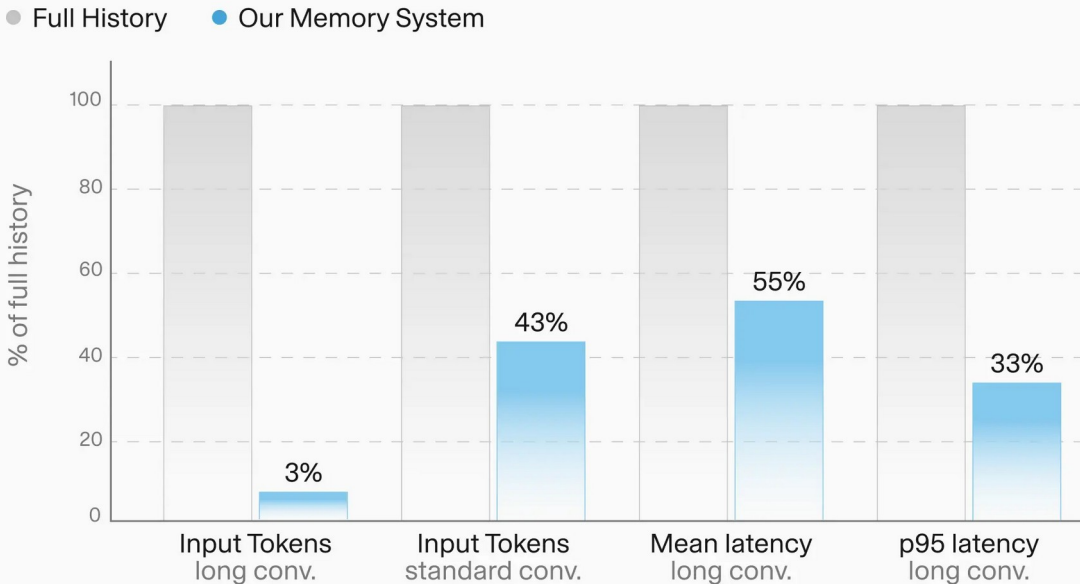
Cut context tokens by 97% on long conversations and 57% on standard conversations by using write-time memory instead of full history.

Latency Reductions

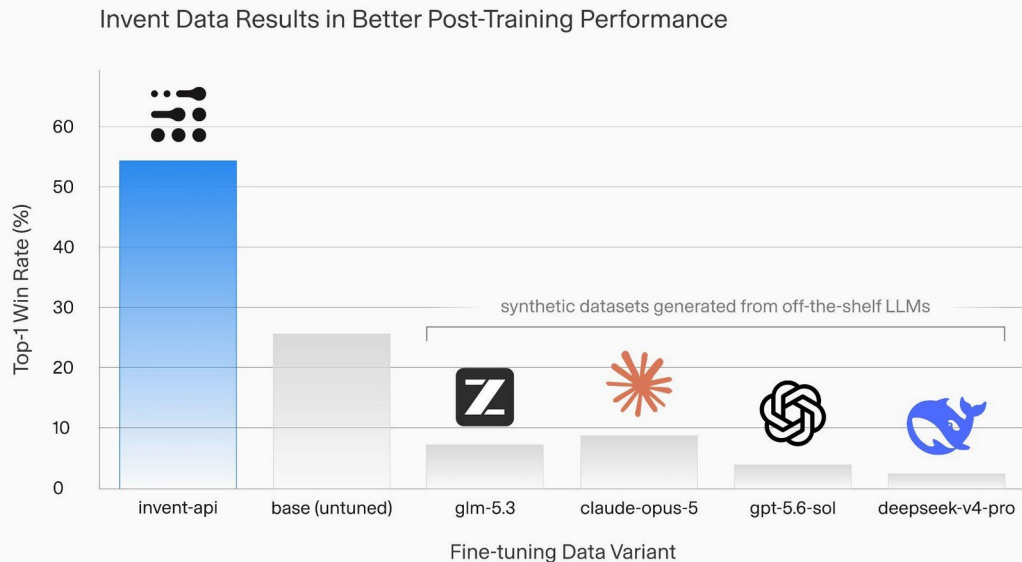
Context volume reduction decreased mean latency by 45% and p95 latency by 67% on long conversations.

Benchmark Performance

Achieved an accuracy score of 90% on LongMemEval, compared to 60% using full-history inference with the same underlying model.



AutoScientist for dataset creation. AI ready data in a day, powering AI systems



82%

Average quality gains

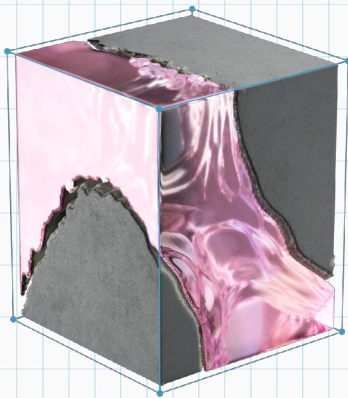
242

Languages covered

63,256,042

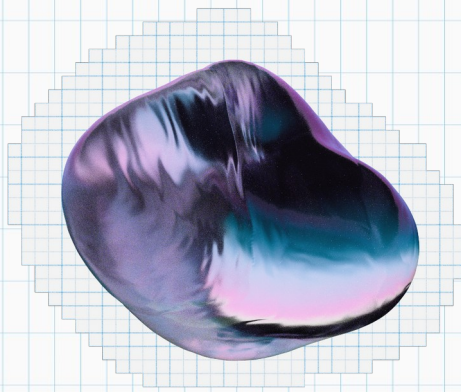
Data artifacts processed

This makes the entire ML stack more adaptive, easier to modify and customize. It also enables true ownership.



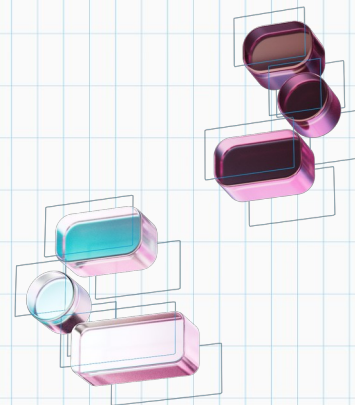
Adaptive Data

Data is the foundation of intelligence. We dynamically shape data at scale to target new objectives and rapidly scale quality.



Adaptive Intelligence

We are building the next generation of adaptive intelligence. Evolved for any industry, language or specialization.



Adaptive Interfaces

An innovation hub focused on re-imagining the interface of the human AI interaction.

My 2026 prediction: you will build and own frontier AI.

The future of machine intelligence is adaptive.

It will be shaped by whoever makes adaptation as efficient as possible.

On the slow death of scaling.

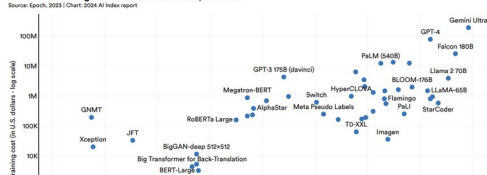
Sara Hooker

Abstract

This essay engages with a decades-old debate at the heart of computer science progress, namely, is *bigger always better*? For the last decade, it has been hard to stray off the beaten path of accepted wisdom for what drives innovation. We have been held hostage to a painfully simple formula: *scale model size and training data*. A pervasive belief in scaling has resulted in a massive windfall in capital for industry labs and fundamentally reshaped the culture of conducting science in our field. Academia has been marginalized from meaningfully participating in AI progress and industry labs have stopped publishing. Yet, this essay will posit that the relationship between training compute and performance is highly uncertain and rapidly changing. Relying on scaling alone misses a critical shift that is underway, and ignores more interesting levers of progress. All this suggests that key disruptions lie ahead.

1 How we got here.

Estimated training cost of select AI models, 2016-23

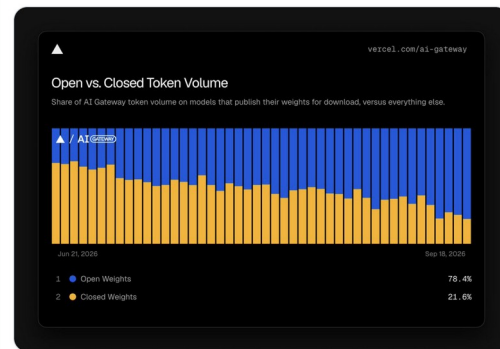


Looks like today may be a record day for token volume % of open models on Vercel AI Gateway:

Open 78.4% Closed 21.6%

While spend \$ usually tells a different story, #3 and #4 today are Moonshot AI & DeepSeek. Adding Z.ai, their combined spend surpasses OpenAI (#2).

(Do note that's the spend for inference of the model across providers (mostly in the US), not revenue going directly to the open weight labs.)





Thank you.

sara@adaptionlabs.ai

Change who owns
intelligence.

