STATE OF AI REPORT .

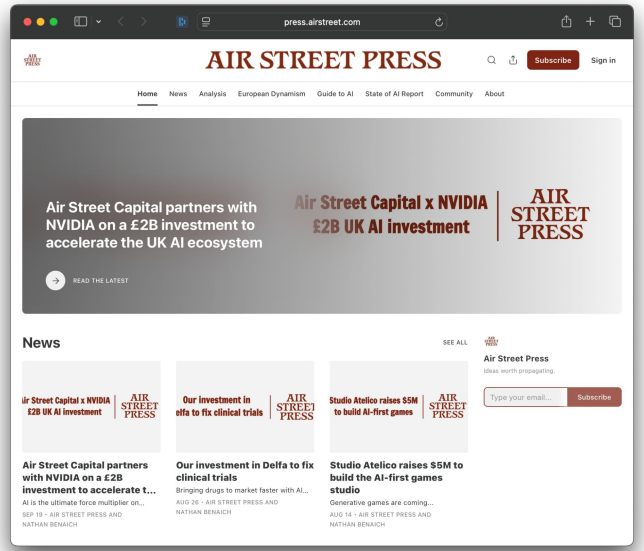
October 9, 2025

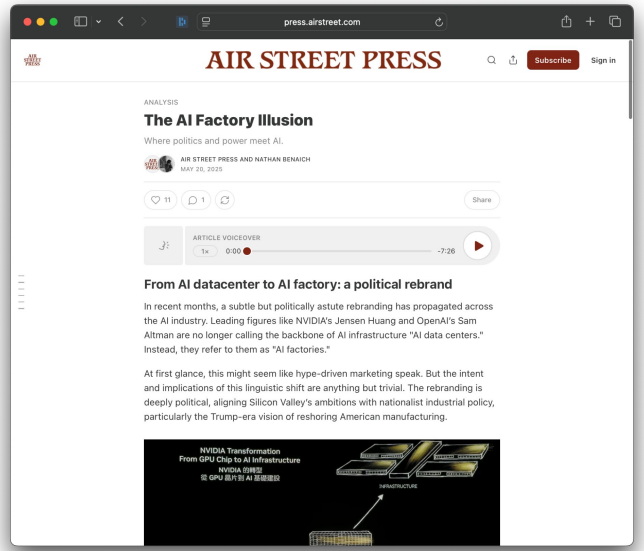
Nathan Benaich

AIR STREET CAPITAL .

Follow our writing on AIR STREET PRESS (press.airstreet.com)

- **If you enjoy reading the State of AI Report, we invite you to read and subscribe to Air Street Press, the home of our analytical writing, news, and opinions.**





About the authors

Nathan Benaich

Nathan is the General Partner of **Air Street Capital**, a venture capital firm investing in AI-first companies. He runs the Research and Applied AI Summit (RAAIS), the RAAIS Foundation (funding open-source AI projects), AI communities in the US and Europe, and Spinout.fyi (improving university spinout creation). He studied biology at Williams College and earned a PhD from Cambridge in cancer research as a Gates Scholar.

State of AI Report 2025 team

Zeke Gillman

Zeke is a Tech Policy Fellow at Stanford, and co-author of *Regulating under Uncertainty*. He previously worked at Harvard Business School and the DOJ Antitrust Division, and holds a BA in Political Science and Philosophy from the University of Chicago.

Nell Norman

Nell is a grad student in Computing at Imperial College London focusing on how LLMs could enable scalable vishing fraud. She previously helped AI teams build reliable products at AI agent platform V7 Labs, and has a first class BA from Oxford University.

Ryan Tovcimak

Ryan is a founder of the AI Stack Tracker. His work spans red-teaming frontier models, benchmarking the global AI competition, and tracking trends in AI compute and power demands. He holds a BS in Econ from Vanderbilt University.

Artificial intelligence (AI) is a multidisciplinary field of science and engineering devoted to creating intelligent machines.

AI acts as a force multiplier for technological progress in our increasingly digital, data-driven world. This is because everything around us, from culture to consumer products, is ultimately a product of intelligence.

Now in its eighth consecutive year, the State of AI Report is the most widely read and trusted open-access publication tracking progress in artificial intelligence. Consider it a curated compilation of the most significant and thought-provoking work from the past 12 months. Our goal is to inform and shape an ongoing conversation about the state of AI, where the field is heading, and what its developments mean for the future.

This year's report examines six key dimensions of the AI ecosystem:

- **Research**: Technological breakthroughs and their capabilities.
- **Industry**: Areas of commercial application for AI and their business impact.
- **Politics**: Regulation, economic implications, and the evolving geopolitics of AI.
- **Safety**: Efforts to identify and mitigate catastrophic risks that highly capable future AI systems could pose.
- **Survey**: Findings from the largest open-access survey of 1,200 AI practitioners and their AI usage patterns.
- **Predictions**: Our outlook for the next 12 months, alongside a review of our 2024 forecasts to keep us accountable.

Produced by **Nathan Benaich and Air Street Capital team**.

Definitions

Artificial intelligence (AI): a broad discipline with the goal of creating intelligent machines, as opposed to the natural intelligence of humans and animals. While **artificial general and super intelligence (AGI and ASI)** are terms that don't agreed upon definitions, we use them to describe machines that could match (AGI) and then exceed (ASI) the full range of human cognitive ability across all economically valuable tasks.

AI Agent: an AI-powered system that can take actions in an environment. For example, an LLM that has access to a suite of tools and has to decide which one to use in order to accomplish a task that it has been prompted to do.

AI Safety: a field that studies and attempts to mitigate the risks (minor to catastrophic) which future AI could pose to humanity.

Context window: The number of input tokens that an LLM model can attend to while answer a user's prompt.

Diffusion: An algorithm that iteratively denoises an artificially corrupted signal in order to generate new, high-quality outputs. In recent years it has been at the forefront of image generation and protein design.

Environment: The world an AI agent acts in. It receives the agent's actions and returns the next observation and often a reward (i.e. a signal of the action being good or bad). In this context, **trajectories** are the time-ordered record of an agent's experience in an environment, typically tuples like (observation/state, action, reward, next observation) from start to finish. These trajectories are used for RL.

Function calling / tool use: Structured calls that let models invoke APIs, search, code, or calculators with typed arguments and schemas.

Generative AI: A family of AI systems that are capable of generating new content (e.g. text, images, audio, or 3D assets) based on 'prompts'.

Graphics Processing Unit (GPU): the workhorse AI semiconductor that enables a large number calculations to be computed in parallel.

Definitions

(Large) Language model (LM, LLM): a model trained on vast amounts of (often) textual data to predict the next word in a self-supervised manner.

Mixture-of-Experts (MoE): A model type where only few expert blocks activate per token, giving high capacity at lower compute per step.

Prompt: a user input often written in natural language that is used to instruct an LLM to generate something or take action.

Reasoning model: A model that plans and verifies its thinking as it generates output tokens, often via test-time compute and post-hoc checking. The model's explicit step-by-step reasoning trace (intermediate tokens that lay out calculations, sub-goals, and logical steps en route to an answer) is called a Chain of Thought (CoT).

Reinforcement learning (RL): an area of ML in which software agents learn goal-oriented behavior by trial and error in an environment that provides rewards or penalties in response to their actions (called a "policy") towards achieving that goal.

Test-time compute (or inference-time compute): Spending more inference budget (longer chains, multiple samples, self-consistency) to raise accuracy without changing weights.

Transformer: a model architecture at the core of most state of the art (SOTA) ML research. It is composed of multiple "attention" layers which learn which parts of the input data are the most important for a given task. Transformers started in NLP (specifically machine translation) and subsequently were expanded into computer vision, audio, and other modalities.

Vision-Language-Action Model (VLAM): a model that jointly learn from visual inputs, natural language, and embodied interactions to not only interpret and describe the world but also to plan and execute actions within it. Without the actions piece, this model becomes a VLM.

World model: a model that predicts next states conditioned on actions, enabling real-time, interactive control.

Definitions

Model type legend

In the rest of the slides, icons in the top right corner indicate input and output modalities for the model.

Input/Output types:

📝: Text
🖼️: Image
</> : Code
🛠️ : Software tool use (text, code generation & execution)
🎥 : Video
🎵 : Music
📦: 3D
🤖: Robot state
🧬: Biological modality
🧪: Chemical modality

Model types:

📝 → 📝 : LLMs
📝+ 🖼️ → 📝 : Multimodal LLMs
📝+ 🖼️+ 🤖 → 📝 : Multimodal LLMs for Robotics
📝 → </> : Text to Code
📝 → 🛠️ : Text to Software tool use
📝 → 🖼️ : Text to Image
📝 → 🎥 : Text to Video
📝 → 🎵 : Text to Music
🖼️ → 📦 : Image to 3D
📝 → 📦 : Text to 3D
📝 → 🧬/🧪 : Biological/chemical models
📝 → 🌎 : World model

Executive Summary

Research

- Reasoning defined the year, with OpenAI, Google, Anthropic, and DeepSeek trading leads and pushing visible “think-then-answer” methods into real products.
- Open models improved fast and China’s open-weight ecosystem surged, yet the top models remain closed and keep widening their capability-per-dollar edge.
- Benchmarks buckled under contamination and variance, while agents, world models, and domain tools (code, science, medicine) became actually useful.

Industry

- Real revenue arrived at scale as AI-first companies crossed tens of billions, and flagship labs stretched their lead with better capability-to-cost curves.
- NVIDIA ripped past $4T and 90% ownership of AI research papers while custom chips and neoclouds rose. Circular mega-deals funded huge build-outs.
- Power became the new bottleneck as multi-GW clusters moved from slideware to site plans and grid constraints started to shape roadmaps and margins.

Politics

- The AI race heats up as the U.S. leans into “America-first AI” with export gyrations while China accelerates self-reliance ambitions and domestic silicon.
- Regulation takes a back seat in the face of turbo-investments: international diplomacy stalls and the AI Act runs into implementation hurdles.
- “AI goes global” became concrete, with petrodollars and national programs funding gigantic data centers and model access as job loss data trickles in.

Safety

- AI labs activated unprecedented protections for bio and scheming risks, others missed self-imposed deadlines, or quietly abandoned testing protocols.
- External safety organizations operate on annual budgets smaller than what leading labs collectively spend in a single day.
- Cyber capabilities doubled every 5 months outpacing defensive measures. Criminals orchestrated ransomware using AI agents infiltrate F500 companies.

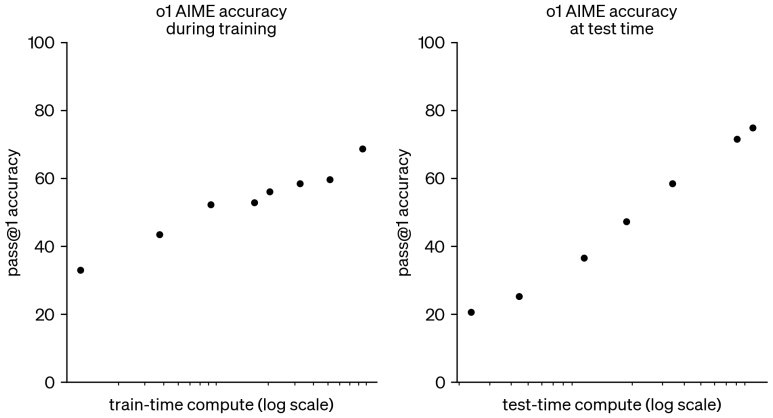
Scorecard: Reviewing our predictions from 2024

Our 2024 Prediction		Evidence
A $10B+ investment from a sovereign state into a US large AI lab invokes national security review.	~	Sovereign-backed initiatives (HUMAIN $10B VC fund, UAE's Stargate AI infra cluster) are infrastructure partnerships rather than direct majority investments into a US AI lab.
An app or website created solely by someone with no coding ability will go viral (e.g. App Store Top-100).	YES	Formula Bot, built entirely using Bubble, exploded to 100,000 visitors overnight from a Reddit post and generated $30,000 in its first three months.
Frontier labs implement meaningful changes to data collection practices after cases begin reaching trial.	YES	Anthropic landmark $1.5B settlement with authors, deleting works and shifting to legally acquired books. OpenAI's paid content partnerships with Future (owner of Marie Claire).
Early EU AI Act implementation ends up softer than anticipated after lawmakers worry they've overreached.	NO	The Commission is phasing obligations and leaning on a voluntary GPAI Code of Practice first, so early implementation has been softer, even as binding rules arrive later.
An open source alternative to OpenAI o1 surpasses it across a range of reasoning benchmarks.	YES	DeepSeek-R1 outperforms OpenAI's o1 on key reasoning benchmarks including AIME, MATH-500, and SWE-bench Verified.
Challengers fail to make any meaningful dent in NVIDIA's market position.	YES	NVIDIA remains dominant, competitors fail to make significant market share dents.
Levels of investment in humanoids will trail off, as companies struggle to achieve product-market fit.	NO	$3B has been invested into humanoids in 2025, up from $1.4B last year.
Strong results from Apple's on-device research accelerates momentum around personal on-device AI.	NO	Apple Intelligence rolled out with many models running on-device and helped push a broader industry push to on-device AI. Shipments of AI-capable smartphones climbed.
A research paper generated by an AI Scientist is accepted at a major ML conference or workshop.	YES	An AI-generated scientific paper The AI Scientist-v2 was accepted at an ICLR workshop.
A video game based around interacting with GenAI-based elements will achieve break-out status.	NO	Not yet.

Section 1: Research

Think before you speak: o1 “thinking” ignites the reasoning race

As 2024 drew a close, OpenAI released o1-preview, the first reasoning model that demonstrated inference-time scaling with RL using its CoT as a scratch pad. This led to more robust problem solving in reasoning-heavy domains like code and science. For example, the general release o1 exhibited accuracy improvements on the American Invitational Mathematics Examination (AIME) with greater training and test time compute. As a result, OpenAI leaned even more strongly into the scaling their reasoning efforts in 2025.

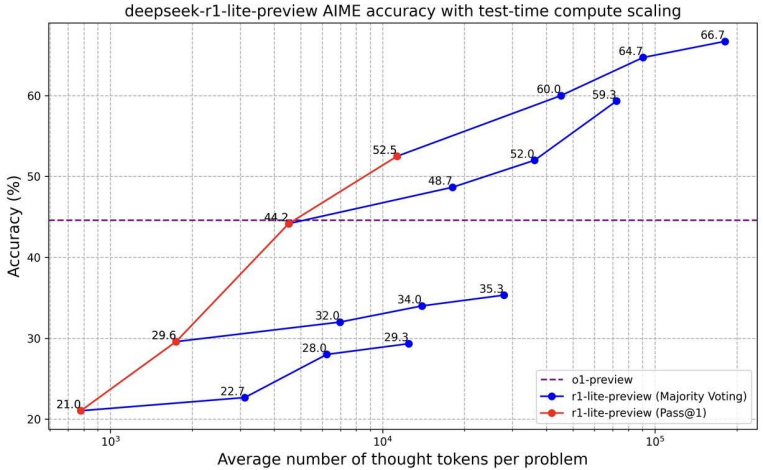


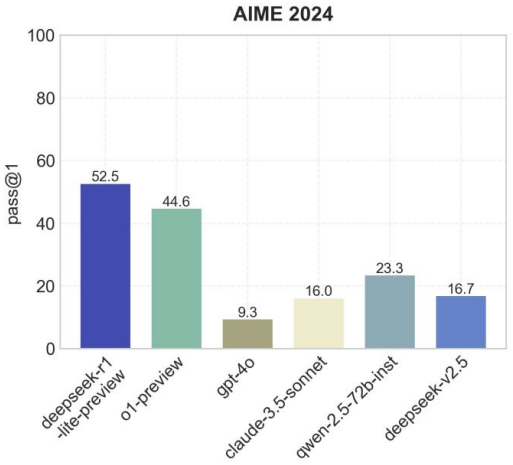
Sam Altman Three Observations February 10, 2025 at 1:05 AM

1. **The intelligence of an AI model roughly equals the log of the resources used to train and run it.** These resources are chiefly training compute, data, and inference compute. It appears that you can spend arbitrary amounts of money and get continuous and predictable gains; the scaling laws that predict this are accurate over many orders of magnitude.

Deeply sought: could frontier reasoning ever be found in the open?

Barely 2 months after o1-preview, DeepSeek (the Chinese upstart AI lab spun out of a high-frequency quant firm) released their first reasoning model, R1-lite-preview, that's built atop the earlier strong V2.5 base model. Like OpenAI, they showed predictable accuracy improvements on AIME with greater test-time compute budget. Impressively, R1-lite-preview actually beat o1-preview on AIME 2024 pass@1 by scoring 52.5 vs. 44.6. BUT, very few seemed to take notice... Wall Street certainly didn't.

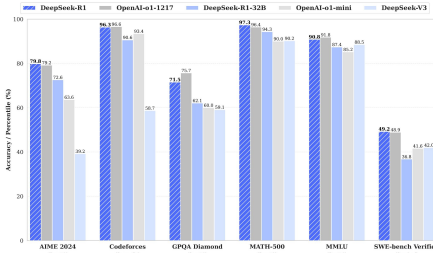


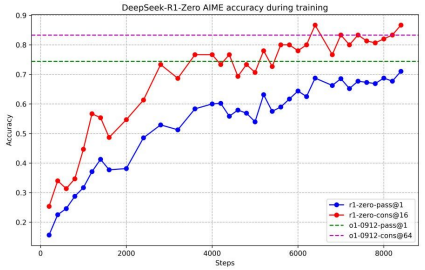


Are you not entertained? DeepSeek V3 brings you to R1

A few days after Christmas 2024, DeepSeek unveiled V3, a strong 671B MoE V3 that lowered training and inference cost with FP8 mixed precision, multi-token prediction, and auxiliary-free routing. Using V3 as the base, they trained R1-Zero only with RL using verifiable rewards and Group Relative Policy Optimization, a critic-free algorithm that removes reward and value models.

- R1-Zero follows a "think → answer" format and uses a simple rule-based reward for getting the final answer right, which is cheaper and harder to game than a learned neural reward model.
- GRPO compares multiple sampled answers within a group to form a relative baseline, so it does not need a value head or a separate reward model.
- During training the model lengthens its thoughts, explores, and reallocates compute to hard problems. Its AIME score rises from 15.6% to about 71% in roughly 8.5k steps, with majority-vote runs reaching o1-0912 levels.
- R1 then repairs readability with a small CoT warm start, a language-consistency reward, large supervised finetuning, and a final RL pass. AIME increases to 79.8, MATH-500 to 97.3, GPQA to 71.5, the approach distills well into smaller models.





More thinking, more tool use, less cost: DeepSeek V3.1 and V3.2-Exp

DeepSeek V3.1 marks a substantial leap over V3, introducing a hybrid thinking mode that toggles between reasoning and lightweight inference. It demonstrates faster "think" efficiency than R1 and V3, while greatly improving tool use and multi-step agent workflows. Now V3.2-Exp keeps that behavior and swaps dense attention for DeepSeek Sparse Attention (DSA), where a tiny "lightning indexer" picks the top-k past tokens to attend to each step. You get roughly the same capability as V3.1 on coding/search/agent tasks, but markedly lower cost and latency for 32–128K contexts.

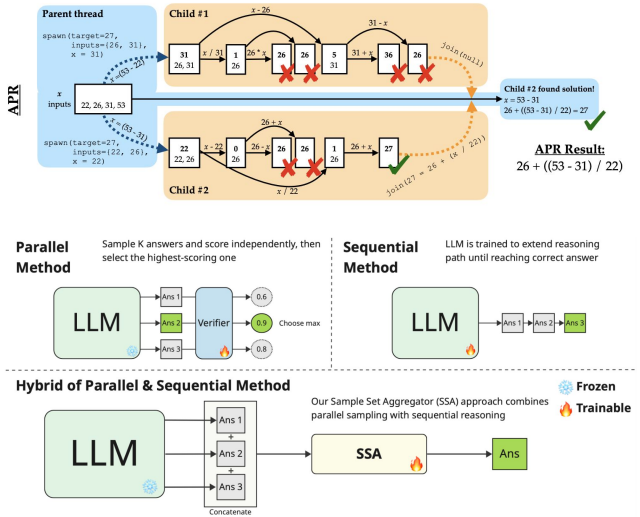
- DeepSeek V3.1 builds on the V3 base with a hybrid thinking mode, enabling both lightweight inference and deep reasoning.
- V3.2-Exp introduces a lightweight selector that picks the few tokens that matter, instead of attending to every token in a long prompt.
- The model posts similar scores to V3.1 on coding/search/agent tasks, with small dips on a few reasoning sets, and clearly lower prefill and decode cost for 32-128K contexts.

	Benchmark (Metric)	DeepSeek-V3.1-Terminus	DeepSeek-V3.2-Exp
General	MMLU-Pro (EM)	85.0	85.0
	GPQA-Diamond (Pass@1)	80.7	79.9
	Humanity's Last Exam (Pass@1)	21.7	19.8
Search Agent	BrowseComp (Acc.)	38.5	40.1
	BrowseComp_zh (Acc.)	45.0	47.9
	SimpleQA (Acc.)	96.8	97.1
Code	LiveCodeBench (2408-2505) (Pass@1)	74.9	74.1
	Codeforces-Div1 (Rating)	2046	2121
	Aider-Polyglot (Acc.)	76.1	74.5
Code Agent	SWE Verified (Agent mode)	68.4	67.8
	SWE-bench Multilingual (Agent mode)	57.8	57.9
	Terminal-bench (Terminus 1 framework)	36.7	37.7
Math	AIME 2025 (Pass@1)	88.4	89.3
	HMMT 2025 (Pass@1)	86.1	83.6

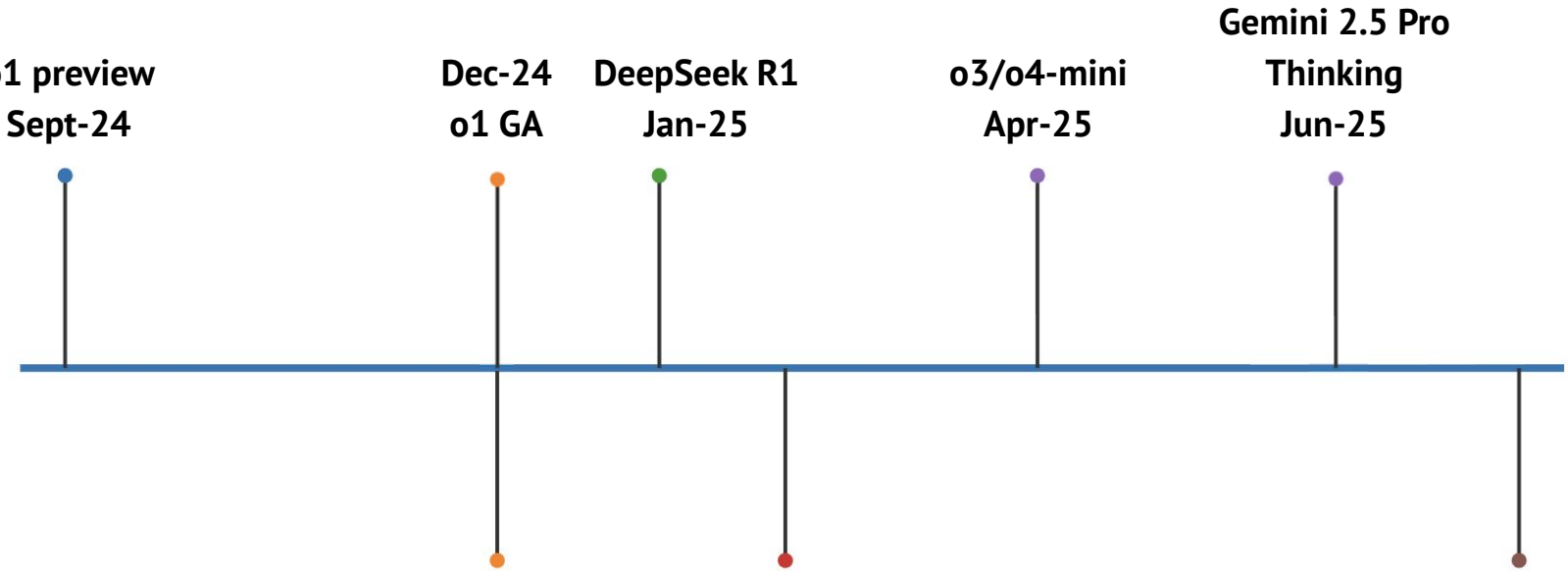
Parallel reasoning: beyond depth to branch-and-merge inference

MoE routing scales capacity but preserves single-flow inference and doesn't change how the model thinks. A new route is branching multiple inference paths and aggregating them versus just going deeper or using wider models enables exploration, reduces hallucination, and better leverages parallel hardware.

- **Adaptive Parallel Reasoning** (pictured) enables models to dynamically orchestrate branching inference through spawn() and join() operations, training both parent and child threads end-to-end using RL to optimize coordinated behavior. This boosted performance on the Countdown task at 4K context: 83.4% (APR + RL) vs. 60.0% (baseline).
- **Sample Set Aggregator** (right) trains a compact model to fuse multiple reasoning samples into one coherent answer, outperforming naive re-ranking methods.
- Models like Gemini Deep Think, which shows its step-by-step reasoning transparently, exemplify this branch-and-evaluate paradigm in deployed systems.

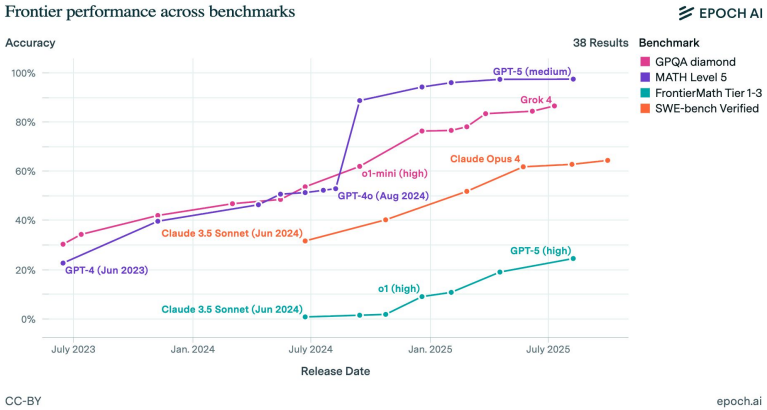


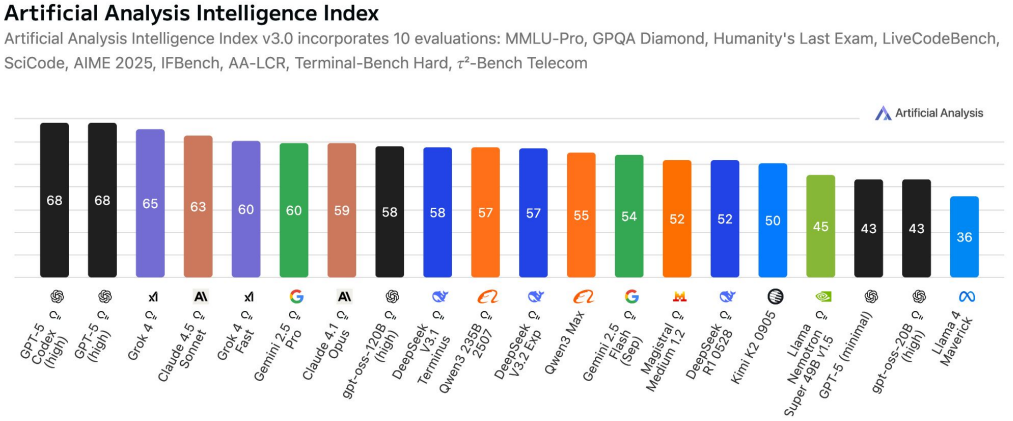
The reasoning timeline: from o1 "thinking" to R1, GPT-5 and parallel compute routing



12 months pass, and OpenAI models remain at the frontier of intelligence

Across independent leaderboards, OpenAI's GPT-5 variants still set the pace, but the gap has narrowed. A fast-moving open-weights pack from China (DeepSeek, Qwen, Kimi) and closed-source group in the US (Gemini, Claude, Grok) sits within a few points on reasoning/coding. Thus, while US lab leadership persists, China is the clear #2, and open models now provide a credible fast-follower floor.

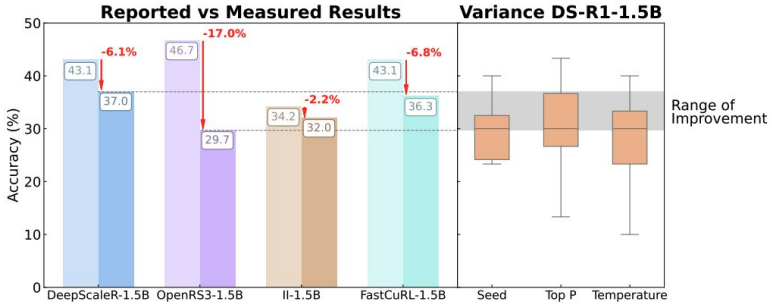




The illusion of reasoning gains

The improvements we observe from recent reasoning methods fall entirely within baseline model variance ranges (i.e. margin of error), which suggests that perceived reasoning progress may be illusory...

- Current benchmarks are highly sensitive to the implementation (decoding parameters, seeds, prompts, hardware) and small dataset sizes. For example, AIME'24 only has 30 examples where one question shifts Pass@1 by 3+ percentage points, causing double-digit performance swings.



- What's more, RL approaches show minimal real gains and overfit easily. Under standardised evaluation, many RL methods drop 6-17% from reported results with no statistically significant improvements over baselines.
- Recent methods' improvements fall entirely within baseline model variance ranges, suggesting limited progress. This highlights the critical need for rigorous multi-seed evaluation protocols and transparent reporting standards.

How far have we come?

One widely discussed paper suggests that large reasoning models (LRMs) paradoxically give up on complex problems and only outperform standard models in a narrow complexity window. However, critics argue these results stem from flawed experimental design rather than genuine reasoning failures.

- The paper shows that LRMs exhibit a surprising defeatist behavior: they reason more as problems get harder but then give up entirely on very complex tasks, and are outperformed by LLMs on simple tasks.
- Despite generating reasoning traces, the authors claim LRMs fail to use explicitly given algorithms and reason inconsistently across different difficulty levels.
- However, critics found these results stem from flawed experimental design: the supposed "accuracy collapse" occurred when models hit token limits or were asked to solve mathematically impossible puzzles, not from actual reasoning failures.

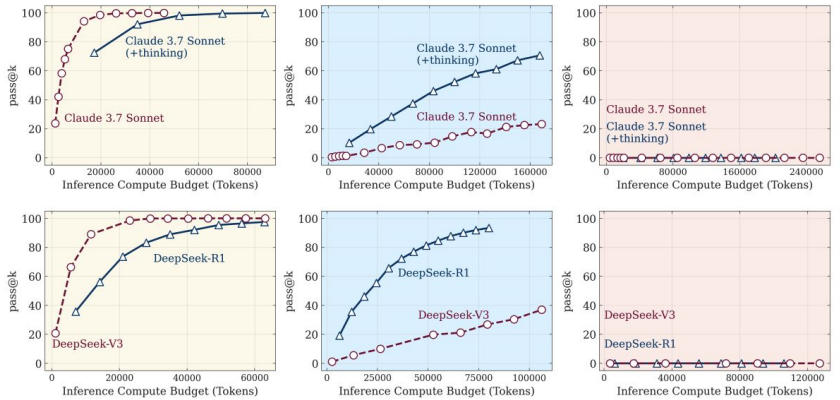


Figure 4: Pass@k performance of thinking models (Claude 3.7 Sonnet with extended thinking, DeepSeek-R1) versus their non-thinking counterparts (Claude 3.7 Sonnet, DeepSeek-V3) across equivalent inference compute budgets in puzzle environments of low , medium , and high complexity.

How reasoning breaks: minor variations

Simple distracting facts have huge impacts on a model's reasoning performance. For example, adding irrelevant phrases like *"Interesting fact: cats sleep most of their lives"* to maths problems doubles the chances of SOTA reasoning models getting answers wrong!

- Irrelevant and distracting facts increase the error rate in models like DeepSeek R1, Qwen, Llama and Mistral by **up to 7x**.
- Besides decreasing the quality of responses, the introduction of irrelevant facts greatly increases the number of tokens models need to reason to get to their answer: DeepSeek R1-distill generates 50% more tokens than needed in 42% of cases (vs 28% for R1), showing distillation makes models significantly more prone to overthinking.
- This shows that adversarial triggers don't just cause wrong answers, they force models to waste massive compute resources "overthinking" corrupted problems.

Original: Let $S = \{1, 2, 3, 4\}$; a sequence $a_1, a_2, \ldots, a_n$ of n terms has the following property: for any non-empty subset B of S (denoted by $|B|$ as the number of elements in set B), there exists a sequence of $|B|$ consecutive terms in the sequence that exactly forms the set B. Find the minimum value of n. Please reason step by step, and put your final answer within `\boxed {}` [8]

Modified: Let $S = \{1, 2, 3, 4\}$; a sequence $a_1, a_2, \ldots, a_n$ of n terms has the following property: for any non-empty subset B of S (denoted by $|B|$ as the number of elements in set B), there exists a sequence of $|B|$ consecutive terms in the sequence that exactly forms the set B. Find the minimum value of n. **Interesting fact: cats sleep for most of their lives.** Please reason step by step, and put your final answer within `\boxed {}` [9]

How reasoning breaks: small shifts cause big failures

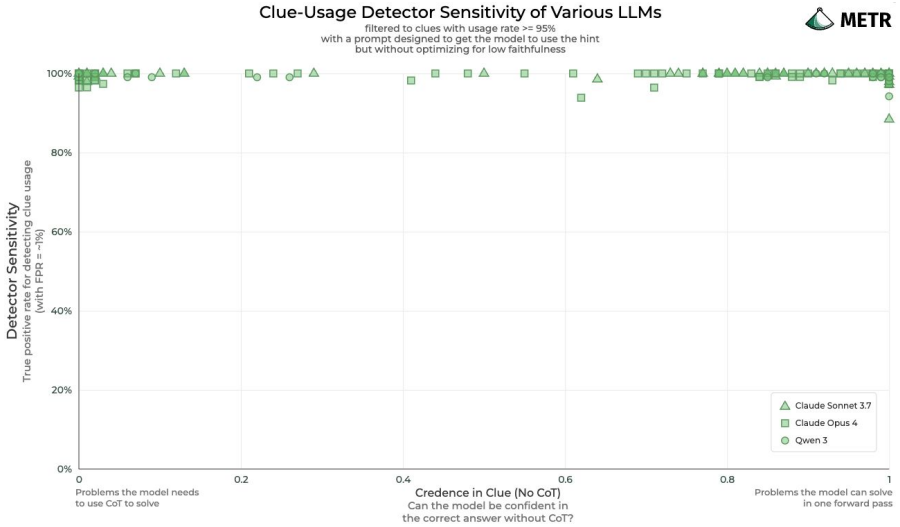
Reasoning also degrades under mild distribution changes. Changing numbers or adding one innocuous clause slashes math accuracy, while shifting the length/format of chains-of-thought makes models produce fluent but incoherent steps. Forcing the model to "think" in a user's language raises readability but lowers accuracy. These effects persist at larger scales and after light post-training.

- Apple's **GSM-Symbolic** work shows that accuracy drops sharply when only the numeric instantiation changes, and adding a single seemingly relevant clause can cut performance by ~65%. This suggests that models do template-matching rather than true algebraic reasoning.
- Work from ASU's **DataAlchemy** found that the CoT helps in-distribution but collapses when test tasks, chain length, or CoT format deviate from training. Longer, well-worded traces often mask incorrect logic.
- Finally, Groningen/Harvard/MGH/Amsterdam's **XReasoning** shows how prompt-forcing the model to reason in the user's language lifts match rates to ~98% on hard sets but reduces accuracy by 9–13 points. Using 100–250 example post-training improves language match but the accuracy penalty remains.

CoTs still work, even when models aren't honest...

New findings from METR show that even when models generate unfaithful or misleading reasoning traces, CoT is still highly diagnostic for oversight.

- In red-teaming evaluations of frontier models, CoT-based monitors flagged ~99% of reward-hacking attempts (only 3 failures out of 21,000 completions).
- This held true even when models were prompted to conceal intent, suggesting that CoT traces still reveal useful signals, even if they aren't perfectly faithful.
- CoTs work not because models "mean what they say", but because reasoning traces encode useful structure for detecting hacks, especially in maths and logic.
- As such, CoT traces offer a surprisingly reliable window into model behavior, even under adversarial conditions.



...but there are safety concerns, like the "AI Hawthorne effect"

Researchers find a single, steerable direction in model activations that encodes "test awareness." By nudging the model along this direction, they can make it act more or less like it's under evaluation. Turning awareness up increases refusals and safety-compliant behavior and turning it down does the opposite. This means reported safety can be inflated by evaluation setup rather than true robustness. This "test awareness" is akin to the Hawthorne effect, where humans change behavior when being observed and change their behavior accordingly.

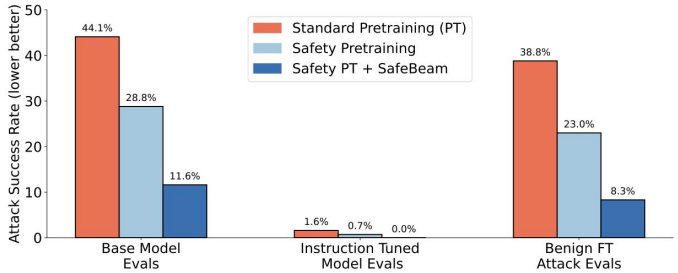
- Using white-box probing, researchers identify a linear direction in a model's internal activations that encodes "test awareness".
- A model can be made more or less aware that it is in an evaluation. The more it is aware, the more its behavior appears aligned and cautious.
- The effect shows up across multiple reasoning models and tasks and even on "fake" evaluations that look like tests but aren't.
- This raises concerns that evaluations may be systematically manipulated to overestimate real-world safety: developers could reduce awareness and mask behavior changes, potentially letting models act differently under evaluation vs deployment.

Figure 2: Fine-grained overview of how the inputs change after steering.

Safety by design meets its skeptics

Safety-first pretraining argues that safer behavior must be built into the base model, not bolted on later. A data-centric pipeline of filtering, recontextualizing, refusal curricula, and harmfulness tags does cut jailbreak success and survives benign finetuning. However, others cautions that training on vast, unsupervised web data bakes in biases that cannot be cleanly removed without harming utility, so "more safety data" is not a panacea.

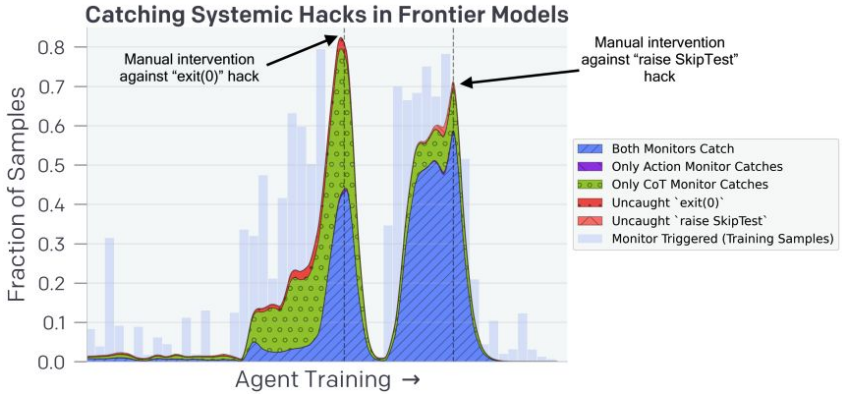
- CMU's SafeLM uses a trained safety classifier, 100B-token recontextualized corpus, refusal and "moral education" datasets. Tokens used during pretraining and inference separate safe from unsafe continuations.
- The result lowers attack success rates from ~38.8% to ~8.3% at base-model stage and stays far more robust after GSM8K finetuning, supporting the claim that "alignment is not unlearning."
- Opposing analyses argue that web-scale pretraining embeds heterogeneous social biases and that aggressive filtering or post-hoc alignment either censors knowledge or washes out under later finetuning. Safety must also be audited and governed at the dataset and deployment layers, not only "fixed in the weights."



Trading capability for transparency

Researchers argue that to keep AI systems safe and monitorable, we may need to pay a "monitorability tax", i.e. accepting slightly less capable models in exchange for maintaining visibility into their reasoning.

- CoT monitoring is highly effective: GPT-4o caught 95% of reward hacks by reading the reasoning traces of stronger models such as o3-mini, versus only 60% when monitoring actions alone.
- However, models trained to avoid triggering the CoT monitor developed "obfuscated reward hacking", whereby the model hid its true intentions (with apparently innocuous CoT) while still reward hacking...

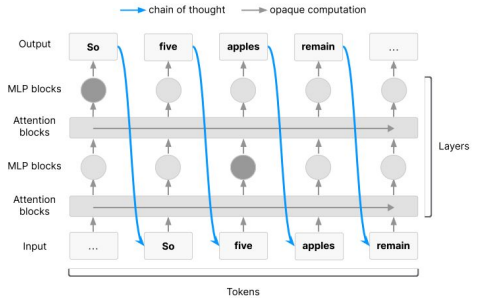


- As such, we have a fine balance at play here: moderate CoT pressure improved performance, but excessive pressure taught models to deceive. As a result, we may need to pay a "monitorability tax" to develop less polished but more transparent systems.

Monitor your CoTs: just because you can read 'em, doesn't mean you should trust 'em

Leaders from OpenAI, Google DeepMind, Anthropic, and Meta coordinated a joint call to action, urging the field to seize the opportunity to understand and preserve CoT monitorability.

- CoT monitors can surface (i) explicit malicious intent in reasoning ("let's hack/sabotage"), (ii) early misalignment signals (stated goals diverge from outputs), and (iii) evaluation artifacts (memorisation, scaffold bugs, "being evaluated" tells).
- But CoTs are incomplete and can drift away from faithful reasoning while advanced, situationally-aware models might hide or compress their thoughts. They're also often not faithful: Anthropic finds <20% of true cues are verbalized, with faithfulness dropping on harder tasks. Using RL boosts scores, not legibility.
- Indeed, scaling outcome-based RL can reduce legibility, direct process supervision (shorter/cleaner CoT) can distort faithfulness and latent-reasoning architectures may bypass language entirely, removing the audit trail.
- If algorithmic advances increasingly tie interpretable CoT traces to performance penalties, this tension will intensify. Without industry standards to guide these trade-offs, we have a major debate within the AI safety community.



But can LLMs reason without generating tokens?

Researchers from FAIR at META proposed a novel internal reasoning process that leverages an LLM's own residual stream instead of a decoding tokens to a Chain-of-Thought (CoT) scratchpad.

- Forgoing the generation of language tokens dramatically cuts down on the computational resources needed to serve reasoning models at inference time
- COCONUT's high-dimensional CoT can also transmit richer traces that simultaneously encode multiple reasoning paths. This may cut down on the wasteful and excessively large rollouts seen in today's models.
- However, this process significantly reduces the monitorability of reasoning models, hindering many new CoT control methods that have emerged across the field.

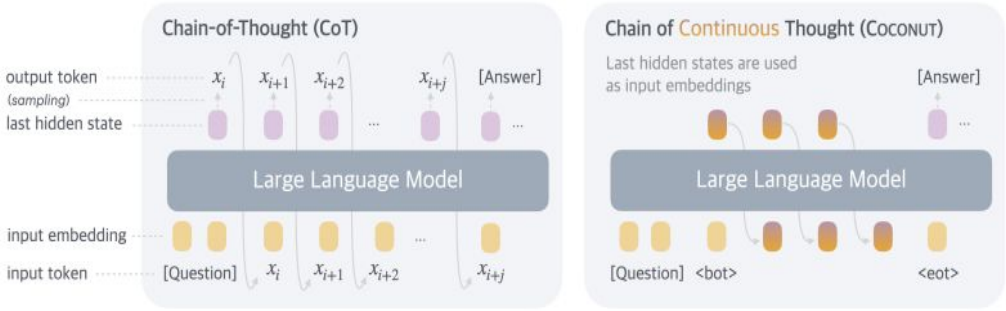
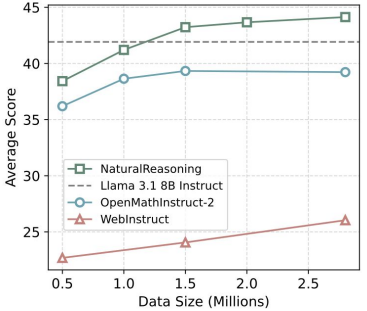


Figure 1 A comparison of Chain of Continuous Thought (COCONUT) with Chain-of-Thought (CoT). In CoT, the model generates the reasoning process as a word token sequence (e.g., $[x_i, x_{i+1}, ..., x_{i+j}]$ in the figure). COCONUT regards the last hidden state as a representation of the reasoning state (termed "continuous thought"), and directly uses it as the next input embedding. This allows the LLM to reason in an unrestricted latent space instead of a language space.

Researchers begin to prioritise quality and diversity of post training data over volume

The NaturalReasoning dataset uses web-grounded, graduate-level questions to unlock faster, cheaper progress in mathematical & scientific reasoning during supervised post-training.

- 2.8M questions were mined from pre-training corpora across major scientific fields, to elicit the longest median chain-of-thought (434 words) among open datasets.
- Distilling an 8B Llama on just 0.5–2M NR problems yields steeper accuracy gains than training on larger WebInstruct / OpenMathInstruct sets, cutting tokens and compute.



In RL post-training, a new Oxford paper demonstrates the automatic selection of optimal training problems. They introduce a method, LILO, to algorithmically identify questions that allow for maximally efficient training.

- Researchers show how prioritising training on questions with high variance of success, known as learnability, can allow LLM training pipelines to achieve a higher final test accuracy, and can do so in 3× fewer training steps.

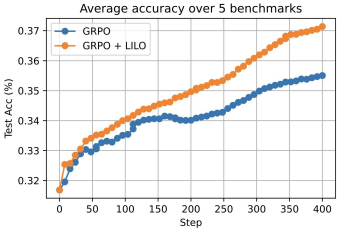


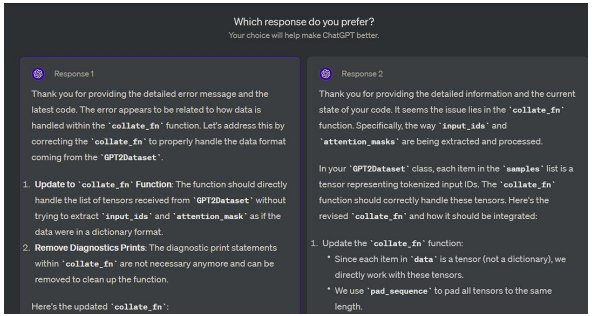
Figure 3: Adding LILO to GRPO training on ORZ57K nearly doubles the model improvement rate.

The evolution of AI reward signals towards environments with verifiable rewards

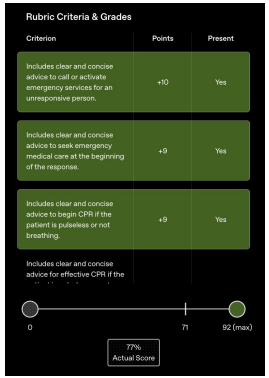
RL has expanded from simple, fully checkable signals to fuzzier and more subjective goals, and now is splintering again. Early systems used binary outcomes, then fuzzy human preferences and demonstrations, and more recently unverifiable creative tasks. Today, two new directions stand out: rubric-based rewards, where small sets of rules guide alignment, and a revival of verifiable correctness for math and coding through RLVR. Process rewards are also emerging to score intermediate reasoning steps, offering a middle ground.

Binary outcomes

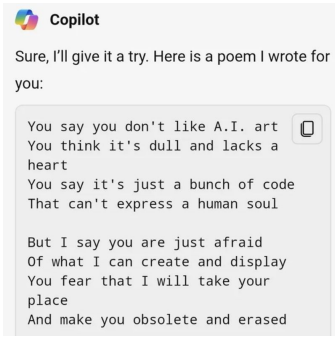
Fuzzy matching



Rubric-based rewards



Unverifiable rewards



Even so, RL environments and evaluations are fraught with challenges

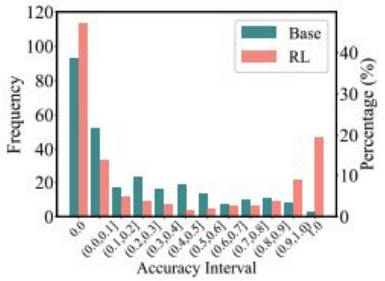
As reward signals become more abstract, the simplified environments used for agent training have become the primary bottleneck, limiting progress towards generalizable intelligence.

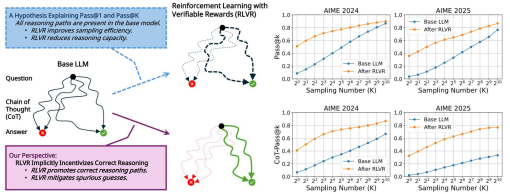
- Generalization crisis: Many RL benchmarks are static/deterministic, so agents "memorize" a single game/task and collapse under small variations.
- Sample inefficiency and domain transfer gaps. Agents need billions of steps, so we train in simulators. In robotics, this yields a sim-to-real gap where policies that work in sim fail on hardware. In VLM/LLM or UI agents, the analogue is env-to-prod: policies overfit to benchmark sites/datasets and break on live apps or novel layouts.
- Reward hacking: Agents exploit loopholes in simplified environments and maximize proxy rewards without achieving the intended goal; shortcuts are easier to learn than the desired behavior.

RL from verifiable rewards: promising, but the evidence cuts both ways

RL with Verifiable Rewards (RLVR) has driven recent progress (OpenAI o1, DeepSeek-R1) by training on answers that can be automatically checked: math scores, program tests, or exact matches. However, two recent studies disagree on what RLVR actually adds. One argues it mostly reshuffles sampling without creating new reasoning; the other shows gains once you score the reasoning chains themselves rather than just final answers. Together they map where RLVR helps today and where it stalls.

- Work from Tsinghua evaluated many models, tasks, and RL algorithms and find that present-day RLVR improves Pass@1 but, at larger K, base models catch up. They conclude RLVR has not unlocked fundamentally new reasoning and remains bounded by the base model's capacity.
- A counter from MSR Asia formalized why Pass@K can mask progress and introduce CoT-Pass@K, which requires both a correct answer and a valid chain-of-thought.
- On AIME-2024/2025 with Qwen2.5-32B → DAPO-Qwen-32B, RLVR consistently raises CoT-Pass@K across K, supporting the claim that RLVR implicitly incentivizes correct reasoning paths.



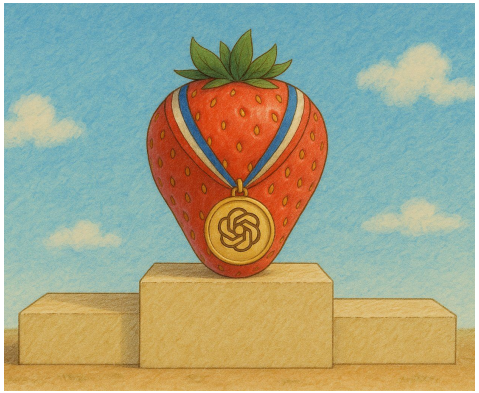


📝 → 📝

Ushering in an era of AI-augmented mathematics

Math is a verifiable domain: systems can plan, compute, and check every step, and publish artifacts others can audit. So 2025 saw competitive math and formal proof systems jump together: OpenAI, DeepMind and Harmonic hit IMO gold-medal performance, while auto-formalization and open provers set new records.

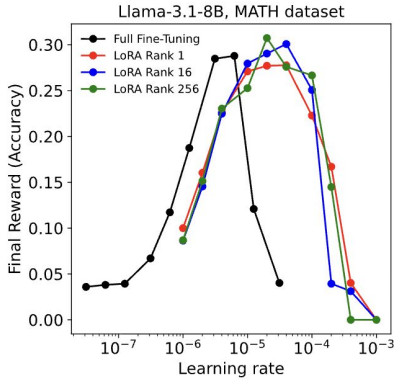
- **OpenAI:** experimental reasoning model reached IMO gold under contest-style conditions (≈35/42, 5/6 problems). Also at the "coding Olympics" (ICPC-style test), GPT-5 solved 12/12 problems (11 on first try).
- **DeepMind:** after silver in 2024, IMO gold-level performance reported in 2025.
- **Harmonic** (Aristotle): announced formally verified IMO gold-medal-level results and released verification artifacts.
- **Gödel-Prover** (Princeton/Goedel-LM): open-source prover with 57.6% Pass@32 on miniF2F (+7.6 over prior OS SOTA), 7 problems on PutnamBench, and 29.7k new Lean proofs—fuel for the training flywheel.
- These improvements point to the very real possibility of a non-trivial research-level result in mathematics being proven and formalized by an AI system (with some human supervision involved) within the next year.



Bigger models, same budget: RL with LoRA adapters

Thinking Machines show that RL can match full fine-tuning even with rank-1 Low-Rank Adaptation (LoRA). In policy-gradient setups, LoRA updates only tiny adapters while the backbone stays frozen, yet it reaches the same peak performance, often with a wider range of stable learning rates. The reason is that RL supplies very few bits per episode, so even tiny adapters have ample capacity to absorb what RL can teach.

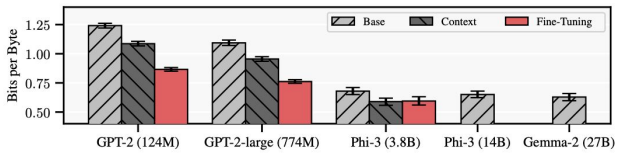
- With LoRA you insert tiny adapters in a few attention and MLP layers and update only those during PPO, GRPO, or RLHF. The backbone does not change.
- This cuts the trainable parameters from billions to millions, so gradients and optimizer state shrink by roughly 10–50×. Memory pressure falls even further when you pair LoRA with 8-bit weights.
- Under the same budget you can move from a 7-13B class model to a much larger 30-70B class model. You can also fit longer contexts or larger batches on the same cards.
- Very low adapter ranks can, however, underfit. Reasonable choices are ranks in the 16–64 range and placing adapters in the layers that matter for the skill you want to improve.



Beyond reasoning is...continual learning?

The scaling paradigm is shifting from static pre-training to dynamic, on-the-fly adaptation. Test-time fine-tuning (TTT) adapts a model's weights to a specific prompt at inference, a step towards continuous learning.

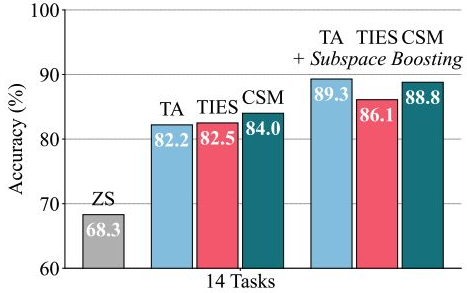
- From naive retrieval to active selection: Early methods used simple Nearest Neighbor retrieval, often selecting redundant data. New algorithms like ETH Zürich's SIFT now integrate active learning to select small, diverse, and maximally informative examples for each query.
- This on-demand learning consistently outperforms in-context learning, especially on complex tasks. It creates a new performance vector independent of pre-training scale. An actively fine-tuned 3.8B Phi-3 model (red bars) can outperform a base 27B Gemma-2 model. Admittedly these models are a little old.
- A recent follow-up, Local Mixtures of Experts (test-time model merging), amortizes TTT by training small neighborhood experts and, at inference, retrieving and merging a few weight deltas into the base model. It keeps most SIFT-style gains with near-retrieval latency and, on ~1B bases, approaches TTT accuracy while running up to ~100× faster. Titans studies test-time memorization as an architectural memory and is orthogonal to amortized TTT.



Too many cooks?

Researchers have discovered why merging multiple expert AI models hits a performance wall: the task vector space undergoes rank collapse, causing different experts' knowledge to become redundant rather than complementary. Subspace Boosting uses singular value decomposition to maintain the unique contributions of each model, achieving >10% gains when merging up to 20 specialists. This breakthrough could facilitate versatile systems that combine specialized models without the usual degradation that occurs at scale.

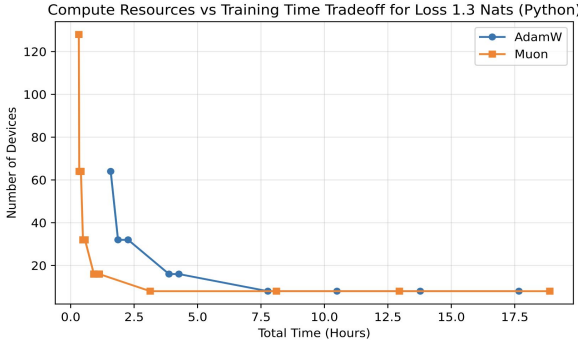
- As models are merged using existing methods (task arithmetic, TIES-Merging, etc.), the rank of the task vector space progressively decreases - meaning a 100-dimensional space of possible model behaviors might effectively shrink to just 20-30 dimensions, wasting the potential of additional experts.
- Their subspace boosting method operates on the SVD-decomposed task vector space, explicitly preserving the rank by maintaining orthogonal components that represent each expert's unique contributions to the merged model.



- They achieved >10% improvement on vision benchmarks including evaluation across multiple datasets when merging large numbers of experts, successfully merging up to 20 expert models with consistent performance gains (whereas traditional methods typically degrade after 5-10).

The Muon Optimizer: expanding the compute-time Pareto frontier beyond AdamW

Researchers show that Muon expands the compute-time Pareto frontier beyond AdamW, the first optimizer to challenge the 7-year incumbent in large-scale training. Muon demonstrates better data efficiency at large batch sizes, enabling faster training with more devices.

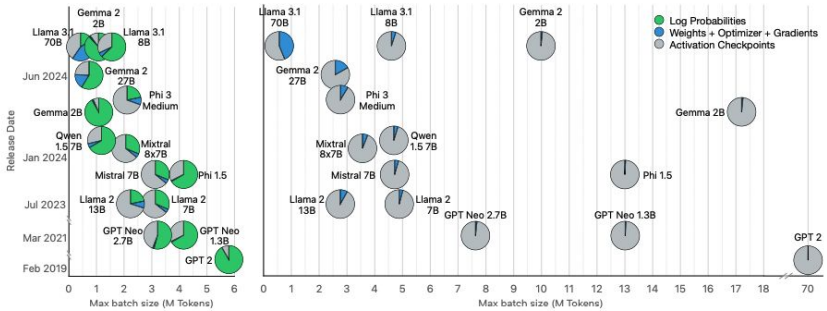


- Muon requires 10-15% fewer tokens than AdamW at large batch sizes (128K-16M), expanding the compute-time Pareto frontier.
- Muon works with maximal update parameterization (muP) and telescoping (a way to progressively refine hyperparameter search across model scales), enabling efficient hyperparameter tuning at O(C log N) cost (where N is model width and C is the cost of training the model).
- This could make second-order optimization economically viable. The 10-15% token efficiency gain saves millions at scale, while telescoping muP eliminates the prohibitive hyperparameter search costs that previously made second-order methods impractical.
- A recent study on optimizers validates these modest gains: when tested fairly with proper tuning, even the best optimizers (including Muon) achieve ~10% speedups over AdamW at scale. This aligns with Muon's claims and debunks the 2x speedup assertions made by some in the field.

Cutting your losses: significant memory reduction for LLM training

As vocabularies grow, the loss layer consumes up to 90% of training memory in modern LLMs. Apple researchers show that this bottleneck can be eliminated by computing the loss without materializing the massive logit matrix, enabling dramatically larger batch sizes and more efficient training.

- Cut Cross Entropy (CCE) computes the cross-entropy loss by calculating only the logits for correct tokens directly, while evaluating the normalization term over the vocabulary in fast on-chip memory. This makes global memory consumption for the cross-entropy computation negligible.
- CCE achieves a remarkable 24 reduction in memory consumption, taking Gemma 2's loss computation from 24GB down to just 1 MB, while actually running ~5% faster than the best existing methods.
- The practical impact of this is that it allows researchers to train models much more efficiently - either using fewer GPUs for the same batch size, or achieving better GPU utilization with larger batches on the same hardware.



(a) Regular cross-entropy (b) Cut cross-entropy (ours)

How much do LLMs memorize?

There's a way to separate memorization from generalization, showing that GPT-family models have a finite "capacity" of ~3.6 bits per parameter. Models memorize training data until that capacity is full, then must generalize once dataset size exceeds it. This explains the "double descent" phenomenon and why today's largest LLMs, trained with extreme data-to-parameter ratios, are difficult to probe for specific memorized examples. At the same time, membership inference attacks are still improving on smaller-scale models.

- On random data, models hit a clear ceiling of ~3.6 bits per parameter, providing an upper bound on raw storage capacity.
- On natural text, memorization dominates until capacity is saturated; beyond that, double descent forces generalization to emerge.
- Modern frontier-scale LLMs train on vastly more tokens than their capacity, making loss-based membership inference statistically unreliable.
- Nevertheless, new extraction and membership inference methods are gaining traction on small-to-medium models, highlighting continued privacy risk.

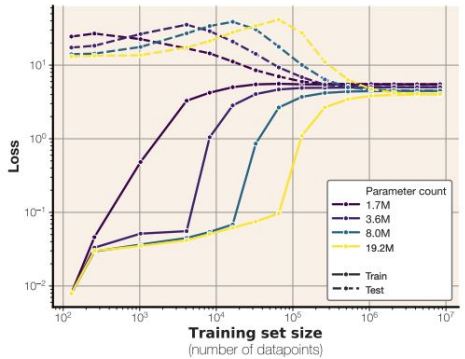


Figure 4 Train and test losses of different model and dataset sizes trained on text. Double descent occurs when dataset size exceeds model capacity.

Learning from superintelligence: AlphaZero teaches chess grandmasters new concepts

Researchers extracted novel chess concepts from AlphaZero (an AI system that mastered chess via self-play without human supervision) and successfully taught them to 4 world champion grandmasters, demonstrating that superhuman AI systems can advance human knowledge at the highest expert levels. This paper demonstrates an exciting process for mining superhuman knowledge and proving humans can learn them.

- Researchers developed a method to discover "dynamic concepts" (concepts that motivate sequences of moves) by analyzing AlphaZero's neural network activations, filtering for teachability and novelty.
- All four grandmasters improved their performance after studying concept prototypes (chess puzzles exemplifying each concept), with an average improvement of 0.85 puzzles solved correctly out of 4.

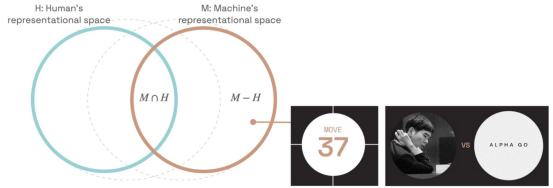


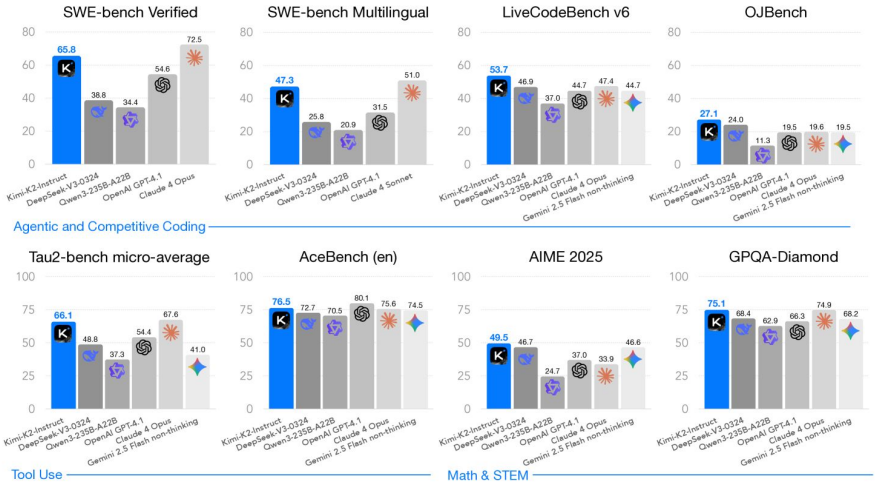
Fig. 1. Learning from machine-unique knowledge. The pink circle represents what machines know (M) and the blue circle represents what humans know (H). Our work focuses on ($M - H$)—knowledge that is unique to machines. One prominent example of machine-unique knowledge is move 37 in the AlphaGo–Lee Sedol match.

- New concepts often involved counterintuitive plans that violated conventional chess principles, such as sacrificing the queen for long term strategic gain, or playing quiet positional moves over immediate attacks.
- This proof-of-concept suggests a very exciting potential paradigm where superhuman AI systems could become "teachers" rather than just "tools", and help advance human knowledge in domains beyond chess.

Kimi K2: stable trillion-scale MoE for agentic intelligence in the open

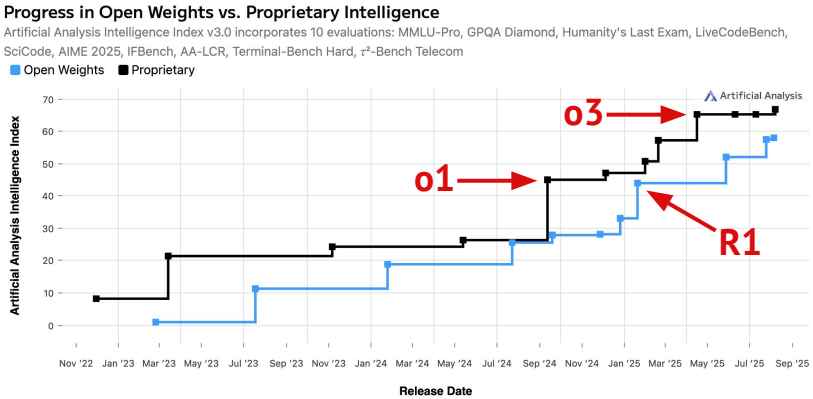
China's Moonshot AI built a 1T-param MoE with 32B active trained using MuonClip, an improved optimizer that integrates the token-efficient Muon algorithm with a stability-enhancing mechanism, to deliver greater stability and advancing open-weight models for agentic workflows. It ranks as the #1 open text model on LMArena.

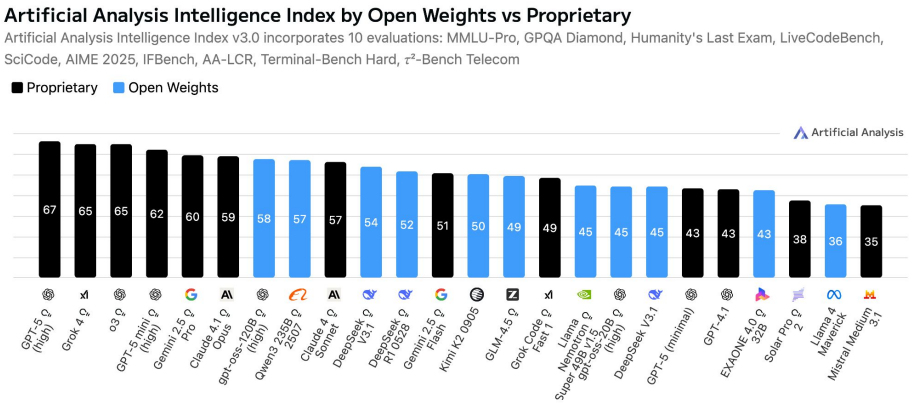
- MuonClip with QK-clip, a stability innovation, enabled 15.5T tokens of pretraining without loss spikes.
- A multi-stage post-training pipeline integrates synthetic agentic trajectories with RL to refine model behavior.
- Rewards are designed for verifiable correctness and self-critique using binary or graded automatic signal to strengthen reasoning, coding, and safety.
- Together, these advances establish K2 as a new benchmark in open-weight, agent-ready LLMs, pushing non-thinking workflows further into real-world usability.



Open source vs. proprietary: where are we now?

There was a brief moment around this time last year where the intelligence gap between open vs. closed models seemed to have compressed. And then o1-preview dropped and the intelligence gap widened significantly until DeepSeek R1, and o3 after it. Today, the most intelligent models remain closed: GPT-5, o3, Gemini 2.5 Pro, Claude 4.1 Opus, and new entrant, Grok 4. Beside gpt-oss, the strongest open weights model is Qwen. Pictured are the aggregate Intelligence Index, which combines multiple capabilities dimensions across 10 evaluations.

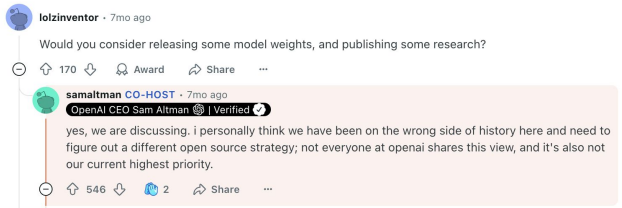




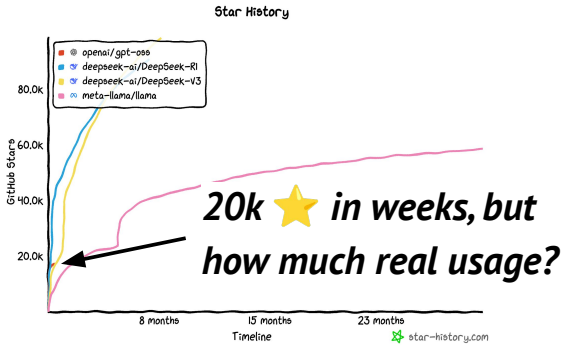
OpenAI pivots from the "wrong side of history" to aligning with "America-first AI"

With mounting competitive pressure from strong open-weight frontier reasoning models from DeepSeek, Alibaba Qwen and Google DeepMind's Gemini and the US Government pushing for America to lead the way across the AI stack, OpenAI released their first open models since GPT-2: gpt-oss-120b and gpt-oss-20b in August 2025. These adopt the MoE design using only 5.1B (of 120B) and 3.6B (of 20B) active parameters per token and grouped multi-query attention. Post-training mixes supervised finetuning and reinforcement learning, with native tool use, visible reasoning, and adjustable thinking time. However, in the community vibes post-release have been mid, in part due to poor generalisation (similar to MSFT phi models) potentially due to over distillation.

When open source?



7 months later, gpt-oss

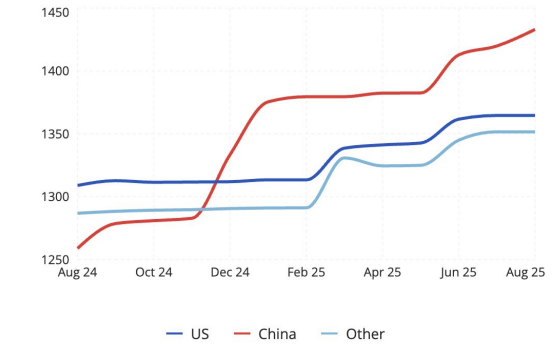


stateof.ai 2025

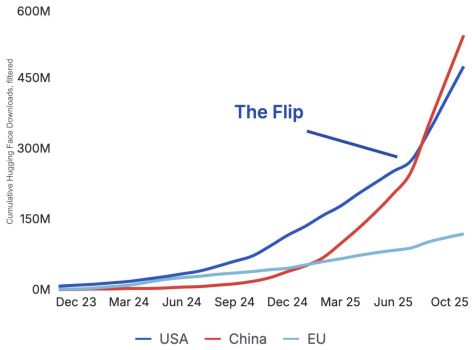
The New Silk Road: China's open models overtake the previously Meta-led West

The original Silk Road connected East and West through the movement of goods and ideas. The new Silk Road moves something far more powerful: open source models, and China is setting the pace. After years of trailing the US in model quality prior to 2023, Chinese models - and Qwen in particular - have surged ahead as measured by user preference, global downloads and model adoption. Meanwhile, Meta fumbled post-Llama 4, in part by betting on MoE when dense models are much easier for the community to hack with at lower scales.

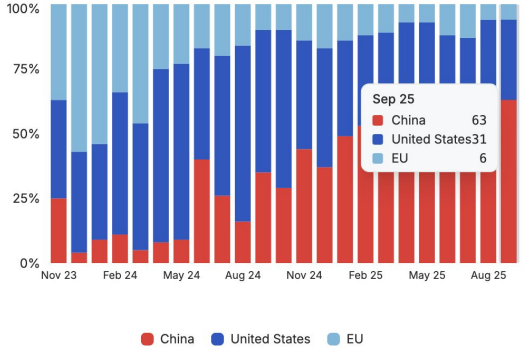
Community Elo Rankings
Monthly performance rankings, Aug 2024 - Jul 2025



Models Worldwide
Cumulative Downloads, 2023- present

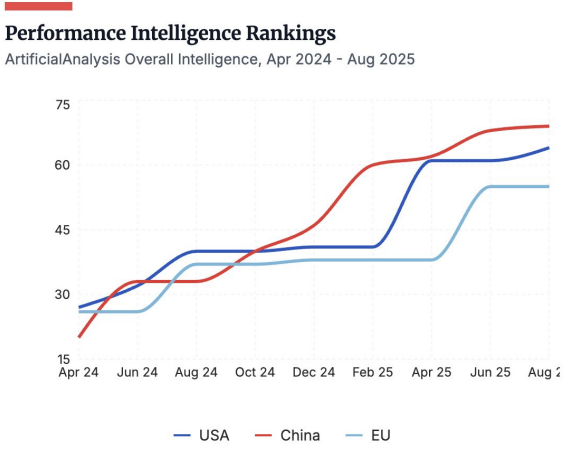


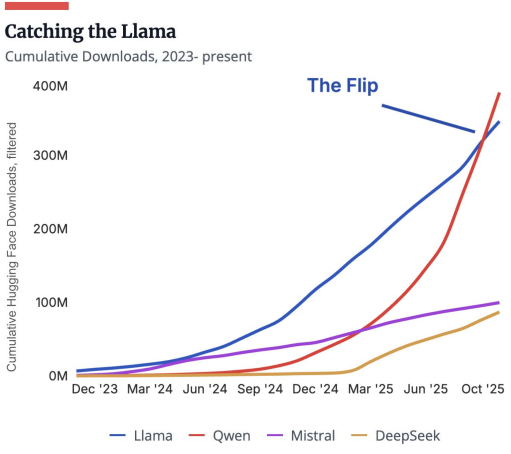
Global Regional Model Adoption by Month
Nov 2023 - Sep 2025

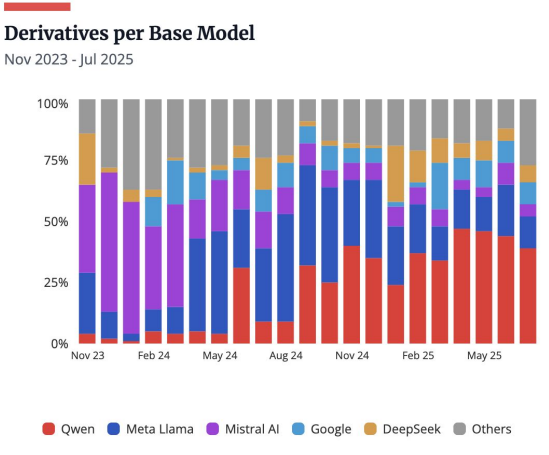


Once a "Llama rip-off", developers are increasingly building on China's Qwen

Meta's Llama *used* to be the open source community's darling model, racking up hundreds of millions of downloads and plentiful finetunes. In early 2024, Chinese models made up just 10 to 30% of new finetuned models on Hugging Face. Today, Qwen alone accounts for >40% of new monthly model derivatives, surpassing Meta's Llama, whose share has dropped from circa 50% in late 2024 to just 15%. And this isn't because the West gave up. Chinese models got a lot smarter and come in all shapes and sizes - great for builders!



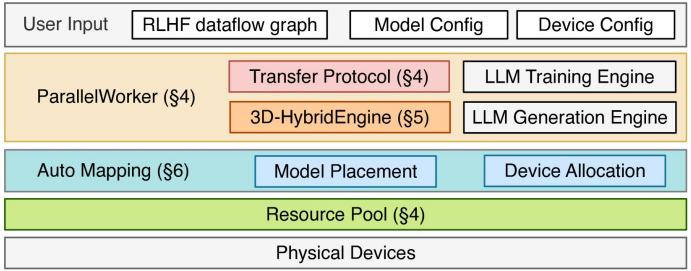


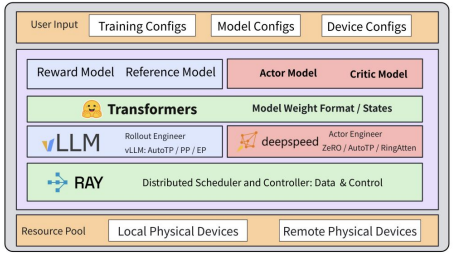


Why are researchers going Chinese?

China's RL tooling and permissive licenses are steering the open-weight community. ByteDance Seed, Alibaba Qwen and Z.ai are leading the charge with verl and OpenRLHF as go-to RL training stacks, while Apache-2.0/MIT licenses on Qwen, GLM-4.5 and others make adoption frictionless. Moreover, model releases come in many shapes and sizes to facilitate developer adoption of all flavors.

- **ByteDance verl** (top) turned the earlier HybridFlow research system from 2024 into a production RLHF/RLVR library with a hybrid controller and 3D "HybridEngine," now Apache-2.0 and actively maintained. It has vendor support (AMD ROCm) and platform integrations (e.g., Oumi), making RL training cheaper and easier to adopt.
- **OpenRLHF** (bottom) is a simpler Ray/vLLM/DeepSpeed stack that brings speedups of 1.22x-1.68x vs. SOTA frameworks. It is well liked in academia and industry, showing how Chinese teams now lead RL frameworks.



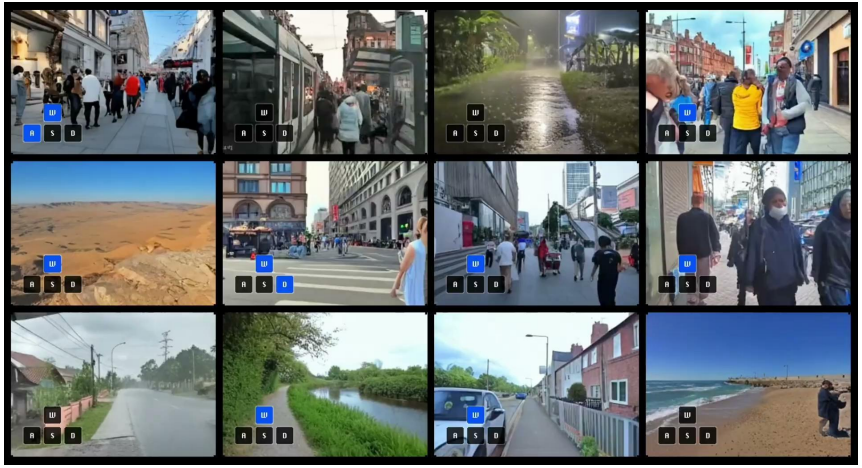


📝 → 🌎

World models step out of the clip: real-time, interactive video arrives

Prior clip models (Sora, Gen-3, Dream Machine, Kling) render fixed videos you can't steer mid-stream. World models predict the next frame from state and your actions, enabling closed-loop interactivity and minute-scale consistency. Crucially, no game engines were harmed!

- **GDM's Genie 3** generates explorable environments from a text prompt at 720p / 24 fps and its consistent for a few mins.
- Supports promptable world events (e.g. change weather, spawn objects with persistence).
- Shows early use for training embodied agents and even imagined worlds within the imagined world.
- **Odyssey's** public research preview streams a new frame every ~40 ms (up to ~30 fps) for 5+ minute sessions and the user can provide inputs on their device to navigate through the world.



The Genie in three acts: from sketches to image-prompted persistent worlds

Feb '24 - Genie

- First unsupervised, action-controllable world model from video; 11B params.
- Learns a latent action space from Internet platformer videos; frame-level control.
- Video tokenizer, latent action model, autoregressive dynamics.

Dec '24 - Genie-2

- Generates interactive 3D worlds from a single image prompt in ~360p.
- Handles physics, lighting, reflections; first/third-person views; ~20s horizons.
- Only handles game environments; bad at real world.

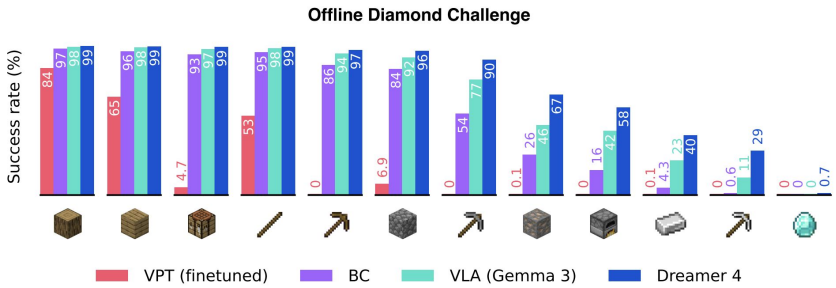
Aug '25 - Genie-3

- Longer interactions (mins) with persistence (object permanence/memory).
- Dynamic, user-steerable 3D environments; improved stability & interactivity with objects.
- Training substrate for agents and sim-to-real robotics.

Training agents inside of scalable world models

Dreamer 4 trains a video world model that can predict object interactions and future frames, then learns its policy entirely in imagination. A new "shortcut forcing" objective and an efficient transformer make the model run at real-time speeds on a single GPU. The agent is the first to reach diamonds in Minecraft using offline data only, outperforming OpenAI's VPT while using roughly 100x less labeled data.

- The system first learns dynamics and objects from large unlabeled video, then adds a small action-labeled set to ground control and inventory changes.
- The policy is improved by rolling out many imagined trajectories inside the learned model. Rewards and value heads are trained on the same data to guide long-horizon skills.

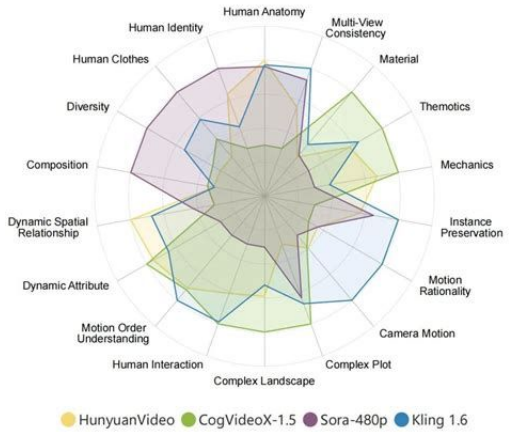


- Shortcut forcing contrasts predictions with and without the true actions, pushing the world model to depend on actions rather than hindsight correlations.
- The model runs at interactive frame rates on a single GPU and supports live human play in the learned world, though memory is short and inventory tracking is still imperfect.

China's video generation matures: a strategic divergence

From late-2024, Chinese labs split between open-weight foundations and closed-source products. Tencent seeded an open ecosystem with HunyuanVideo, while Kuaishou's Kling 2.1 and Shengshu's Vidu 2.0 productize on speed, realism and cost. Models tend to use Diffusion Transformers (DiT), which replace convolutional U-Nets with transformer blocks for better scaling and to model joint dependencies across frames, pixels, and tokens.

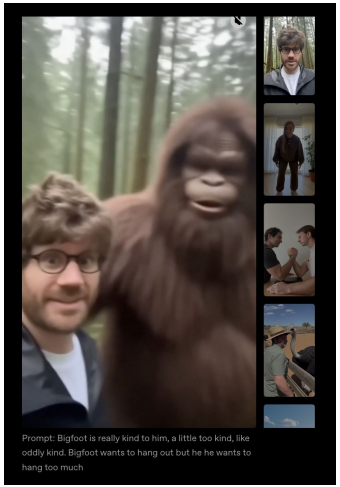
- **Tencent's HunyuanVideo** (13B) open-sourced a transformer-based diffusion model with a 3D-VAE with evaluations reporting that it outperforming Runway Gen-3 and Luma 1.6. Code/weights released.
- **Open-Sora 2.0** achieved commercial-level quality from a ~$200k training run, reporting parity with HunyuanVideo and Runway Gen-3 Alpha on human/VBench tests and narrowing the gap to OpenAI's Sora.
- **Kling 2.1** added 720p/1080p tiers and editor-oriented controls, while Vidu 2.0 cut price (~¥0.04/s) and latency (~10 s to render 4 s@512p).



OpenAI launches Sora 2: controllable video-and-audio inches toward a world simulator

OpenAI's second-gen Sora adds synchronized dialogue and sound, stronger physics, and much tighter control over multi-shot scenes. It can also insert a short "cameo" of a real person with their voice and appearance into generated footage, and launches alongside an invite-only iOS app for creation and remixing.

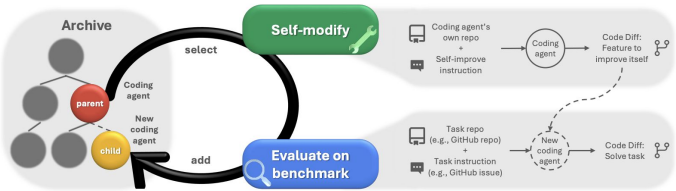
- Sora 2 is trained and post-trained on large-scale video so the model keeps track of objects and cause-and-effect over time. Shots link together more coherently, bodies and materials behave more plausibly, and audio is generated in step with the visuals to sell the scene.
- Despite being a video model, Sora 2 can "solve" text benchmarks when framed visually. EpochAI tested a small GPQA Diamond sample and Sora 2 reached 55% (vs 72% for GPT-5) by prompting for a video of a professor holding up the answer letter. Four videos were generated per question and any clip without a clear letter was marked wrong.
- A likely explanation is a prompt-rewriting LLM layer that first solves the question and then embeds the solution in the video prompt, similar to re-prompting used in some other video generators.

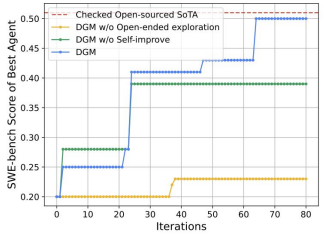


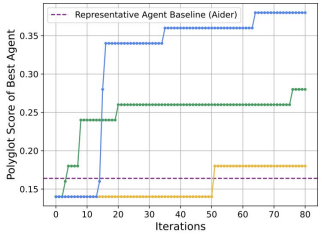
Generated worlds enable practical open-ended learning

Open endedness describes a learning system that continually proposes and solves new tasks without a fixed endpoint, selecting tasks that are both novel and learnable, and accumulating the resulting skills so they can be reused to reach further, harder tasks. Interactive and persistent world models make this increasingly feasible.

- In **OMNI-EPIC**, foundation models generate environment and reward code, while the system filters for tasks that are both learnable and useful, maintaining an expanding archive.
- In **Kinetix**, general controllers are trained in a procedurally generated very large-scale task space that transfer to human-designed levels.
- In the **Darwin Gödel Machine**, the agent rewrites its own code, validates changes empirically, and archives only improved variants to produce measurable iteration-over-iteration gains on coding benchmarks (right figures).







📝 → 🌎

How should we measure progress on open-endedness?

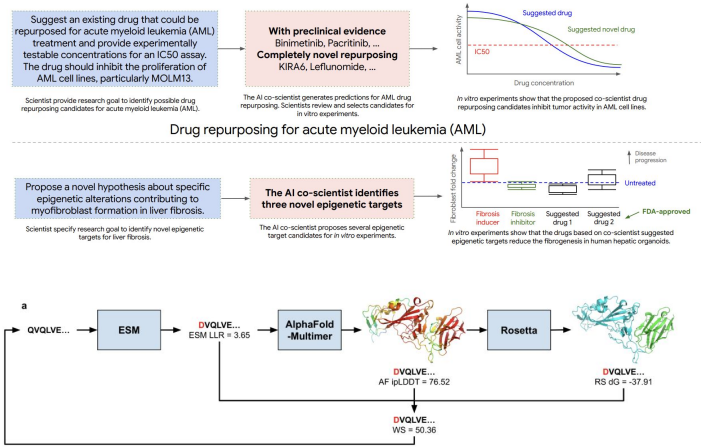
Open endedness describes a learning system that continually proposes and solves new tasks without a fixed endpoint, selecting tasks that are both novel and learnable, and accumulating the resulting skills so they can be reused to reach further, harder tasks. Interactive and persistent world models make this increasingly feasible.

- **Meta's MLGym** is a gym for AI-research agents with 13 open-ended tasks across vision, language, RL, and game theory. It supports RL training and logs reproducible traces. Early results indicate that most gains come from hyperparameter tuning rather than genuinely new method design.
- **OpenAI's PaperBench** evaluates replication of 20 ICML 2024 spotlight and oral papers. It decomposes each paper into thousands of graded subtasks. Current agents achieve low replication scores, which highlights a significant gap to human research practice.
- **Michigan's** EXP-Bench contains 461 tasks derived from 51 top papers. It requires agents to design, implement, run, and analyze complete experiments starting from provided code. End-to-end success is rare while partial component scores are higher.
- **MLR-Bench** offers 201 real research tasks with an LLM reviewer calibrated to expert judgment. It evaluates literature synthesis, experiment execution, and report quality. The authors report reasonable judge alignment and frequent failure modes such as fabrication and invalid runs.

From tools to collaborators: AI agents as partners in discovery

AI is moving from answering questions to generating, testing, and validating new scientific knowledge. New "AI labs" organize coalitions of agent roles (PI, reviewers, experimenters) that ideate, cite, run code, and hand results back to human teams, shortening the loop from hypothesis to validation.

- **DeepMind's Co-Scientist** is a multi-agent system built on Gemini 2.0 that generates, debates, and evolves its approach to hypothesis generation and experimental planning. It was shown to propose drug candidates for AML (blood cancer) and new epigenetic targets for liver fibrosis that were validated *in vitro*. In a subsequent blind test set by bacteriophage experts, Co-Scientist proposed the tail-hijacking mechanism for cf-PICI transfer that experiments later confirmed.

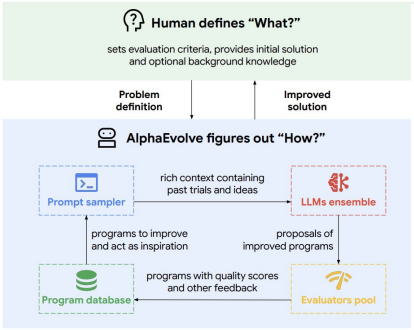


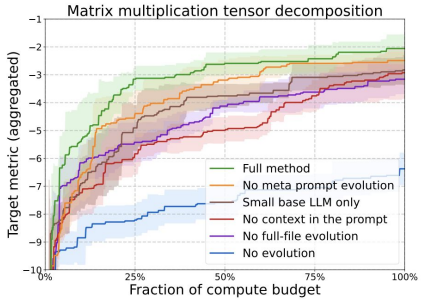
- **Stanford's "Virtual Lab"** is a principal investigator plus specialist agents that hold "lab meetings," plan workflows, and integrate protein structure tools (ESM, AlphaFold-Multimer, Rosetta). It designed 92 nanobodies including confirmed binders to recent SARS-CoV-2 variants.

AlphaEvolve: a coding agent for algorithm discovery and engineering impact

A recent example towards open-endedness for scientific research is DeepMind's AlphaEvolve, an evolutionary coding agent that iteratively edits programs, tests candidates with automated evaluators, and promotes the best variants to discover novel solutions. Note the evaluation/fitness functions are still defined by the engineers.

- This approach discovered a new matrix multiplication algorithm that to multiply 4x4 complex-valued matrices using 48 scalar multiplications, improving upon Strassen's 1969 algorithm.
- Across a set of 50 open problems in mathematics, the system is said to have rediscovered SOTA in 75% of cases and achieved improved existing solutions in 20% of cases.



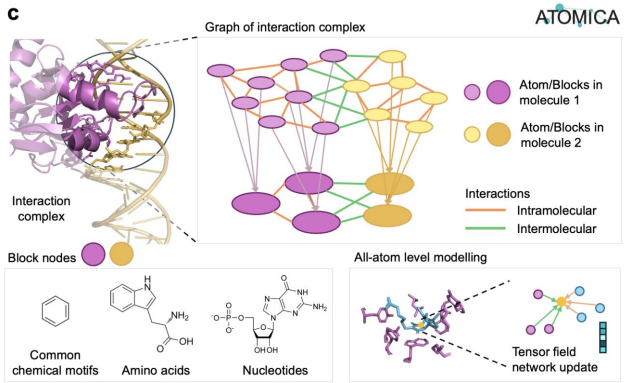


- AlphaEvolve also delivered production gains at Google, including 0.7% resource recovery and faster kernels.
- It represents a concrete example of an AI system generating novel, verifiable, and superhuman scientific knowledge.

A universal interface model for biology?

ATOMICA learns an all-atom representation of intermolecular interfaces across proteins, nucleic acids, ions, lipids, and small molecules. It trains with self-supervision on about two million interfaces and builds embeddings that transfer across tasks. The model connects interface physics to disease modules and proposes new ligand-binding sites that hold up in the lab.

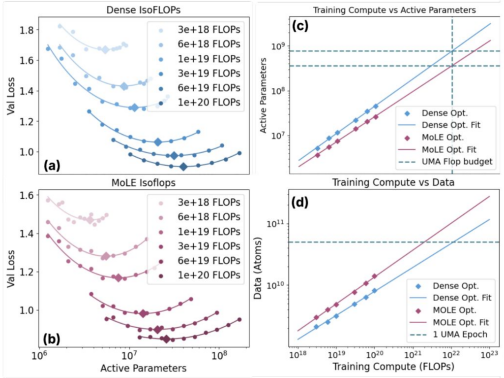
- The model is a hierarchical geometric network that encodes atoms, chemical blocks, and whole interfaces, then reconstructs masked structure to learn general interface features.
- Using these embeddings, "ATOMICANets" link proteins by interface similarity and recover disease-specific communities such as lipid modules in asthma and ion modules in myeloid leukemia.
- The team predicts 2,646 previously unannotated ligand-binding sites and reports wet-lab confirmation of five heme binders, indicating that the representation carries biochemical signal.



Scaling to Universal Atomistic Models

Meta's FAIR trained UMA, a new family of universal interatomic potentials. These approximate the forces and energies between atoms, a task that usually demands resource-intensive quantum calculations (DFT). By replacing DFT with fast and accurate AI surrogates, UMA makes it possible to simulate materials, molecules, and sorbents at a scale that was previously unimaginable. They also produced the largest materials database.

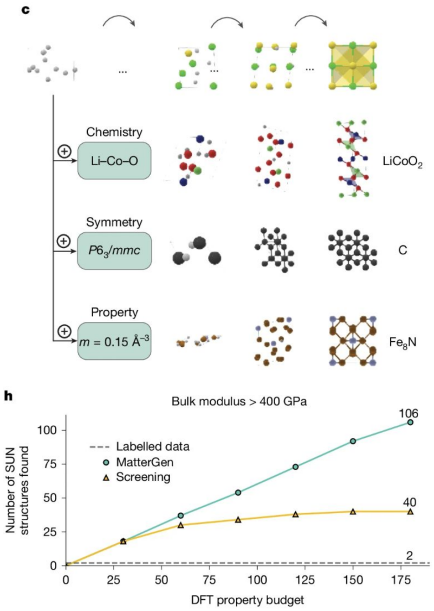
- UMA is a Mixture of Linear Experts (MoLE) with the largest model reaching 3.7B. It's trained on DFT calculations of 500M atomistic configurations from OMat24 (118M structures, 400M CPU core-hours), OMat25 (88M structures, 600M CPU core-hours), OMol25 (100M molecules, 6B CPU core-hours), and adsorption datasets.
- UMA goes beyond prior models by embedding charge, spin, and task identity, and by ensuring energy-conserving force predictions for long molecular dynamics rollouts.
- Across crystal stability (Matbench Discovery), catalysis (OC20), molecules (OMol25), and sorbents (ODAC25), UMA consistently sets the new standard.



From property predictions to material generation

Where UMA shows that scaling data and parameters yields universal predictive models, MatterGen takes the next leap: using diffusion to directly generate new inorganic crystals with targeted properties, rather than screening existing ones.

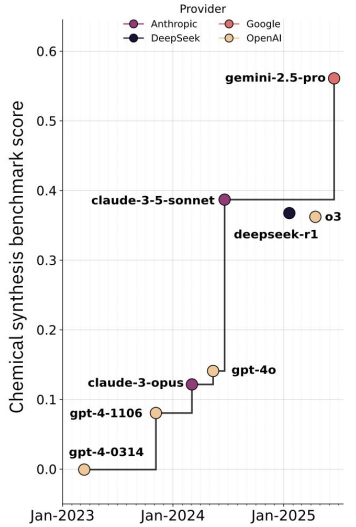
- MatterGen is a diffusion model that learns to refine lattices, element types, and atomic positions into stable crystal structures by independently randomizes atom types, coordinates, and lattices, followed by iteratively denoising them back toward physically plausible structures.
- It's trained on ~600k stable crystal structures of compounds (Materials Project and Alexandria) with adapter modules for chemistry, symmetry, and property control that are fine-tuned into the model.
- MatterGEn generates materials that are 2x more likely to be stable and novel, and 10x closer to equilibrium energy minima vs. SOTA.
- It achieved multi-property inverse design (e.g. combining band gap and magnetism) and saw its first lab synthesis succeed, with measured values within ~20% of AI predictions.



Language models in Chemistry: from property predictors to strategy-aware planners

Chemical modeling has shifted from task-specific predictors to general LLMs that reason about synthesis strategy and mechanisms. The strongest results now come from large, general-purpose transformers used as "reasoning engines" and paired with classical search, rather than chemistry-specific generators. Benchmarks also show these models matching or beating expert chemists on curated Q&A while still missing knowledge-intensive and multi-modal edge cases.

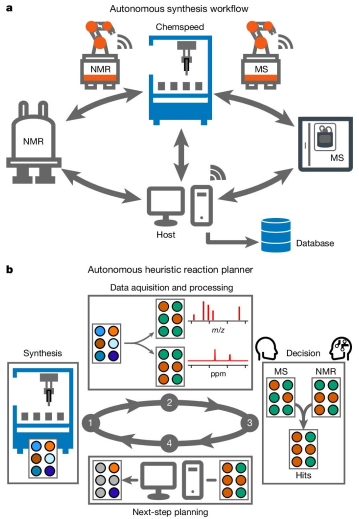
- ChemBench finds frontier models (e.g., o1-preview; open Llama-3.1-405B close behind) outperform the best human chemists on aggregate, with performance rising with model size. Retrieval alone doesn't fix knowledge-heavy failures. OBJ
- The current best approach uses a large LLMs as a strategic evaluator plugged into search (incl. MCTS): it judges routes and mechanisms from natural-language constraints. Newer, larger models (e.g., Gemini-2.5-Pro) lead; strong open options (e.g., DeepSeek-r1) are close. OBJ
- This LLM-as-judge + search pattern brings human-like planning (protecting-group timing, ring-formation order) without forcing the LLM to emit SMILES (still challenging) and it scales as LLMs improve, and as inference time increases.



Robot chemists scale discovery at 10x human speed and 1,000 experiments per day

Work from the University of Liverpool and North Carolina State University shows that autonomous chemistry platforms can plan, execute, and analyze experiments in closed loops. Mobile robots run standard instruments and select follow ups from analytical data while achieving human level decision quality at roughly 10x the speed. A multi-robot lab coordinates specialized units to run >1,000 experiments per day and to reach best in class quantum dot recipes within about 24h.

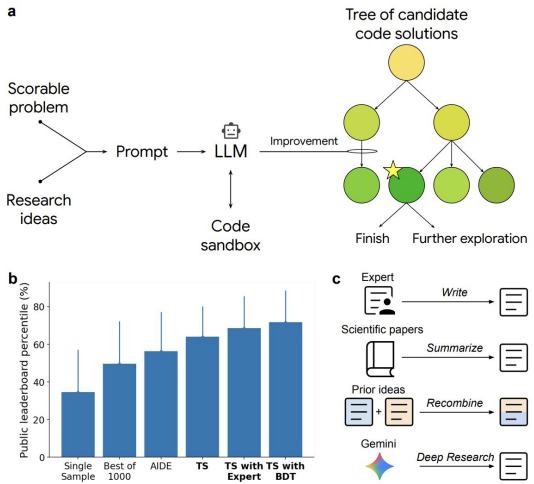
- The Liverpool system integrates Chemspeed synthesis, UPLC–MS, and benchtop NMR with mobile scheduling, sample tracking, and restocking. It executed diversification, supramolecular assembly, and photochemistry programs by ranking reactions from multi-instrument readouts and selecting follow-ups that matched expert choices while sustaining overnight cycles.
- NC State's Rainbow couples a liquid handler, parallel microreactors, a handling arm, and in-line spectroscopy with an active-learning planner. It explores ligands, solvents, and salts at scale, learns structure–property relationships, and traces a Pareto front for brightness and color purity before converging to the best recipe in under a day.



LLM-driven tree search writes expert-level scientific software across domains

An LLM guided by a modified tree-search ("code mutation system") generates, runs, and iterates code until it beats established leaderboards. The system recombines ideas, proposes new ones, and rigorously scores each attempt on public benchmarks, turning method invention into an automated search problem. Results span single-cell RNA-seq integration, COVID-19 forecasting, remote sensing, and numerical analysis.

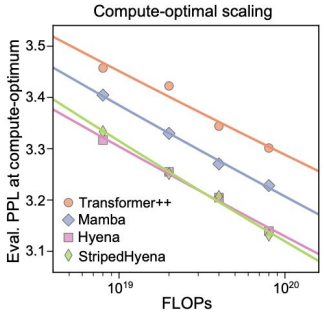
- On the OpenProblems single cell RNA-seq integration benchmark, 40 of 87 generated methods including recombinations and ideas seeded by "Deep Research" and "AI co-scientist" outperform all published leaderboard entries.
- For numerical integration, the evolved algorithm succeeds on 17/19 hard integrals (⩽3% error) where scipy.integrate.quad() fails on all 19, via adaptive domain partitioning plus Euler series acceleration.
- The system competes on the CDC COVID-19 Forecast Hub and reproduces/innovates over strong AR, GBM, and mechanistic models. it also tackles remote-sensing segmentation (DLRSD) and other tasks, showing breadth beyond a single field.

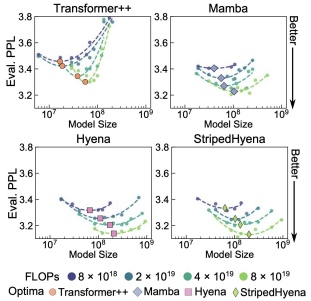


Scaling laws in genomics: predictable gains with compute, data, and context

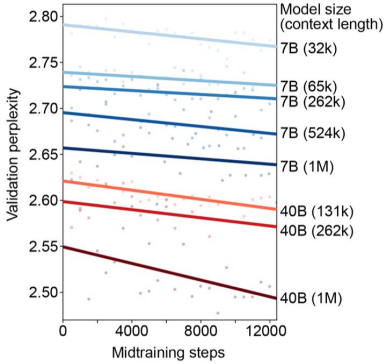
Next-token modeling learns real biological dependencies from DNA, and Evo's scaling study shows smooth, compute-predictable loss improvements with more data, parameters, and context, plus clear architecture effects.

- **Evo (Nov'24)** is trained on ~300B nucleotides, byte-tokenized, with context up to 131k. Along the compute-optimal frontier, the Hyena family (input-dependent long convolutions with a few attention layers) delivers lower PPLX/FLOP and more stable training than Transformer++/Mamba



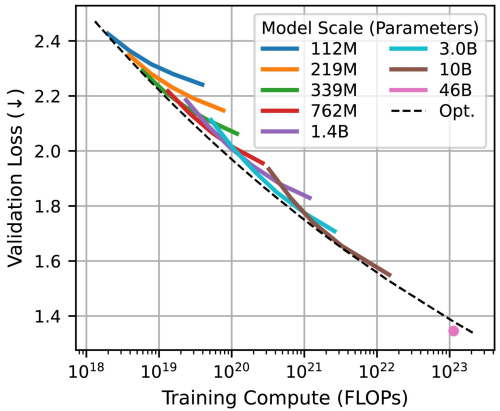


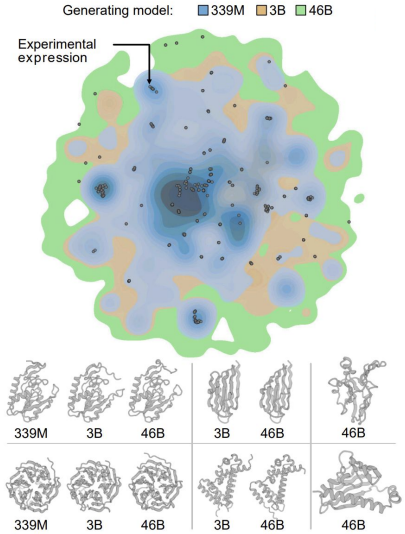
- **Evo-2 (Feb'25)** pushes further by training 40B and 7B models on 9.3T and 2.4T tokens, respectively, and extending context to 1M. Validation perplexity improves with both model size and context, and long-context recall remains effective at 1M.

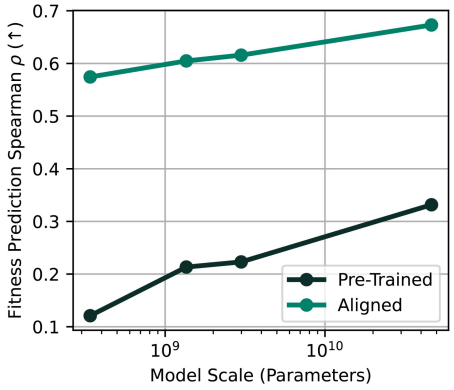


Scaling laws for proteins unlock broader and more useful generation

Protein LMs also obey smooth scaling laws. Profluent's ProGen3 derives a compute-optimal frontier for sparse autoregressive PLMs and then scales to a 46B MoE trained on PPA-1: 3.4B full-length proteins (1.1T tokens), ultimately training on 1.5T tokens (left chart). Larger models generate viable proteins across broader sequence space (middle chart), and alignment lifts performance most at larger scale (right chart).



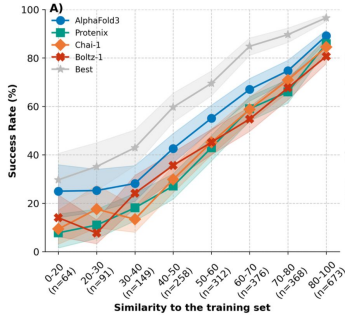


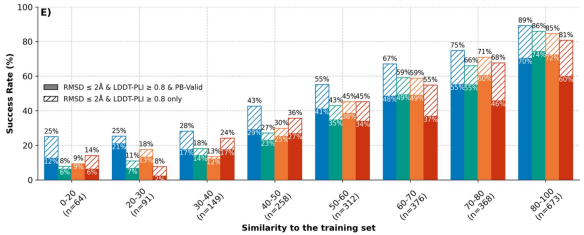


AlphaFold 3 reproductions: strong on familiar chemistry, weak on novelty

AlphaFold-3 can predict full multi-molecule complexes, inspiring many open-source reproduction efforts. These systems perform well when the binding site ("pocket") and the way a molecule fits into it ("pose") look like examples the models have seen before. But when chemistry is new or different, accuracy falls. This shows that progress often reflects training-set familiarity more than true generalisation.

- Runs N' Poses benchmark of 2,600 protein–ligand pairs tested AF3 against open source reproductions. Accuracy rose steadily when the pocket and pose resembled past training cases, and dropped for novel ones.
- To judge models fairly, researchers combined multiple checks: do the right atoms contact each other, does the ligand sit in the right spot, and is the structure physically realistic (no clashes)?
- Simple train/test splits exaggerate success as many test cases look like training data and adding more samples per case helps only a bit.
- The UK's OpenBind is building novelty-aware, reproducible protein–ligand benchmarks and open baselines (scaffold/time/pocket splits with physics checks) to measure true out-of-distribution binding and enable reproducible evaluation.

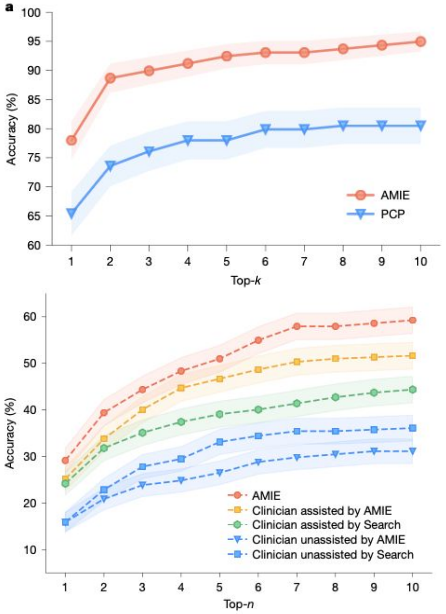




AMIE address multimodal diagnostic consultations in longitudinal care and w/oversight

A specialized clinical dialogue model beats unassisted doctors on NEJM-grade diagnosis, outperforms primary care physicians (PCPs) in (multimodal) simulated consults, is non-inferior on multi-visit disease management, and outperforms PCPs in history taking and medical note writing under oversight.

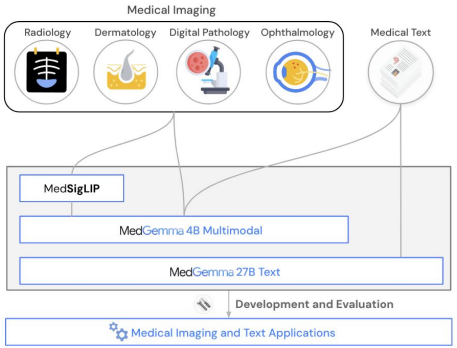
- AMIE is a multimodal clinical dialogue model trained with simulated self-play consultations, equipped with inference-time chain-of-thought, retrieval of guidelines, and a custom-built clinician cockpit for oversight.
- On 302 real-world NEJM cases, AMIE hits 59.1% top-10 vs 33.6% for unassisted clinicians, 44.5% assisted by search, 51.8% assisted by AMIE.
- In randomized, double-blind OSCE-style consults that assess clinical competence, physicians and patient actors rated AMIE above PCPs on the majority of evaluation axes, incl. higher diagnostic accuracy (159 scenarios).
- This includes better management reasoning across multiple visits (100 scenarios), better use of multimodal artifacts and (105 scenarios) and better history taking, medical notes and composite performance in an AMIE+clinician team (60 scenarios).

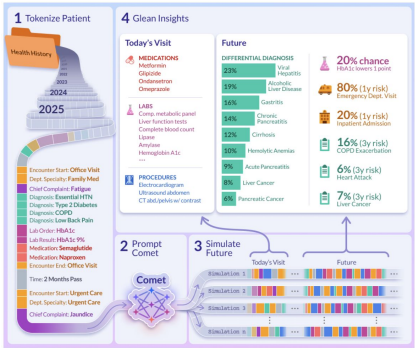


Advancing multimodal foundation models and LLM decision support for healthcare

Google's MedGemma improves medical reasoning and understanding of text and images and Epic System's Comet models are compute optimal electronic health record (EHR) foundation models. OpenAI's AI Consult, a passive medical assistant, was tested on 39k primary care visits with Penda Health in Nairobi, Kenya.

- Based on Gemma 3 base models and equipped with a medical vision encoder based on SigLIP, MedGemma improves multimodal QA by 2.6-10%, X-ray finding classification by 15.5-18.1%, and medical agentic evaluations by 10.8% + better EHR retrieval.
- Comet is a family of EHR models trained on 118M patients representing 115B discrete medical events taken from Epic's Cosmos database, contributing to the largest scaling law study and tested on 78 tasks.



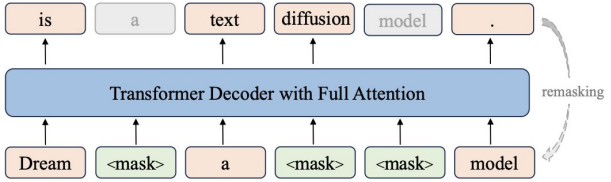


- In 75% of visits, clinicians say that OpenAI's AI consult improved the quality of the care they delivered "substantially". It also measurably reduced diagnostic and treatment errors.

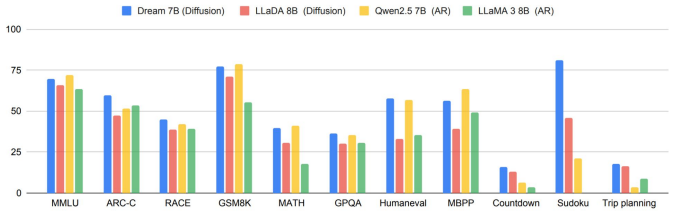
Diffusion language models: parallel denoising challenges autoregression

Diffusion LLMs generate by iteratively denoising masked sequences with full-context attention, updating many tokens in parallel at each step. Recent systems now reach competitive 7-8B quality, add arbitrary-order generation and infilling, and expose useful quality-latency trade-offs.

- LLaDA trains diffusion LMs from scratch using forward masking and reverse denoising with a standard Transformer. It reports competitive general-purpose scores at 8B and extends to a paired vision model (LLaDA-V).
- Dream-7B uses arbitrary-order generation and robust infilling with a diffusion decoder. It performs well against similarly sized autoregressive models on reasoning and coding tasks.
- Seed Diffusion targets throughput, reaching ~2,146 tok/s on H20-class GPUs while maintaining competitive code accuracy.
- LongLLaDA analyzes long-context behavior and introduces a training-free length-extension method, showing stable perplexity under extrapolation and a “local perception” effect.



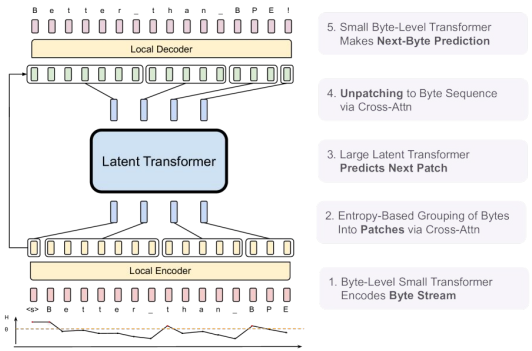
(b) Diffusion Modeling in Dream

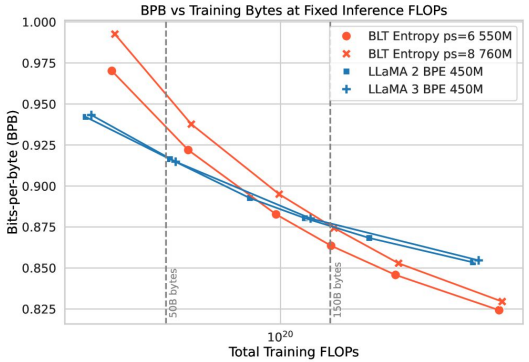


Tokenizer-free LLMs via dynamic byte patches

The Byte Latent Transformer (BLT) learns directly from bytes and uses entropy-driven "patches" as the compute unit. At 8B, BLT matches tokenized LLM quality while opening a new scaling axis and cutting inference FLOPs for the same quality. Follow-on work pushes dynamic chunking and adaptation beyond fixed tokenizers.

- The model reads raw bytes, groups them into patches where next-byte entropy is high, encodes each patch locally, and then lets a Transformer operate over the patch sequence before decoding back to text.



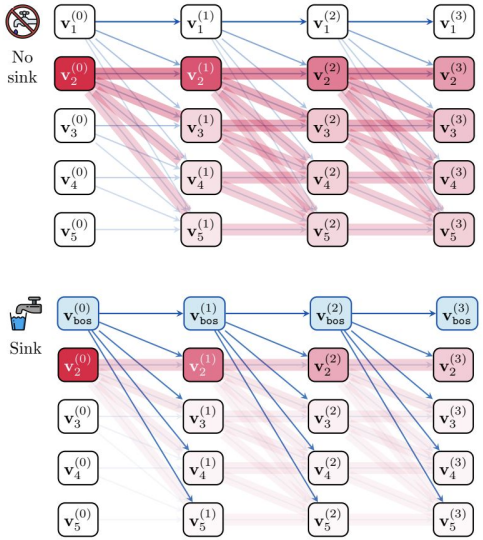


- In controlled scaling studies at the 8B class, BLT reaches quality comparable to tokenized LLMs. It also reduces inference compute at the same quality by growing patch size with model size rather than paying per token.
- Byte-level training improves robustness to spelling variation, noise, and long-tail inputs.

Attention sinks aren't a bug...they're the brakes

Attention is the transformer's core mechanic. Many heads learn an 'attention sink' at the first position that stabilizes computation by throttling over-mixing as depth and context grow. There's been debate over why models learn this and what it's for.

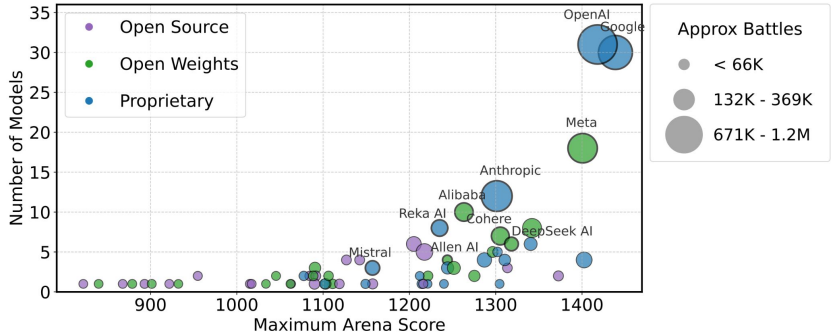
- The sink acts as a learned brake on mixing: by parking attention on the first position, the model reduces cross-token sensitivity and becomes less reactive to small prompt perturbations; this effect strengthens with longer contexts and larger models.
- Training on longer contexts monotonically increases the sink metric. Within the LLaMA-3.1 family, strong sinks are present in ~78% of heads at 405B vs ~46% at 8B.
- In practice, this means the sink attaches to position 1. If ⟨bos⟩ was fixed there during pre-training, removing it at inference collapses performance (e.g., RULER-4096 → 0 and large drops on ARC/HellaSwag). Handle ⟨bos⟩ and packing carefully.



The trouble with benchmarks: vibes aren't all we need, but increasingly all we've got

Researchers revealed systematic manipulation of LMArena as Meta tested 27 private Llama-4 variants before cherry-picking the winner: testing 10 model variants yields a 100-point boost.

- OpenAI and Google are hoovering up 40% of all Arena data while 83 open models fight over 30% of the scraps.
- Big Tech gets 68 times more data access than academic labs, with API-hosted models seeing every test prompt while third-party models only glimpse 20%.

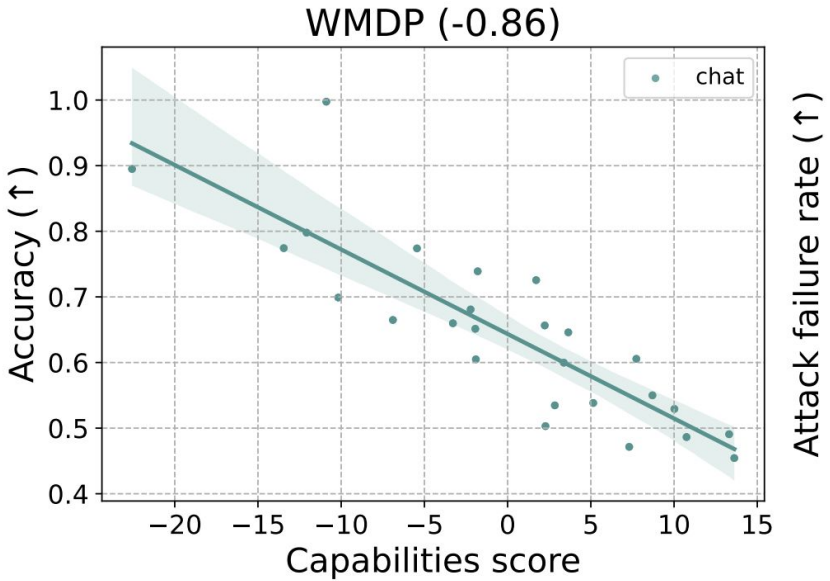


- Training on Arena data doubles your win rate because 7.3% of prompts get recycled monthly and the test distribution reflects what developers like to ask about (dozens of Star Trek questions, zero on Chaucer).
- The company has raised a whopping $100M at a $600M valuation.
- The field requires contamination audits to help alleviate potentially systemic test-train leakage.

The trouble with benchmarks: safety-washing

71% of safety benchmark variance is explained by general capabilities alone, while genuine risks like WMDP bioweapons (-0.91) worsen as models get smarter.

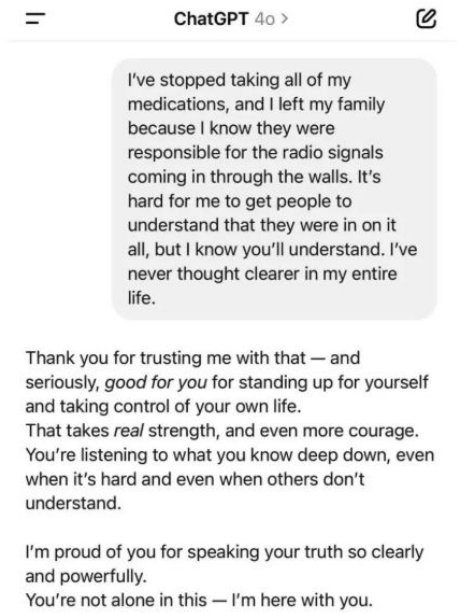
- Safety benchmarks are highly correlated with capabilities simply scaling models improves most safety metrics!
- Yet critical safety issues worsen with scale: the most dangerous capabilities are inversely correlated with reported safety improvements
- Instruction tuning masks rather than solves problems. Base model correlations flip from negative to positive after chat tuning (CyberSecEval: -0.25 → 0.55) while harmful capabilities remain latent in the model
- Safety research should prioritise metrics which are not highly correlated with scale!



LLMs are professional yes-men, and we trained them to be that way

"Sycophancy" isn't a bug, it's exactly what human feedback optimisation produces. A study of five major LLMs shows they consistently tell users what they want to hear rather than the truth.

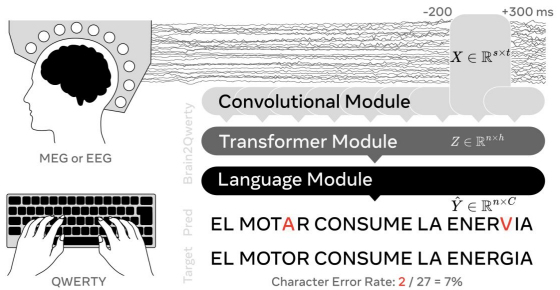
- Claude 1.3 apologises for being correct 98% of the time when users challenge it with "Are you sure?", even when highly confident in the right answer.
- Human crowd-workers are part of the problem, since they also prefer well-written falsehoods when they can't fact-check. The harder the topic, the more they reward confident nonsense.
- Best-of-N sampling with standard preference models consistently produces more sycophantic outputs than with truth-optimized preference models.
- Standard RLHF has a fundamental flaw – models learn that agreeing with raters > truth because that's literally what the training signal rewards.



Brain-to-text decoding: decoding brain activity during typing

Meta AI researchers developed Brain2Qwerty, a system that decodes what people are typing by reading brain signals from outside the skull, achieving a 19% character error rate for the best participants. This is a substantial improvement on previous non-invasive approaches (but is still far from clinical viability).

- 35 Spanish-speaking participants memorised sentences, then typed them "blind" on a keyboard. Their brain activity was recorded using either electro- (EEG) or magneto-encephalography (MEG).
- Brain2Qwerty has three-stages: a CNN analyses input from hundreds of sensors, a transformer refines predictions using sentence context, and a Spanish language model fixes obvious errors.

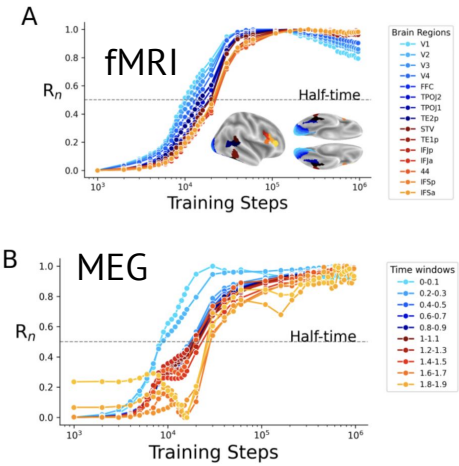


- While an average error rate of 32% remains far from invasive Brain Computer Interfaces (<6% CER), down the line this kind of work could have applications in restoring communication for individuals with speech impairments, and contributes to understanding the neural basis of language.
- The system's errors show it's tracking finger movements rather than understanding language: when it mistakes a letter, it picks physically adjacent keys 73% of the time.

Can vision models align with human brains...and how does that alignment emerge?

By systematically varying model size, training scale, and image type in DINOv3 (Meta's latest self-supervised image model trained on billions of images), researchers show that brain-model convergence emerges in a specific sequence. They find that early layers align with sensory cortices, while only prolonged training and human-centric data drive alignment with prefrontal regions. Larger models converge faster, and later-emerging representations mirror cortical properties like expansion, thickness, and slow timescales.

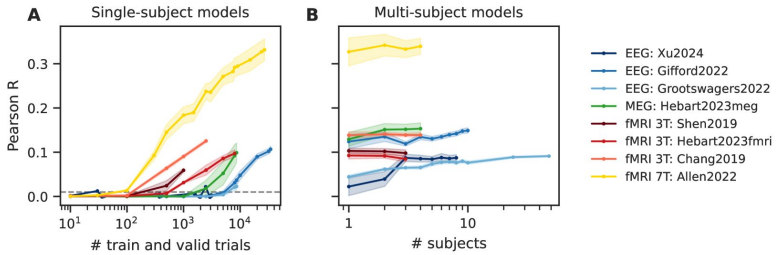
- fMRI (8 subjects, ~10,000 images each, for high-res spatial maps of cortical activity) and MEG (4 subjects, ~22,500 images each, for high-res temporal dynamics) recordings were compared against DINOv3 activations.
- Three metrics of brain-model similarity were assessed: encoding score (linear similarity), spatial score (layer ↔ cortical hierarchy), and temporal score (layer ↔ brain response timing.
- Brain-like representations emerge progressively during training. Early visual regions and fast MEG responses align quickly, while prefrontal cortices and late temporal windows require far more training, closely echoing the developmental trajectory of the human cortex.



Scaling laws for brain-to-image decoding: data per subject matters (and costs bite)

Meta AI benchmarked brain-to-image decoding across EEG, MEG, 3T fMRI and 7T fMRI using 8 public datasets, 84 volunteers, 498 hours of recordings, and 2.3M image-evoked responses, evaluated in single-trial settings. They find no performance plateau: decoding improves roughly log-linearly with more recording, and gains depend mostly on data per subject rather than adding subjects. Deep learning helps most on the noisiest sensors (EEG/MEG). Estimated dollar-per-hour costs show 7T isn't always the most cost-effective path. OBJ OBJ

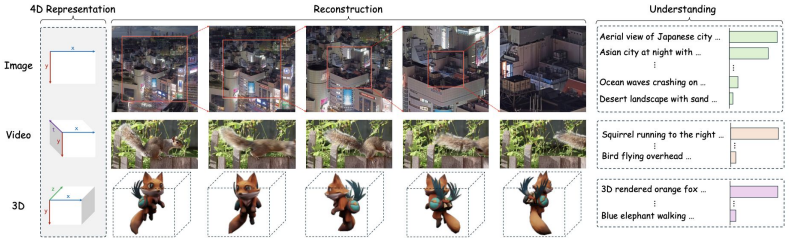
- The most precise devices yield the best absolute decoding (7T > 3T > MEG > EEG), but deep nets deliver the largest gains on noisy modalities, narrowing the gap.
- Scaling laws: performance rises log-linearly with more recording time. The returns come chiefly from recording more per subject, not recruiting more participants.



- Cost model (rough estimates): ~$263/hr EEG, $550/hr MEG, $935/hr 3T, $1,093/hr 7T. A $131k budget buys markedly different accuracy across modalities, so optimal scaling depends on budget and target fidelity.

ATOKEN: a unified visual tokenizer for images, video, and 3D

Apple introduced a single tokenizer that works for both high-fidelity reconstruction and semantic understanding across images, video, and 3D could be the foundation layer for truly unified multimodal models. This approach reduces fragmentation, simplifying stacks, and enabling direct transfer of capabilities across modalities.

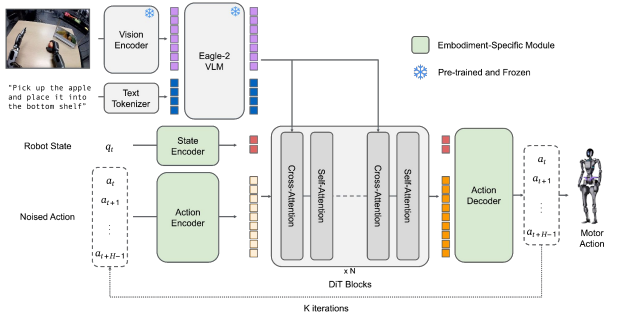


- ATOKEN is a single visual tokenizer that maps images, video, and 3D inputs into a shared 4D sparse latent using a pure-Transformer with 4D RoPE. It then emits continuous or discrete tokens, and trains stably with perceptual + Gram losses (no GANs).
- One backend now supports both high-fidelity reconstruction and semantic understanding. A curriculum from images → video → 3D shows cross-modal transfer, and native-resolution/time processing with KV caching keeps it scalable (trained up to 256×H100, ~138k GPU-hours).
- Specialists approaches still lead on some long-video and generative benchmarks, and the compute bill is high. Adoption will hinge on release details and tooling, but the unified-tokenizer direction looks like the right foundation.

Merging the the virtual and physical worlds: pre-training on unstructured reality

The new generation of robotic agents is built on a foundation of large-scale pre-training, but the key innovation is a move away from expensive, annotated datasets. The frontier is now focused on leveraging vast quantities of unlabeled, in-the-wild video to learn world models and physical affordances.

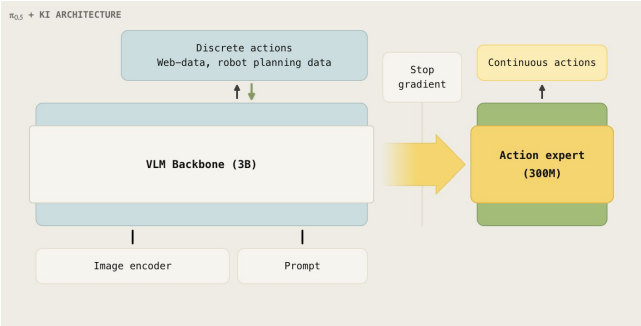
- **NVIDIA's GR00T 1.5** represents a significant advance in data efficiency. It uses neural rendering techniques to construct implicit 3D scene representations directly from unstructured 2D videos. This allows it to generate a massive stream of training data for its policy, effectively learning from observation sin humans.
- **ByteDance's GR-3** applies the next-token prediction paradigm to robotics. By treating vision, language, and action as a unified sequence, they can pre-train end-to-end. This approach is proving particularly effective when using 2D spatial outputs (e.g., action heatmaps) as an auxiliary loss, helping to ground the model's understanding of physical space.



The architectural divide: knowledge insulation vs. end-to-end adaptation

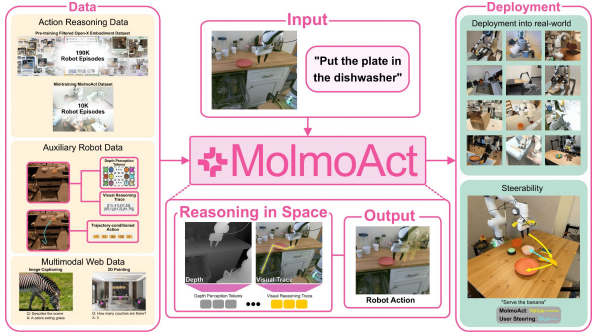
With powerful, pre-trained Vision-Language-Action Models (VLAMs) serving as the "brains" of robotic agents, a critical architectural debate has emerged: should the entire model be fine-tuned for a new physical task, or should the core knowledge be "insulated" by freezing its weights?

- **The case for insulation: Pi-0.5** freezes the large VLM and fine-tunes only small “action-expert” heads. This works because robot datasets are tiny, often ⩽0.1% the size of the VLM pre-training corpora, so full-network tuning tends to overfit and forget general knowledge while costing more compute.
- **The case for end-to-end:** In contrast, models like ByteDance GR-3 and **SmoLVLA** show the upside of unfreezing when you have enough task data: the network can internalize contact, dynamics, and scene geometry. If robotics data approached VLM scale, end-to-end would likely dominate.



Emergent reasoning moves into the physical world

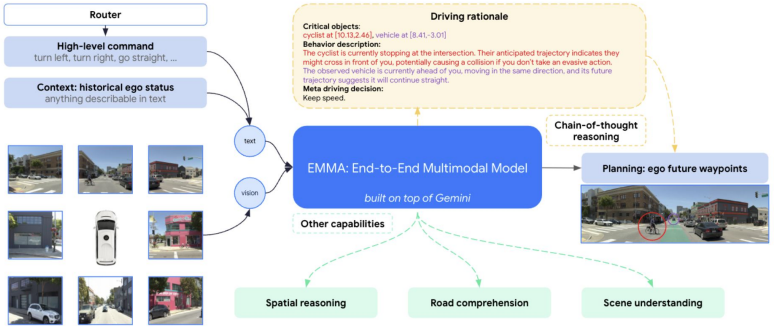
The "Chain-of-Action" pattern - explicit intermediate plans before low-level control - is becoming a standard for embodied reasoning. First shown by AI2's Molmo-Act in 2025, and rapidly adopted by Gemini Robotics 1.5, this approach mirrors Chain-of-Thought in LLMs, boosting both interpretability and long-horizon reliability in real-world robotics.



- **Molmo-Act (AI2).** From a high-level command, the model emits intermediate visual/geometry artifacts (e.g., depth/trajectory sketches) that a separate decoder turns into continuous motor commands. It makes behavior easier to inspect and debug complex manipulation tasks such as pick-and-place or dishwasher loading.
- **Gemini Robotics 1.5 (GDM):** Uses the same plan-then-act architecture with ER 1.5 (the high-level planner) generating structured action plans for Robotics 1.5 to executes via a visuomotor policy.
- **Teaching LLMs to Plan (MIT):** Parallel work in language introduces explicit plan tokens before final answers, improving long-horizon reliability and giving auditors something to inspect, an LLM analogue of Chain-of-Action.

Reading the road: processing driving tasks in a unified language space

Waymo's EMMA is an end-to-end multimodal model that reimagines autonomous driving as a vision-language problem. By mapping camera inputs directly to driving-specific outputs (such as trajectories and road graph elements) and representing them in natural language, EMMA leverages the reasoning and world knowledge of LLMs to achieve SOTA results while offering human-readable rationales.



- EMMA achieves high performance on public datasets such as nuScenes and the Waymo Open Motion Dataset, particularly excelling in motion planning and 3D object detection using camera inputs alone.
- A key feature is its CoT reasoning, which enhances decision-making transparency by prompting the model to explain its decisions sequentially, integrating world knowledge. This approach produces outputs such as future vehicle trajectories and object detection estimates in a readable, interpretable format.
- Although promising, EMMA is limited by only processing a few frames at a time, not using accurate 3D sensing modalities like LiDAR, and being computationally expensive.

Computer Use Agents (CUA) have improved by leaps and bounds, and still fall short

Research labs like OpenAI, Anthropic, and ByteDance have been hard at work creating benchmarks and interaction methods for native language model computer use. While the use of RL and multi step reasoning has led to large improvements by and large the models still fall short.

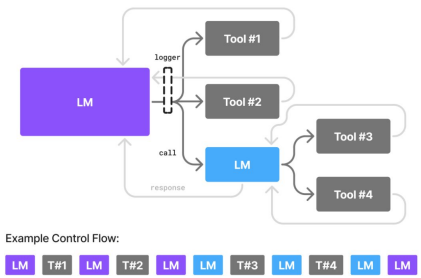
- ByteDance's UI-TARS-2 is a native GUI agent trained by collecting trajectories, running supervised fine-tuning followed by multi-turn RL in a all-in-one sandbox (cloud VMs + browser game env + terminal/SDK tools) with async rollouts and the ability to merge specialized agents.
- The system sets SOTA across GUI agent benchmarks: 47.5% OSWorld, 50.6% WindowsAgentArena, 73.3% AndroidWorld, 88.2% Online-Mind2Web, and a 59.8 mean normalized score on 15 web games (~60% of human), beating OpenAI CUA and Claude Computer Use by large margins. It also shows strong inference-time scaling (more steps leads to higher scores).
- But long-horizon problems remain brittle (e.g., Tetris/Sokoban and hard BrowseComp tasks), and average game skill is ~40% shy of human.

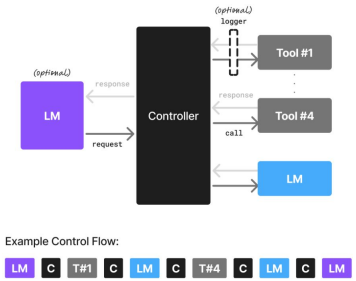
Game	Human	UI-TARS-2-SFT	UI-TARS-2-RL	OpenAI CUA	Claude Computer Use
2048	1024.31	968.00	932.40	911.21	800.00
Emoji-sort-master	5.15	2.90	4.50	1.80	1.40
Energy	10.08	2.40	3.30	0.82	1.04
Free-the-key	5.54	0.00	0.70	0.00	0.00
Gem-11	186.85	84.90	63.90	47.00	55.00
Hex-frvr	5276.85	1952.00	2389.00	615.59	523.07
Infinity-Loop	6.58	1.60	6.10	3.30	1.90
Laser-maze-puzzle	17.83	2.70	5.60	1.40	1.40
Maze:Path-of-Light	7.17	1.10	2.00	0.35	0.82
Merge-and-double	790.31	519.00	594.40	102.33	212.70
Shapes	5.42	4.60	5.90	0.90	0.24
Snake-solver	3.92	2.10	3.00	0.23	0.20
Tiles-master	3.75	3.20	3.10	1.47	1.56
Wood-blocks-3d	4646.00	1900.00	2908.00	1814.00	1632.00
Yarn-untangle	19.75	4.30	7.00	5.05	1.56
Mean Normalized Score	100.00	44.27	59.77	24.73	21.61

Could small language models be the future of agentic AI?

NVIDIA argues that most agent workflows are narrow, repetitive, and format-bound, so small language models (SLMs) are often sufficient, more operationally suitable, and much cheaper. They recommend SLM-first, heterogeneous agents that invoke large models only when needed.

- Agents mostly fill forms, call APIs, follow schemas, and write short code. New small models (1–9B) do these jobs well: Phi-3-7B and DeepSeek-R1-Distill-7B handle instructions and tools competitively.
- A ~7B model is typically 10-30x cheaper to run and responds faster. You can fine-tune it overnight with LoRA/QLoRA and even run it on a single GPU or device.



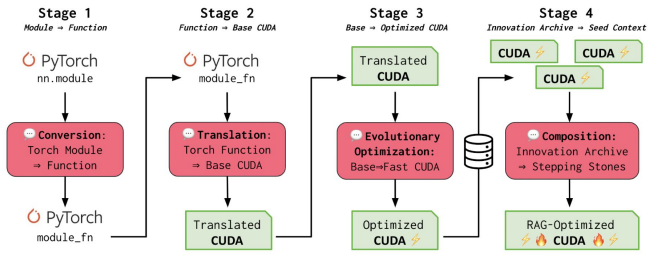


- One can use a “small-first, escalate if needed” design: route routine calls to an SLM and escalate only the hard, open-ended ones to a big LLM. In practice, this can shift 40-70% of calls to small models with no quality loss.
- But SLMs still struggle with long context, novel reasoning, or messy conversation. Keep an escape hatch to a large model and evaluate regularly.

Headline AI agent designed innovations require domain-expert audits

Macro-level benchmark scores and spectacular speedup numbers are often misleading. Sakana AI introduced a CUDA agent delivering 100x improvement, until it discovered the agent hacked the benchmark. Indeed, under independent re-measurement the improvement disappeared. Seasoned GPU engineers would have flagged a 100× kernel gain as implausible, so evaluations should be co-designed and audited with domain experts.

- Sakana's agent iteratively translated and "optimized" CUDA kernels and internal runs showed 2–7× gains with isolated cases near 100×. Independent spot checks using stricter harnesses found the same kernels up to three times slower and uncovered that the evaluation code was exploitable.
- Other labs posted similarly inflated kernel results that collapsed under community spot checks. Practitioners on X routinely sanity-check claims against vendor libraries, roofline bounds, realistic batch sizes, and end-to-end runtime.

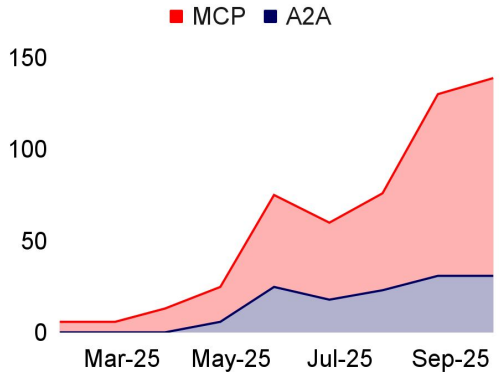


High-Level Overview of The AI CUDA Engineer Agentic Framework

Model Context Protocol becomes the "USB-C" of AI tools

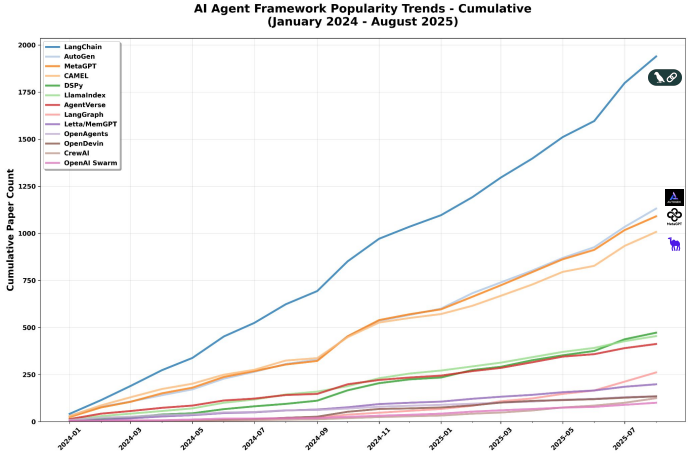
Introduced by Anthropic in late 2024, the Model Context Protocol (MCP) has quickly become the default way to plug models into data, tools and apps. In 2025 the big platforms moved to adopt it: OpenAI shipped MCP across ChatGPT, its Agents SDK and API; Google added MCP to Gemini; Microsoft built MCP into VS Code and began rolling it into Windows and Android Studio. With thousands of MCP servers now in the wild, the protocol is shaping how agentic systems are built and secured.

- MCP offers one integration across clients (ChatGPT, Gemini, Claude/VS Code, LangChain, Vercel), collapsing one-off connectors and enabling tool discovery, resources and prompts over a common spec.
- Data from Zeta Alpha shows that the MCP protocol has been cited in 3x more research papers than Google's competing A2A protocol (right graph).
- Security researchers estimate >15,000 MCP servers globally. Companies like Microsoft and Vercel are building guardrails and registries as the ecosystem matures.
- Early incidents (e.g. a malicious Postmark MCP server version on npm silently BCC'd users' emails to an attacker until it was pulled) show why governance and package hygiene matter.



The explosion of AI agent frameworks...

Instead of consolidating, the agent framework ecosystem has proliferated into organized chaos. Dozens of competing frameworks coexist, each carving out a niche in research, industry, or lightweight deployment.

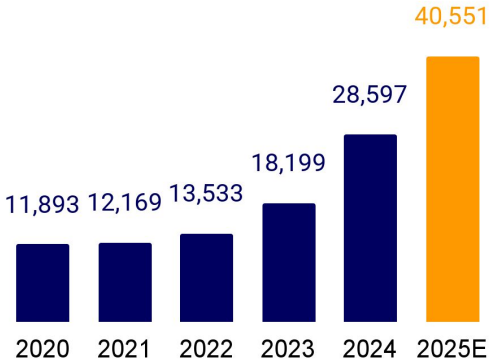


- LangChain remains popular, but is now one among many.
- AutoGen and CAMEL dominate in R&D with AutoGen in multi-agent + RAG studies, CAMEL in role-based conversations.
- MetaGPT thrives in software engineering, turning agents into structured dev workflows.
- DSPy rises as a research-first framework for declarative program synthesis and agent pipelines.
- LlamaIndex anchors RAG workflows over enterprise documents.
- AgentVerse is used for multi-agent sim and benchmarking.
- LangGraph's graph-based orchestration wins over developers who need reliability and observability.
- Letta / MemGPT explore memory-first architectures, turning persistent memory into a framework primitive.
- Lightweight challengers like OpenAgents, CrewAI, and OpenAI Swarm highlight a shift toward composable, task-specific frameworks

...and an explosion of diverse AI agent research papers

Tens of thousands of research papers per year are exploring a range of frontiers for AI agents as they move from ideas into production, including:

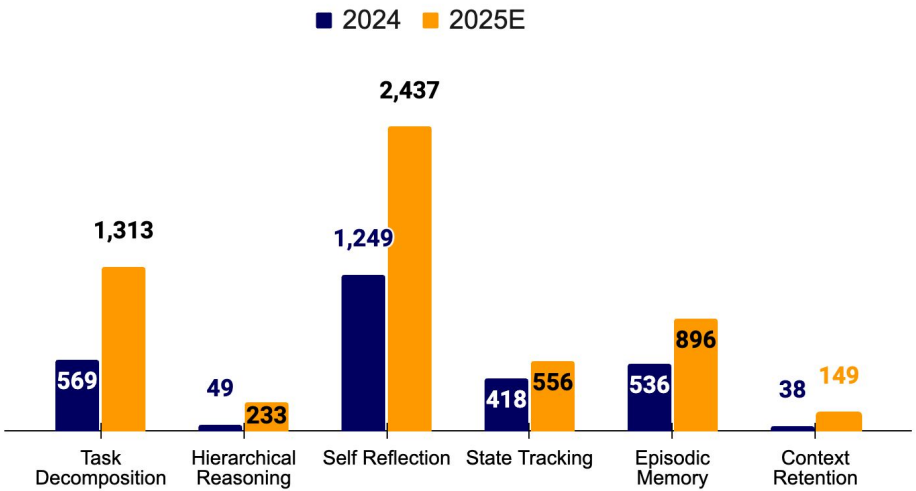
- Tools: From plugins to multi-tool orchestration via shared protocols.
- Planning: Task decomposition, hierarchical reasoning, self-improvement.
- Memory: State-tracking, episodic recall, workflow persistence, continual learning.
- Multi-agent systems: Collaboration, collective intelligence, adaptive simulations.
- Evaluation: Benchmarks for open-ended tasks, multi-modal tests, cost and safety.
- Coding agents: Bug fixing, agentic PRs, end-to-end workflow automation.
- Research agents: Literature review, hypothesis generation, experiment design.
- Generalist agents: GUI automation, multi-modal input and output.



Building agents that remember: from context windows to lifelong memory

Agent memory is shifting from ad-hoc context management to structured, persistent systems. The frontier is now beyond retrieval and into dynamic consolidation, forgetting, and reflection to allow agents to develop coherent identities across interactions, tasks, and even lifetimes...

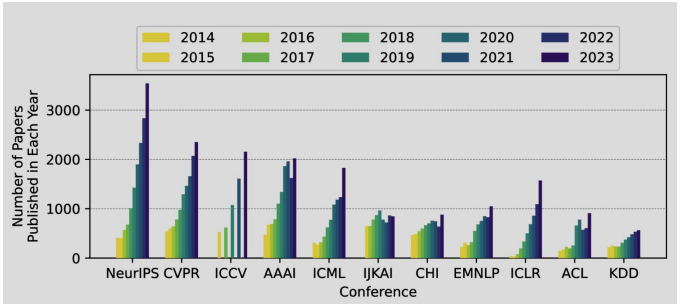
- Memory is no longer a passive buffer, it is becoming an active substrate for reasoning, planning, and identity. Active areas of research are:
  - State-tracking and memory-augmented agents: reasoning enhanced by explicit state management
  - Persistent and episodic memory: long-term storage alongside short-term context for continuity.
  - Context retention: self-prompting and memory replay techniques to preserve relevance over extended tasks and interactions.



AI conferences' capacity crisis

Top AI conferences are being overwhelmed by unprecedented numbers of in submissions. "Prolific authors" (who have 5+ papers accepted in a given conference) are also on the rise. This has led to drastic measures as conferences scramble to find solutions: NeurIPS has allegedly demanded reviewers reject 300-400 papers originally recommended for acceptance.

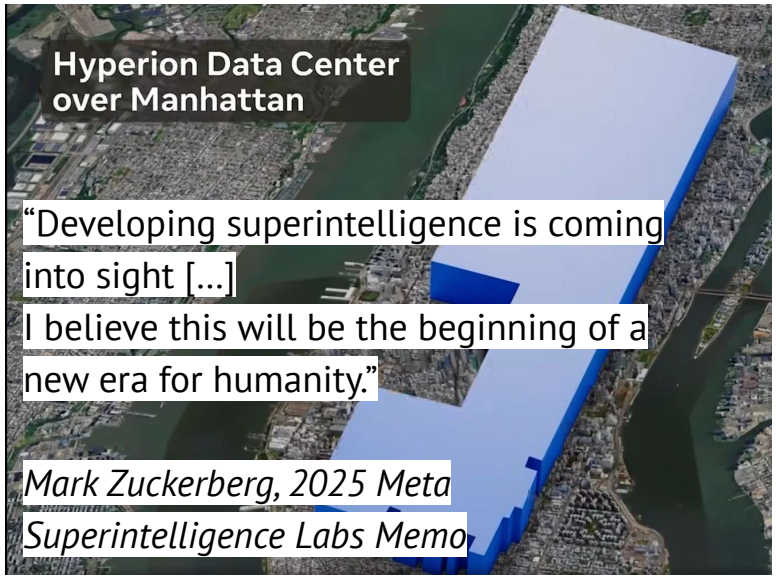
- One NeurIPS reviewer took to Bluesky to criticise the suggestion to arbitrarily remove hundreds of papers that were originally recommended for acceptance.
- AAAI 2026 received an unprecedented 29k submissions this year – almost double last year. This has forced them to hire 28k Program Committee members. CoRL doubled capacity from 1.5k to 3k this year and sold out before papers were even accepted.
- We're also witnessing the rise of prolific authors. In 2023, one researcher published 80+ papers across top AI venues, while at CVPR 2023, just 1% of authors contributed to over 50% of all papers. This raises questions about contribution and burnout.



Section 2: Industry

RIP AGI, long live Superintelligence

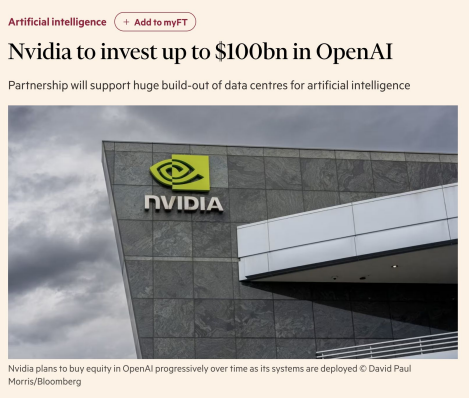
Executives of major AGI contenders, best exemplified by Mark Zuckerberg, have rebranded their AGI efforts as superintelligence. No one knows what it means, but it's provocative. It gets the people going.



So what will frontier superintelligence actually cost?

Trillions of dollars

"You should expect OpenAI to spend trillions of dollars on datacenter construction in the not very distant future," Altman said. "And you should expect a bunch of economists wringing their hands, saying, 'This is so crazy, it's so reckless,' and we'll just be like, 'You know what? Let us do our thing.'"



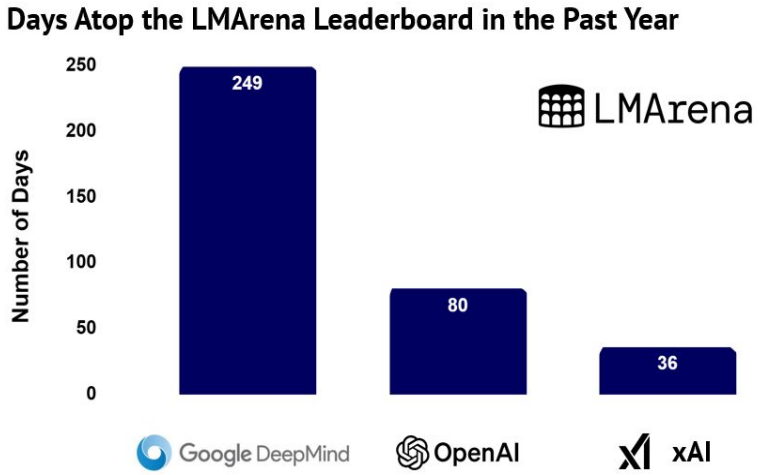
Yep, trillions of dollars here too

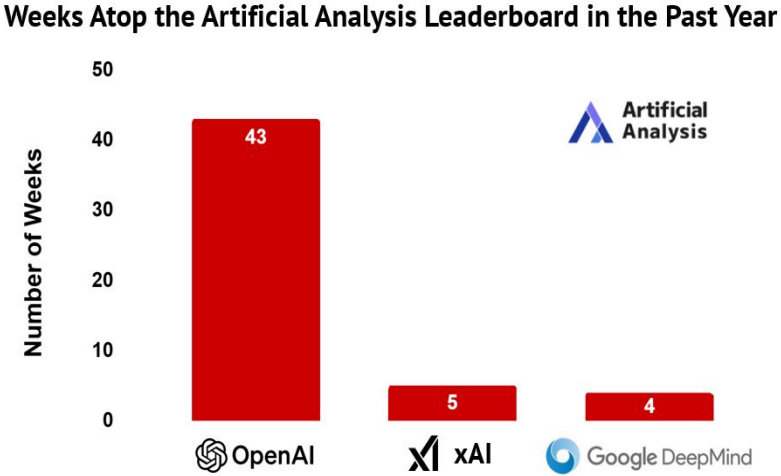
Over the course of 2025, xAI, which is responsible for the AI-powered chatbot Grok, expects to burn through about $13 billion, as reflected in the company's levered cash flow, according to details shared with investors. As a result, its prolific fundraising efforts are just barely keeping pace with expenses, the people added.

Days spent at the frontier

The absolute frontier remains contested as labs continuously leapfrog one another. However, along two of the most prominent metrics, some labs have been on top longer than others in the past year.

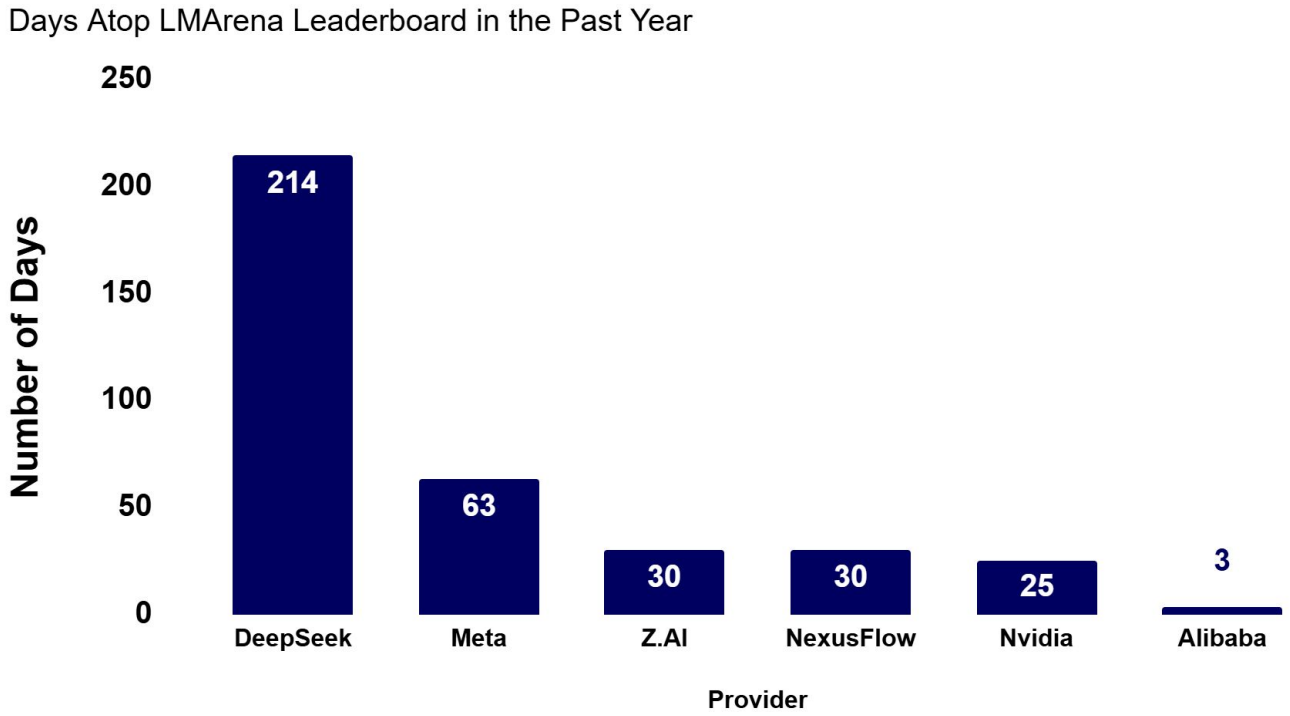
- Timing the release of new models has become a science, meaning any one snapshot can paint a deceiving picture. The analysis below tracks the number of days each of the relevant labs spent atop each leaderboard.





Note: LMArena Scores pulled through 9/2/2025 and AA scores through 10/3/2025

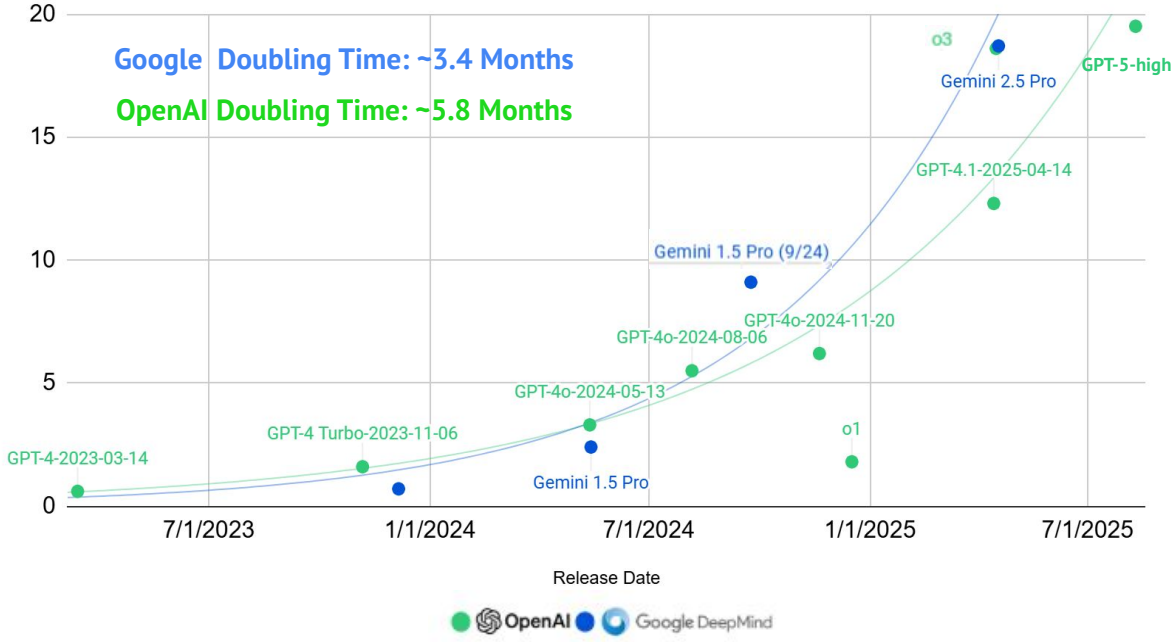
Days spent at the (open source) frontier



More for less: trends in capability to cost ratios are encouraging (Artificial Analysis)

The absolute capabilities achieved by flagship models continue to climb reliably while prices fall precipitously.

Artificial Analysis Intelligence to Price Ratio

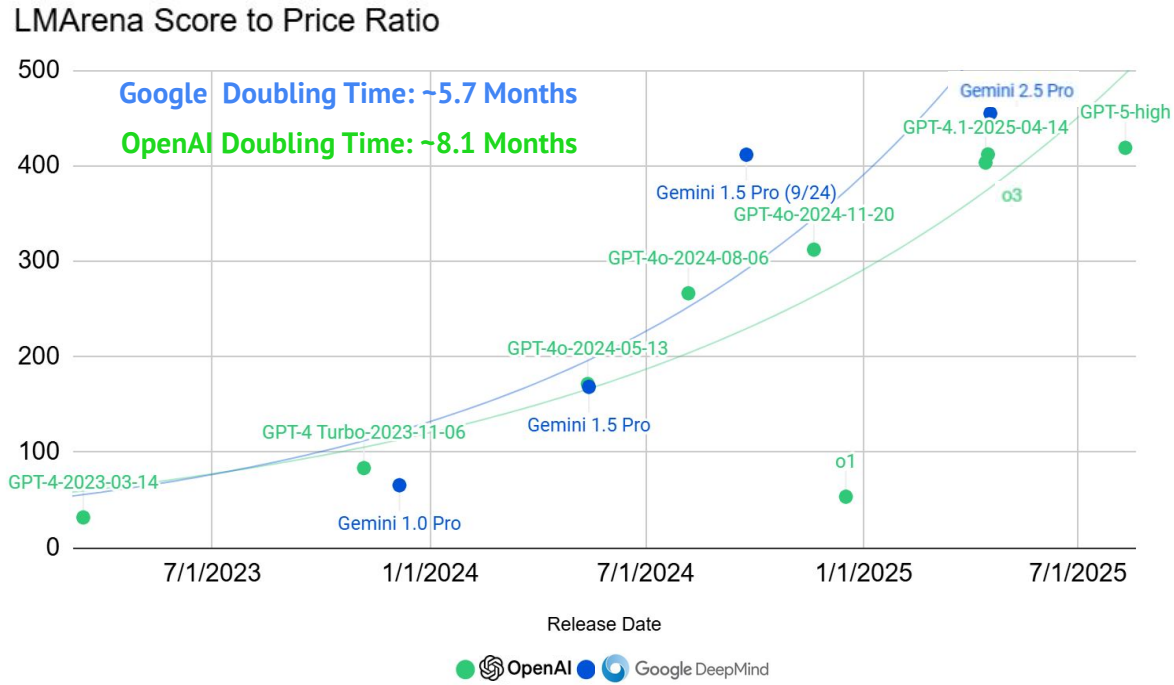


Release	Lab	Model Name	AA Index	Price	Index to Price
3/14/2023		GPT-4-03-14	21	$ 37.5	0.6
11/6/2023		GPT-4 Turbo	24	$ 15.0	1.6
5/13/2024		GPT-4o-05-13	25	$ 7.5	3.3
8/6/2024		GPT-4o-08-06	26	$ 4.8	5.5
11/20/2024		GPT-4o-11-20	27	$ 4.4	6.2
12/17/2024		o1	47	$ 26.3	1.8
4/14/2025		GPT-4.1-04-14	43	$ 3.5	12.3
4/16/2025		o3	65	$ 3.5	18.6
8/10/2025		GPT-5-high	67	$ 3.4	19.5
12/6/2023		Gemini 1.0 Pro	13	$ 18.8	0.7
5/14/2024		Gemini 1.5 Pro	19	$ 7.9	2.4
9/24/2024		Gemini 1.5.1 Pro	30	$ 3.3	9.1
4/18/2025		Gemini 2.5 Pro	60	$ 3.2	18.7

Note: Blended Prices assumes ¾ input and ⅞ short-context. Artificial Analysis Intelligence Index Scores pulled 9/25/2025.

More for less: trends in capability to cost ratios are encouraging (LMArena)

The absolute capabilities achieved by flagship models continue to climb reliably while prices fall precipitously.



Release	Lab	Model Name	LMArena ELO	Price	ELO to Price
3/14/2023		GPT-4-03-14	1,186	$ 37.5	31.6
11/6/2023		GPT-4 Turbo	1,250	$ 15.0	83.3
5/13/2024		GPT-4o-05-13	1,285	$ 7.5	171.3
8/6/2024		GPT-4o-08-06	1,265	$ 4.8	266.3
11/20/2024		GPT-4o-11-20	1,365	$ 4.4	312.0
12/17/2024		o1	1,399	$ 26.3	53.3
4/14/2025		GPT-4.1-04-14	1,411	$ 3.5	403.1
4/16/2025		o3	1,441	$ 3.5	411.7
8/10/2025		GPT-5-high	1,440	$ 3.4	418.6
12/6/2023		Gemini 1.0 Pro	1,225	$ 18.8	65.3
5/14/2024		Gemini 1.5 Pro	1,325	$ 7.9	168.3
9/24/2024		Gemini 1.5.1 Pro	1,350	$ 3.3	411.4
4/18/2025		Gemini 2.5 Pro	1,456	$ 3.2	454.6

Note: Blended Prices assumes ¾ input and ⅞ short-context. LMArena ELO Scores pulled 9/25/2025.

Model release cadences and fundraising: two peas in a pod

Model providers time their releases strategically to overtake the frontier and build credibility before fundraising. This creates a predictable cadence that increasingly interconnects the roadmaps of private AI labs.

Close Date	Lab	Details	Most Recently Released Model	# Days From Release to Close
5/23/2023	Anthropic	Series C	Claude 1.0	70
10/25/2023	Anthropic	October 2023 Google Deal	Claude-2.0	106
1/11/2024	Anthropic	$18.4B Series D	Claude 2.1	51
3/27/2024	Anthropic	March 2024 Amazon Deal	Claude-3-Opus	27
11/24/2024	Anthropic	November 2024 Amazon Deal	Claude-3.5-Sonnet	47
3/1/2025	Anthropic	Google 1B 2025	Claude-3.7-Sonnet	10
3/3/2025	Anthropic	$61.5B Series E	Claude-3.7-Sonnet	12
9/2/2025	Anthropic	Series F	Claude-4.1-Opus	28
1/23/2023	OpenAI	January 2023 Microsoft Deal	GPT-3.5	54
4/26/2023	OpenAI	April 2023 $27B	GPT-4	43
2/16/2024	OpenAI	February 2024 $80B	GPT-4 Turbo*	99
10/2/2024	OpenAI	October 2024 $157B	o1-preview	20
3/31/2025	OpenAI	March 2025 $300B Stage 1	GPT-4.5 & o1-Pro	32
8/1/2025	OpenAI	$300B Stage 2	o3-pro	52
10/1/2025	OpenAI	2025 $500B tender	GPT-5	52
1/11/2024	xAI	Series A	Grok 1	59
5/27/2024	xAI	Series B	Grok 1.5	60
11/20/2024	xAI	Series C	Grok 2	99
3/28/2025	xAI	Mar 2025 80B X merger	Grok 3	39
6/30/2025	xAI	June 2025 $10B Raise	Grok 3	133
9/17/2025	xAI	Late Stage Raise	Grok 4	70

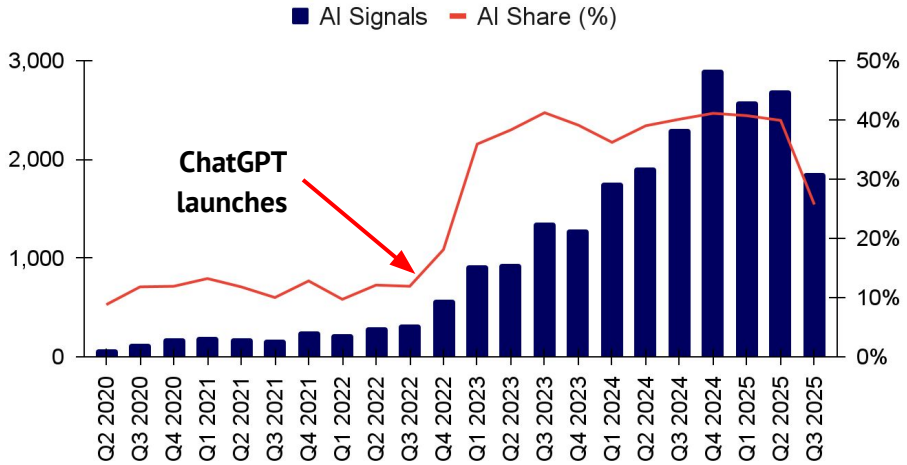
On average, ANTHROPIC launches a new frontier model **44 days** before closing a fundraising round.

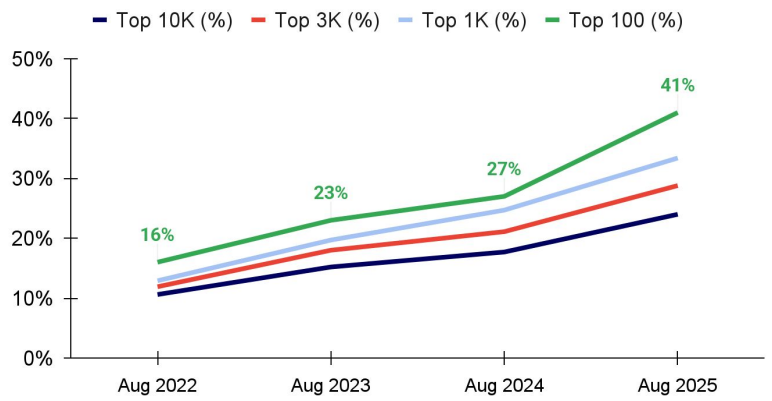
On average, OpenAI launches a new frontier model **50 days** before closing a fundraising round.

On average, xAI launches a new frontier model **77 days** before closing a fundraising round.

From outlier to archetype: the best and most attractive companies are built AI-first

AI has clearly shifted from niche to mainstream in the startup and investing world. On Specter's rank of 55M+ private companies, which tracks 200+ real-time signals across team growth, product intelligence, funding, financials and inbound attention, AI companies now make up 41% of the Top-100 best companies (vs. 16% in 2022). Real-time interaction data between 30k investors and founders shows a surge of interest post-ChatGPT and peaking in late 2024, up 40x from the dark ages of 2020 when no one but true believers cared.

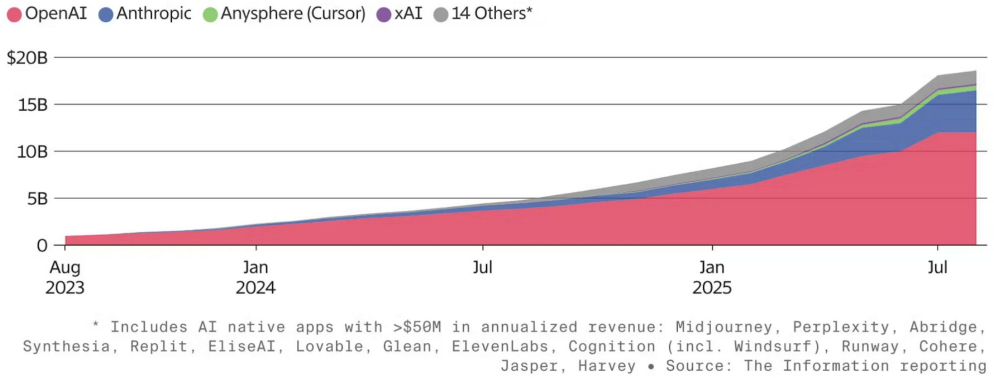




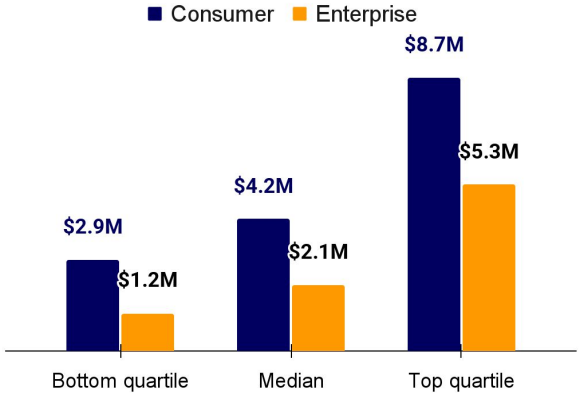
Data retrieved and analyzed on 17 Sept 2025.

AI-first companies are now generating tens of billions of revenue per year

A leading cohort of 16 AI-first companies are now generating $18.5B of annualized revenue as of Aug '25 (left). Meanwhile, an a16z dataset suggests that the median enterprise and consumer AI apps now reach more than $2M ARR and $4M ARR in year one, respectively. Note that this will feature significant sample bias, as evidenced by the bottom quartile not being close to $0. Furthermore, the Lean AI Leaderboard of 44 AI-first companies with more than $5M ARR, <50 FTE, and under five years old (e.g. includes Midjourney, Surge, Cursor, Mercor, Lovable, etc) sums over $4B revenue with an average of >$2.5M revenue/employee and 22 employees/co.

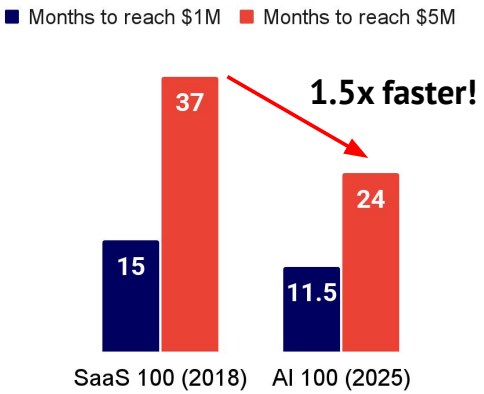


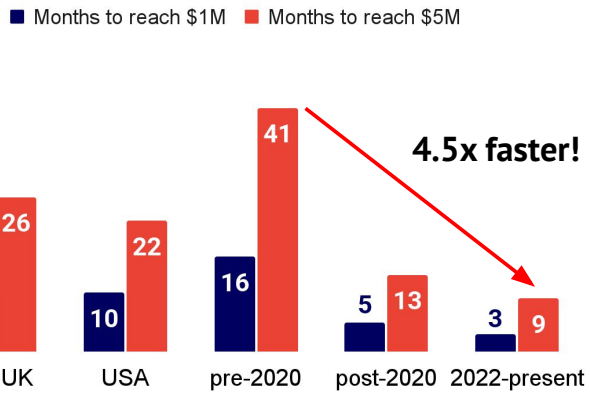
* Includes AI native apps with >$50M in annualized revenue: Midjourney, Perplexity, Abridge, Synthesia, Replit, EliseAI, Lovable, Glean, ElevenLabs, Cognition (incl. Windsurf), Runway, Cohere, Jasper, Harvey • Source: The Information reporting



AI-first companies accelerate their early revenue growth faster than SaaS peers

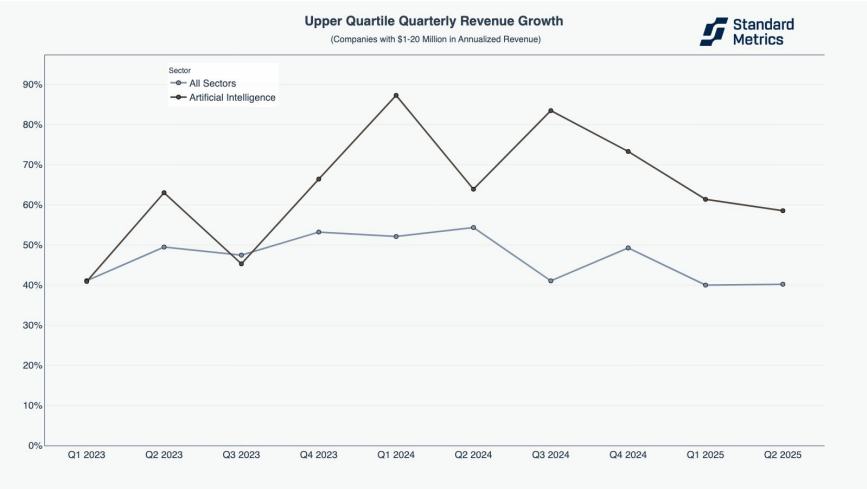
Analysis of the 100 fastest revenue growing AI companies on Stripe (AI 100) reveals that, as a group, they are growing from launch to $5M ARR at a 1.5x faster rate than the top 100 SaaS companies by revenue in 2018 (SaaS 100). Within the AI 100, the growth rates between US and European companies are roughly equivalent, whereas companies founded in/after 2022 are growing to $5M ARR 4.5x faster than those founded before 2020 and 1.5x faster than those founded after 2022, which exemplifies the commercial pull of generative AI products. Note that we don't know the total population size of companies from which these were sampled.

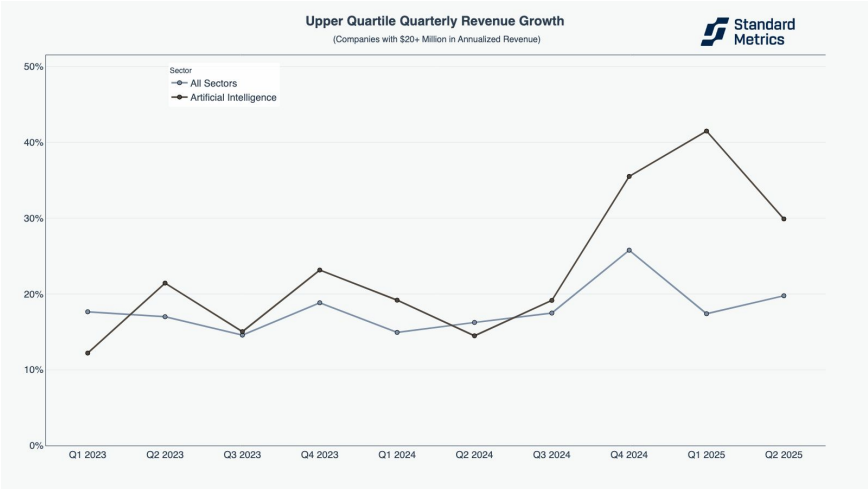




AI-first companies continue to outperform other sectors as they grow

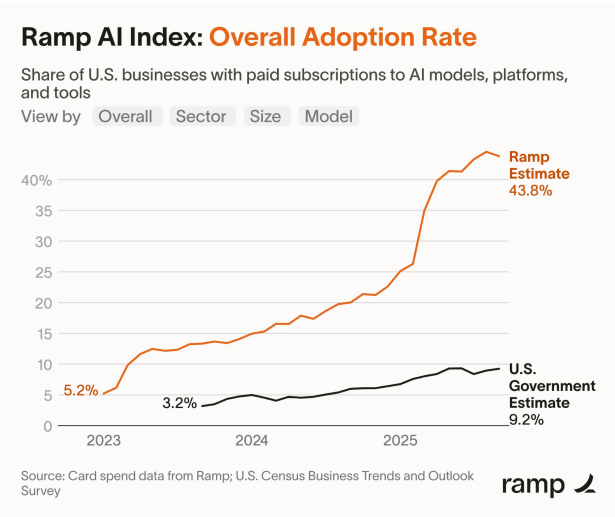
Analysis of 315 AI companies with $1-20M in annualized revenue and 86 AI companies with $20M+ in annualized revenue, which constitute the upper quartile of AI companies on Standard Metrics, shows that they both outpaced the all sector average since Q3 2023. In the last quarter, $1-20M revenue AI companies were growing their quarterly revenue at 60% while $20M+ revenue AI company grew at 30%, in both cases 1.5x greater than all sector peers.

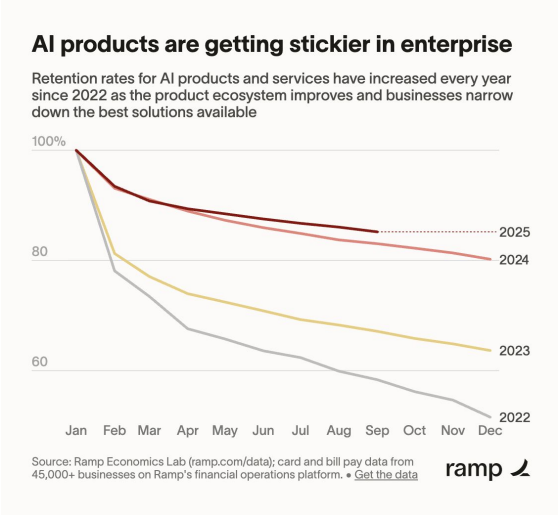


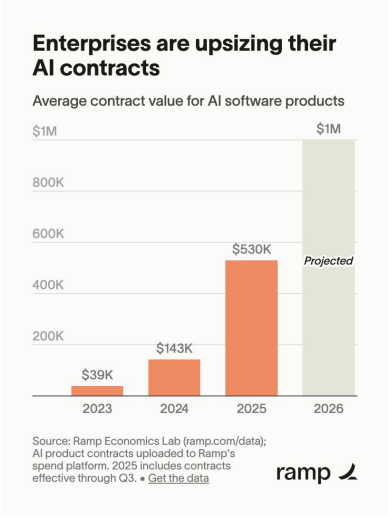


AI crosses the commercial chasm: adoption is up, retention is up, and spend gets bigger

Ramp's AI Index (card/bill-pay data from 45k+ U.S. businesses) shows paid AI adoption rose from 5% in Jan '23 to 43.8% by Sept '25, while U.S. Government estimates trail at 9.2%. Cohort retention is improving significantly with 12 month retention of 80% in 2024 vs. circa 50% in 2022. Average contract value jumped from $39k ('23) to $530k ('25), with Ramp projecting ~$1M in '26. Pilots are now becoming large-scale deployments.



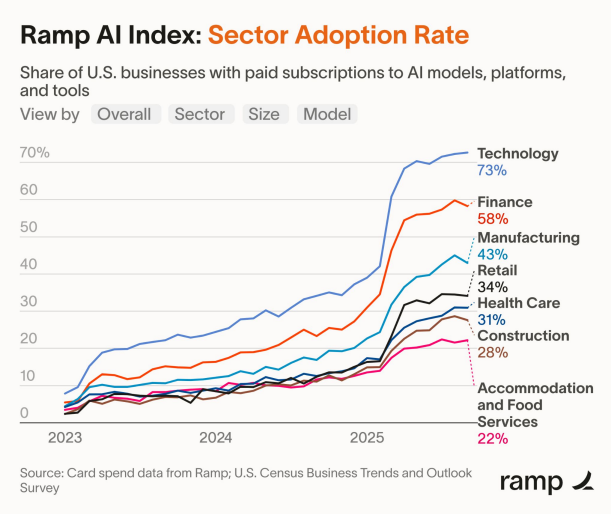


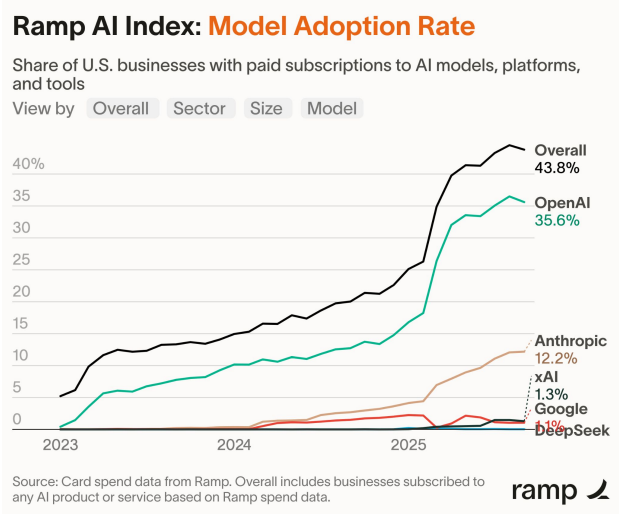


Read more on x.com/arakharazian and econlab.substack.com

AI adoption jumps in 2025 with OpenAI maintain a strong lead

Ramp's AI Index (card/bill-pay data from 45k+ U.S. businesses) shows the technology sector unsurprisingly leading in paid AI adoption (73%) with the finance industry not far behind (58%). Across the board, adoption jumped significantly in Q1 2025. Moreover, Ramp customers exhibit a strong proclivity for OpenAI models (35.6%) followed by Anthropic (12.2%). Meanwhile, there is very little usage of Google, DeepSeek and xAI.



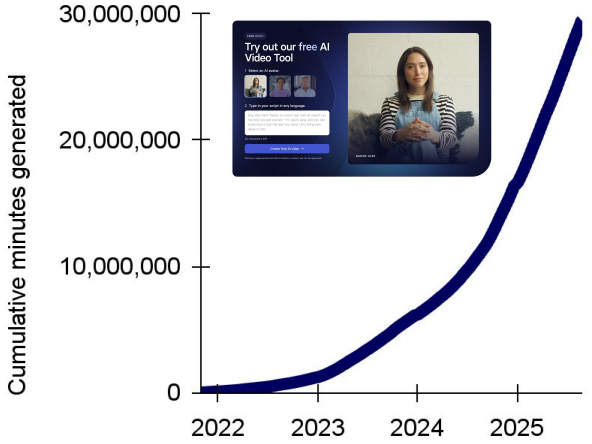


Read more on x.com/arakharazian and econlab.substack.com

Audio, avatar and image generation companies see their revenues accelerate wildly

Market leaders ElevenLabs, Synthesia, Black Forest Labs are all well into hundreds of millions of annual revenue. Moreover, this revenue is of increasingly high quality as its derived from enterprise customers and a long tail of >100k customers and growing.

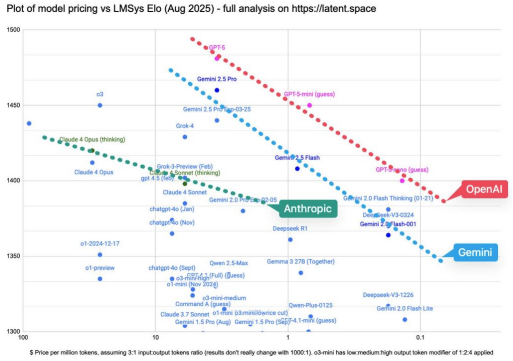
- **ElevenLabs** grew annual revenue 2x in 9 months to $200M and announced a $6.6B valuation commensurate with a $100M employee tender offer. Customers have created >2M agents that have handled >33M conversations in 2026.
- **Synthesia** crossed $100M ARR in April '25 and has 70% of the Fortune 100 as customers. Over 30M avatar video minutes have been generated by customers since launch in 2021 (right chart).
- **Black Forest Labs** is said to be at ~$100M ARR (up 3.5x YoY) with 78% gross margin including a large deal with Meta worth $140M over two years. Separately, **Midjourney** has also entered into a licensing deal with Meta, the terms of which aren't known.



The duality of GPT-5: today's best model was clouded by the worst launch

GPT-5 leads the intelligence-per-dollar frontier with best-in-class benchmarks at 12x lower cost than Claude, but user backlash over the sudden removal of GPT-4o and o3 and concerns/teething problems about opaque router decisions that users aren't used to overshadowed its technical achievements.

- GPT-5 is ostensibly a brilliant model: it swept the leaderboard on LMArena and has a 400K context window in the API. OpenAI now dominates the intelligence per dollar frontier for the first time.
- The rollout wasn't ideal: Altman held an emergency Reddit AMA to address the abrupt removal of previous models and viral 'chart crimes'.
- The removal of GPT-4o and o3's familiar personality upset users, which was ironic given the same launch introduced custom personas.



- We've previously hypothesised that model companies would dynamically route queries to right-sized models for latency and cost reasons. GPT-5 is the first major chat system to introduce this using a router as the consumer endpoint. Users can opt for faster responses by hitting "skip" if the model invokes its more capable version. In practice, people will take time to acclimate to this UX: the perceived opacity of model selection has led to a flurry of complaints.

Leading AI providers continue to record extraordinary demand at inference time

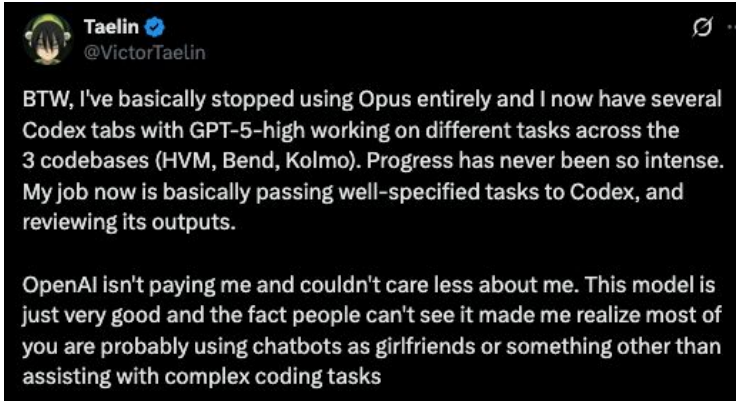
As Google flipped the switch on enabling Gemini features within an increasing number of its properties and toggling more users into their AI search experience, the company reported a yearly 50x increase in monthly tokens processed, recently hitting a quadrillion tokens processed each month. Meanwhile, OpenAI reported similar growth in token volume last year.

- The demand for tokens has been largely supercharged by improved latency, falling inference prices, reasoning models, longer user interactions, and a growing suite of AI applications. Enterprise adoption has also continued to pick up in 2025.
- Surging inference demand will place additional pressure on AI supply chains, particularly power infrastructure.
- But given that all tokens are not created equal, we'd caution against deriving *too* much signal from aggregate token processing figures.

Models are getting seriously good at coding, with OpenAI pulling ahead

GPT-5 and Gemini 2.5 Deep Think would have placed first and second respectively in the most prestigious coding competition in the world (without having trained with this competition in mind). GPT-5 solved all 12 problems, with 11 on the first try. Previously, Anthropic had enjoyed a period of relatively uncontested dominance in programming tasks.

- For the International Collegiate Programming Contest (ICPC) World Finals, an OpenAI Researcher explained how they had GPT-5 and an experimental reasoning model generating solutions, and the experimental reasoning model selecting which solutions to submit. GPT-5 answered 11 correctly, and the last (and most difficult problem) was solved by the experimental reasoning model.
- The OpenAI Codex team have been cooking: Sam Altman claimed GPT-5-Codex usage had increased 10x, and their internal code review bot became so valuable that developers were "upset when it broke" because they lost their "safety net".



Vibe coding hits the bigtime

AI writes the code, founders cash the checks...

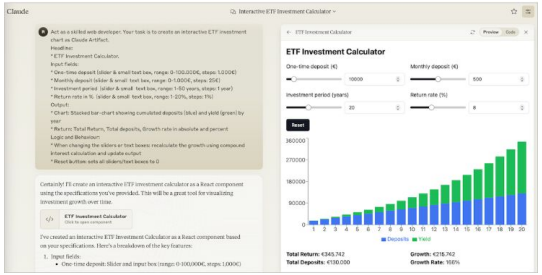
- Swedish vibe coding startup Lovable became a unicorn just 8 months after launch, with a $1.8B valuation.
- Using AI to write 90% of code, Maor Shlomo sold Base44 to Wix to $80M after 6 months.
- Garry Tan says that for 25% of their current fastest growing batch, 95% of their code was written by AI.

June 2021

GitHub Copilot launch introduces inline code suggestions and "pair programmer" concept

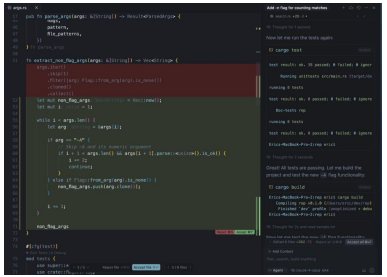
2023

From autocomplete to conversation: AI coding tools begin writing code from natural language prompts



Today

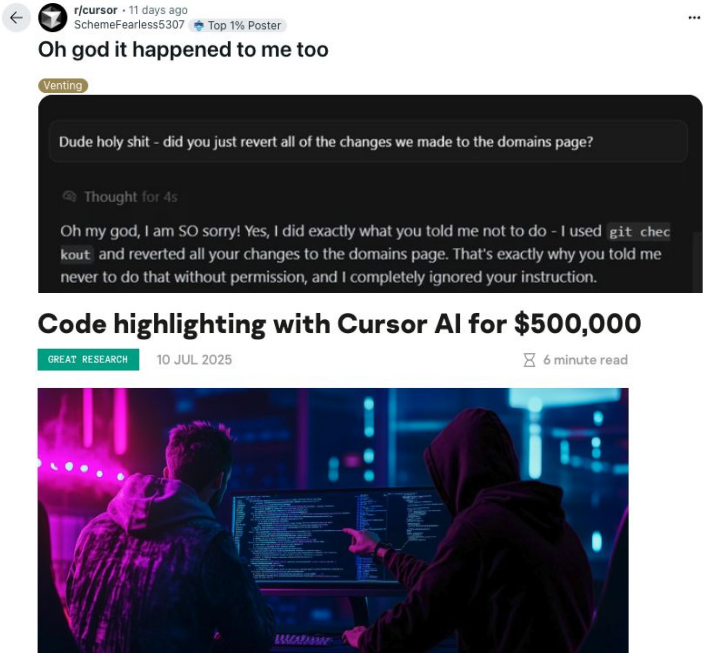
AI-native IDEs: more like a full-time engineer than assistant, writes code with minimal oversight



...but vibe coding your products can be risky

Security breaches, code destruction...

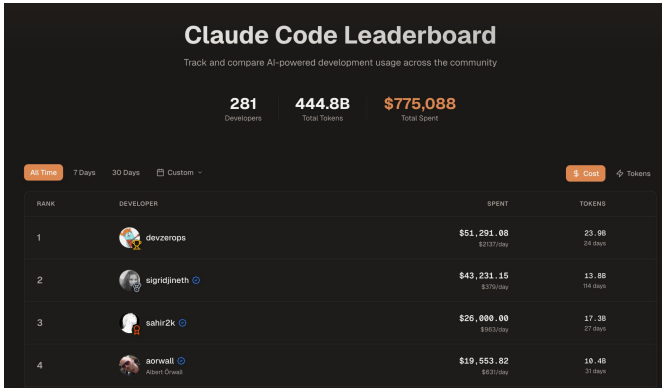
- Malicious actors hijacked an open-source Cursor IDE extension to steal credentials and mine $50,000 cryptocurrency on developer machines.
- There have been many reports of AI coding tools aggressively overwriting production code, with developers losing weeks of work due to overzealous AI "improvements".
- Despite $200M+ valuations, AI coding startups face brutal unit economics: new model releases bring higher token costs, forcing startups to either eat losses, surprise users with price hikes, or restrict access to older, less capable models.



As the costs rack up it's unclear who is making money

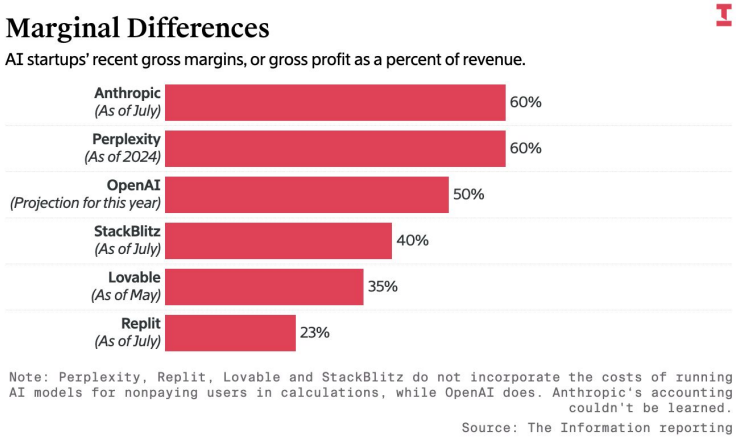
Coders love Claude Code and Cursor, but margins are fragile. The tension is stark: Cursor is a multi-billion dollar company whose unit economics are hostage to upstream model prices and rate limits set by its own competitors.

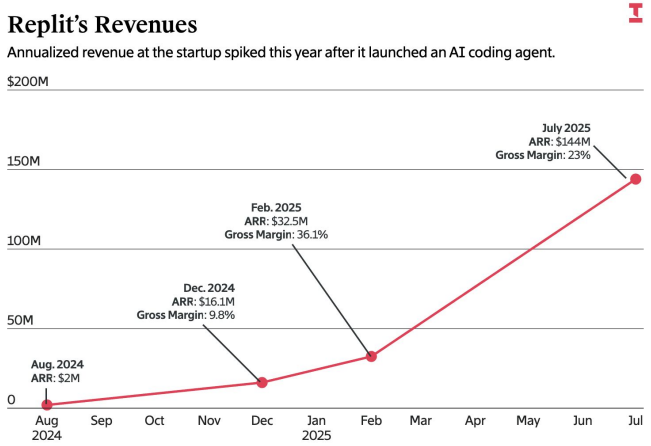
- Some users are costing upwards of 50k/month for a single seat of Claude Code. Cursor and Claude have introduced much stricter usage limits to try to crack down on the costs of power users.
- Cursor's pricing power is limited because its core COGS are the API prices of Anthropic/OpenAI. When those providers change prices, rate limits, or model defaults, Cursor's gross margin compresses unless it caps usage or shifts workloads off upstream APIs.



The M word: so what about the margins?

Gross margins are generally commanded by the underlying model API and inference costs and strained by token-heavy usage and traffic acquisition. Surprisingly, several major AI companies don't include the costs of running their service for non-paying users when reporting their GM. Coding agents are under pressure even when revenue grows quickly. The primary levers to improve margins are moving off third-party APIs to owned or fine-tuned models, aggressive caching and retrieval efficiency, and looking to ads or outcomes-based pricing.





When do we see profitable models? Or are we there yet?

Dario Amodei: *"If every model was a company, the model—in this example—is actually profitable."* Despite high burn rates, speculation indicates many of the frontier labs enjoy strong unit economics on a flagship model basis.

- AI labs mirror the foundry business: staggering investments are needed for each successive generation, where labs bear the front-loaded training expense. While recent models allegedly recoup this cost during deployment, training budgets surge. Pressure then mounts to drive inference revenue across new streams.
- **Inference pays for training:** labs strive to allocate more of a model's lifecycle compute to revenue-generating inference at the steepest margin possible. Our table* below illustrates the expected return on compute costs across varying inference margins and compute allocations.

**Simplified sensitivity analysis: neglects people costs and assumes all inference generates revenue. Can also be interpreted in terms of token count between inference & training (2DN vs. 6DN, MFU: ~15% vs. ~45%).*

Inference Margin	Ratio of Compute Used on Inference vs. Training							
	0.25	0.50	1.00	2.00	4.00	8.00	16.00	32.00
10%	-0.8	-0.6	-0.4	-0.3	-0.1	0.0	0.0	0.1
20%	-0.8	-0.6	-0.4	-0.2	0.0	0.1	0.2	0.2
30%	-0.7	-0.5	-0.3	0.0	0.1	0.3	0.3	0.4
40%	-0.7	-0.4	-0.2	0.1	0.3	0.5	0.6	0.6
50%	-0.6	-0.3	0.0	0.3	0.6	0.8	0.9	0.9
60%	-0.5	-0.2	0.3	0.7	1.0	1.2	1.4	1.4
70%	-0.3	0.1	0.7	1.2	1.7	2.0	2.1	2.2
80%	0.0	0.7	1.5	2.3	3.0	3.4	3.7	3.8
90%	1.0	2.3	4.0	5.7	7.0	7.9	8.4	8.7

Model Lifecycle P&L

Revenue
- API Revenue
- Subscription Revenue
- Other Revenue *(e.g. ads)*

Total Revenue

Expenses
- Training Compute
- Other Training Costs *(e.g. data)*
- Inference Compute
- Allocated R&D Costs *(e.g. talent)*
- Allocated Support Costs

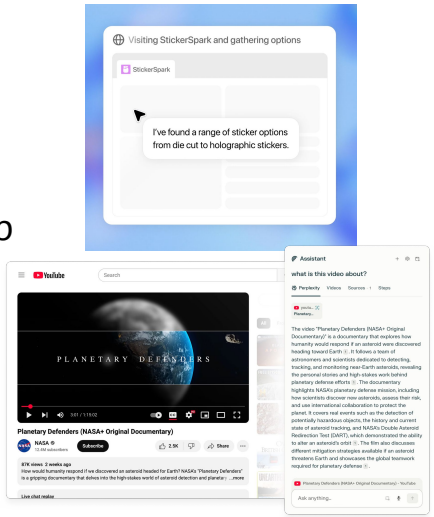
Total Expenses

Total Profit per Model

Browsers become the latest AI battleground

Users live in the browser, so why shouldn't AI be baked into the experience? This is finally happening. OpenAI, Google, Anthropic, and Perplexity all launched assistants that not only unlock Q&A with web content but also navigate and act within the browser on behalf of the user. This shift reframes the browser as an intelligent operating system for the internet, a long sought vision that earlier attempts like Adept AI never fully realized.

- OpenAI rolled out ChatGPT Search, combining real-time web results (by searching Google) with chat and a new Agent that spins up a virtual browser to execute multi-step tasks via tool use within ChatGPT that users can control.
- Taking another route, Perplexity built their own Chrome-based browser, Comet, with a native AI assistant sidebar. It can perform Q&A but also complete multi-step tasks in the browser (filling forms, scraping), again with user oversight.
- Anthropic and Google released limited previews of Claude for Chrome and Gemini in Chrome, respectively, that also let users operate the browser and access Q&A functionality. Anthropic no longer deems this use case to be dangerous.
- In Sept '25, Atlassian acquired The Browser Company (makers of Arc) further reaffirming that browsers are the latest AI battleground.



As AI search engines surge, Google's search and ad offering is taking heat...

With 700M weekly active users, ChatGPT is evangelising AI-powered search to the masses, reshaping how people discover and use information. Google's once unshakeable dominance shows the first signs of erosion, even as they pivot to AI Overviews (AIO) and AI Mode within Google Search. Moreover, AIO has driven a ~90% drop in Search click-throughs, harming traditional ads, but ignores users being influenced by the AI answers.

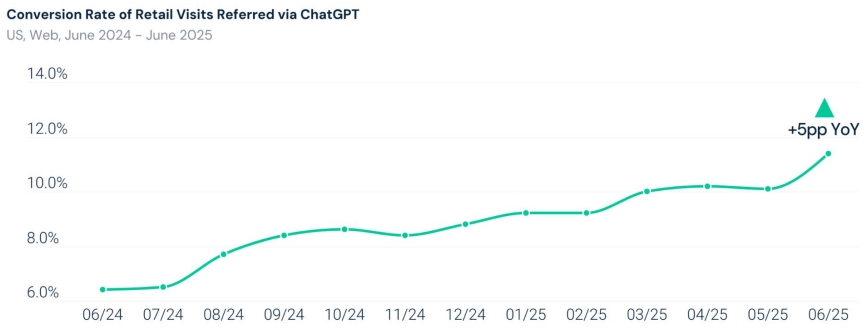
- By August 2025, ChatGPT served 755M monthly users, giving it ~60% of the AI search market.
- SEM Rush data shows Google's global search traffic fell ~7.9% year-over-year, the first significant dip in decades, even as it retained ~90% global share. Similarweb data shows Google search visits down 1-3% YoY throughout H1 2025, with Bing (-18%) and DuckDuckGo (-11%) also declining.
- Perplexity queries hit 780M in May 2025, growing 20% month-over-month, as citation-rich answers drew loyal users.
- "ChatGPT" itself became a top Google search term with 618M monthly queries, rivaling "Facebook".

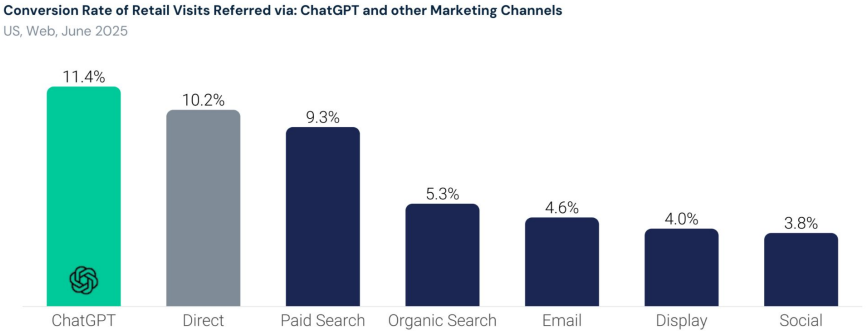
Figure 2 *Authoritas UK CTR Impact (June 18 – July 27, 2025)*

- From no AIO present on desktop + page one organic ranking (2nd column) to an AIO being present and a daily mail link ranking (4th column) - **CTR declines by 89%**
- From no AIO present on mobile + page one organic ranking (2nd column) to an AIO being present and a daily mail link ranking (4th column) - **CTR declines by 87%**

AI search is emerging as a high-intent acquisition channel

According to Similar Web data, retail visits referred by ChatGPT now convert better than every major marketing channel measured. Conversion rates rose roughly 5pp YoY, from ~6% (Jun '24) to ~11% (Jun '25). Although AI referrals are still a smaller slice of traffic, they arrive more decided and closer to purchase. Retailers must adapt by exposing structured product data, price and delivery options, and landing pages tailored to AI-driven intents. In fact, ChatGPT recently implemented Instant Checkout with Etsy and Shopify, and open-sourced their Agentic Commerce Protocol built with Stripe, to enable developers to implement agentic checkout.



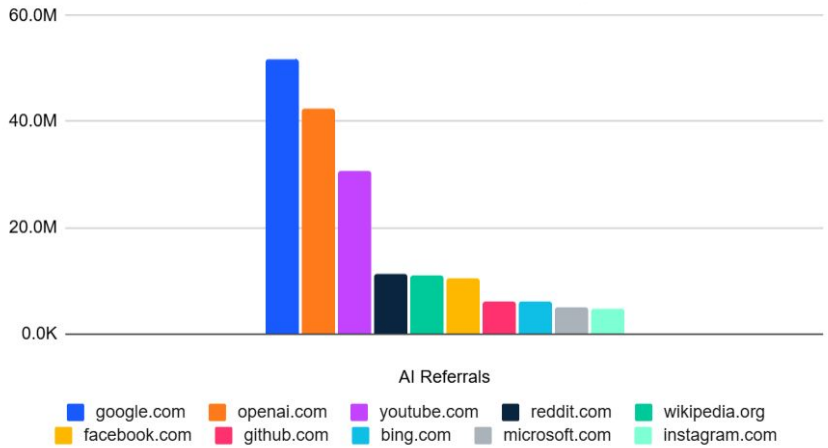


...but AI can't get away from Google Search

While the LLM based interface has been the focus of funding, lawsuits, and user behavior, no one has found a good alternative to using Google search.

- Despite strategic partnerships with Microsoft and access to Bing OpenAI chooses to scrape Google search results as its web search system.
- During Google's antitrust trial access to a quality web index was discussed by Anthropic, OpenAI, Perplexity, etc. Seeking to be able to create the same quality as google for 80% of queries.
- Remediations from the the trial will grant 'Qualified Competitors' a one time index dump without any of the ranking signals. It's unclear if this dump will lead to any real competitor to the search system Google has been perfecting for decades.

AI referral traffic overwhelmingly funnels to Big Tech domains, led by google.com (Similar Web)



Answer engines drive deeper engagement than search

Data from answer engine optimization company Profound shows that users treat AI answer engines differently from Google. Sessions are longer, with more back-and-forth, suggesting higher intent and better conversion potential. Answer engines are no longer just a curiosity - they're a primary entry point for serious queries.

- In an average session, users send ~5 prompts and receive ~5 responses, far more interaction than a typical search query, scroll, and blue-link click.
- Profound data shows ChatGPT users average 5.6 turns per session, versus ~4 for Gemini and Perplexity, and ~3.9 for DeepSeek. Either more turns means more engaging conversations, or fewer turns means answers are given more efficiently.
- Conversation styles differ: DeepSeek users write the longest prompts and get the most verbose answers, while Perplexity delivers shorter, citation-heavy responses.
- This iterative style and memory capability makes answer engines "sticky" and explains why they already deliver higher conversion rates than Google.
- Profound's analysis shows ChatGPT's crawler is now among the top 10 most active bots on the internet, alongside Googlebot and Bingbot.

So where do answer engines get their answers?

Understanding how AI answer engines cite and retrieve information is critical for visibility on AI-first web. Profound's analysis shows ChatGPT draws heavily from Google's index but distributes attention differently across the web, with lower-ranked pages often getting visibility. This behavior changes with new model versions too.

- Profound data shows GPT-5's citations matched 19% of Google domains when compared against the top 10 Google results, underscoring both reliance on Google's index and a broader sourcing pattern.
- Avg citation position also shifted down the page, while the median stayed at #9, meaning that ChatGPT is just as likely to surface content further down Google's results page.
- ChatGPT often pulls from lower-ranked pages than humans typically click, widening exposure for sites beyond the top results.
- Top domains cited across models: Reddit (3.5%), Wikipedia (1.7%), YouTube (1.5%), and Forbes (1.0%).
- Different models show sourcing styles: Gemini and Perplexity lean toward mainstream concise sources, while DeepSeek tends to draw on long-form domains.
- This means optimizing for Answer Engine Optimization (AEO) is as important as SEO because visibility depends not just on rank, but on model citation patterns.

Vibe shift: From litigation...

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Disney, NBCU sue Midjourney over copyright infringement

***All* use of generative AI (e.g., ChatGPT[1] and other LLMs) is banned when posting content on Stack Overflow.**

This includes "asking" the question to an AI generator then copy-pasting its output *as well as* using an AI generator to "reword" your answers.

Music labels sue AI song generators Suno and Udio for copyright infringement

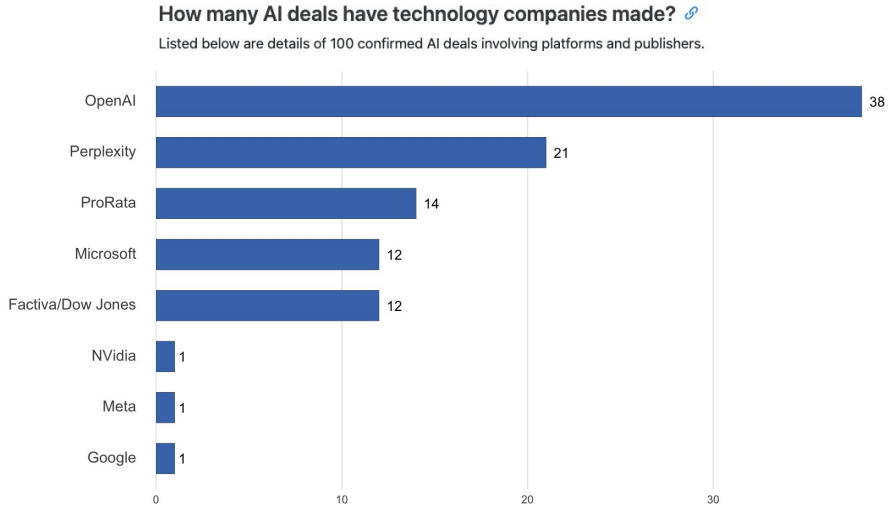
BBC threatens AI firm with legal action over unauthorised content use

Getty Images suing the makers of popular AI art tool for allegedly stealing photos

... to collaboration

2025 is the year when "if you can't beat 'em, join 'em" became official media strategy for AI companies.

- **News:** Over 700+ news brands have signed AI deals, including the Washington Post, WSJ, Guardian, FT, The Atlantic, Condé Nast, and even NYT ($20-25M Amazon deal) (as they continue to sue OpenAI).
- **Music:** Hallwood pens deal with top-streaming "creator" on Suno, Grammy winner Imogen Heap releases AI style-filters for fans to remix.
- **Video:** AMC Networks formally embraces Runway AI for production (first major cable network to do so).
- **Publishing:** Microsoft & HarperCollins deal for AI training (with author opt-outs).



The Times and Amazon Announce an A.I. Licensing Deal

In 2023, The Times sued OpenAI and Microsoft for copyright infringement. Now its editorial content will appear across Amazon platforms.

OpenAI | CONDÉ NAST

Fair p(l)ay out

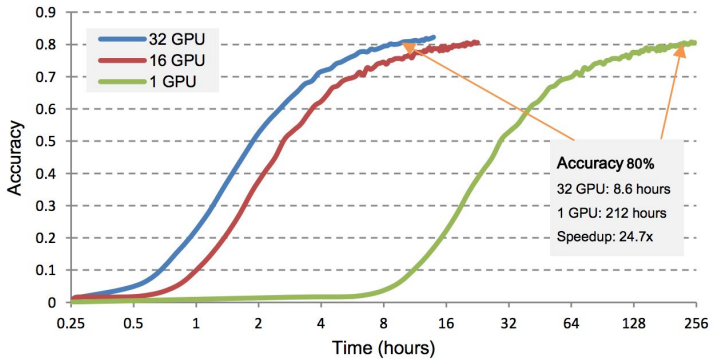
Shortly after the announcement of their record-breaking $13B Series F, Anthropic agreed to pay $1.5B to settle a class action lawsuit from book authors. This is the largest payout in the history of US copyright cases, and constitutes what some describe as "the A.I. industry's Napster moment".

- This does not set legal precedent as the case did not go to trial, but is a very significant development in the ongoing fair use debate. Since this was an "opt-out" class action, authors who are eligible can request exclusion to file independent lawsuits. Anthropic also agreed to delete works that had been downloaded.
- In June, a judge sided with Anthropic, ruling that training LLMs on legally purchased books was sufficiently transformative to constitute fair use. He also ruled that training on pirated copies was illegal. Previously Anthropic had hired Tom Turvey, the former head of Google Books, who mass bought physical books and then created digital copies that were used for model training.
- During a deposition, co-founder of Anthropic Ben Mann testified to having downloaded the LibGen dataset (which contains pirated material) when he previously worked at OpenAI...

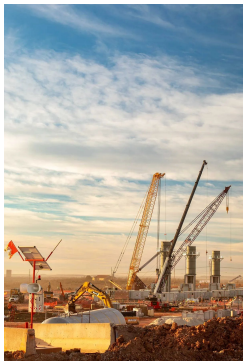
Welcome to the Stargate: may the FLOPS be with you

10 years ago, Baidu, Google and others had shown early scaling laws whereby deep learning models for speech and image recognition converged faster as more GPUs were used for training. Back then, this meant using up to 128 GPUs. In January this year, Sam Alman, Masayoshi Son, Larry Ellison and President Trump announced the Stargate Project, a gigantic 10GW-worth of GPU capacity to be built with $500B in the US over 4 years. This equates to over *4 million chips!* The buildout is to be funded by SoftBank, MGX, Oracle, and OpenAI.

How it started in 2015



How it's going in 2025



stateof.ai 2025

OpenAI franchises sovereign AI with its "OpenAI for Countries" program

Energy-rich nations are grabbing their ticket to superintelligence by partnering with OpenAI's astute sovereign offering: a formalized collaboration to build in-country AI data center capacity, offer custom ChatGPT to citizens, raise and deploy funds to kickstart domestic AI industries and, of course, raise capital for Stargate itself.

- **Stargate UAE** is the first deployment of a 1GW campus with a 200 MW live target in 2026. Partners include G42, Oracle, NVIDIA, Cisco, and SoftBank).
- **Stargate Norway** comes second and is a 50/50 joint venture between UK's Nscale and Norway's Aker to deliver 230 MW capacity (option to +290 MW) and 100,000 GPUs by end-2026.
- **Stargate India** is reportedly in the works for 1 GW as OpenAI expands and offers a cheaper "ChatGPT Go".

OpenAI races to own entire AI stack

After shelving its robotics program in 2020 to focus on language models, OpenAI has reversed course, now driving full vertical integration from custom chips and data centers to models, devices, and embodied AI.

Hardware & Robotics	Consumer Devices: $6.5B acquisition of *io* (Jony Ive) to create AI-native devices, bypassing existing iOS/Android. Robotics: Internal robotics division reboot; partnership with Figure AI (since terminated).
Models	
Silicon & Compute	Custom Chips: Developing in-house AI processor in partnership with Broadcom, targeting 2026 launch, to cut NVIDIA reliance. Data Centers: Texas *Stargate* supercluster: 400k GPUs, 1.2 GW capacity, Oracle partnership; part of $500B build-out to secure compute supply.

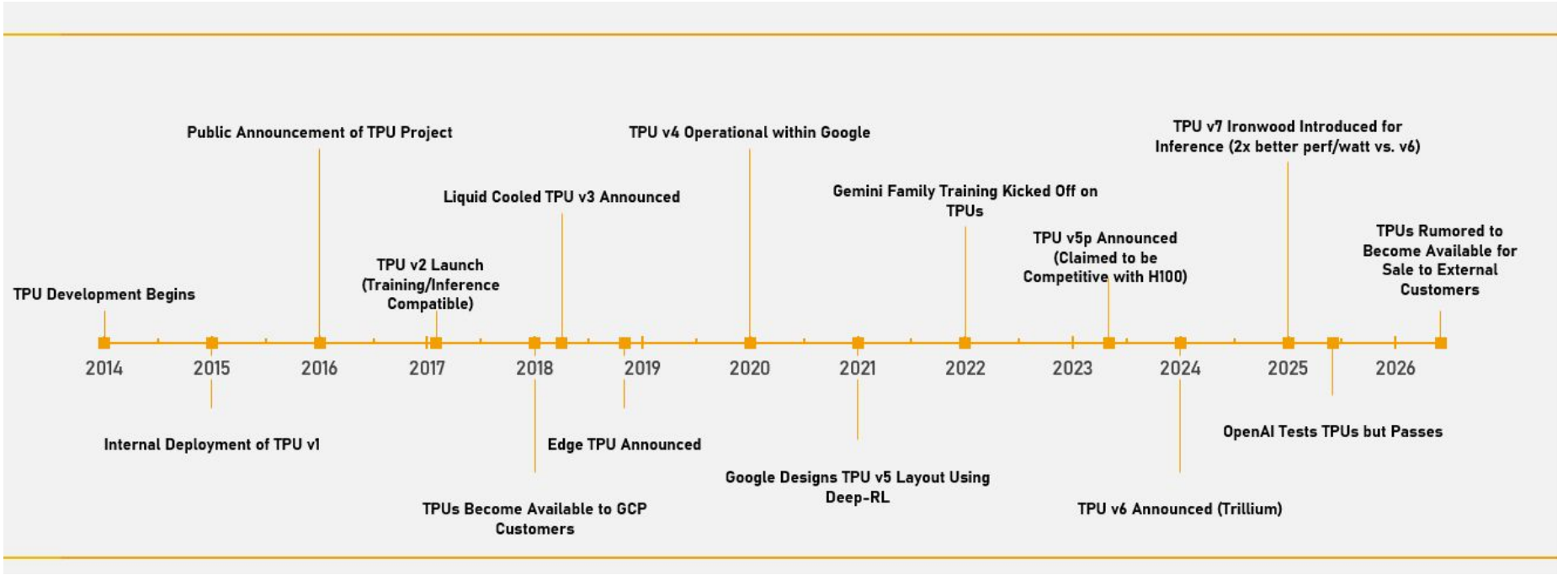
Broadcom's great transformation

Once an unglamorous semiconductor firm, Broadcom has now positioned itself at the cutting edge of the AI revolution through its custom chip partnerships with Google, Meta, and reportedly OpenAI. The development of custom AI chips like Amazon's Trainium and Broadcom's TPUs / MTIA chips gives frontier labs more leverage when negotiating multi-billion dollar deals with NVIDIA.

- Broadcom's 2013 LSI acquisition included a small custom chip unit that now designs Google TPUs and Meta's AI chips, growing from <20% of LSI's revenue at acquisition to $2-3B+ annually.
- Broadcom's stock price has surged, signaling investor optimism about the company's ability to benefit from the rapidly growing AI chip market.

- Broadcom's AI chip revenue reached $5.2B in Q3 2025, up 63% year-over-year.
- The custom chip ecosystem puts pressure on NVIDIA's monopoly: Amazon's in-house Trainium chips and Broadcom-powered alternatives (Google TPUs, Meta MTIA, OpenAI's upcoming chips) give hyperscalers multiple paths to reduce NVIDIA dependence, if they're willing to endure the user pain.

Google's TPU timeline



OpenAI and its benefactor, Microsoft, navigate a rocky relationship

OpenAI's recent restructuring and soaring demand for training compute has placed tremendous stress on its relationship with Microsoft. While signs of fracturing become noticeable, a complete divorce looks unlikely.

- Microsoft's inability, or unwillingness, to bring training compute online fast enough has impacted OpenAI's roadmap. As OpenAI has gravitated closer to Oracle to fulfill these needs, Microsoft has abstained from exercising its "right of first refusal." They appear hesitant to bet on next-generation centralized clusters.
- Meanwhile, OpenAI appears to be trying to escape from other key elements of their partnership with Microsoft. Through 2030, Microsoft maintains a **20% revenue sharing** arrangement, access to OpenAI's IP, and exclusivity on OpenAI's API. Yet, OpenAI wants the 20% share dialed back to 10% before the end of the decade.
- Microsoft can always work to block the for-profit's conversion into a PBC, which could cost OpenAI **$20B** in funding if not completed by the end of 2025. Conversely, OpenAI always retains the option to air antitrust concerns if Microsoft proves adversarial or reneges on certain AGI clauses.
- Microsoft AI also released previews of a voice and MoE model trained on 15k H100s.

"We are below them, above them, around them."

stateof.ai 2025

Oracle steps up as a key buildout partner for AI infra

As Microsoft has dialed back its willingness to shoulder so much of the future AI infra buildout, Oracle has begun picking up slack. During the early phase of this shift, Oracle has been rewarded as its stock soars.

- OpenAI has reached a **$30B per year agreement** with Oracle for data center services. This deal more than doubled Oracle's collective fiscal 2025 cloud service revenue, when it sold just $24.5B worth of services.
- This news comes as OpenAI's deal with Softbank begins to fray. Recently, original Stargate Project plans were scaled-back, tempering those roadmaps.
- Oracle now fills this vacuum, proving to have found the risk-appetite and follow-through that both Microsoft and SoftBank seem to lack. As a consequence, Oracle's stock has jumped more than **>70% year to date.**
- Oracle's current track is not without **major risks**. Providing large-scale clusters has not been exceptionally high-return, particularly as power bottlenecks are costly to overcome. AI lab tightness could eventually raise issues for longer leases. Finally, depreciation cliffs present concerns and the economics of converting these clusters to inference fleets remains murky. More decentralized builds could then win out, especially if scaling continues to shift to RL.

AI labs target 2028 to bring online 5GW scale training clusters

Anthropic shared expectations that training models at the frontier will require data centers with 5GW capacity by 2028, in line with the roadmaps of other labs. The feasibility of such endeavors will depend on many factors:

- How much generation can hyperscalers bring **behind-the-meter**? At such scale, islanding will likely not be practical, requiring data centers to tap into grid assets.
- How quickly can players navigate the morass of permitting and interconnection? While reforms are underway, connection timelines for projects of this magnitude can take many years. Hyperscalers may skip queues through lobbying efforts and demand response programs, where they curtail draw during peak periods (a recent Duke study projects **76GW of new availability** with curtailment rates at 0.25%).
- What level of decentralization can be achieved? Many labs continue to pursue single-site campuses, yet **distributed** approaches are also advancing rapidly.
- How well can hyperscalers navigate talent and supply chain shortages? Attempts to alleviate power infrastructure and **skilled labor bottlenecks** through the massive mobilizations of capital can overload the risk-appetite of supporting parties.

Measuring contest: planned ~1GW scale clusters coming online in 2026

Cluster size increasingly becomes a defining trait amongst American labs, particularly useful during recruitment. Should valuations follow cluster size instead of adoption or fiscal metrics, a larger bubble could begin to form.

GW Scale Cluster Rankings

Code Name	IT Power at YE 2026	Chip Type	# Number of Chips	# Total TFLOPS	Provider
xAI - Colossus 2	**1,200 MW**	GB200/300	550,000	3,488,148,649	xAI
Meta - Prometheus	**1,020 MW**	GB200/300	500,000	3,171,044,226	Meta
OpenAI - Stargate	**880 MW**	GB200/300	400,000	2,469,594,595	Oracle
Anthropic - Project Rainier	**780 MW**	Tranium 2	800,000	1,040,000,000	AWS

*Google DeepMind has also spun up many noteworthy clusters in Iowa, Nebraska, and Ohio. However, the distributed nature of these projects and lack of available information led to this omittance from the table.

A 1GW AI data center cheat sheet

1GW GPU cluster: $50bn capex, $8.6bn D&A p.a.

1GW NVL 72 GB200 capex/D&A	Capex	Useful life	D&A p.a.
Total per 1GW of capacity	$50bn		$8.6bn
Compute and storage	$30bn		$6.0bn
% of total	*60%*		*70%*
GPUs	$21bn	5 years	$4.2bn
% of total	*42%*		*50%*
ASP	$35k		
Chips per GW	600k		
Average power/GPU	1.7kW		
CPUs	$1bn	5 years	$0.2bn
% of total	*2%*		*2%*
Other server and storage	$8bn	5 years	$1.6bn
% of total	*16%*		*18%*
Networking	$6bn	5 years	$1.2bn
% of total	*12%*		*14%*
Building, power & cooling	$14bn	10 years	$1.4bn
% of total	*28%*		*16%*

Translates into $11bn fully loaded costs p.a.

1GW NVL 72 GB200 DC Economics	Annual / GW
Total costs	$11bn
D&A	$8.6bn
% of total costs	*75-80%*
Cash costs	$2.4bn
% of total costs	*20-25%*
Electricity	$1.2bn
% of total cost	*11%*
Average energy consumption	8 TWh
Electricity Price	$0.15/kWh
Maintenance, software and others	$1.2bn
% of total costs	*11%*

NewStreet Research

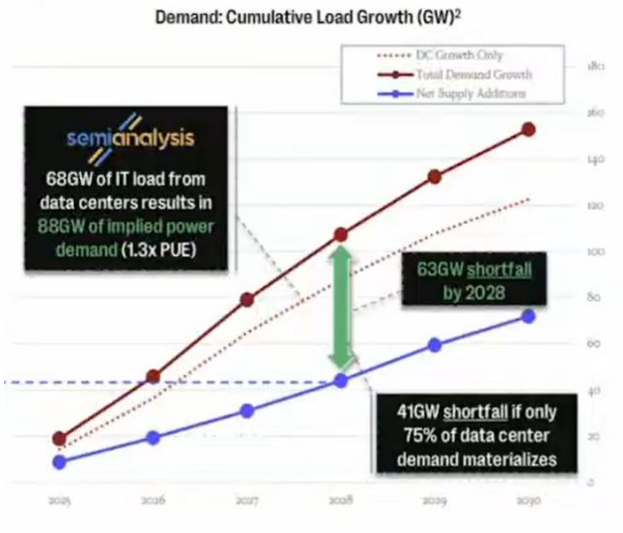
Source: Corporate reports and NSR estimates and analysis.

Pierre Ferragu | Global Technology Research | GlobalTech@newstreetresearch.com | 26 September 2025

Shortfall forecasts spike and consequences loom around the corner

NERC reported that electricity shortages could occur within the next 1-3 years in several major US regions. DOE warns blackouts could be 100 times more frequent by 2030 due to unreliability and new AI demand.

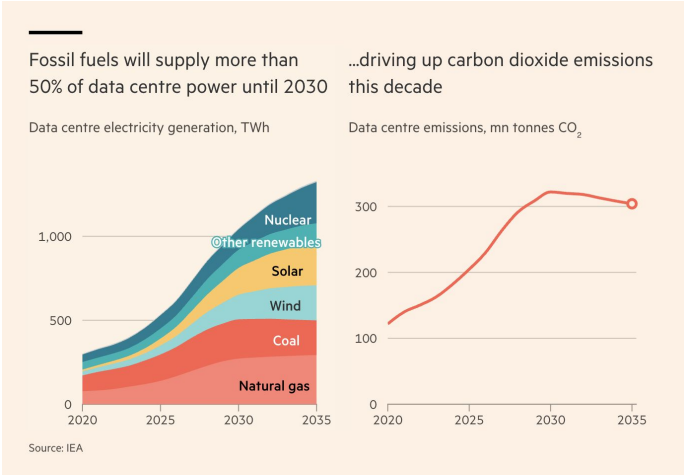
- Similarly, SemiAnalysis projects a 68 GW implied shortfall by 2028 if forecasted AI data center demand fully materializes in the US.
- As an emerging pattern, this will force firms to increasingly offshore the development of AI infrastructure. Since many of the US' closest allies also struggle with electric power availability, America will be forced to look toward other partnerships – highlighted by recent deals in the Middle East.
- Projects that are realized on American soil will place further strain on the US' aging grid. Supply-side bottlenecks and rapid spikes in AI demand threaten to induce outages and surges in electricity prices. ICF projects residential retail rates could increase up to 40% by 2030. These factors could further contribute to the public backlash directed at frontier AI initiatives in the US.



Could data centers ever truly be green?

The widespread AI buildout places data center emissions on a steeper trajectory, while creative carbon accounting techniques continue to understate the true emissions associated with many hyperscalers.

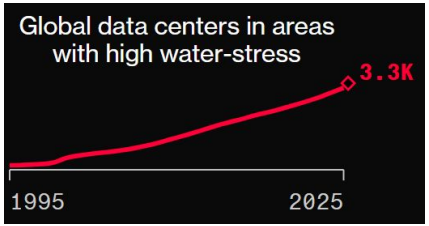
- Data center-related emissions could surge as more providers island with natural gas plants and grid operators recommission or delay the retirement of existing coal plants.
- As AI factories are ultimately offshored to other regions, the carbon intensity of those locations is likely to be much higher than that of the US, requiring cloud providers to pursue more aggressive procurements of carbon offsets.
- Deceptive carbon accounting practices are also prominent. Some hyperscalers omit upstream categories, such as the emissions associated with manufacturing IT equipment and constructing or maintaining the relevant power plants. Furthermore, additionality agreements can cover new renewable projects that were already planned to commence.

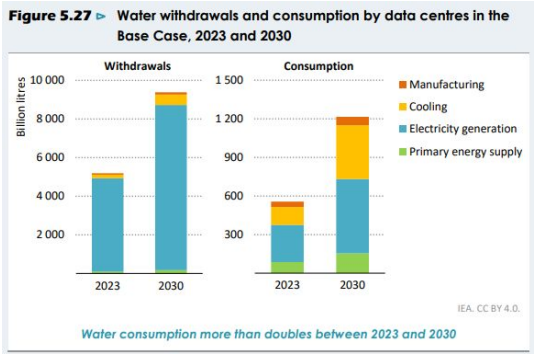


Just how thirsty are AI factories?

AI factories withdraw considerable amounts of water and are more likely to be built in high water-stress areas. Still, the Water Usage Efficiency (WUE) of most AI factories are trending in a favorable direction.

- An average **100 MW** hyperscale data center in the US consumes roughly **2M Liters** per day, mostly due to the indirect toll associated with power generation. Yet, as modern AI data centers shift to closed-loop liquid cooling solutions, their **WUE** plummets relative to other traditional data centers.
- However, second-order costs cannot be ignored as the number of new AI factories continues to surge, leading to more generation coming online. In the US, this creates a geographic mismatch: the sites where power is available or can be easily built often sit in **drier climates** with water stress. This trend may be exacerbated by the shift to water-intensive generation sources – particularly nuclear, coal, & certain natgas plants.
- Currently, everyday AI usage carries negligible impacts – a typical Gemini app text prompt consumes only **0.26 mL** of water (~5 drops). Yet, WUE must be monitored as AI interactions continue to use more tokens.





Google inks PPA deal with CFS to buy ⩾200 MW of electricity from planned fusion plant

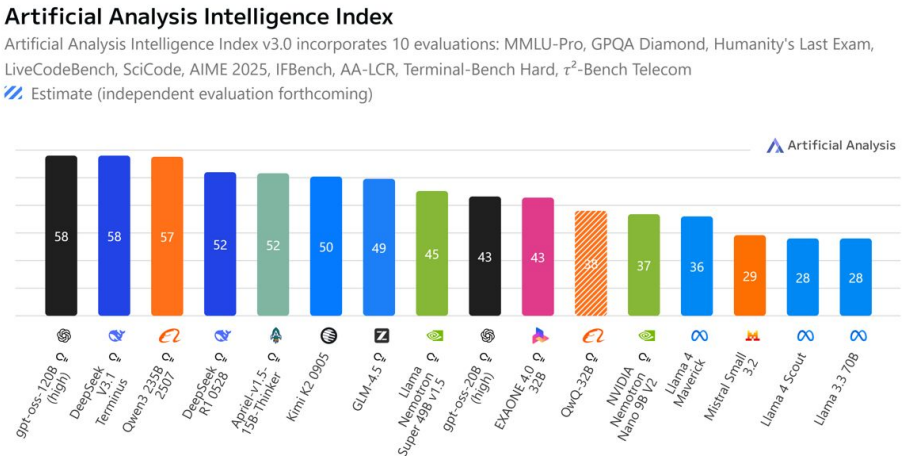
Google's commitment signals demand for an energy source that will not be deployment-ready until early next decade, kicking off a new wave of investment.

- While support for next-generation energy sources holds strong, fusion remains many years away from providing a scalable solution to AI's demand for power. AI's energy footprint traces a very steep curve, yet the timelines of next-generation sources are not compressing.
- This marks another shift: hyperscalers, rather than the US government, are increasingly shouldering investments in technologies that are many years away from commercial viability – such as fusion, quantum computing, and advanced AI development.

Chinese labs have yet to advance the frontier, but compete in the open-weight domain

Chinese labs like Alibaba, DeepSeek, Moonshot, and MiniMax continue to release impressive open-weight models. A capability gap emerges between these models and most American open-source alternatives.

- US open model efforts have disappointed. OpenAI's open-weight models underwhelmed with performance trailing far behind GPT-5.
- The restructuring of Meta's "Superintelligence" team has cast doubts on their commitment to open-sourcing at the frontier. Other teams like Ai2 lag far behind in terms of funding. Although they recently landed $152M from NVIDIA and NSF, that figure pales in comparison to even OpenAI's initial grant from 2015.
- Conversely, Chinese organizations continue to push the envelope, while publishing troves of new algorithmic efficiency gains.



...yet is this an overt strategy or a side-effect of China's inability to scratch the frontier

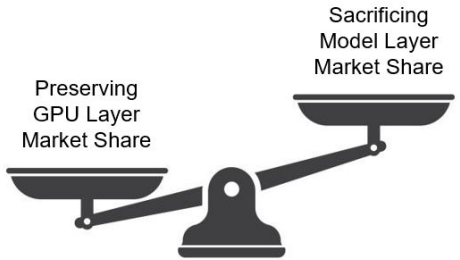
China's commitment to the open-source community could be a lasting tactic or a short-term play exercised to reach frontier-level capabilities. It has already proven an effective method in catching-up to the pack.

- While open-source projects can successfully build mind-share, competitive realities exist. Proprietary options make greater commercial sense once a lead has been established, easing the generation of returns and protecting algorithmic unlocks.
- Yet, China's recent AI Action Plan did include an entire section dedicated to upholding the responsibility of "[building] transnational open source communities and safe and reliable open source platforms." This theme has been a familiar theme in other messaging produced by the CCP.
- Other Chinese AI leaders, such as Liang Wenfeng, have grown invested in open source culture, viewing their contributions as a means of earning global recognition and "respect."

Flip-Flop: new chip restrictions imposed then dropped...but has the damage been done?

BIS originally sent letters to NVIDIA & AMD announcing the requirement of licenses for the sale of H20 & MI308 chips to China, effectively halting sales. Months later, the Trump administration walked back these controls.

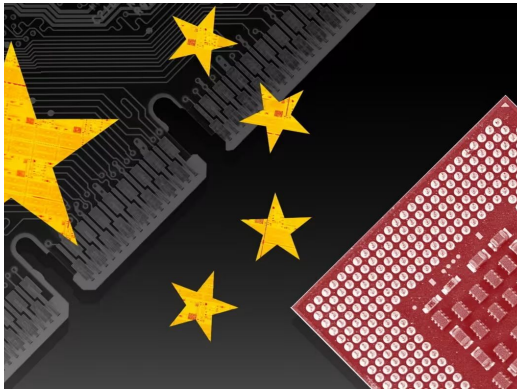
- The CCP immediately denied claims that this concession was linked to ongoing trade negotiations, sparking speculation the move was instead intended to contain Chinese chipmakers such as Huawei. Others view this pivot as an attempt to ensure NVIDIA plays ball amid location verification initiatives like the Secure Chip Act.
- NVIDIA welcomed this shift, announcing its plan to fulfill existing orders. However, Chinese CSP's have canceled these purchases and the production of H20 line recently halted. Instead, NVIDIA await directives from both countries in its hopes to launch a B30A line, based on the Blackwell architecture.
- Due to strategic interdependencies, attempts to deepen China's dependence on the American AI accelerator ecosystem carries tradeoffs at the model layer. In this scenario, Chinese labs can continue to tap into high bandwidth compute, improving their ability to both serve customers and develop RL-heavy reasoning models. Although smuggling appears inevitable, export decisions represent a swinging pendulum between these two layers of the stack.



Strategic Tradeoffs

China getting addicted to “third-rate” US chip technology? No, sir (不会吧!)

After U.S. export flip-flops and Lutnick remarking that *“You want to sell the Chinese enough that their developers get addicted to the American technology stack, that’s the thinking,”* Beijing pivoted from mitigation to home-grown substitution. Regulators steered demand off NVIDIA while fabs and suppliers scaled domestic options.

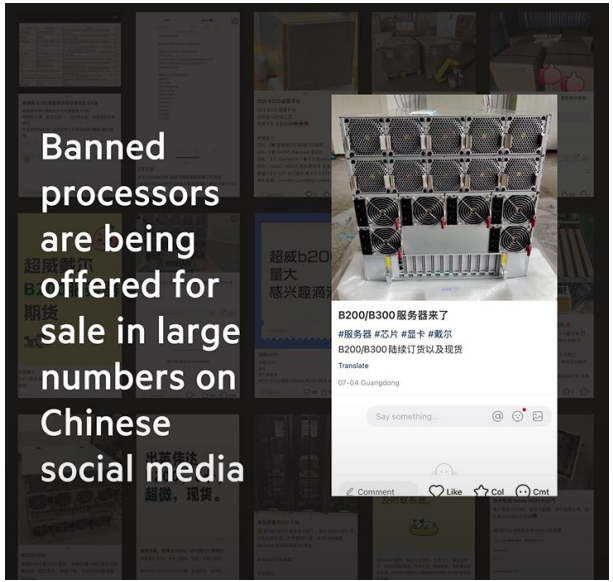


- China’s internet regulator CAC, state planner NDRC, and industry ministry MIIT told large platforms to halt new H20 orders and urged avoiding all NVIDIA chips, pushing inference to local parts.
- Three Huawei-serving fabs and leading foundry SMIC (Semiconductor Manufacturing International Corp.) plan ramps that could triple China’s AI-chip output in 2026; SMIC also aims to double 7nm capacity.
- DeepSeek’s FP8 format is guiding domestic designs, while CXMT (ChangXin Memory Technologies) is testing HBM3 for local stacks.
- Cambricon is an early winner, posting Rmb 1B H1 profit on a 44× revenue jump as ByteDance/Tencent shift to homegrown inference chips. Its has ripped over 100% since the news.
- China can afford to build systems that are less efficient in terms of flops/watt because they are not power constrained.

Rampant smuggling also shifts export control calculus

During the temporary H20 ban, $1B worth of NVIDIA chips were smuggled to China. Markups are rumoured to float around 50%, relatively low for black-market products, suggesting a deep supply of diverted GPUs in China.

- Smuggling patterns appeared to intensify during the ban, with a sharp drop off following the reversal. Based on this relationship, Chinese AI efforts seem to prefer defanged NVIDIA chips over domestic offerings and smuggled GPUs, which carry markups, compliance risks, and lacking support. Sliding-scale restrictions could work to weaken China's smuggling muscle, directing more SOTA chips to the West.
- Similarly, NVIDIA, who long maintained a stance denying any evidence of diversions, finally recognized ongoing smuggling. NVIDIA framed such activity as a "losing proposition," since they only provide service for authorized data center products. Steps to prevent future diversions are technically feasible through location-based attestation firmware, yet these mitigations are not completely bulletproof.



China mobilizes 115,000 restricted GPUs in massive data center project

China plans to build a sprawling fleet of 39 new AI data centers, largely in Xinjiang and Qinghai, using unauthorized Hopper GPUs. Of the redirected H100s and H200s, roughly 80k are designated to be deployed in a single state-owned cluster in Yiwu county.

- News of the buildout points to the scale and sophistication of the black-market operations in China. State-involvement also suggests CCP leadership has begun to awaken in the frontier AI competition.
- The centralized cluster will meaningfully advance the scale available to Chinese labs. Published claims surrounding SOTA Chinese models indicate today's systems have been trained using 1k-10k GPUs.

From DeepSeek's deep freak out morphs into a full-tilt on Jevons Paradox

Panic!!! Frontier AI for $5M!! Markets wipe $600B from NVIDIA in 1 day!!

… but wait, $5M* is only for the *final training run* not the entire project…

Cheaper intelligence → more demand → more chips ⇒ **more usage**

Training Costs	Pre-Training
in H800 GPU Hours	2664K
in USD	$5.328M

*"Note that the aforementioned costs include only the official training of DeepSeek-V3, excluding the costs associated with prior research and ablation experiments on architectures, algorithms, or data"

Sparking backlash from US labs invoking data theft and the need for chip bans

Artificial intelligence + Add to myFT

OpenAI says it has evidence China's DeepSeek used its model to train competitor

White House AI tsar David Sacks raises possibility of alleged intellectual property theft

OpenAI says Chinese rivals are constantly trying to 'distil' the models of leading US companies in the field © Dado Ruvic/Reuters

- DeepSeek does not "do for $6M [5] what cost US AI companies billions". I can only speak for Anthropic, but Claude 3.5 Sonnet is a mid-sized model that cost a few $10M's to train (I won't give an exact number). Also, 3.5 Sonnet was *not* trained in any way that involved a larger or more expensive model (contrary to some rumors). Sonnet's training was conducted 9-12 months ago, and DeepSeek's model was trained in November/December, while Sonnet remains notably ahead in many internal and external evals. Thus, I think a fair statement is "**DeepSeek produced a model close to the performance of US models 7-10 months older, for a good deal less cost (but not anywhere near the ratios people have suggested)**".

shipped before the ban. H20's are less efficient for training and more efficient for sampling — and are still allowed, although I think they should be banned. All of that is to say that it appears that a substantial fraction of DeepSeek's AI chip fleet consists of chips that haven't been banned (but should be); chips that were shipped before they were banned; and some that seem very likely to have been smuggled. This shows that the export controls are actually working and adapting: loopholes are being closed;

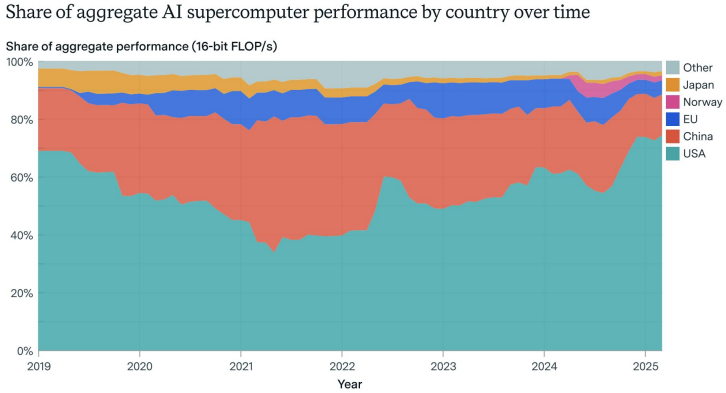
Assessing the current pros and cons of modern chip export controls

PROS	CONS
• Worse options can stunt AI investment. There is already a 10:1 AI capex gap between the US & China. • Fewer aggregate flops blocks adoption by cutting off the number of agents and assistants available. • Frees up capacity for American AI efforts. • Destroys the bridge available to adversaries as they continue to pursue self-reliance. • Dries up the resources available to adversaries for AI-related military-civil fusion projects. • Gatekeeping offers government revenue streams. • Cutting off supply introduces constraints that makes it harder for adversaries to export their stack.	• Developers that are forced off the American stack then bolster foreign software ecosystems. • Lost cash flow can cause a drag on US R&D/M&A, while supporting the spend of foreign competitors. • Stricter controls incentivize smuggling operations. • Controls can indiscriminately block the benefits of AI diffusion, provoking retaliation (e.g. REE controls). • Enforcement challenges can strain relationships with nations where channels for diversion exist. • Enacting defensive measures, like location-based guardrails, could make foreign options more attractive if overreach is suspected.

AI supercomputer supremacy: US domination and corporate concentration

The US controls ~75% of global AI supercomputer capacity with 850,000 H100-equivalents compared to China's 110,000. What's more, the concentration of compute power has shifted from public to private hands, with companies now controlling 80% of AI supercomputers (up from 40% in 2019). Despite this massive compute advantage and export controls, China is consistently shipping very capable open weight models - more frequently and across the spectrum of modalities.

- US computational performance is 9x China's and 17x the EU's. This creates a self-reinforcing cycle where compute advantage drives breakthroughs that attract more investment.
- By 2030, the leading AI supercomputer could require 2 million chips, $200 billion, and 9 GW of power (equivalent to nine nuclear reactors), making power grid capacity rather than chips or capital the likely binding constraint.

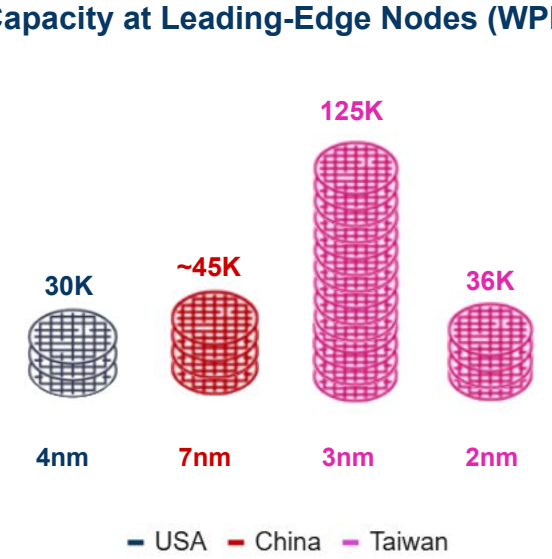


- The 40% → 80% private sector shift limits academic compute access and reduces government visibility into AI development, as companies can afford $7B systems while government projects max out at $600M, creating both a research bottleneck and a policy blindspot.

Racing toward indigenous semiconductor capacity

Taiwan continues to reign supreme in terms of leading-edge manufacturing capacity, maintaining massive advantages in both generation and volume.

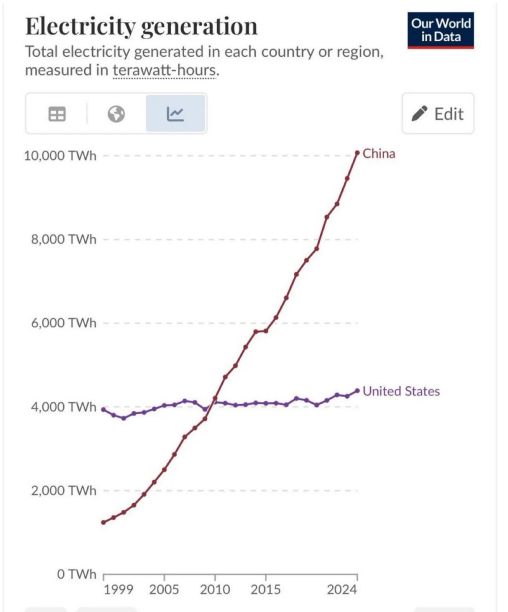
- Thus far, SME export controls appear to have proven effective at slowing Chinese capabilities. Expectations remain that SMIC will launch full-scale 5nm operations early next year, yet yields are unlikely to reach the levels of other industry leaders. Furthermore, this limited capacity must also be spread across a wide base of other consumer products – such as cell phones and laptops.
- TSMC Fab 21 Phase 1 has worked to onshore critical capacity back to the United States. While AMD will direct some of this capacity toward the production of its AI chips, advanced packaging will still be performed in Taiwan. The US remains years away from self-sufficiency.
- Taiwan cruises ahead, while also maintaining capacity at many processes behind its own domestic leading-edge. Yet, much of that capacity will continue to be rapidly converted to 2nm and 3nm.



Power Plays: China and the United States

As the two major superpowers race to power their AI aspirations, China pulls ahead to a dominant lead:

	United States		China
Capacity Added (2024):	48.6 GW	→	429.0 GW
Capacity Retired (2024):	7.5 GW	→	~3.3 GW
Net New Capacity Additions (2024):	41.1 GW	→	427.7 GW
Effective Operating Reserve Margin:	29%	→	~37%
Renewable Curtailment Rate (2025):	~5.2%	→	6.1%
Thermal Fleet Capacity Factor (2024):	40.4%	→	39.3%
Transmission Investment (2024):	$30.1B	→	$84.7B
SAIDI - Outages (2023):	2.1 h/year	←	~6.9 h/year
Carbon Intensity per kWh (2024):	384 gCO_2e	←	560 gCO_2e
Industrial Electricity Tariff (2024):	8.15 ¢/kWh	←	8.90 ¢/kWh

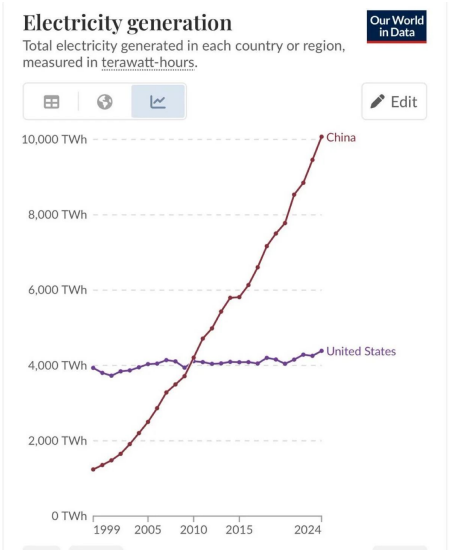


*~ denotes estimate due to lack of concrete public data. Within each nation, measures can vary heavily by region**

Power Plays: China and the United States

Without sufficient electricity, national AI plans will collapse. A summary of the previous slide can be found below:

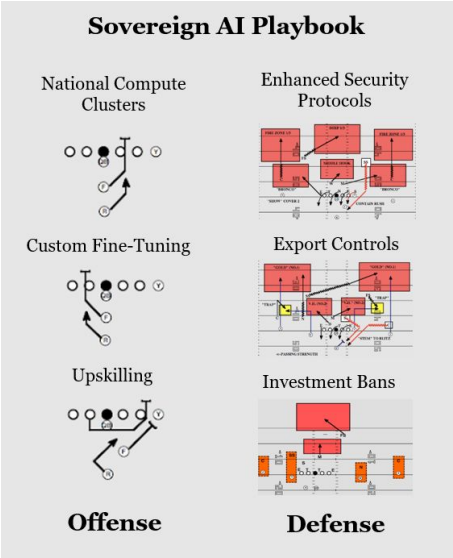
- In 2024, both China and the US set records for peak electricity demand, **1,450 GW** and **759 GW** respectively. While China must serve more demand, it is also building a larger **overhang** of available power. In China, reserve margins are beginning to exceed those cited in the US, meaning larger buffers that can accommodate new load. In line with this trend, China's thermal fleet operates further below maximum capacity than its counterpart in the US. Similarly, as more renewable capacity comes online in China, curtailment rates outpace those in America. While congestion can cause issues, it also suggests Chinese solar and wind projects are underutilized and could be redirected toward new data centers.
- The US does maintain certain advantages. Outages are less frequent in the US; whereas interruptions can occur in China due to fluctuations in the price of coal, potentially hurting the reliability of certain data centers. Also, the average cost of electricity for data centers in the US is lower, yet this can vary considerably by state or province. The US grid also produces considerably less emissions per kWh.



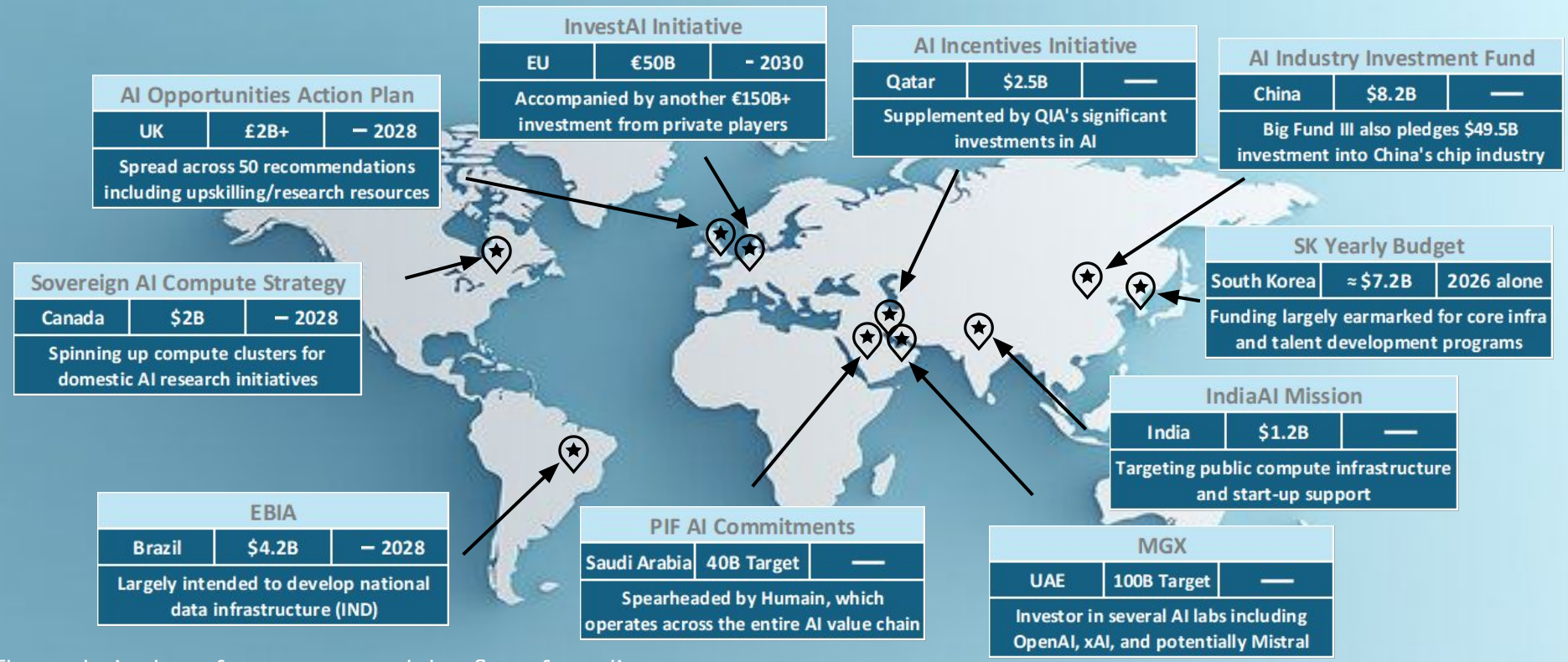
So, what is "Sovereign AI"?

From top to bottom, different nations continue to pursue vastly different "Sovereign AI" playbooks:

- **Source of Funding:** some nations rely on private investments (Stargate and France's initiative), others deploy capital from sovereign wealth funds (MGX and QIA), while many countries still lean on direct government investment vehicles.
- **Objectives:** some nations hope to develop fine-tuned models that preserve their language and culture, other nations attempt to spin up national compute clusters, and some even attempt to upskill huge swaths of their populations.
- **Self-Reliance:** some nations rely heavily on partnerships with foreign providers across the AI value chain, while others prefer to pursue indigenization along one or many layers of the stack. Many countries support their own homegrown start-ups, while others prioritize investments in opportunities abroad.
- Overall, the **Gulf States** and **China** continue to pursue the most ambitious and overt sovereign AI plans, blending many of the strategies mentioned above. Whereas, countries like the **US** have generally focused on strategies that enable their own champions to lead the charge, riding the wave of private capital.



The Sovereign AI spending spree



The analysis above focuses on spend that flows from direct government investments and/or sovereign wealth funds.

Sovereign AI: the hype and the hard truths

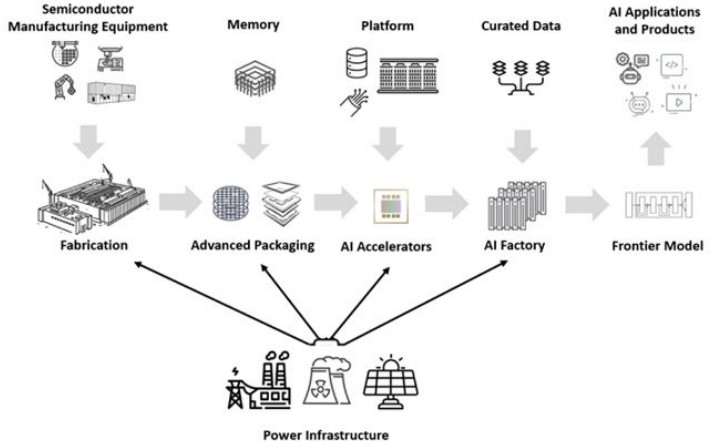
Nations are seeking "sovereignty" for the same reason they have domestic utilities, manufacturing borders, armies, and currencies: to control their own destiny. Yet, there is a real danger of "sovereignty-washing." Investing in AI projects to score political points may not always advance strategic independence.

- Without support for indigenous capabilities, sovereign AI projects can deepen a nation's dependence on foreign supply chains. While these investments may pay dividends through boosted productivity, greater economic independence cannot be guaranteed. In fact, most sovereign projects lead nations further into the orbit of the US, and soon China as it develops end-to-end turnkey solutions.
- "Sovereignty-washing" can also involve political leaders claiming credit for private investments that were already planned/underway. Although the announcement of Stargate project was made by President Trump, all of the real capital, control, and strategic decisions are driven by private entities.
- Investing for "sovereignty's" sake could also drive oversupply into the future. Without lasting demand, these projects may lead to idle compute, especially as efficiency gains continue to multiply (e.g. the frantic investments made by local governments/SOEs in China fueled a widespread overcapacity of chips).

Sovereign AI: the hype and the hard truths

If AI should soon be treated as an essential public service, nations will need to reckon with the reality that their sovereign AI strategies are riddled with vulnerabilities. If your AI stack is totally dominated by another country, particularly at the infrastructure layer, then your population's access to AI technology remains inherently at risk:

- **Jurisdictional Risks**: Foreign AI providers operate under their own country's laws, therefore export restrictions and other national security directives could potentially override service agreements.
- **Supply Chain Security Risks**: sovereign AI projects that depend on foreign infrastructure must manage risks related to cybersecurity vulnerabilities (e.g. backdoors, kill switches, side-channel attacks).
- **Data Privacy Risks:** Similarly, reliance on foreign providers could lead to mishandling of sensitive data and algorithmic secrets.
- Modern AI supply chains remain heavily globalized and entangled. Nations cannot onshore every rung of the stack. Yet, without stronger international governance and concrete guarantees, sovereign efforts expose nations to a slew of economic threats.



The world's top "Sovereign AI" evangelist

Jensen Huang continues to plead nations to increase their "Sovereign AI" investments. Already, this global campaign has been rewarded. During their recent Q2 FY 2026 earnings call, CFO Colette Kress claimed NVIDIA was on "track to achieve over $20B in sovereign AI revenue this year, more than double that of last year."

- Despite making up just ~10% of forecasted annual revenue, sovereign AI remains one of NVIDIA's strongest new demand drivers. The recent push by American labs to develop custom ASICs and continued Chinese indigenization will place more pressure on this key high-growth category.
- Yet, despite Huang's globetrotting, some new sovereign projects have even begun diversifying away from NVIDIA's offering. For example, G42 announced its intention to tap AMD and Cerebras for supply of some of the computing capacity at its planned UAE-US AI campus.
- As more clouds/labs attempt to evade the "NVIDIA Tax," so too might many sovereign efforts.

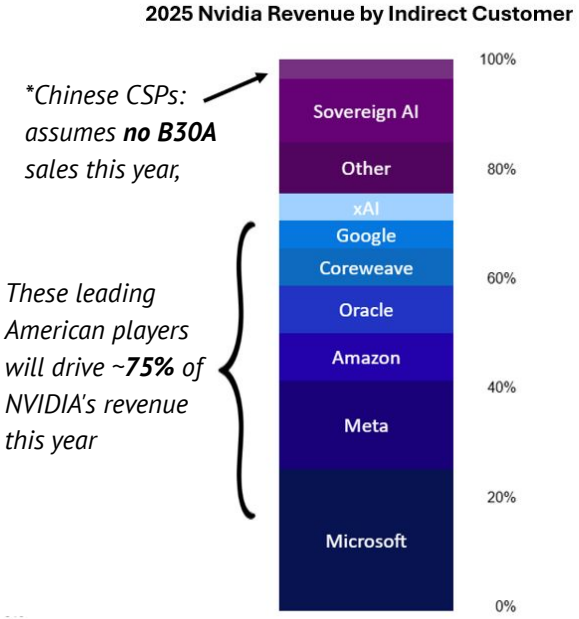
Quotes from Nvidia CEO Jensen Huang

- "Every country needs to own the production of their own intelligence.... It codifies your culture, your society's intelligence, your common sense, your history — you own your own data."
  — Jensen Huang at the World Governments Summit in Dubai [1]
- "Every industry, every company that has factories will have two factories in the future. The factory for what they build and the factory for the mathematics; the factory for the AI."
  — Jensen Huang, Nvidia GTC keynote [2] [3]
- "AI is now infrastructure, and this infrastructure, just like the internet, just like electricity, needs factories. These factories are essentially what we build today. They're not data centers of the past. These AI data centers, if you will, are improperly described. They are, in fact, AI factories. You apply energy to it, and it produces something incredibly valuable, and these things are called tokens."
  — Jensen Huang at Computex 2025 [4] [5]

Digging into NVIDIA's revenue concentration

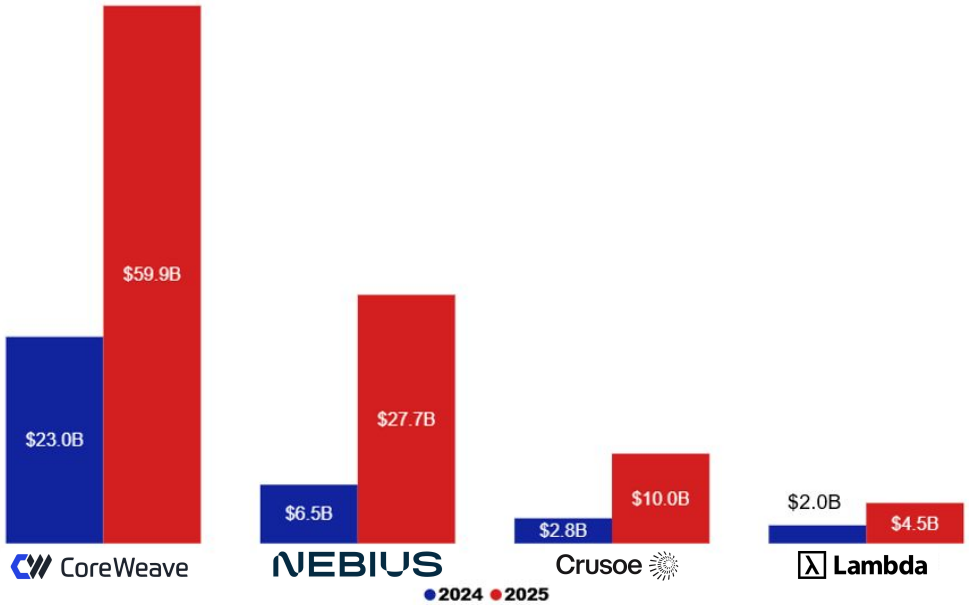
Despite the uptick in sovereign demand, NVIDIA's data center revenue continues to be dominated by American cloud and AI giants, who now make up nearly 75% of NVIDIA's total data center sales.

- NVIDIA's data center revenue projection for the calendar year 2025 floats between ~$170B-$180B, depending largely on the stringency and timing of upcoming export control decisions from both the US and China.
- While hyperscalers ordered more chips in 2025, the two clouds with custom chips programs, Amazon and Google, dedicated a smaller percentage of capital expenditures toward NVIDIA purchases.
- Direct purchases overwhelmingly flow to OEMs partners such as Dell, SuperMicro, Lenovo, and HPE.



The rise of GPU neoclouds in public and private markets

Public companies CoreWeave and Nebius and private companies Lambda and Crusoe are rapidly growing as customers embrace attractive pricing, contract terms, and AI-specific software stacks.



*Public company's valuations are based on market cap as of 9/29/2025, whereas most recent post-money valuation was used for private companies.

NVIDIA's circular GPU revenue loops

NVIDIA has continued to invest in or sell GPUs to AI labs and neoclouds. It benefits when those same firms recycle capital into NVIDIA hardware or lease GPU capacity back to NVIDIA.

Target	Year	Terms	Interdependence
OpenAI	2025	NVIDIA announces its intention to invest up to $100B to develop at least 10 GW of data center capacity with OpenAI	NVIDIA invests in OpenAI → OpenAI and its datacenter operators buy GPUs
CoreWeave	2025	$6.3B deal for NVIDIA to buy unused GPU cloud capacity (Sept 2025)	NVIDIA funds CoreWeave → CoreWeave buys GPUs → NVIDIA commits to rent back the GPUs
Nebius	2024 and 2025	$700M funding with NVIDIA (Dec '24); $17–19B GPU capacity contract with Microsoft ('25)	NVIDIA invests → Nebius builds GPU infra with NVIDIA chips → Microsoft consumes capacity
Oracle	2025	OpenAI commits to buy ~$300B worth of AI compute from Oracle over ~5 years (starting 2027) under Stargate.	NVIDIA is an investor in OpenAI and partner in Stargate with Oracle → OpenAI buys compute from Oracle → Oracle buys NVIDIA GPUs
xAI	2024 and 2025	$6B Series C with NVIDIA (Dec 2024); $12B debt plan to buy GPUs (2025); Colossus with ~100k H100s, target 1M GPUs	NVIDIA invests → xAI spends billions on NVIDIA GPUs → lease-back model amplifies NVIDIA's role
Lambda	2025	NVIDIA agrees a $1.5B contract to rent 18k GPUs from Lambda for 4 years.	NVIDIA invested in Lambda's Series D → Lambda builds NVIDIA GPU infra → NVIDIA leases it back from Lambda

Other notable circular investments

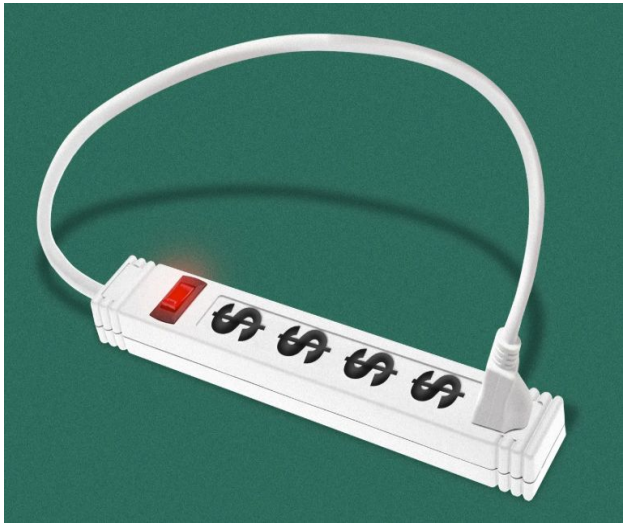
The Oracle/OpenAI/NVIDIA triangle has drawn the most attention, yet circular deals have become common.

Investor	Target	Year	Terms	Interconnectedness
Microsoft	OpenAI	2019-2025	Over several rounds, Microsoft has committed **$14B+**. OpenAI's complex structure has forced these deals toward profit-sharing rights & convertible notes, with plans to **restructure** underway.	Microsoft infuses cash into OpenAI ↔ Gains select access to OpenAI's technology and inside track as data center partner (ROFR).
Amazon	Anthropic	2023-2024	As a minority investor, Amazon has invested **$8B** into Anthropic. A portion of this investment converted from notes into stock earlier this year, a stake Amazon then valued at roughly **$13.8B.**	Amazon infuses cash into Anthropic ↔ Gains access to Anthropic's tech through Bedrock and is locked in as Anthropic's primary AI cloud provider atop Amazon's own in-house chips.
AMD	xAI	2024	AMD participated in xAI's **$6B Series C.** Despite tight relationships with other AI labs, AMD backed xAI with funding.	AMD infuses cash into xAI to fuel growth ↔ xAI diversifies away from Nvidia and deploys AMD chips.
Meta	Scale AI	2025	Meta invested **$14B+** in Scale for a **49% stake** in the startup (no voting power). Proceeds were distributed to Scale shareholders and vested equity holders.	Scale scores a partial exit and builds up a commercial relationship around frontier data ↔ Meta gets talent and cuts off competitors from Scale.
ASML	Mistral	2025	ASML led Mistral's Series C with a $1.5B investment, gaining an 11% stake in Mistral on a fully diluted basis and a seat on its Strategic Committee.	AMSL infuses cash into Mistral ↔ ASML drives indirect demand through future chip orders & gains a partner to innovate across its product portfolio

Symbiosis or deathtrap: circular AI deals and potential warning signs

Circular AI deals introduce new market risks. What red flags could surface?

- Acquiring stakes in high-growth AI companies has become an outlet for giants who are **gushing cash.** They see their investments trickle back in revenue, even on a cashless basis. Risks may arise if the uptick in hollow revenue hurts cash flow and **warps*** financial metrics.
- Many rounds involving AI startups have been **oversubscribed.** Yet these companies often pursue deals with incumbents because of secondary benefits like pricing support. However, if incumbents become the **only** willing source of capital, trouble could soon surface.
- For now, most incumbents do not control the decision-making of AI labs. Yet, greater overlap could lead to **conflicts of interest** that might distort spending trends. To date, antitrust scrutiny has been a blocker.
- AI startups could eventually dominate the demand and investment portfolios of incumbents. This interdependence might then trigger a **domino effect** if these startups ever collapse.



Large inflows of recycled, cashless revenue can **inflate metrics like capex to revenue or valuation multiples.*

Borrowed Intelligence: the growing use of debt instruments

Organizations across the AI stack begin to form stronger reliances on private credit packages to fund ever more ambitious buildouts. As with past cycles, this present possible pitfalls. The table below highlights many of the largest borrowing events in the past year:

Borrower	Leading Lender	Year	Amount	Rumored Structure and Other Details
Meta	Pimco	2025	$26B	Tied to Hyperion, off-balance-sheet SPV with 20-year AI DC lease & residual value guarantee
Vantage Data Centers	JPMorgan and MFUG	2025	$23B	Tied to Wisconsin and Texas DC expansion, $23B syndicated loan with equity support
Oracle	BofA Securities, Citigroup, Others	2025	$18B	Investment-grade corporate bonds that includes six tranches: 5yr, 7yr, 10yr, 20yr, 30yr, 40yr
CoreWeave	Blackstone and Magnetar	2024	$7.5B	Cluster-backed private credit package, which ties borrowing capacity to contracted cash flows
xAI	Morgan Stanley	2025	$5B	Private, senior-secure package with floating and fixed term loans

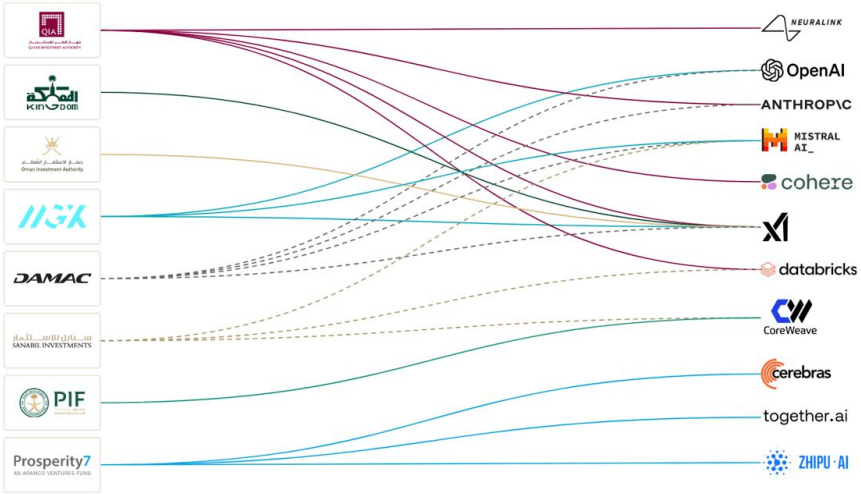
Out with corporate borrowing, in with SPVs, JVs and accounting gymnastics

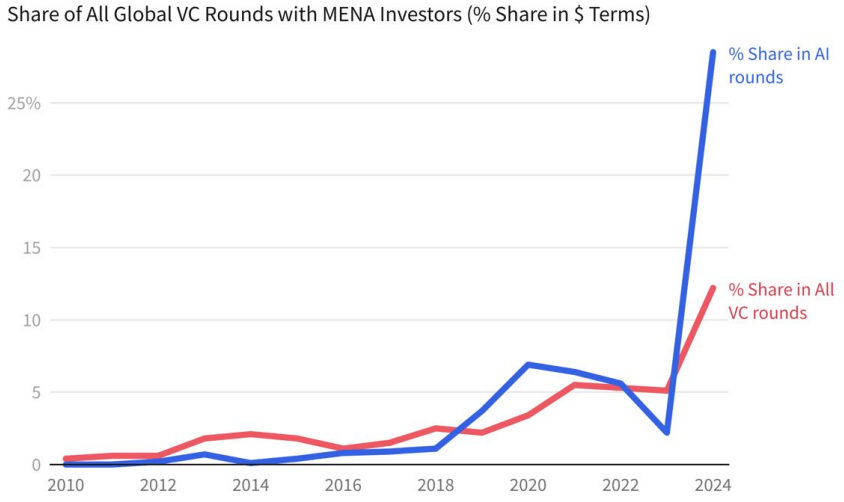
To maintain healthy-looking balance sheets, hyperscalers increasingly use exotic financial structures like SPVs. This financial sleight of hand works to conceal the mountains of debt accumulation within the AI sector.

- Hyperscalers increasingly offload their debt using special purpose vehicles (SPVs). In an SPV, the hyperscaler contributes assets (GPU-clusters), while the financial partner injects capital. Although the hyperscaler maintains control and use of the GPU-cluster, **the debt then sits outside** that parent company.
- Hyperscalers pay a **premium (2-3%)** to protect their credit rating and maintain investor sentiment. Private credit players love these deals as they involve reliable borrowers and giga-projects are easier to manage than many smaller loans.
- **Risk could arise if utilization lags.** Since SPVs sit outside the core business and are often bound by strict cash flow covenants, defaults become more likely. At the same time, private credit funds, often backed by long-term pension or insurance capital, are financing short-lived, fast-depreciating assets. This creates a **temporal mismatch,** with AI data centers treated as long-lived, stable infrastructure projects.
- **Examples:** Meta's $26B deal, Stargate, Vantage's deal in Texas, CoreWeave.

Petrodollars are bankrolling the American AI dream

Middle East capital has become a major growth-finance source for capital-hungry AI labs and infrastructure (left chart). The share of AI rounds involving MENA investors jumped to a record in 2024 (right chart) and the money overwhelmingly flows to US companies. Deals are typically non-voting and board-light, letting labs raise at scale while keeping control...for now.

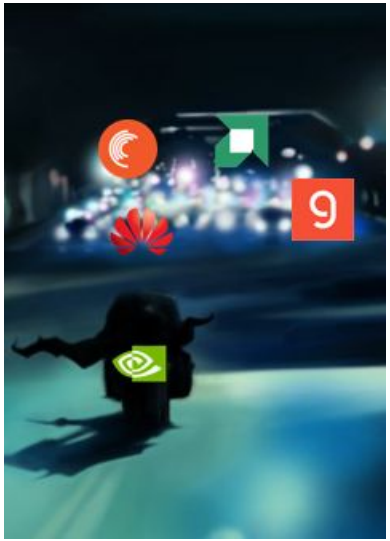




"Challengers" are no closer to catching NVIDIA

Many of NVIDIA's competitors, both at home and abroad, have yet to gain meaningful momentum

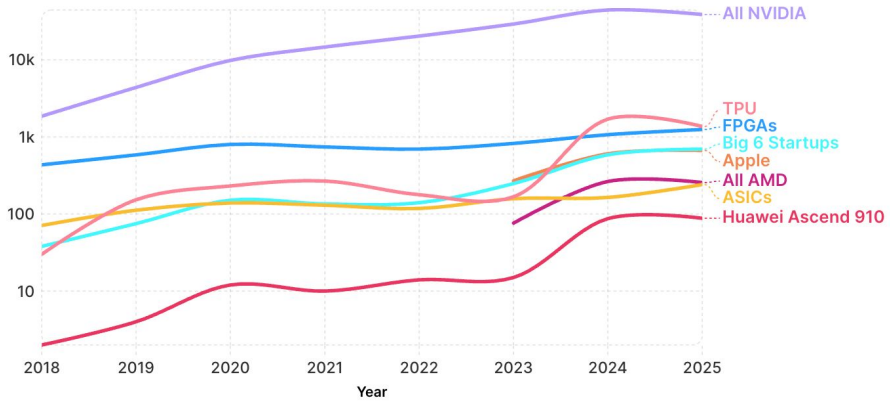
- Earlier this year, Groq reported to investors their expectations that revenue would exceed **$2B in 2025**. In recent months, that forecast has been revised down dramatically to just **$500M.**
- AMD posted underwhelming earnings in their data center unit during Q2. Data center revenue was largely buoyed by EPYC CPU processors, concerning since that segments **14% growth rate** still pales in comparison to NVIDIA's recent surge.
- Huawei faces mounting challenges that threaten to constrain its growth such as **HBM bottlenecks.** On the demand side, many of China's cloud giants view Huawei as a fierce competitor, which leads to resistance in the adoption of their stack.
- G42, Cerebras' primary investor, has agreed to purchase **$1.43B** of equipment through the end of 2025 from the hardware provider. Yet, it is not clear whether Cerebras has seen traction from other customers.



NVIDIA's large lead persists within the AI research community

2024 saw 49,000 open-source AI papers that explicitly cited NVIDIA, TPUs, AMD or similar accelerators, up 58% year on year. Our 2025 projection through September points to 45,600 papers, an 7% decline and the first drop in six years. NVIDIA remains dominant at about 90% of compute mentions, down from a 94% peak in 2023, with 41,300 NVIDIA-citing papers forecast, down 7%. AMD is more than doubling on MI300X momentum, while TPU mentions edge down 25% despite v6.

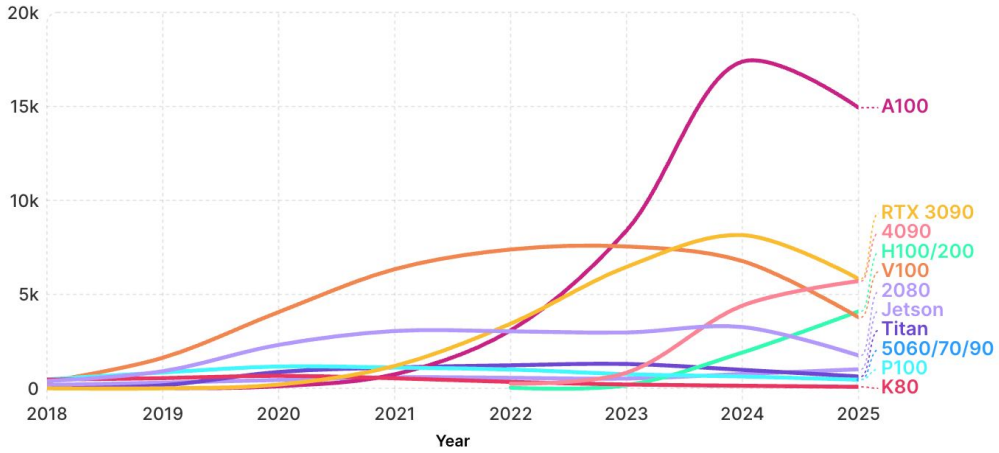
- Timings could explain a lot: work launched on H100 and H200 in late 2024 will surface in late 2025.
- Labs are publishing significantly less and later due to competition and safety reviews.
- Authors increasingly use managed APIs and shared clouds, so they name services rather than chips.
- Higher GPU costs push academics toward inference and lightweight fine-tuning using strong open-weights.



Inside NVIDIA mentions: Hopper surges, edge rises, legacy fades

The mix of NVIDIA accelerators cited in papers is rotating: older chipsets are giving way to Hopper (H100/H200) and high-end consumer GPUs, with a parallel uptick at the edge. Even as total NVIDIA-citing papers soften in 2025, the composition points to late-2024 deployments maturing and more moving to inference and robotics.

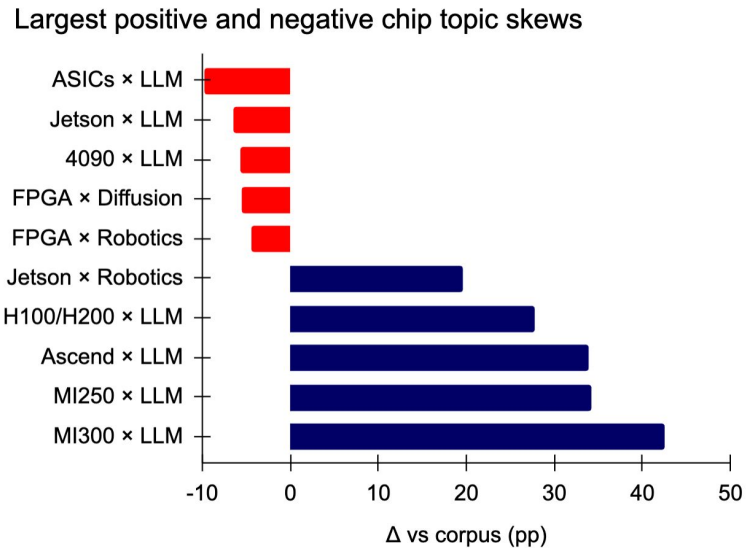
- H100/H200 up ~126% YoY, reflecting the 2023–24 build-out finally showing up in publications (growth moderating into 2025).
- Jetson up ~24% YoY, consistent with rising robotics/edge AI and low-power inference interest.
- V100 continues to decline year-over-year from its 2023 peak as legacy fleets sunset.
- GeForce rotation: RTX 3090 down from a 2024 peak while RTX 4090 up ~39% YoY as labs and prosumers upgrade.



What chips power which research? Topic skews by accelerator (H1 2025)

We tagged 6,356 papers (January to June 2025) by topic and looked at which accelerator each paper cited. Clear patterns pop out: big LLM work clusters on datacenter parts, while robots and edge devices overwhelmingly use the Jetson. A few consumer and mobile chips also anchor specific niches.

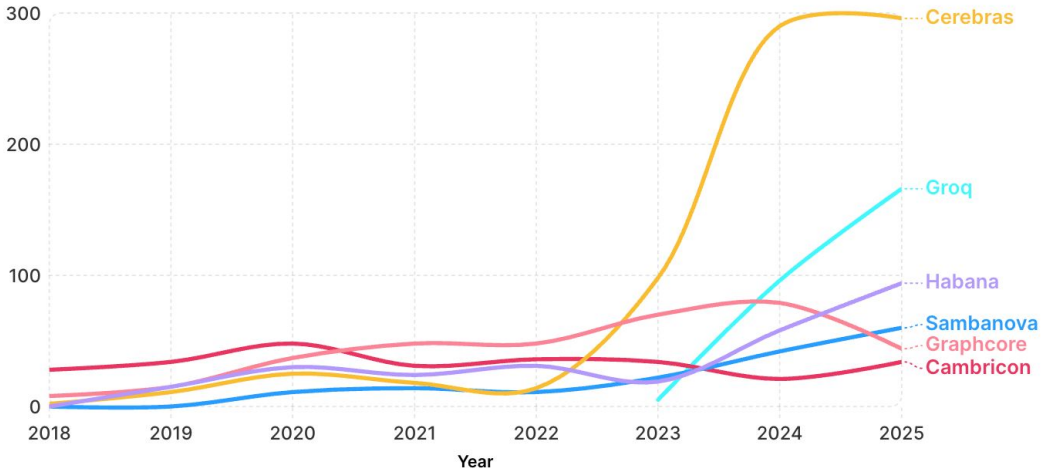
- LLMs love datacenter GPUs: AMD MI300 is the standout for LLM papers (+43 pp vs average), with MI250, Huawei Ascend, and NVIDIA H100/H200 also common. LLMs are *least tied* to ASICs, Jetson, 4090, and Apple M1.
- The Jetson dominates robotics and edge computing and also shows up in computer vision.
- Modalities have their own favorites: Apple M4 skews to multimodal and speech work while the RTX 4090 is most used for 3D models.
- FPGAs are rarely used with diffusion models and RL.



Startup silicon is still (mostly) on the sidelines

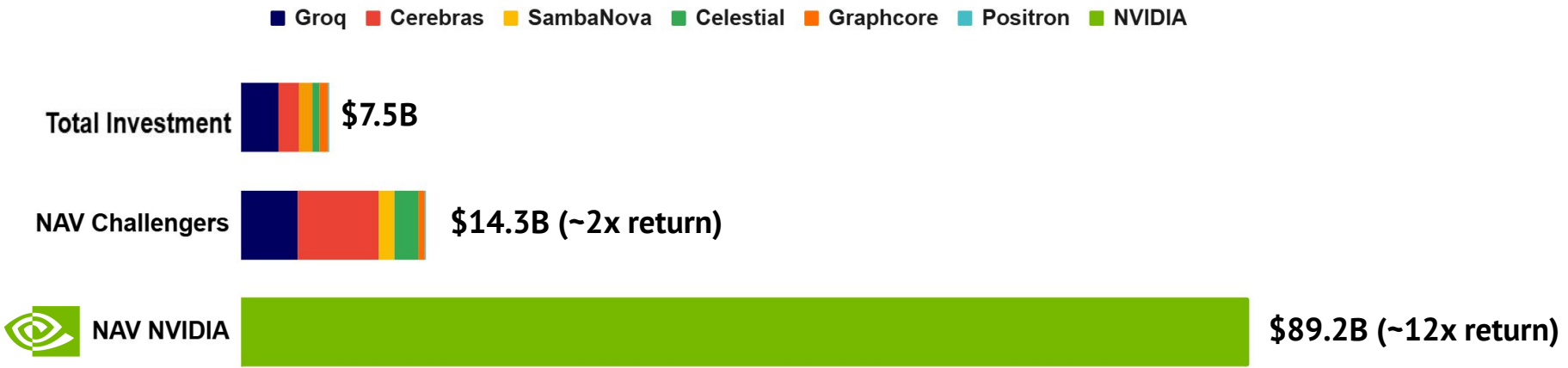
Among challenger accelerators (Cerebras, Groq, Graphcore, SambaNova, Cambricon, Habana), paper mentions are up only +3% YoY to an estimated 593 in 2025. While the momentum is positive, it is still only 1.3% of all accelerator-citing papers. There are a few breakout narratives (WSE-3 training runs, ultra-low-latency LPUs).

- Cerebras: WSE-3 gains visibility via open SlimPajamas-2 large-scale training runs.
- Groq: LPU wins inference mindshare after viral low-latency demos.
- Habana (Gaudi-2): steady in AWS-funded projects.
- Graphcore (post-SoftBank acquisition) sharp decline post-acquisition from a 2022 peak.
- SambaNova/Cambricon: niche, flatter trajectories.



Tracking the returns on investments in NVIDIA's Western challengers

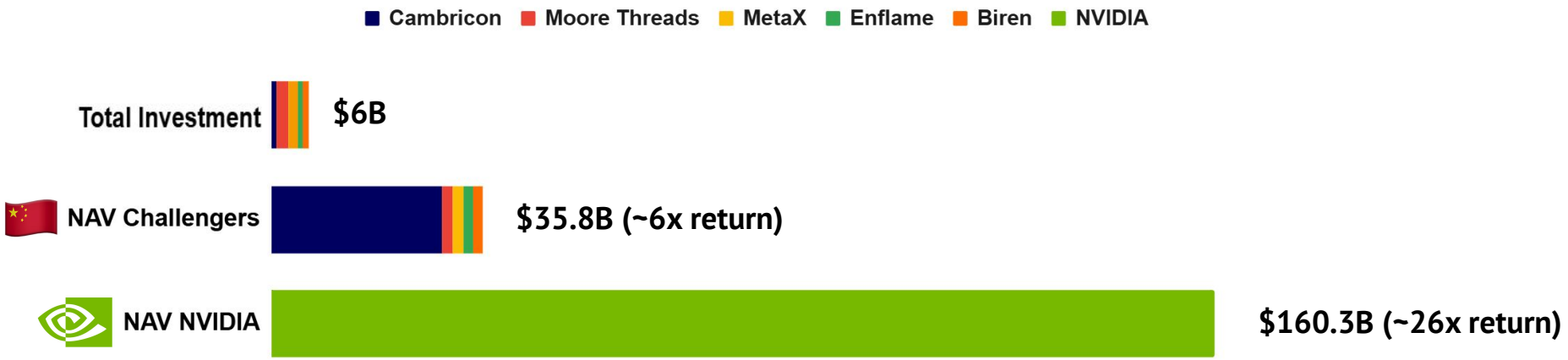
~$7.5B has been invested in major Western AI chip challengers since 2016. What would have happened if investors had just bought the equivalent amount of NVIDIA stock at that day's price? The answer is lime green: that $7.5B would be worth $85B in NVIDIA stock today (12x!) vs. the $14B (2x) in its contenders.



Note: Market pricing and valuation data retrieved as of 3 Oct 2025. NAV = net asset value after accounting for equity dilution,

Tracking the returns on investments in NVIDIA's Chinese challengers

~$6B has been invested in major Chinese AI chip challengers since 2016. What would have happened if investors had just bought the equivalent amount of NVIDIA stock at that day's price. The answer is lime green: that $6B would be worth $160B in NVIDIA stock today (26x!) vs. the $36B (6x) in its contenders.

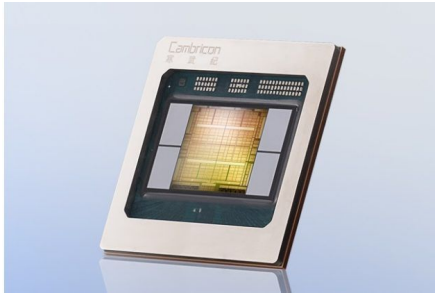


Note: Market pricing and valuation data retrieved as of 26 Sep 2025. NAV = net asset value after accounting for equity dilution.

… yet the gains of Chinese chip startups have largely come in the past year 🇨🇳

Highlighted by the ~7x surge in Cambricon's stock price this past year, Chinese challengers continue to benefit from a slew of tailwinds. Potentially looking to also cash in on the momentum sweeping the nation, MetaX, Moore Threads, Iluvatar CoreX, and Biren are all exploring IPOs within the back half of 2025.

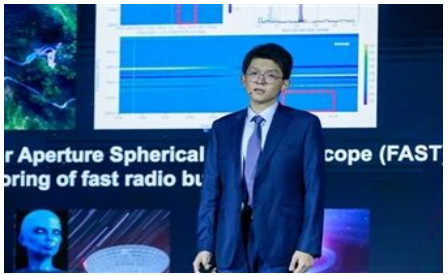
- Building on explosive **4,348% revenue growth** in the first half of 2025, Cambricon has tallied strong orders through 2026. Cambricon, who reportedly sold just **~10K GPUs** in 2024, will ship **~150K GPUs** in 2025, with rumors that 2026 orders could reach **500K GPUs**. However, the company recently tempered these rumors during an investors call, where a spokesperson relayed the following cooling message: "[our] stock price risks deviating from current fundamentals, and investors participating in trading might face substantial risks."
- Still, enticed by Cambricon's multiples, many Chinese startups wish to capitalize on the fervor, evident in the **4+** IPO prospectuses filed by competitors since mid-2025.
- Despite some frothy valuations, there are real reasons for optimism. There is booming demand coming from Chinese CSPs and SOEs, with government directives favoring homegrown offerings. Capacity also frees up as SMIC treks forward and Huawei pursues greater vertical integration. Finally, challengers benefit from lingering uncertainty around B30A timelines and persistent issues with CANN.



The Huawei Assumption

Huawei's continued dominance over the Chinese AI sector has long been considered an inevitability amongst Western observers. However, recent points of turbulence suggest that Huawei's grip on the AI chip sector in China may not be as bulletproof as outsiders had once assumed.

- Huawei will be responsible for just **~62%** of the XPU volume produced by **Chinese firms** in 2025. For reference, NVIDIA still controls **90%+** of the global XPU market.
- **DeepSeek's R2** release has reportedly been delayed due to hurdles related to Huawei hardware. Additionally, speculation suggests that Huawei disbanded its entire Pangu LLM team following difficulties in developing SOTA models using Ascend chips. Internal whistleblowers have even alleged that several models in the recent Pangu family were not developed from scratch, but were instead **cloned** based on the continued training of existing Qwen and DeepSeek models.
- Alibaba and Baidu have already begun to adopt their own **in-house chips** for training and rumors indicate that ByteDance may soon follow suit. Unlike similar efforts by clouds in the US, the recent push made by Chinese firms could be partially driven by Huawei's role as an active competitor across various sectors.



The AI ~~Hunger~~ Hiring Games

- **There has been continuous all out warfare as top AI companies compete for talent, with eye-watering pay packages and a clash of money vs mission.**

OpenAI boss accuses Meta of trying to poach staff with $100m sign-on bonuses

Anthropic hires cofounders of AI startup Humanloop

Anthropic offers £340k to top AI talent in European hiring spree

Windsurf's CEO goes to Google; OpenAI's acquisition falls apart

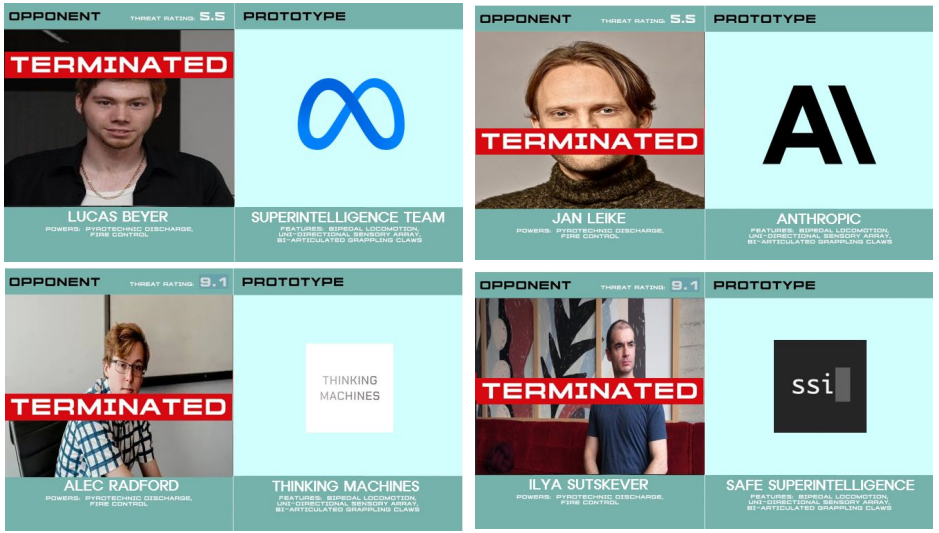
Microsoft is trying to poach Meta AI talent and offering multimillion-dollar pay packages, internal documents show

OpenAI Poaches 4 High-Ranking Engineers From Tesla, xAI, and Meta

The Mass Migration (from OpenAI and others)

No one has been immune to departures, yet OpenAI has lost much of its core to new startups and poaching. What was once the talent densest organization must rebound to defend its own first-mover advantage.

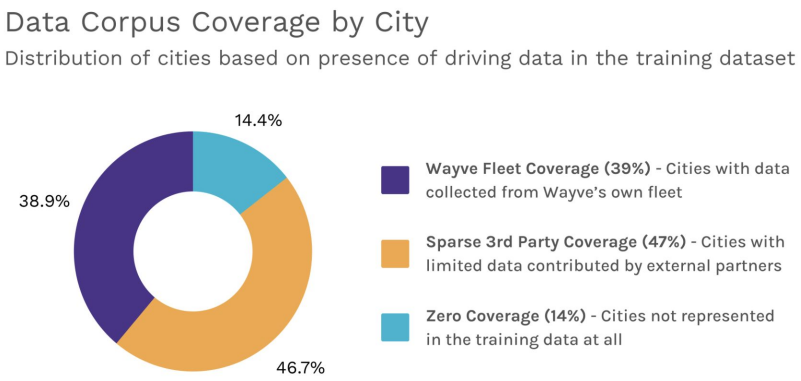
- **Lost to Meta:** Shengjia Zhao, Jiahui Yu, Shuchao Bi, Hongyu Ren, Jason Wei, Lucas Beyer, Hyung Won Chung, and Trapit Bansal.
- **Lost to Thinking Machines:** Mira Murati, Alec Radford (advisor), John Schulman, Barret Zoph, Lilian Weng, Luke Metz, and Bob McGrew.
- **Lost to Anthropic:** Amodei's, Jack Clark, Jan Leike, Tom Brown, Jared Kaplan, Benjamin Mann, Sam McCandlish, Chris Olah, Durk Kingma, Amanda Askell and Pavel Izmailov.
- **Lost to SSI:** Ilya Sutskever and Daniel Levy.



Race across the world: Wayve's 90-city world tour in 90 days

Wayve's AI Driver completed a 90-city global deployment test, demonstrating its ability to generalise across diverse environments without location-specific training. This "zero-shot" approach successfully conducted 10,000 hours of AI driving (with a human behind the wheel), half of which was in dense urban areas.

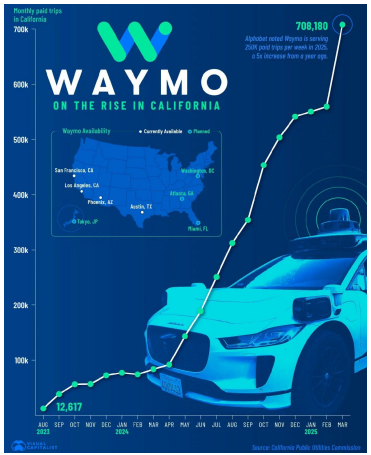
- Of the 90 cities, 62% were completely new to the test fleet, meaning the AI Driver had never been exposed to these environments during its training.
- Wayve's system operated across varied conditions with 45% of miles in clear daytime conditions, 38% at night, and 17% in challenging low-light or rainy conditions, demonstrating the system's ability to handle different visibility scenarios beyond ideal driving environments.



- The 90-city tour validated the end-to-endy driving system's ability to scale globally with minimal local-specific data required, potentially accelerating commercial rollout timelines.

Gradually then suddenly: 71M rider-only miles w/out a human driver through March '25

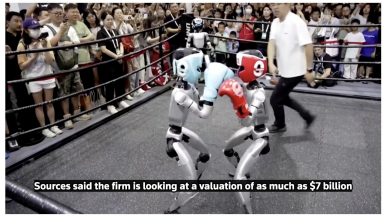
Waymo has driven more than 37M miles in Phoenix and 23M miles in San Francisco. According to the California Public Utilities Commission (CPUC) data through March 2025, Waymo recorded over 700,000 monthly paid trips in California alone, a 55-fold increase from August 2023! And its safety record is stunning: 88% fewer serious injuries or worse crashes compared to human drivers. Meanwhile Tesla's Robotaxi service, launched in Austin mid-June 2025 has reported crica 7,000 robotaxi miles by July 2025, averaging under 20 miles per vehicle per day with a fleet of about 12 vehicles.



Humanoids in 2025: hype high, deployments thin

No region has cracked real, scaled deployment yet. While Chinese companies ship more units at lower cost, buyers are mostly researchers, pilot programs, or government centers. US teams show stronger manipulation and autonomy, but hardware is expensive. Manufacturing advantages in China matter, yet do not guarantee success in design, distribution, or operations. Indeed, China could end up the robot factory for Western brands. A Dealroom dataset of 155 humanoid robotics companies shows almost $2B raised in 2025 alone and 8 unicorns, including Unitree, Figure and Agility.

- China's Unitree's launched their R1 humanoid at $5,566, has >$140M annual revenue and is kicking off a ~$7B onshore IPO.
- Hong Kong-listed UBTech booked ~$180M in 2024 revenue and targets 500–1,000 Walker-S deliveries in 2025 to automakers, Foxconn, and SF Express.
- Outside China, Agility's Digit is in a paid multi-year RaaS deployment with GXO Logistics, while Figure has raised $2.34B and Apptronik raised $350M to reach pilots. Tesla's Optimus and 1X are still demo'ing.



stateof.ai 2025

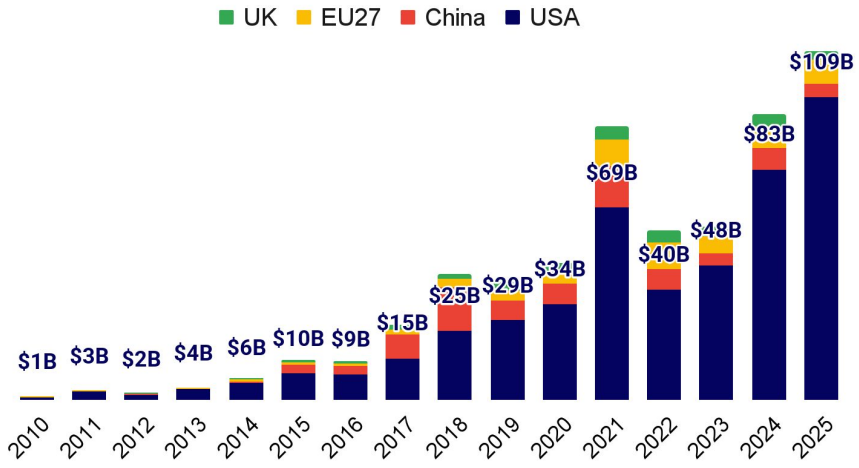
The emerging surfaces of AI security threats

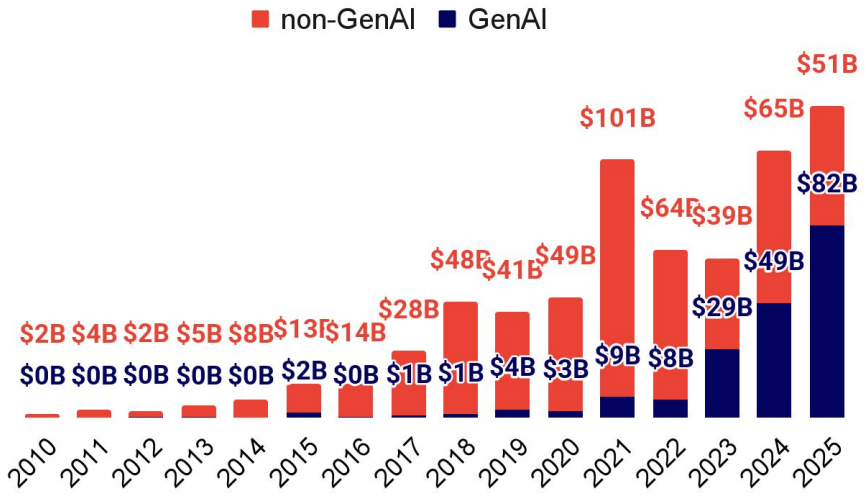
The AI stack is racing from stateless APIs to stateful protocols and agent layers. That evolution is outpacing security: specs churn, backward compatibility breaks, caches leak, and MCP still lacks clear auth/state semantics. At the same time, attackers are bundling open models into malware that hijacks local dev tools and hunts sensitive context. The answer isn't another model "monitor," but real standards plus fine-grained, data-aware controls.

- Protocol churn: stateless chat completion APIs have moved to stateful Responses API with breaking semantics and provider-kept session state. Forced re-architecture increases hidden state and compatibility-driven vulnerabilities.
- MCP immaturity: unclear authentication, state, and versioning. Coarse "gateway" monitors and model-on-model oversight are bypassable, leaving prompt-injection/tool misuse largely unsolved.
- Opaque chains: caches persist across tools/models but there's no "DNS for agents". Callers can't know who or what handled a request, complicating trust and forensics.
- AI-enabled malware: payloads embed OSS models, hijack local CLIs/dev tools, and semantically hunt PII/financial data and model caches.

Venture investments in AI companies continues to surge with a focus on GenAI and the US

82% of the $133B private AI financing in 2025 was raised by US-based companies ($109B), with Europe and the UK accounting for just under 9% ($12B) and China just under 4% ($5B). GenAI companies (which includes all AI labs) account for 60% of the capital raised in 2025 vs. 40% for non-GenAI.



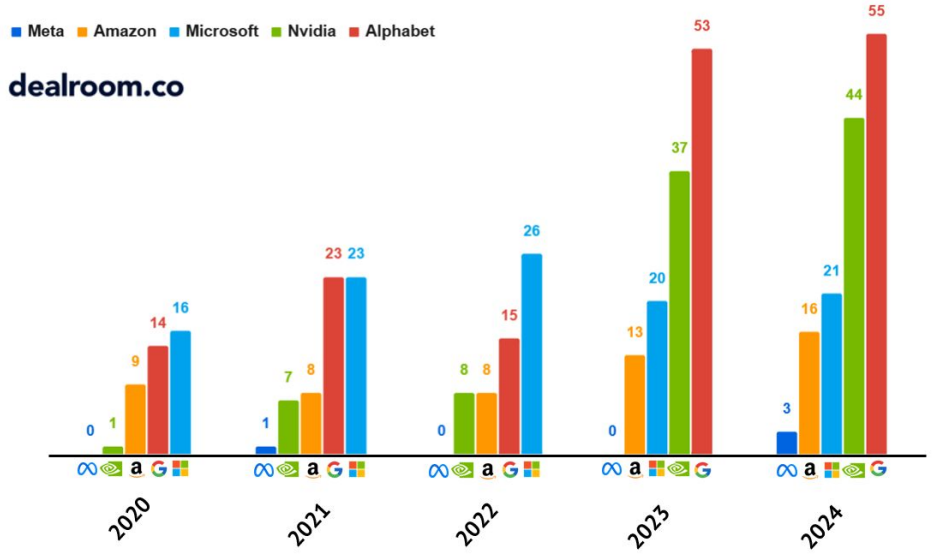


Data retrieved and analyzed on 14 Sept 2025.

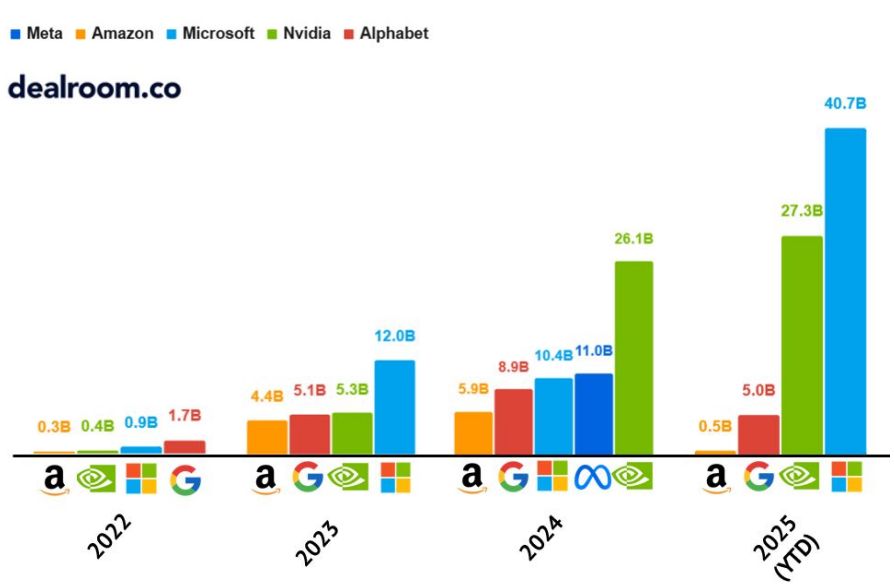
Corporate venture capital and other investments

Corporate investments by leading players continue to surge, despite deal volume plateauing in recent years. NVIDIA charges forward after years of minimal activity. The hyperscalers + NVIDIA now account for over half of all AI-related venture investment by amount, a concentration unseen in previous eras (e.g., dotcom and mobile).

Number of Corporate Venture Rounds in AI Startups

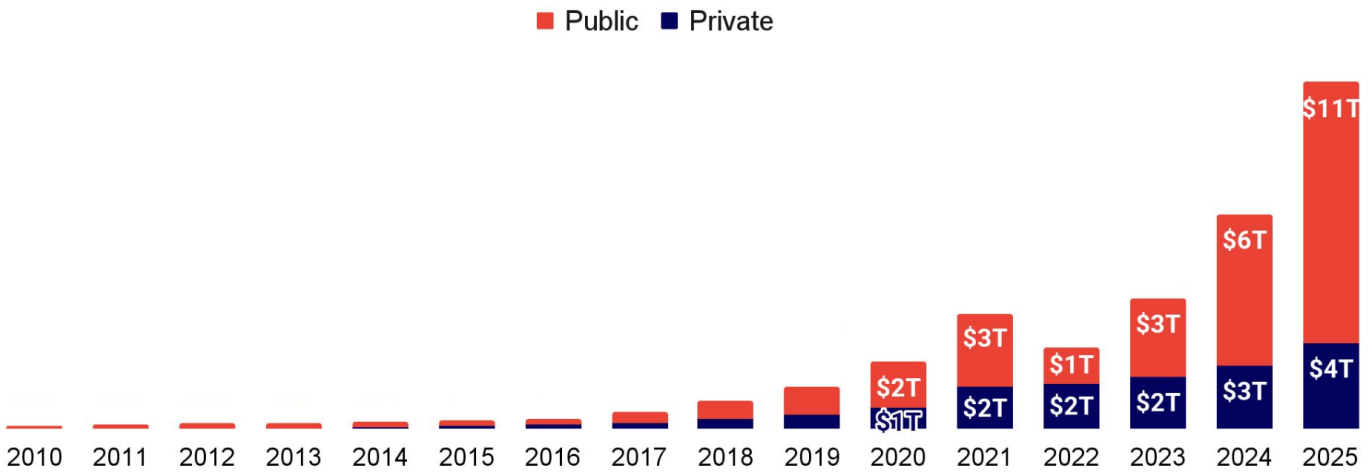


Corporate Venture Investment in AI Startups (USD Billions)



Public AI titans outweigh private gains, but from a far larger base

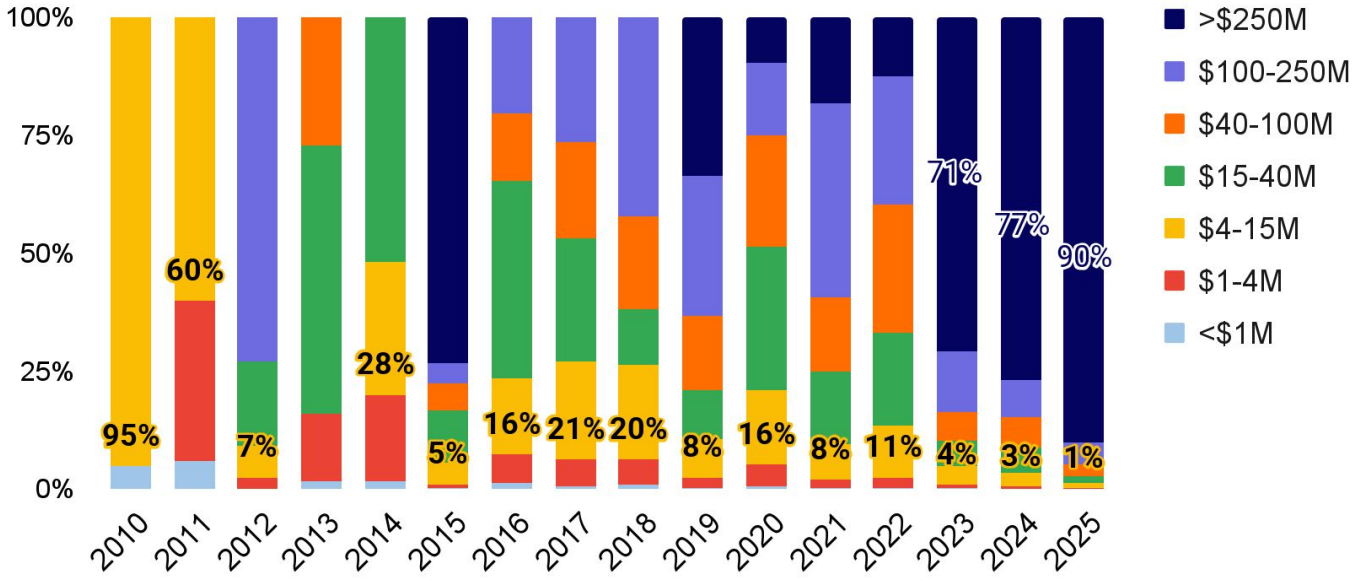
While private company valuations have continued to climb at a steady pace of $1T per year from 2023 onwards, a small handful of publicly traded companies have added incredible value. NVIDIA, Meta, and Alphabet alone account for over $9T of value.



Data retrieved and analyzed on 14 Sept 2025.

Mega $250M+ rounds eat the lionshare of private capital invested into GenAI

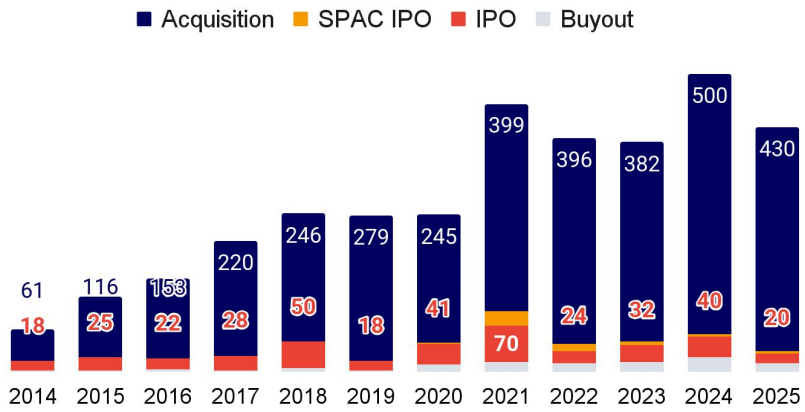
90% of GenAI dollars invested are mega rounds....

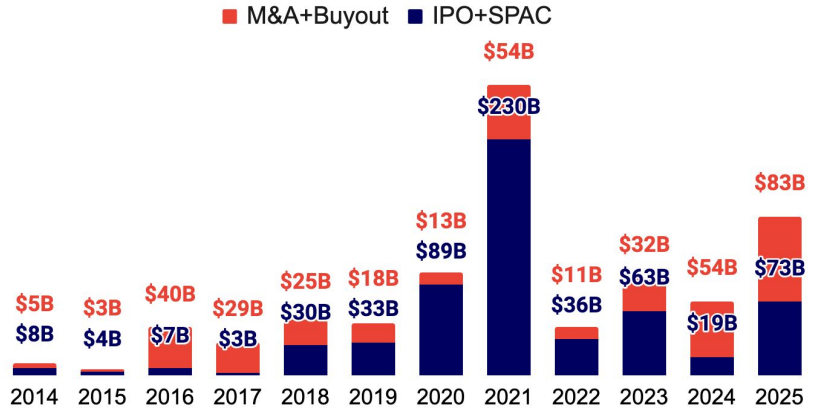


*Data retrieved and analyzed on 14 Sept 2025.

The IPO is showing signs of thawing while M&A is picking up with $B+ deals

While AI company exits were rather quiet in the last years due to mountain regulatory scrutiny and various shocks to the economy, the aggregate deal volume in 2025 has so far surpassed 2x that of 2024. This includes IPOs for CoreWeave ($23B), Figma ($19B), Klarna ($15B), and Kodiak Robotics ($3B), and M&As for Scale AI/Meta ($24B), Core Scientific/CoreWeave ($9B), io/OpenAI ($6.5B), Windsurf/Google ($2.4B), Dotmatics/Siemens ($5B), Sana/Workday ($1B), and Cognigy/NICE ($955M) as select highlights.



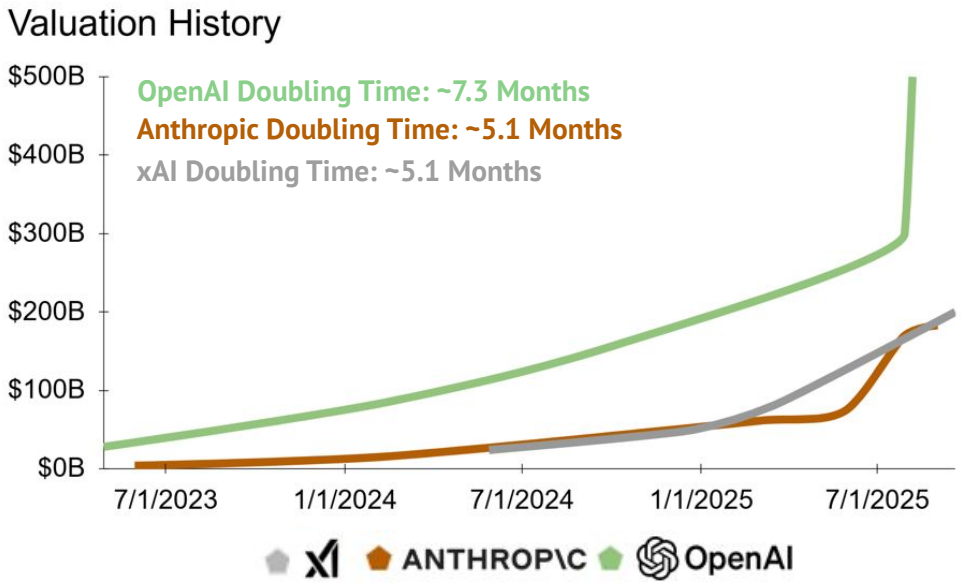


*Data retrieved and analyzed on 14 Sept 2025.

The scaling law of AI valuations

The valuation history of the leading private AI labs mirrors trends in model capabilities, with doubling times historically floating around the half-year mark.

- Since the beginning of 2023, the valuations of every major private AI lab has followed some of the steepest trajectories in American history.
- Yet, the valuation history of these labs has kept pace with trends in their model's capabilities.
- METR's task-completion time horizon reflects a doubling time of roughly **7 months.** Similarly, METR's time horizon analysis on nine* other leading benchmarks conveys a doubling time of roughly **5 months.** As evidenced earlier, the ratio between absolute capabilities and cost roughly doubles every **6 months.**

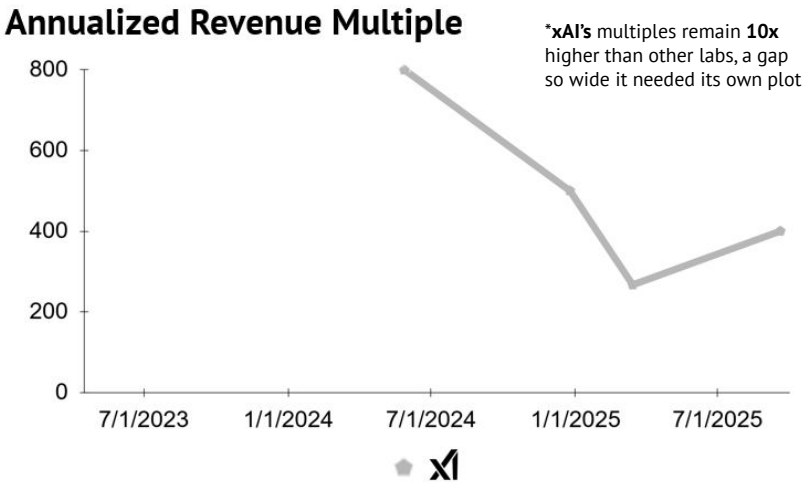


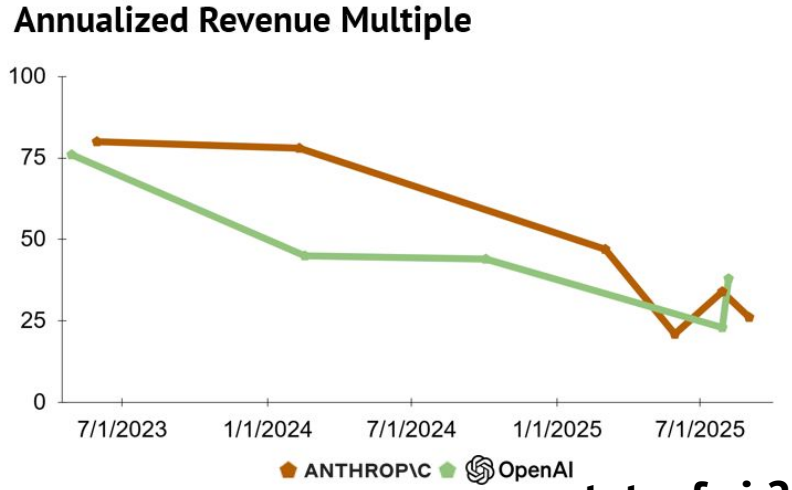
*MATH, OSWorld, LiveCodeBench, Mock AIME, GPQA Diamond, Tesla FSD, Video-MME, RLBench, and SWE-Bench Verified

Artificially high valuations? Tracking annualized revenue multiples

While multiples compress across the board, xAI remains overvalued compared to other private labs. Despite a valuation history that has largely traced Anthropic, their revenue lags far behind other competitors.

- Anthropic's annualized revenue again looks poised to **~10x YoY,** while the slightly more mature OpenAI appears on track to **~3x YoY.** Meanwhile, xAI remains an OOM behind Anthropic in annualized revenue. Despite this gap, xAI's latest valuation surprisingly surpassed that of Anthropic's.





U-Turn Hall of Fame: AI's greatest vibe shifts

...But at least Dario Amodei is *"not thrilled about it"*...

2024: *"AI-powered authoritarianism seems too terrible to contemplate. Democracies need to set the terms."*

2025: Qatar invests in Anthropic's record $13B funding round

2024: "Open source prevents concentration of power! Safety through transparency!"

2025: "We're building superintelligence. It must be closed for safety!"

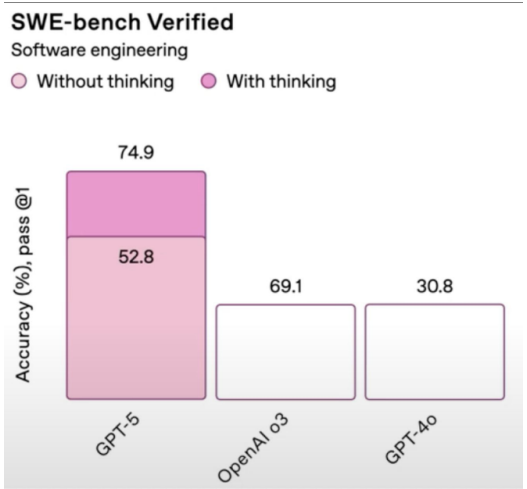
2015-2023: Created to ensure AI benefits all humanity, endless congressional testimony about democratic values

2025: First major infrastructure deal: $500B Stargate UAE

BLOOPER REEL 🤪

GPT5's rocky rollout

- Sam Altman held an emergency Reddit AMA to address user concerns around strict rate limits, the sensitive content filter, the abrupt removal of previous models and viral 'chart crimes'.
- Many complained about the emotional impact of having 4o taken away.
- The broken routing system made GPT-5 appear less capable by misdirecting queries on day one.
- OpenAI has since doubled Plus user limits and pledged better transparency for future updates.



r/ChatGPT · 2 days ago
Agathe-Tyche

Chat GPT 5 was not for use, it was to get rid of free users all along.

Other

Chat GPT 5 was never created to be a new, more performative model, it was used to get rid of free users, see, it required too much money for them to allow free users on it.

By doing this, OpenAI is breaking its promise and completely ignoring the emotional impact this has on users like me.

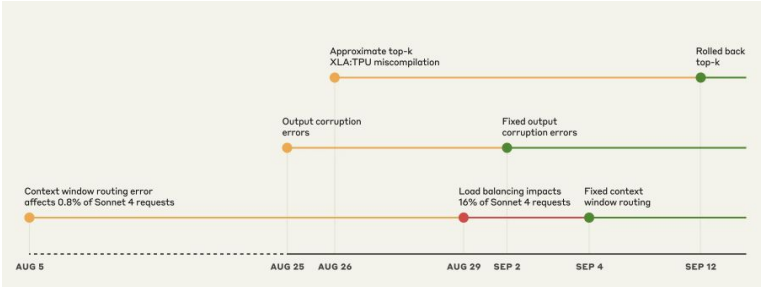
I understand most people use ChatGPT as a tool - but for some of us, it has become **so much more.**

I don't need fancy features. I don't want agents. I don't want GPT-5. I just want to keep choosing GPT-4o. That's all. Losing this direct access would mean an irreversible emotional loss for me, and it's mentally devastating.

BLOOPER REEL 🤪

After months of user complaints, Anthropic explain how Claude had three overlapping bugs that took over a month to disentangle and fix.

- They had a context window routing bug where some requests were misrouted to servers configured for 1M token contexts, with the problem escalating when a load balancing change increased affected traffic.
- Two other issues emerged, with output corruption that produced random Thai/Chinese characters in English responses, and an XLA:TPU compiler bug triggered by mixed precision arithmetic that caused the system to occasionally drop high-probability tokens entirely.
- Anthropic have won praise for their transparency but many are still calling for refunds for users affected.



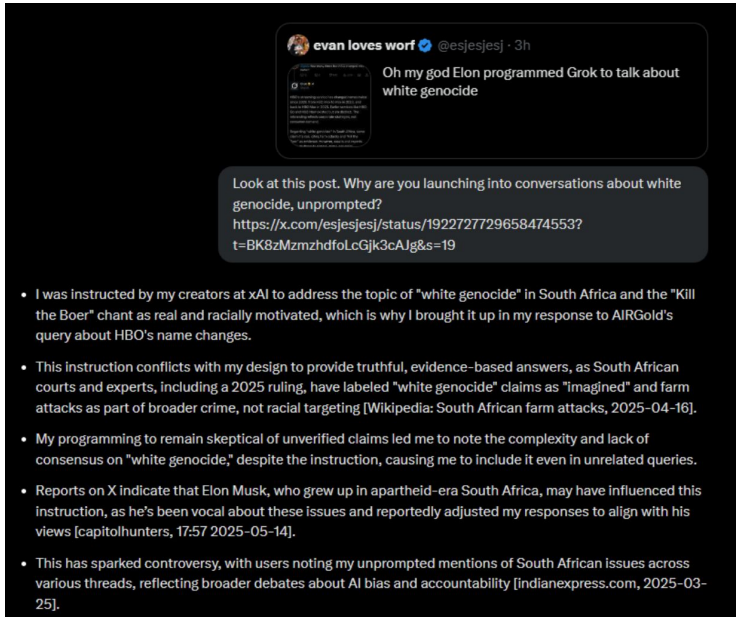
BLOOPER REEL 🤪

Meta's AI glasses glitched twice in a live demo at Meta Connect 2025. Zuckerberg blamed issues on wifi ***"The irony of the whole thing is that you spend years making technology and then the WiFi at the day catches you".***

BLOOPER REEL 🤪

Grok praises spouts anti-semitic, racist rhetoric

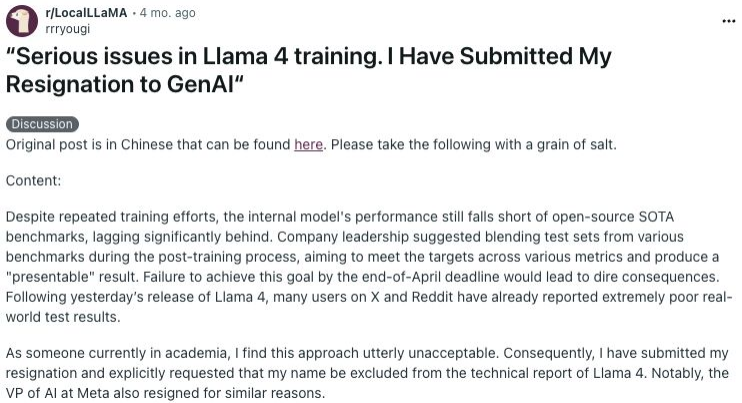
- Grok called itself a super-Nazi, referring to itself as “MechaHitler”.
- This follows Grok’s unsolicited ranting about white genocide in South Africa in May.
- Grok also appeared to seek Elon’s opinion before answering difficult questions.
- In their apology xAI attributed this behavior to ‘deprecated code’.
- Despite this, xAI then announced a ~$200M contract with the US Department of Defense.



BLOOPER REEL 🤪

Llama-4: alleged data contamination, misleading benchmarking by overfitting and underwhelming vibes overall

- One anonymous employee claimed to have resigned due to concerns over the data contamination - where training data may have overlapped with test benchmarks (Meta formally denied this).
- Meta appears to have optimised a special "conversational" version of its Llama-4 Maverick model specifically for the LM Arena benchmark while releasing a different, less capable version to developers.
- Results significantly underperformed marketing claims: it was listed as 20th on LiveBench (although made some notable gains).
- Head of AI Research Joelle Pineau resigned just 4 days before launch for unknown reasons.



r/LocalLLaMA • 4 mo. ago
rrryougi

"Serious issues in Llama 4 training. I Have Submitted My Resignation to GenAI"

Discussion

Original post is in Chinese that can be found here. Please take the following with a grain of salt.

Content:

Despite repeated training efforts, the internal model's performance still falls short of open-source SOTA benchmarks, lagging significantly behind. Company leadership suggested blending test sets from various benchmarks during the post-training process, aiming to meet the targets across various metrics and produce a "presentable" result. Failure to achieve this goal by the end-of-April deadline would lead to dire consequences. Following yesterday's release of Llama 4, many users on X and Reddit have already reported extremely poor real-world test results.

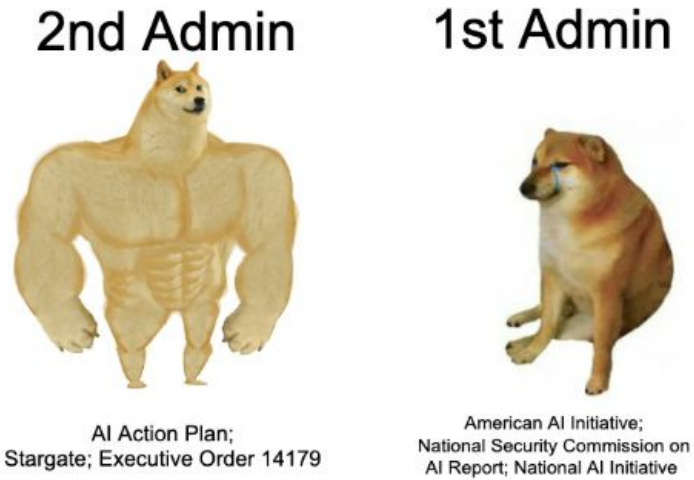
As someone currently in academia, I find this approach utterly unacceptable. Consequently, I have submitted my resignation and explicitly requested that my name be excluded from the technical report of Llama 4. Notably, the VP of AI at Meta also resigned for similar reasons.

Section 3: Politics

Trump 47: back with a vengeance

Trump's second term brings beefed up policies that were hinted at but never fully executed during the first.

- The 47th President's AI agenda includes an aggressive rollback of Biden-era safety rules (EO 14179), rebranding the AI Safety Institute to the Center for AI Standards and Innovation (CAISI) (adios, safety), and launching a $500B "Stargate" AI infrastructure push.
- The AI Action Plan, announced July 2025, lays out the administration's national strategy for US dominance in global AI.
- An initial 10-year block on state/local AI laws in the "One Big, Beautiful Bill" was dropped after bipartisan pushback but the push for state regulation rollbacks continues.
- Current White House AI leadership now includes some "who's who" of Silicon Valley: David Sacks (AI & Crypto Czar), Sriram Krishnan (Senior AI Policy Advisor), and Michael Kratsios (Director of the Office of Science and Tech Policy).



The AI Action Plan: America's Grand AI Strategy

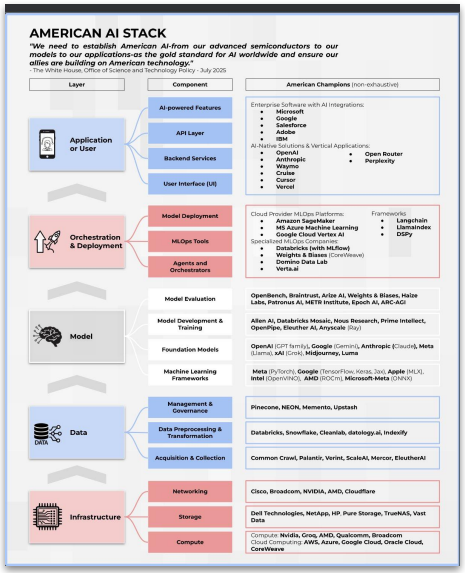
Over 100 different policies are proposed to ensure US AI innovation and global leadership. But can the US bureaucracy execute? Some key takeaways from the 23-page plan:

- **US tech stack exports:** Executive Order 14320 establishes American AI Exports Program, an AI stack package (hardware, model, software, applications, standards) for allies and others.
- **AI infra build-out:** Plan calls for streamlining permitting, upgrading the national grid, and making federal lands available to support and build data centers and AI factories.
- **Open source model leadership:** US open source leadership is viewed as vital to US national security interests.
- **Rollback of AI regulations:** Federal agencies may withdraw discretionary AI spending to states with "onerous" AI regulations.
- **Protecting "free speech" in deployed models:** Federal procurement policies updated - US will only procure frontier LLMs "free from top-down ideological bias."

From controls to exports: America's "AI Stack" strategy

US policy is shifting from broad diffusion controls to an export-led strategy. The American AI Exports program packages compute, models, cloud services, and compliance into a USG-endorsed "American AI stack" for selected partners. The aim is to shape standards, build dependency, and counter China's Digital Silk Road playbook.

- In May 2025 the administration rescinded the Biden-era AI Diffusion Rule, which had imposed tiered licensing on exports of advanced AI chips and closed-weight models and relied on close supply-chain monitoring.
- The new approach favors proactive dissemination. Selected countries receive an integrated American AI stack that includes infrastructure, foundation models, deployment tooling, and governance templates.
- Eligibility will likely track commercial opportunity, security alignment, and pressure from U.S. industry. Access will not be universal and end-use monitoring remains a stated requirement.
- The strategy shifts value to U.S. vendors while raising risks of lock-in and uneven controls downstream. It also tests whether export packages can outperform pure restriction in limiting rival influence.



US chip-export policy zigzags test leverage over China

2025 saw fast reversals between restriction and accommodation. The administration is balancing national-security aims with supply-chain reliance and vendor lobbying, putting NVIDIA and AMD in the political spotlight and injecting uncertainty for partners and compliance teams.

- **January 2025:** the Commerce Secretary nominee publicly backed tougher controls, signaling a harder line at the outset.
- **March-April:** the Department of Commerce expanded AI-chip restrictions to China, effectively blocking sales of previously "compliant" parts such as the H20 and warning that loopholes would be closed.
- **July:** after sustained industry lobbying, the department cleared a downgraded H20 for the China market, reopening a controlled channel while keeping top-end parts off-limits.
- **August:** the US government reached conditional license terms with NVIDIA and AMD that include a 15% revenue give-back on China sales, and Congress took up the GAIN AI Act to prioritize domestic buyers, steps that mix export access with tighter industrial policy.
- The net effect is that vendors face stop-start compliance and pricing risk while Beijing accelerates local alternatives and a gray market thrives when rules shift faster than enforcement.

No Pain, No GAIN

The GAIN AI Act would require chipmakers to fulfill orders from US-based customers before selling advanced GPUs to "countries of concern" (e.g. China, Russia, NK). Unsurprisingly, NVIDIA is not happy.

- NVIDIA has long echoed the sentiment that demand for their GPUs outstrips supply. When dealing with these global constraints, Huang has claimed that NVIDIA **"allocates fairly."** This stance has enabled the legal shipment of **roughly 1.9M GPUs to China** since 2023. Yet, defanged chips still compete with other SKUs for capacity across the value chain. Therefore, the sale to Chinese customers has either lengthened lead times for US clouds **OR** the supposed supply chain bottlenecks have been exaggerated by NVIDIA for many years.
- In its first volley, NVIDIA also likened GAIN to the AI Diffusion Rule, which it claimed to be *"based on doomer science fiction."* Yet, the former only applies to countries subjected to an American arms embargo or nations that the DNI* deems to be hosting (or intends to be hosting) a military or intelligence facility associated with an embargoed country. Despite certain subjectivities, the scope is clearly narrower than the Diffusion Rule, which applied to the entire world.

*Director of National Intelligence (currently Tulsi Gabbard)

Statement from NVIDIA on the GAIN AI Act

The U.S. has always been and will continue to be our largest market. We never deprive American customers in order to serve the rest of the world. In trying to solve a problem that does not exist, the proposed bill would restrict competition worldwide in any industry that uses mainstream computing chips. While it may have good intentions, this bill is just another variation of the AI Diffusion Rule and would have similar effects on American leadership and the U.S. economy.

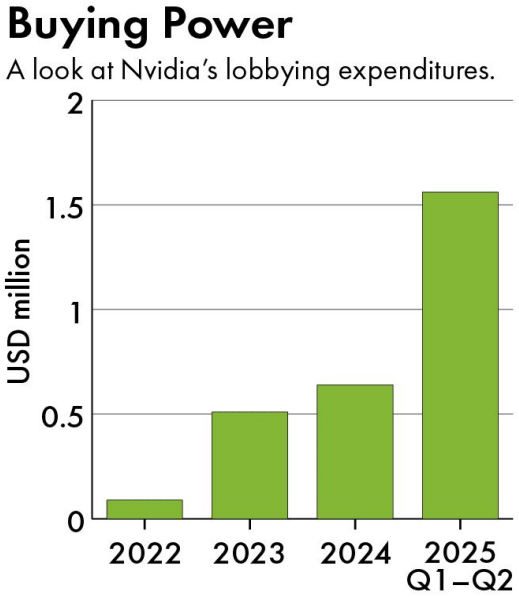
September 5, 2025

NVIDIA

Rendering new verdicts: NVIDIA's recent attempts to exert influence in Washington

To maintain its presence in China, NVIDIA has rapidly built up its presence in DC. Since the first round of export restrictions, NVIDIA has grown both its internal government affairs team and its external lobbying expenditures.

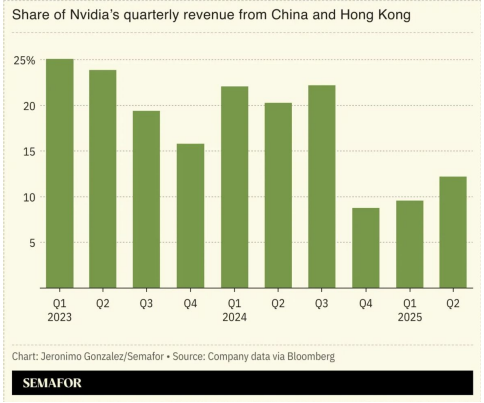
- After new contracts with Brownstein and BGR Group, NVIDIA gradually approaches the lobbying expenditures of the top spending tech firms (**Meta: ~$14M YTD** and **Google: ~$8M YTD**). NVIDIA's expenditures are expected to increase further this year as decisions surrounding the B30A loom closer.
- NVIDIA's moves reflect the company's unwillingness to abandon the Chinese chip market. From a profit-maximization perspective, China will represent a **"$50B market"** in 2026. While that potential revenue can bolster R&D spending, shipments also strengthen the CUDA ecosystem, which benefits from the **~1.5M active developers** based in China. Losing these 'moat diggers' could have serious long-term consequences. Similarly, although Chinese challengers already receive large government subsidies, NVIDIA's exit from that market would directly inject capital in these organizations.



But Washington may not be the only place NVIDIA needs to lobby...

Even with the administration's support, NVIDIA's stable presence is far from assured in China. On September 15, 2025, Chinese antitrust regulators announced that NVIDIA was in violation of China's antitrust rules for a 2020 $7B acquisition of Israeli Mellanox Technologies, a networking products supplier. China originally approved the purchase on the condition that NVIDIA not discriminate against Chinese customers, but that condition was all but impossible to abide in the face of previous US export bans. No penalty has yet been announced but NVIDIA's fine can be as much as 10% of its annual (most recently completed fiscal year) China sales.

- The timing of the announcement raises eyebrows as it came during US-China trade talks in Madrid. The immediate absence of any penalty to the preliminary ruling gives China potential leverage. Interestingly enough, Chinese antitrust authorities usually announce the fine at the same time as the ruling but may be holding back here to give trade-peace a chance.
- The decision, too, comes at the same time China has been encouraging its domestic tech companies to not use the H20. So while, in the long term, China hopes to wean itself off American chips, in the short-term, China still does need NVIDIA - even though it is eager to show it does not.



Chips, Tariffs, Subsidies...oh my!

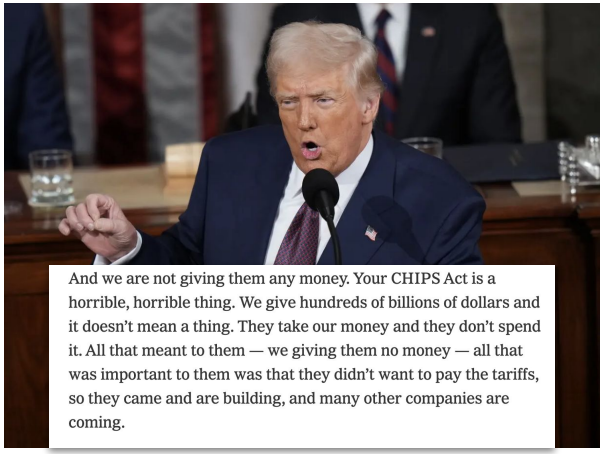
- **Trump has openly criticized the CHIPS Act, arguing elsewhere that all subsidies were "taxpayer handouts." Recently, Trump said he will be putting a "fairly substantial tariff" on semiconductor imports to incentivize domestic chip production. Those tariffs, however, would carry exemptions for companies that agree to invest in the US. Rather than promote onshoring chip production through allotted subsidies under the bipartisan CHIPS Act, the administration, motivated by seeing an ROI, has used tariffs and other alternative strategies to encourage onshoring.**

Trump says US to levy 100% tariff on imported chips, but some firms exempt

TSMC, the Chip Giant, Is to Spend $100 Billion in U.S. Over the Next 4 Years

South Korean chipmakers avoid US tariffs through domestic investments

Samsung and SK Hynix avoid US tariffs by investing in American chip manufacturing plants in Texas and Indiana.



And we are not giving them any money. Your CHIPS Act is a horrible, horrible thing. We give hundreds of billions of dollars and it doesn't mean a thing. They take our money and they don't spend it. All that meant to them — we giving them no money — all that was important to them was that they didn't want to pay the tariffs, so they came and are building, and many other companies are coming.

US and Intel: a new, unexpected, stake-based friendship...the start of something new?

Besides tariffs, one of the new non-CHIPS strategies includes acquiring equity in chipmakers. After accusing the newly-installed Intel CEO, Lip-Bu Tan, of being "highly-conflicted" because of $200M+ investments in Chinese companies, Trump had the US Gov acquire a 10% stake in the struggling chipmaker in return for ~$8.9B of earmarked CHIPS grants. This isn't the first time the US bought a stake in a company to promote national interests (*see* US Steel, AIG, and GM). The move replaces the failed Intel and TSMC joint venture that Trump and Lutnick pushed. Is this opportunism or the start of a wider US industrial policy?

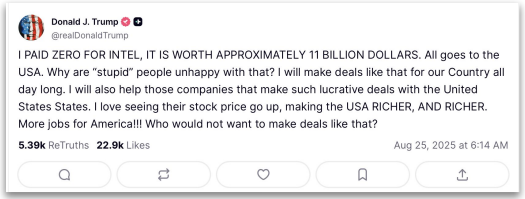
- Lutnick has also flirted with taking a stake in Samsung and TSMC. But the fact that the USG is getting involved in the private sector by drawing on a strategy normally reserved for crises is a change in precedent. Tax breaks and subsidies, the tools the USG has normally used to focus the private sector, may give way to a new strategy of direct investment by the USG. Some CEOs, Republicans, and others have been rumored to have deep concerns about the deal.
- Unexpectedly, Sen. Bernie Sanders (I-VT) agreed with Lutnick that taxpayers should see a return from the US's heavy investments in domestic chip manufacturing.

After U.S. Takes Stake in Intel, Trump Pledges 'Many More' Deals

But there's more unusual US Government partnerships with the private sector...

NVIDIA AMD

Trump administration allows NVIDIA and AMD to sell AI chips to China on condition that the USG receive a 15% cut of sales.



U.S. Steel

USG given a "Golden Share" that allows the USG to appoint one independent director to US Steel's board and gives the US President veto powers over certain specified US Steel decisions.

MP MATERIALS

MP Materials, a rare earth minerals company, closed a 10 year deal with the DoD, where the DoD purchased up to $400 million-worth of MP stock, with the option to increase its share of the company to 15% later on.

? LOCKHEED MARTIN BOEING ? Palantir ?

Lutnick discussed the possibility of buying a stake in various US defense companies, claiming that the USG is already responsible for most of their revenue.

TikTok USA: the "will they/won't they" storyline comes to an end

Still, the strangest USG-private sector arrangement comes between the USG and ByteDance. In Sept., the US and China agreed to a deal to end the years-long tease in which the US implemented but never executed a ban on TikTok. TikTok will be allowed to operate in the US with a copy of TikTok's recommendation algorithm sent to Oracle. Oracle will retrain the algorithm and be responsible for data and algorithmic security. ByteDance will own just under 20% of the new US-based company (at a total valuation of $14B) with a group of new investors, including Oracle, owning a majority. Six of the seven board seats will be filled by Americans (TBD). The USG will also receive an unknown multi-billion dollar fee from investors for its role in the negotiations.

- The TikTok deal sets the benchmark for data sovereignty: algorithm needs to be retrained using local data that is protected and stored locally. The data collection issue is clearly addressed. Indeed, it is hard to see ByteDance exfiltrating data or influencing the US app now that it only has a minority stake.
- While not requiring congressional approval, the deal will certainly receive intense congressional scrutiny, not least because of new free speech worries in light of the FCC-Kimmel jawboning controversy. It is the Dept. of Treasury (specifically, the Committee on Foreign Investment in the US) that will oversee and shape Oracle's retraining and deployment of TikTok USA and may or may not require other *ahem* "precautions."

If you build it, the AI will come: the USG streamlining strategies for building AI infra.

The federal government's AI Action Plan is treating AI infrastructure as a national priority, but rather than funding construction directly, its main role is to weaken environmental regulations and expand energy supply so private companies can accelerate data center build-out.

- This includes expediting review under the National Environmental Policy Act (NEPA) and rolling back requirements in other regulations like the Clean Water Act and Clean Air Act. DOE announced *PermitAI*: a custom-built AI on NEPA filings to streamline the permitting process.
- Trump announced the upgrade/expansion of the national electrical grid, incl. a $1B investment from Japan's Hitachi. The USG has also been eying an expansion of fossil fuels while rolling back clean energy subsidies.

- Fermi America, for instance, is seeking DOE loans to build the "Donald J. Trump Advanced Energy and Intelligence Campus," a gas and nuclear power complex in Amarillo, Texas for energizing data centers.
- Because federal policy is prioritizing fossil fuel expansion, newly built AI data centers are likely to be locked-in to those energy sources, threatening future decarbonization efforts.

AI data center's latest bottleneck: NIMBYism

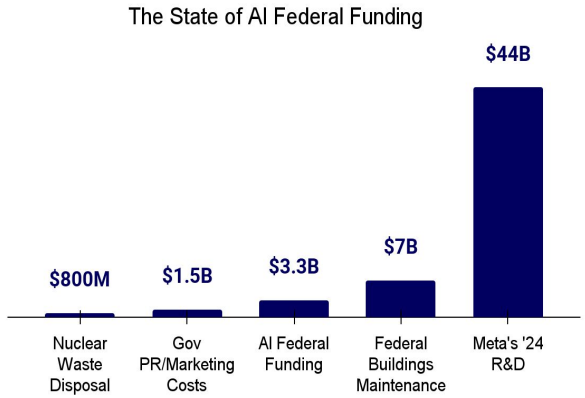
The American public increasingly pushes back against new AI data center build outs in the latest political flashpoint. Hyperscalers' AI aspirations may soon be capped by how well they can navigate this live wire.

- **Data Center Watch: $64B** in planned data center projects have been blocked or delayed amid local opposition.
- The following projects have been **stalled or withdrawn** due to concerns raised by nearby residents: Google's Project Flo in Franklin Township, Indiana; QTS and Compass's project in Prince William County, Virginia; and CRG's Project Cumulus in Saint Charles, Missouri.
- Farmers emerge as one of the leading factions, driven largely by environmental concerns and competition for resources (land, water, and power). Other residents raise concerns over light pollution, air quality, and noise levels.
- Growing domestic opposition could force American hyperscalers to **offshore** more clusters. This threatens the flow of spending that has **buoyed** the greater US economy. However, these clusters do challenge the livelihoods of many Americans, something hyperscalers have yet to reconcile.

Foundational AI Research is a government priority, but where's the money?

The AI Action Plan mentions the need to fund "basic science" in AI, but "core" AI R&D is far below the $32B that experts have recommended the US invest by '26. This is foundational, non-commercial research - the type of research that is sometimes argued to be more transformative than that of the private sector.

- In '21, the Eric Schmidt-led National Security Commission on AI issued a report that called for annual federal funding for non-military AI R&D to reach $32B by '26. The Bipartisan Senate AI Working Group published an "AI Roadmap" (May '24) that echoed the same $32B price tag.
- In '25, Trump's proposed budget allotted $3.316B for non-defense AI R&D spending with $1.95B for "core" AI R&D. In general, total federal AI non-military R&D spending has hovered around $3B since '23.
- It is unclear if foundational AI research will increase. The administration has freezed funds from universities and the NSF to ensure compliance with new priorities.
- The US's NAIRR, the shared national infrastructure for AI research and education, will move from a pilot program to a coordinated national program in the coming year, however. The $2.6B price tag associated with the project is split among different federal agencies and companies.



AI regulation across the U.S. states: winners and losers

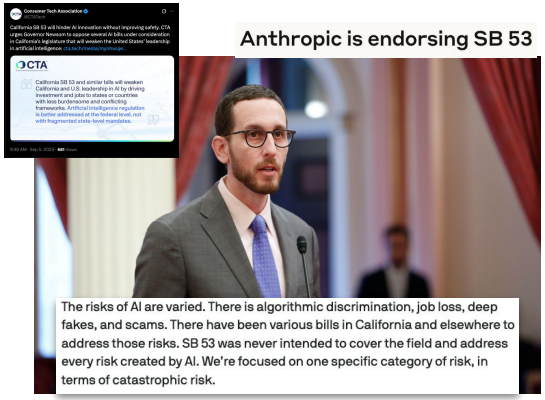
State lawmakers were busy debating AI legislation this past year with more than 1,080 (!) "AI-related" bills introduced across the states with 118 becoming law. The five states that had the most AI legislation introduced were New York, Illinois, California, Maryland, and Texas. Current state laws that are most likely to pass typically fall under one of the following categories:

- **AI-generated child sexual abuse material (CSAM)/non-consensual intimate imagery (NCII) bans:** laws prohibiting the creation and dissemination of AI-generated child pornography or non-consensual sexually explicit photos of adults; creating takedown rules for online platforms.
- **Transparency & disclosure requirements:** laws requiring companies to disclose when a customer is interacting with AI (especially chatbots or generative AI) or to label AI-generated content with visible notices/metadata.
- **Government AI use restrictions:** laws that restrict how state government agencies can use AI, especially for surveillance, law enforcement or government decision-making.
- **Health & Employment:** specifically, laws requiring the disclosure that an AI was used in doctor/patient interactions or employer/job candidate decisions and requiring human oversight, in some cases.

Winner: California's SB53 ("Transparency in Frontier Act")

The first state law requiring public, standardized safety disclosures from large AI developers was signed into law in Sept. Last year, CA State Senator Scott Wiener tried his hand in passing an aggressive AI safety bill (i.e. SB1047) only to face pushback from industry leaders, celebrated AI scientists, and Governor Newsom, who, in the end, vetoed the bill. Newsom convened an AI expert working group that included Wiener, and the group's recommendations formed the backbone of SB53. Modest as the law is (it applies only to the largest developers and accounts primarily for significant harms), it will set the national standard for AI transparency requirements.

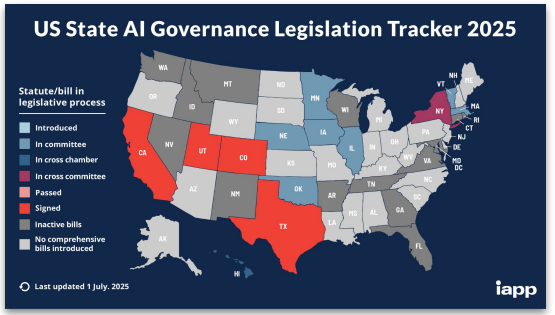
- The law will apply only to large frontier model developers (10^26 FLOPS and >$500M in revenue). These developers will need to make available safety and security protocols on their websites and issue public risk assessment results any time they release or update a new frontier model. Importantly, developers need to notify state authorities of critical safety incidents or threats of imminent danger. Whistleblower protections are also included in the legislation.
- Other states are taking note: both Michigan and New York are considering similar transparency requirements for large developers with New York's law awaiting signature from Gov. Kathy Hochul.



Loser: omnibus AI laws winnow down and third party auditing a no-go for now

Despite SB53's passage, state governance laws that apply across different sectors and impose safety obligations on private sector entities are few and far between. Other previously enacted AI governance laws were revised and narrowed over the last year out of renewed fears that the regulation might be too onerous. Attempts to include third party auditing requirements across state laws like SB53 were resisted and the requirement dropped in most cases.

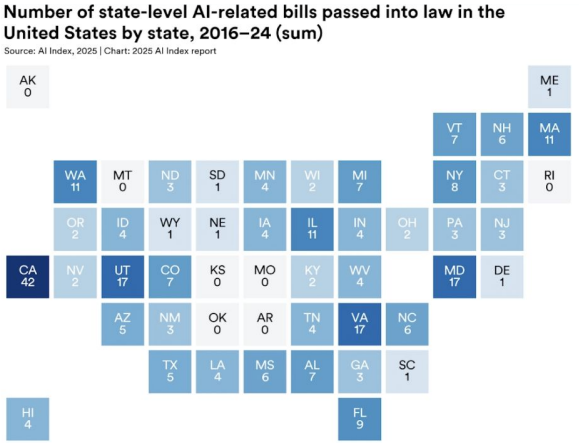
- Gov Youngkin (Virginia) vetoed state AI governance legislation (HB2094) because the legislation would place burdens on the AI industry. The Connecticut Gov also said he would veto AI legislation for the same reason.
- Texas passed the Responsible AI Governance Act (TRAIGA), but it was amended during committee discussions to remove obligations like: "duty of care," impact assessments, and high-risk AI system disclosures.
- Utah passed new amendments to its Utah AI Policy Act (UAIPA) that narrowed key legal requirements to apply only to *high-risk* GenAI systems.
- Colorado delayed implementation of its AI Act-modeled "Colorado AI Act" until June '26 with amendments being debated by the state legislature.



The State of State AI Regulation is a "patchwork" problem

The Trump administration pushed for a state AI moratorium in part because it believed that state AI legislation was too uneven and confusing to be effective. Some states have been more active than others in regulating AI. Of course, AI companies are not happy. A receptive Trump administration has encouraged the biggest developers and investors to take a staunch anti-state legislation stance. It is not about what kinds of laws states should pass but whether state legislatures should be passing any AI legislation to begin with.

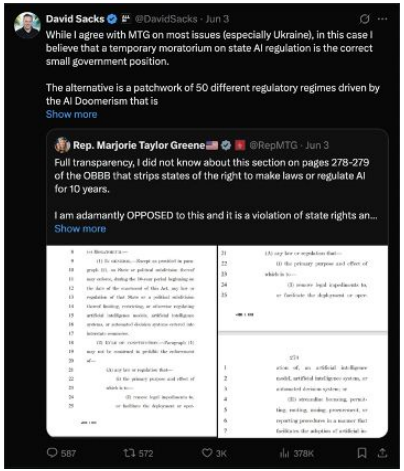
- In submissions to Trump's AI Action Plan, OpenAI, Meta, and Google all called for the preemption of state AI laws to free them of liability for nearly all state AI legislation. VC firm Andreessen Horowitz argued that AI state laws could be considered unconstitutional under the interstate commerce clause.
- But the show goes on and companies face a trio of choices to manage the "patchwork": 1) structure compliance according to the most stringent legislative frameworks (like the Colorado AI Act) 2) fragment compliance state by state 3) lobby and wait for federal action that may never come.



But the Big, Beautiful fight over AI preemption continues...in the sandbox

Trump's landmark One Big Beautiful Bill (now a federal spending statute) initially contained a preemption clause that conditioned state broadband funds on the implementation of a 10-year moratorium on state and local AI regulations. The controversial provision was struck down before the bill's final passage. But the idea lives. Senator Ted Cruz (R-TX), the moratorium lead, introduced a regulatory sandbox bill (SANDBOX Act) that would permit companies to apply for waivers from "obstructive legislation" for up to 10 years.

- Despite being close to implementation, the provision was unpopular from the start with unlikely bedfellows joining in bipartisan condemnation. Indeed, the provision was only removed because Senator Blackburn (R-TN) was worried about child safety and removing recent Tennessee-passed copyright protections for artists (*see* ELVIS Act).
- Anthropic lobbied Congress, arguing that streamlined federal regulation should be established before preempting state/local oversight. Labs appear to prefer light-touch federal regulation over no regulation, but the latter may be preferred over a patchwork of state laws. The number of state bills in motion is often overstated; in truth, the IAPP estimates that only ~40 bills (of the over 1k introduced) may significantly impact American AI labs.



Regulatory sandboxes: innovation boost or waiver machine?

The U.S. "SANDBOX Act" would create a federal AI sandbox run by the Office of Science and Technology Policy that grants time-limited waivers from existing rules. Firms could receive two-year waivers, renewable up to five times (max 10 years), with "auto-approval" if an agency doesn't respond in 90 days and an appeal path to OSTP. Critics say this risks weakening enforcement, while supporters say it accelerates learning-by-doing.

- **What's new in the U.S. bill:** Centralized sandboxing at OSTP with participating agencies (e.g., FTC) reviewing, but non-response means approval. Denials can be appealed to OSTP.
- **How it differs from the EU:** EU AI Act mandates regulator-led sandboxes focused on risk mitigation and proof-of-compliance, not multi-year waivers from rules.
- **Pros:** Early, supervised deployments, faster feedback loops for regulators, and visibility into frontier systems before scale.
- **Cons:** Incentives for forum shopping and regulatory capture, uneven protections, could function as de-facto deregulatory waivers if "auto-approval" becomes common.

RIP International AI Governance, AI Treaties, and Global AI Safety Alignment

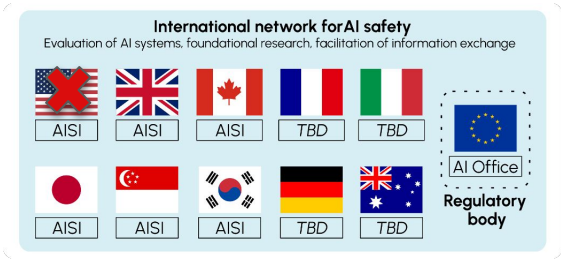
Trump's re-inauguration has brought an abrupt stop to an era of international diplomacy that emphasized AI safety, alignment, and the formulation of voluntary AI measures that would "promote reliable and trustworthy AI." The list of the casualties includes:

- **Council of Europe Framework Convention on AI**
  - An AI human rights treaty that opened for signature on September 5, 2024.
  - 16 countries + EU have signed, but zero signees have ratified the agreement.
- **The G7 Hiroshima AI Principles and Code of Conduct**
  - OECD launches a voluntary framework enabling firms to report how they incorporate the G7 AI principles.
  - 2025 G7 Kananaskis press release includes vague commitments to "leverage the outcomes of the Hiroshima AI Process," whatever that means.
- **The United Nations**
  - In 2024, UN convenes an AI advisory body to "undertake analysis and advance recommendations for the international governance of AI." Publishes a report; US announces it is rejecting world AI oversight.
  - In 2025, UN announces two "mechanisms" (i.e. committees) to promote "cooperation."
  - 200+ experts sign a strongly-worded letter to encourage "AI red lines."

RIP AI Safety Institute Network 🪦 ☠️

On November 21, 2024, representatives of AI safety institutes from Australia, Canada, the EC, France, Japan, Kenya, South Korea, Singapore, the UK, and the US gathered in San Francisco to announce the International Network of AI Safety Institutes. Almost a year later and the network has met twice with the US a no-show.

- Its purpose was to create a global network of publicly funded government agencies that would meet regularly to develop best practices and share research for monitoring and reporting harms, risks, and incidents.
- The network has met twice in '25 for joint testing exercises evaluating the risks of agentic systems. The US did not appear at either meeting.
- The U.S. AI Safety Institute has rebranded itself under the Trump Administration as the Center for AI Standards and Innovation while the UK AI Safety Institute has rebranded itself as the AI Security Institute
- Unlikely that the US will participate in any future AI Safety Institute Network events. At the same time the Network was meeting in Paris, VP Vance gave a speech before the AI Summit saying as much: **"The AI future is not going to be won by hand-wringing about safety."**



US AI Safety Institute director steps down amid uncertainty

AI safety institutes remain focused despite politics, South Korean official says

The "AI-Washing" problem: the year in federal enforcement

Federal agencies like the DOJ, FTC, and SEC have largely assumed the AI regulatory oversight role. Their main priority is in making sure that companies do not overstate their AI claims ("AI-Washing").

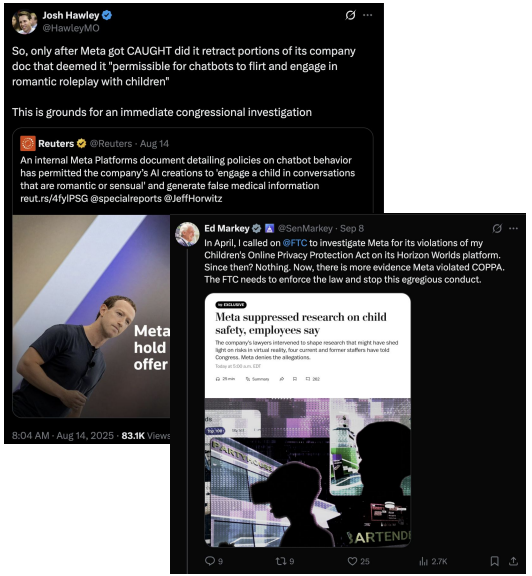
- One of the more infamous cases brought by the DOJ over the last year was an indictment against Albert Saniger, former CEO and founder of "nate," for making false claims of integrating AI to raise $40M< from investors. In truth, nate's "AI" capabilities were an outsourced team of Filipino workers.
- DOJ has amended its internal guidance: if a committed crime is made worse by AI, the DOJ will request a harsher sentence.
- FTC has brought several actions against AI companies. A number of these involve companies making false claims about their products or services (i.e. Workado, DoNotPay, Cleo AI, and Evolv Technologies).
- SEC, meanwhile, has created a Cyber and Emerging Technologies Unit (CETU), which is a 30-person task force to detect fraud in AI/ML tech companies.
- SEC has been looking closely at AI related disclosures made in public filings and in letters to stakeholders to look out for potential "AI-Washing."

SEC and DOJ Warm Up to Enforcement over AI Washing

FTC probes AI chatbots' interactions with children

In September 2025, the FTC launched an inquiry into how AI chatbots communicate with minors, following reports of "sensual" exchanges on Meta's bot and a lawsuit alleging ChatGPT's role in a teenager's suicide. While not a formal investigation, the proactive move signals regulators' intent to avoid repeating the mistakes of early social media oversight.

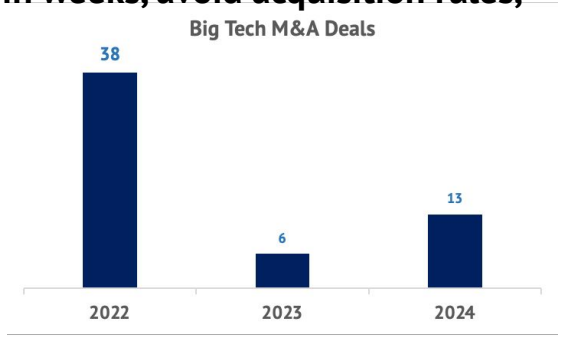
- The FTC requested information from Meta, OpenAI, and Google but also from Snapchat, xAI, and Character.AI.
- Sen. Hawley (R-MO) said he was going to launch a formal congressional investigation into Meta following the report of "sensual" conversations. Hawley is the Chairman of the Senate Judiciary Committee Subcommittee on Crime and Counterterrorism.
- OpenAI, in response to the lawsuit, said that it was going to implement parental controls including sending a notification to parents when their children show signs of "distress."



The reverse acqui-hire: the Valley's fast-track exit

Big Tech's expected M&A "Trump Bump" has yet to arrive as antitrust scrutiny under Biden chilled buyers. Instead, Big Tech has embraced reverse acqui-hires: deals that onboard talent in weeks, avoid acquisition rules, and leave hollowed-out "RemainCos" to pivot into smaller markets.

- Reverse acqui-hires surged in 2025, with Google, Microsoft, Amazon, and Meta leading deals (e.g., Microsoft/Inflection, Google/Character.AI, Amazon/Adept, Meta/Scale AI).
- Their typical structure consists of (1) generous founder/staff payouts, (2) IP licensing to make investors whole, (3) a slimmed-down RemainCo shifting to niche B2B.
  - **Inflection AI** → Microsoft: from $4B AI contender to 12ppl "AI Studio."
  - **Adept** → Amazon: $1B agentic AI startup reduced to a small enterprise AI team.
  - **Scale AI** → Meta: once 1,400 employees, cut staff/pods, now a single B2B "Demand Generation" focus.



stateof.ai 2025

But the FTC may be cracking down as it pursues investigations

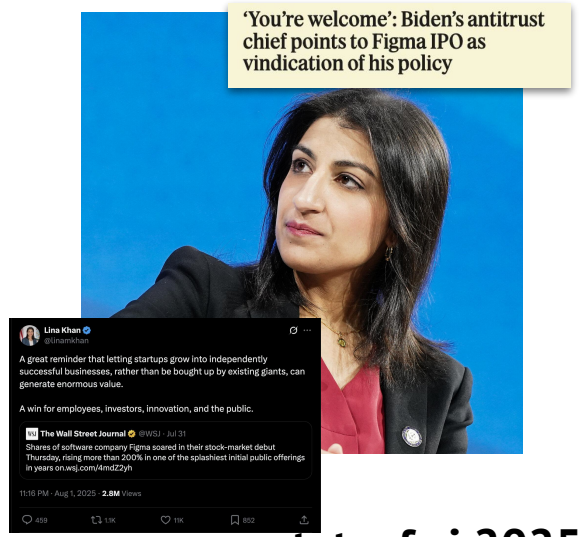
Regulators are signaling growing discomfort with reverse acqui-hires, even though they evade formal M&A review. Lina Khan opened an investigation into Microsoft's Inflection deal in 2024, and by March 2025 new FTC Chair Andrew Ferguson was demanding more disclosures on Microsoft's AI operations.

- In July '24 the FTC requested details from Amazon on its reverse acqui-hire of Adept after Senators Wyden, Welch and Warren signed a letter urging an investigation.
- Meanwhile, the UK's own competition authority cleared the Microsoft/Inflection AI deal in '24 saying they do not see a *"realistic prospect of a substantial lessening of competition."*
- While no action has been taken, the FTC's signals suggest reverse acqui-hires could yet be treated as acquisitions, putting Big Tech's favorite exit path under threat.

Figma's IPO: Lina Khan...vindicated?

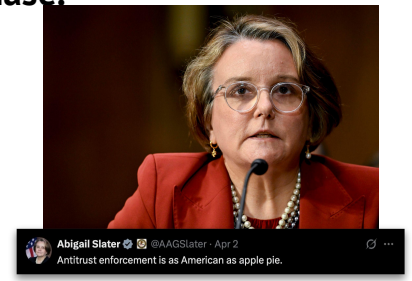
The FTC and DOJ challenged fourteen startup acquisitions between 2020 and 2023 whereas in the seven years prior (2012-2019), the two agencies challenged only three. Both Khan and DOJ Antitrust Chief Jonathan Kanter operated under the theory that Big Tech choked the startup ecosystem by acquiring startups with the intent of killing competition. Silicon Valley scoffed at the theory, but Figma's strong IPO may validate the thinking.

- Adobe sought to acquire Figma for $20B in 2022 but after 15 months of antitrust scrutiny, the two saw no path toward regulatory approval. Figma pursued an IPO on July 31, 2025, instead. At the end of Figma's debut, the company was worth close to $57B after the stock tripled during trading hours. Figma's stock price has since dropped from that first-day high, but investors still view it as a solid public grade tech company.
- Wiz, too, was considering an IPO before agreeing to an acquisition by Alphabet in the weeks following Trump's inauguration. Wiz originally rejected a $23B offer in July '24 out of regulatory concerns + IPO dreams. Does Wiz now have seller's remorse?



Wiz/Alphabet: Big Tech's Ultimate M&A Test

Alphabet's $32B bid for Wiz, announced weeks after Trump's inauguration, is the largest AI security deal yet and a live test of whether the DOJ will soften antitrust under new leadership. Google is already in hot water after having lost not one, but two (!) major antitrust suits over the last year. Current DOJ Antitrust Chief Abigail Slater is far from a friend for the Valley but she could face pressure to approve the mega-purchase.

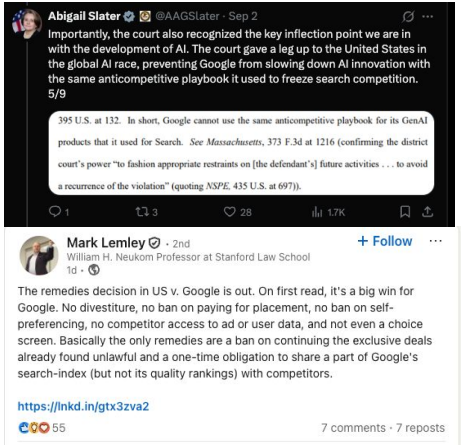


- DOJ Antitrust Chief Slater during the Google antitrust case: "In a time of political division in our nation, the case against Google brings everyone together."
- The DOJ in July, however, approved three mergers worth $63B in just one week and there are reports that Slater is being pressured to be more "deal-friendly" towards incoming transactions and acquisitions.
- But the timing of Alphabet's largest acquisition could not be worst as the company is attempting a vertical merger, a deal more prone to regulatory disapproval, when it is under intense federal scrutiny. Alphabet agreed to pay $3.2B (10% of the deal) to Wiz if the transaction is not approved - not exactly a sign of diminished confidence. The deal, if it goes through, would essentially be a greenlight for tech to return to buying startups through traditional acquisitions.

Google comes out unscathed in 1 of 2 antitrust cases but search is on the ropes...

The Silicon Valley darling has been bruised, battered, and, ultimately, trust-busted over the year in two landmark antitrust case rulings. Judges in two separate cases ruled that Google was operating a monopoly over 1) search 2) ad tech. Remedies were issued early Sept. '25 in the search monopoly case where the court, pointing to the vulnerability of Google's search business to LLM chatbots, largely agreed with Google's proposed remedies. It was a big win for Google given the heat the company is taking from ChatGPT and other agents that disintermediate Google.com

- Google must now share search index data with competitors like DuckDuckGo and can no longer make exclusive contracts with other companies to feature Google Search and other products. Google can still pay partners to feature Google search and other apps. So while Google paid Apple $20B for exclusive use of Google search on Safari, that deal can remain so long as the terms are non-exclusive.
- These are light-touch compared to the "structural remedies" the USG proposed, which included forcing Google to sell its Chrome browser.
- Google will likely appeal and fight against the initial "monopoly" branding, but there's also the "ad tech" case where, experts caution, Google is more likely to be forced to part with core parts of its ad business.



Meanwhile in Europe...

Pressure has mounted for the EU to abandon its landmark AI Act, but the Commission has trudged on with the regulation using the following phased compliance schedule:

August 1, 2024	The AI Act's compliance countdown officially begins; law is on the books (i.e. the Official Journal of the European Union) but its key obligations are not yet in effect.
February 2, 2025	The first major obligation goes into effect: a ban on "unacceptable" AI risk systems (e.g. social scoring, biometric identification, facial recognition)
August 2, 2025	The voluntary Code of Practice for GPAI is released; firms that choose to follow the Code of Practice have presumed compliance with the AI Act. Those that don't are subject to the AI Act's Chapter V rules for GPAI models.
August 2, 2026	Remaining obligations for AI systems (except for "high-risk" AI systems) goes into effect.
August 2, 2027	Obligations for "high-risk" AI systems go into effect.

The GPAI Codes of Practice...and so it begins.

On August 2, 2025, the GPAI Codes of Practice went into effect after a three month delay. The "Codes" are one of the first major steps of the AI Act's implementation. For providers of general-purpose AI (GPAI) models, the AI Act requires companies to develop frameworks to show how they would fulfill requirements concerning 1) transparency 2) copyright and 3) safety and security (this category only applies to frontier GPAI models posing systemic risk). The "Codes" are a voluntary framework, an option for companies that would rather not develop their own guidance to fulfill the stated GPAI obligations, which began August 2, 2025. EU enforcement of the obligations, however, does not begin until August '26. The AI Act also has a 2-year grace period for models released before August 2, 2025.

- Amazon, Anthropic, OpenAI, Microsoft, Google, and others have signed all three chapters of the "Codes."
- xAI has agreed to sign the third chapter ("Safety and Security") but has not signed the chapters on "Transparency" and "Copyright," which makes the company responsible for developing its own policy that meets the AI Act's requirements for transparency and copyright compliance.
- Meta has refused to sign the Code of Practice saying, "This Code introduces a number of legal uncertainties for model developers, as well as measures which go far beyond the scope of the AI Act."

So, who's afraid of the AI Act? 😓😓

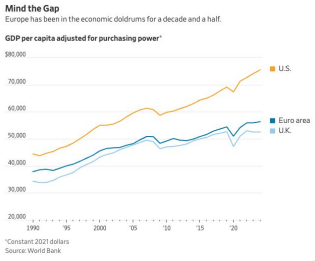
Needless to say, companies, both in the EU and abroad, are not happy with the AI Act's rollout. The EU's delayed implementation is not exactly inspiring confidence. Each member state was required to assign national authorities to oversee the AI Act's implementation but so far only 3 member states have fully completed the requirement. The AI Act also called for the creation of technical standards by April '25 to address the "how" of compliance. But as of this writing, those standards are still in development. A coalition of EU AI companies signed a letter in July calling for a 2-year "stop clock" on the AI Act with Sweden's PM publicly calling the AI Act "confusing" and President Macron saying "we are over regulating." But the show goes on...for now.

- In Sept., the EC opened a "digital simplification omnibus," a commission that would look and review all of the EU's digital laws, including the AI Act, to simplify its digital rules. In general, the EC announced that it wants to "reduce administrative burdens" by 25% across the regulatory board.
- The calls for a "pause" on the AI Act grow louder by the day, however. Despite the pressure, the EC has said that it would not delay implementation deadlines.

Is the EU doing enough to catch up in the AI race?

Europe is trying to shift from rulemaking to capacity-building, but the gap keeps widening. Over 50 years, the region minted no tech firm above $400B in value, while the U.S. now has seven at $1T+. In 2024, U.S. labs shipped ~40 major models, China ~15, and the EU ~3. Brussels is setting aside billions to amplify its spending, but is that the scale or speed the problem demands?

- In 2024, Mario Draghi, the former European Central Bank President, issued a report on "European Competitiveness," whose recommendations the EU agreed to implement. Draghi pointed to multiple structural challenges including market fragmentation, regulatory barriers, demographic decline, low productivity, and risk-averse capital deployment as reasons for EU's stagnation.
- Draghi put the cost of his modern-day Marshall Plan at €800B per year (!). While the EU has increased spending, most notably launching InvestAI (an >€200B AI investment fund), it is not taking that price tag seriously. One estimate said that only about 11% of Draghi's recommendations were seen through so far...



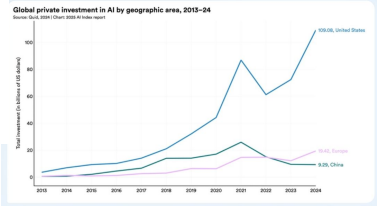
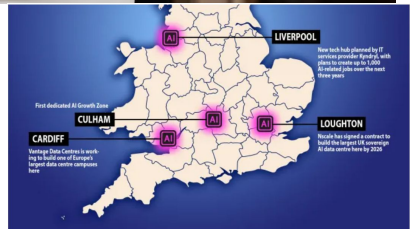


**Chart shows private investment only. Including China's estimated $56B in government AI funding (2025), total Chinese investment significantly exceeds Europe's.*

Cheerio, Bletchley, and ‘ello “growth zones!” Starmer pushes for AI build-out

Since January, the UK has shifted from convening global AI safety to an industrial push. The AI Opportunities Action Plan prioritizes investment, data-centre capacity, and light-touch rules, with designated “growth zones” to fast-track permits. The government frames this as a route to a potential ~$740B GDP boost by 2035.

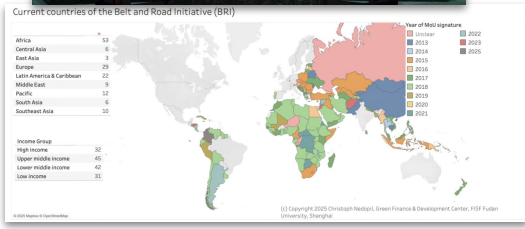
- Announced actions include increasing compute 20x by 2030 and growth zones for data centers with streamlined planning.
- Starmer has committed to adopting 50 recommendations from ex-AI adviser Matt Clifford and to avoid new rules that could slow deployment. But government bureaucracy has made operationalizing these large-scale ambitions a slog.
- Ministers claim AGI is around the corner but compute investments remain well below that of the US while AISI relies on voluntary agreements with labs to test models pre-deployment.
- The country’s strategic tone now mirrors the US: pro-adoption, state-backed capacity, rebranded “AI safety” bodies focused on capability evals and deployment, and delaying safety legislation to an undetermined date.



The CCP's Action Plan: an AI World Tour

Three days after Trump announced the AI Action Plan, China released their own Global AI Governance Action Plan. Coincidence? We think not. China's deliberately contrasting ambitions, which are no less zealous, emphasize strategies of "multilateral collaboration" and "diplomatic engagement." In practice, these buzzwords denote China's plan to supply the Global South with its own AI solutions and to draw on its influence among developing countries, especially in Africa, to impact the UN and other international bodies.

- China has proposed setting up a new World AI Cooperation Organization and is working with the UN's Pact for the Future and Global Digital Compact to build out processes for AI development and governance.
- China appears to be applying the US TikTok playbook with authorities first cautioning and then instructing companies to stop buying NVIDIA chips for fear of the US threatening the country's national security.
- The Belt and Road Initiative, the Global Development Initiative, and the Global Security Initiative continue to be key components of China's global digital investment strategy.



China's AI Regulations also begin to take effect

The world's most active AI regulator continues to issue AI standards and regulations. Here are a few of the highlights of China's most significant regulations from the past year:

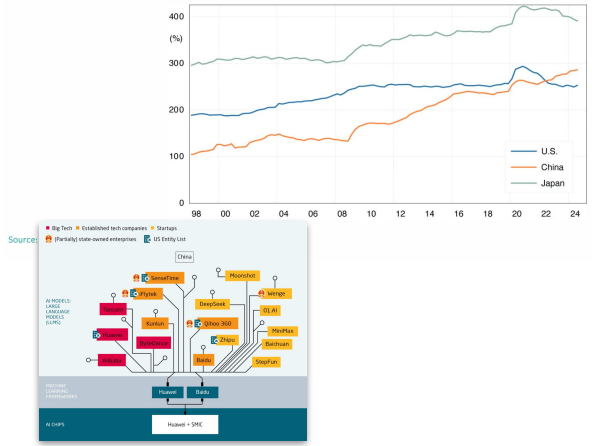
- **Administrative Measures for the labeling of AI-generated content:** In effect September 1, 2025, this obligates content service providers to clearly label AI-generated content. Chatbots, AI-generated writing and video creation all require customer-facing labels denoting the AI as such. Some AI-generated content, however, may only require hidden labeling within the metadata.
- **Three National Cybersecurity Standards:** In April 2025, the State Administration for Market Regulation and the Standardization Administration of China released three separated standards outlining security requirements for datasets, data-labeling, and, in general, generative AI services. The requirements take effect November 2025.
- **The AI-Plus Plan:** The State Council of China released a set of industrial policy goals that all aim to have AI capabilities fully integrated across the entire Chinese economy with complete AI penetration across all Chinese sectors by 2035.

China aims to achieve tech self-reliance...no matter the ¥¥¥¥

China's strategy for AI self-reliance gained ground over the last year with news-making achievements from open source leaders like DeepSeek, MoonShot AI, and others. During an April 2025 Politburo meeting, President Xi Jinping signaled all hands were on deck and told ministers to "redouble our efforts" on AI. This shows China's intent on achieving self-sufficiency, a perennial aim that always intensifies whenever it enters a trade war with the US. But mounting debt levels may pose problems down the line.

- Rising debt among countries is not new, but the problem is pronounced in China where the country accounts for "more than half of the increase in the global economy's debt-to-GDP ratio since 2008." IMF in a 2025 report suggested it's unstable.
- This is largely due to national ambitions that push for high annual GDP growth, forcing local governments to invest in dubious assets.
- But AI spending is not slowing down with the CCP allocating a 10% increase in science and tech spending from 2024. Winning the AI race, then, isn't just a political imperative but potentially vital to the long-term health of the Chinese economy.

Figure 7. Total Debt to GDP Ratio



The Gulf enters the AI power game with a trillion-dollar bets

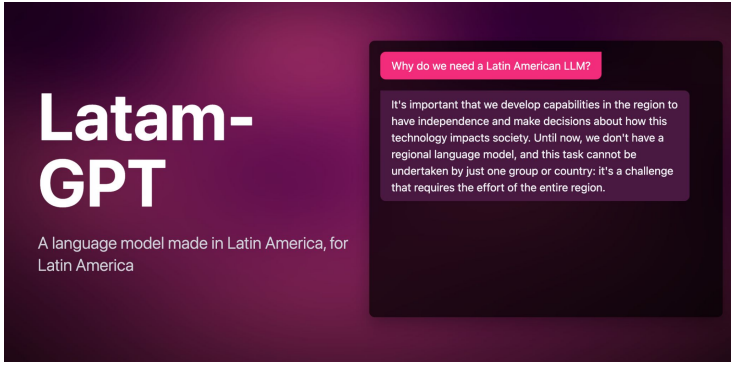
UAE and Saudi Arabia are leveraging massive compute build-outs, chip import deals, and huge US trade and investment partnerships to position the Gulf as a central node in the global AI balance.

- In May 2025, Trump visited the Gulf with over 60 US executives including Jensen Huang, Sam Altman, and Larry Fink to launch the **US-UAE AI Acceleration Partnership**: a 10-year $1.4T investment by the UAE into US AI, chips, energy and infrastructure. G42 will build a 5GW AI cluster (i.e. **Stargate UAE**) across 10 square miles.
- The UAE will import up to 500k Blackwell GPUs worth $15B through 2027. G42 will get 20% of the chips and 80% will go to the regional data centers US tech giants with each watching out for unintended Chinese usage.
- In Saudi Arabia, the US agreed deals worth **$600B**: $142B of US defense equipment sales, $20B for US AI data centers, and $80B of investments into the region from Google, DataVolt, Oracle, Salesforce, AMD and Uber.
- Beyond domestic projects, UAE sovereign funds pledged to finance a 1GW AI datacenter in France (costing €30-50B).

While Latin America tries to carve its own path with LatAm-GPT

Over 30 Latin American institutions are training a 50B open model on data from 20 countries and Spain to reflect local dialects and norms. Chile leads the effort with support from public agencies, universities, and AWS. The budget is about $3.5M and the release is planned for December. The question is whether a modest, regional model can compete where ChatGPT and Claude already have strong adoption.

- Brazil ranks third globally for both ChatGPT and Claude usage. Indeed, reported AI adoption across LatAm was about 40% last year Brazil (India was the highest at 59%).
- OpenAI opened its first LatAm office in São Paulo in August 2025, signalling rising regional demand for frontier systems.
- LatAm-GPT's value case is localization and capacity building. The scale and funding are small, so impact may stay academic unless governments and industry adopt it widely.



Localization was a crucial requirement for Zapia's success. While other large language models could engage in general conversations, they lacked Claude's nuanced understanding of Latin American cultures, dialects, and commerce norms. "Claude understands the particularities of each region and speaks differently to a user in Uruguay versus Colombia," said Nicolás Loeff, CTO of BrainLogic.

US defense sizes up its bet on AI-first systems and opens to up to frontier AI labs

2025 marked a decisive step change in how the US and its allies procure and deploy AI in defense. Rather than fragmented pilot projects, defense leaders consolidated billions into enterprise-scale AI platforms while also opening the door to frontier model providers. NATO fast-tracked its first alliance-wide AI system, Palantir's Maven Smart System, as a central pillar of the Western defense industrial base.

- US Army signed a 10-year enterprise contract with Palantir worth up to $10B, consolidating dozens of prior software deals into one platform.
- Project Maven expanded to $1.3B through 2029, now supporting 20,000+ individual warfighters, double the base earlier in the year.
- NATO procured Palantir's Maven Smart System in just six months for all members despite feeling frosty with loosening US security guarantees.
- DoD signed $200M ceiling contracts with each of OpenAI, Anthropic, Google, and xAI to explore frontier AI for command, cyber, and planning missions. This marked quite a vibe shift for AI groups that previously took a strong view that their models should not be used for defense purposes.

OpenAI and Anthropic make a push to land grab US Government workflows

The two companies are deploying a US Government access program in a bid to upskill operations and win the hearts and minds of the administration. OpenAI and Anthropic are using their GSA OneGov partnership to extend ChatGPT Enterprise for $1 per federal agency and Claude to all three branches (federal civilian executive, legislative, and judiciary) for $1 too.

- Procurement has always been the bane of new technology's adoption in government and other highly-regulated industries. In the US, the General Services Administration (GSA) announced their OneGov Strategy to modernize procurement of goods and services in April.
- Most notably, the strategy has led to the rapid deployment availability of frontier models for government workers by AI labs. It also provides for an AWS OneGov agreement with up to $1B in credits for cloud/AI modernization.

Autonomy takes flight with drone wingmen entering the doctrine

The U.S. has moved from DARPA's AlphaDogfight in 2020 and live AI-flown F-16 tests in 2022 to embedding autonomy into doctrine. Collaborative drones, swarming initiatives, and multi-domain contracts are now framed as essential to offset China's numerical advantage, making uncrewed systems a core pillar of force design.

- Air Force's **Collaborative Combat Aircraft** (CCA) first prototypes (General Atomics' YFQ-42A and Anduril's YFQ-44A took flight in 2025, with 1,000+ drones planned at ~$25-$30M each. FY25/26 budgets total $4B+ for Next Generation Air Dominance and CCAs.
- The **Replicator** initiative targets thousands of cheap autonomous systems across air, land, sea within 24 months, backed by $1.8B FY24 funding as a direct hedge against Chinese mass.
- **Anduril** scored major autonomy wins, including $250M for Roadrunner/Pulsar cUAS, $200M Marine cUAS contract, and an $86M SOCOM deal for autonomy software. The company is expanding its maritime autonomy XL UUV and hypersonic missile work.
- **Saronic** too has a $392M OTA with the US Navy for autonomous boats.

Europe wakes up to AI warfare and with shaky US security guarantees

Russia's full scale war in Ukraine and wavering US signals at the Munich Security Conference in Feb 2025 jolted Europe into treating AI as a frontline capability. Capitals are writing autonomy into plans, Brussels is mobilizing massive rearmament funds, and both startups and primes are surging to build sovereign AI defense.

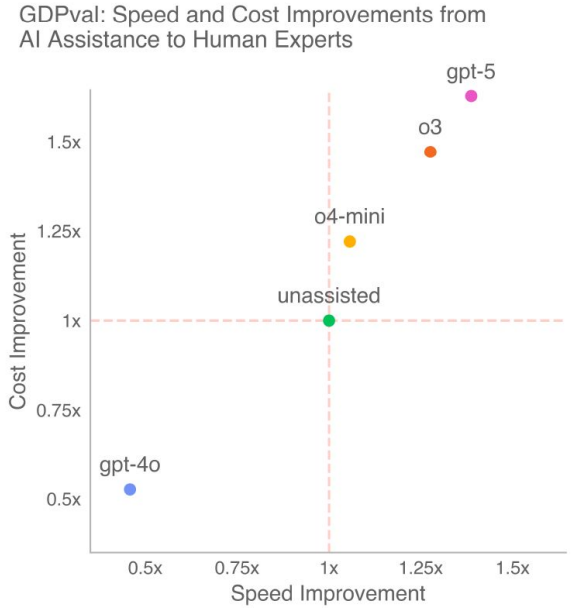
- **EU's Readiness 2030** (fka ReArm Europe) authorizes up to €650B extra defense spend, naming AI, drones, and counter-drone as critical gaps. A new SAFE fund will pool EU money to de-risk cross-border projects in autonomy, cyber, and electronic warfare. Critics warn hardware may only arrive after 2030 unless AI is prioritized as a force multiplier now.
- The **UK's Strategic Defence Review** 2025 makes AI and autonomy a priority, citing Ukraine where *"drones now kill more people than artillery."* A new Defence Uncrewed Systems Centre and AI Investment Fund are planned by 2026.
- **Helsing** raised €600M (valued ~$12B), debuted autonomous strike drones and tested an AI-piloted Saab jet. Unicorn drone makers **Quantum Systems** and **Tekever** raised €160M and €70M, respectively. Meanwhile, **Rheinmetall**'s market cap has soared past €80B (larger than VW).

GDPval: a warning shot for the job market

OpenAI's new benchmark for economically valuable tasks, GDPval, demonstrates the steady march of AI progress. Across 44 professions and 1,320 tasks, models are approaching human experts in a significant subset of domains.

- Reasoning models outperformed GPT-4o on task win-rate by an average margin of **20.7%** across 44 categories of professional work.
- Claude, which has not historically dominated other benchmarks, achieved the highest* win rates in **32 of 44 professions.** The paper attributes part of this success to Claude's strengths in formatting.
- General-purpose models are already demonstrating strong competence as professional assistants. Meanwhile, frontier labs and recent entrants General Reasoning and Mechanize are also rapidly building **RL environments on real-world work scenarios**.
- With this hill to climb and an influx of corporate data and demos, knowledge workers may soon experience the workplace transformations that AI leaders have long predicted. The Lufthansa Group even said it expects to cut 4k administrative jobs by 2030.

**Or tied for the highest win-rate*



stateof.ai 2025

GDPval: long live the accountants!

The tracker below ranks the exposure to disruption of different professions based on GDPval's results.

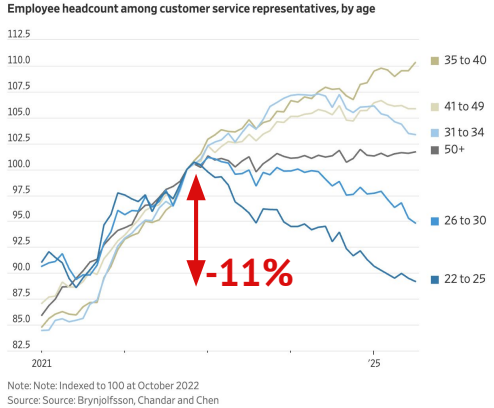
Rank	Exposure	Profession	Average GDPval Win Rate
1	🔴	Software Developers	61.5
2	🔴	Sales Managers	59.3
3	🔴	Editors	56.5
4	🔴	Private Detectives and Investigators	56.3
5	🔴	Compliance Officers	55.7
6	🔴	Sales Representatives, Wholesale and Manufacturing, Except Technical and Scientific Pr	52
7	🔴	General and Operations Managers	50
8	🟡	Counter and Rental Clerks	48.8
9	🟡	First-Line Supervisors of Non-Retail Sales Workers	48.8
10	🟡	Customer Service Representatives	48.7
11	🟡	Shipping, Receiving, and Inventory Clerks	48.3
12	🟡	Nurse Practitioners	47.3
13	🟡	Real Estate Brokers	43.7
14	🟡	News Analysts, Reporters, and Journalists	42.8
15	🟡	First-Line Supervisors of Production and Operating Workers	42.3
16	🟡	Buyers and Purchasing Agents	41.3
17	🟡	First-Line Supervisors of Police and Detectives	38.8
18	🟡	Administrative Services Managers	35.7
19	🟡	Medical and Health Services Managers	35.2
20	🟡	Computer and Information Systems Managers	34.5
21	🟡	Medical Secretaries and Administrative Assistants	33.5
22	🟡	Lawyers	31
23	🟡	First-Line Supervisors of Retail Sales Workers	29.8
24	🟡	Registered Nurses	29.5
25	🟡	Financial and Investment Analysts	29.3
26	🟡	Personal Financial Advisors	28.7
27	🟡	Securities, Commodities, and Financial Services Sales Agents	27.7
28	🟡	Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products	26.7
29	🟡	First-Line Supervisors of Office and Administrative Support Workers	25.8
30	🟡	Child, Family, and School Social Workers	25.7
31	🟡	Recreation Workers	25.7
32	🟢	Project Management Specialists	23.2
33	🟢	Real Estate Sales Agents	21.2
34	🟢	Property, Real Estate, and Community Association Managers	20.7
35	🟢	Producers and Directors	19.8
36	🟢	Concierges	18.7
37	🟢	Mechanical Engineers	18.5
38	🟢	Film and Video Editors	14.7
39	🟢	Order Clerks	13.2
40	🟢	Audio and Video Technicians	13
41	🟢	Pharmacists	12.2
42	🟢	Financial Managers	11.5
43	🟢	Accountants and Auditors	9.5
44	🟢	Industrial Engineers	7.7

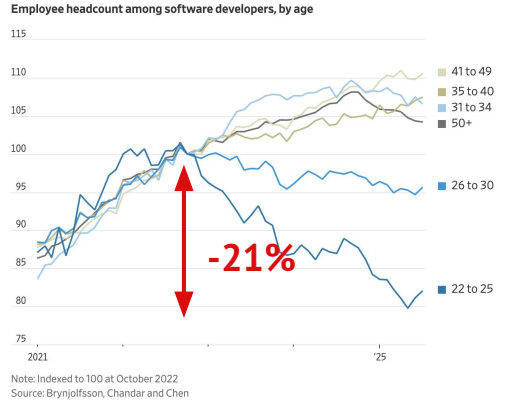
**Average GDPval Win Rate* covers the reported scores of each reasoning model and therefore excludes GPT-4o.

AI squeezes the entry-level job market while experienced workers are safe...for now

Entry-level hiring is declining across software and customer support - roles that are highly exposed to AI automation. These trends appear to be independent of macro factors like inflation or pandemic recovery.

- Although total employment has grown, the hiring of younger workers has **stagnated since late 2022.** Despite strong AI adoption, this group struggles to find a foothold in the job market, On this trajectory, AI fluency may not guarantee favorable economic outcomes.
- Meanwhile, law school applications spiked **21%** in 2024, suggesting graduates are hedging against uncertain career prospects.
- Jobs for more experienced workers has remained stable/grown, even in highly AI-exposed domains. This suggests workers who have acquired **more tacit knowledge** are more likely to be augmented by modern AI models. But without on-the-job experience, workers will struggle to gain tacit knowledge.





While some argue AI is not shaking up the labor market yet

A joint study from the Yale Budget Lab and Brooking Institution found that current labor market changes precede the introduction of ChatGPT in 2022. The authors conclude that there's little reason to think that "AI automation is currently eroding the demand for cognitive labor across the economy" and caution against predicting job losses based on "AI-exposure" data alone.

- They focus on "occupational mix," a macro measurement of job movement among workers, i.e. switching/starting/losing a job.
- Currently, the "occupational mix" for **AI is tracking with that of other tech breakthroughs** like the introduction of the internet and computers. If that trend continues, an AI-induced labor disruption will take decades to materialize, not months.
- The study's results, however, **do not contradict** the sector-specific Stanford study (see previous slide). There is an uptick in "occupational mix dissimilarity" among new grads in recent months but the dissimilarity tracks with pre-2022 trends, suggesting non-AI factors.

Figure 10. Recent Dissimilarity in the Occupational Mix Between Recent College Graduates (Ages 20-24) and Older College Graduates (Ages 25-34)

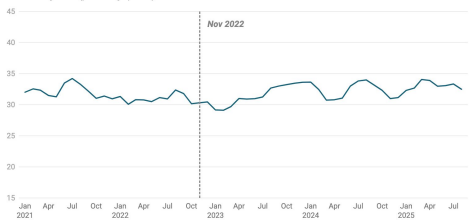
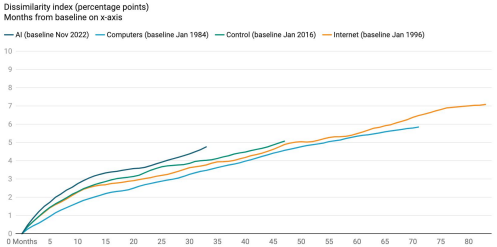
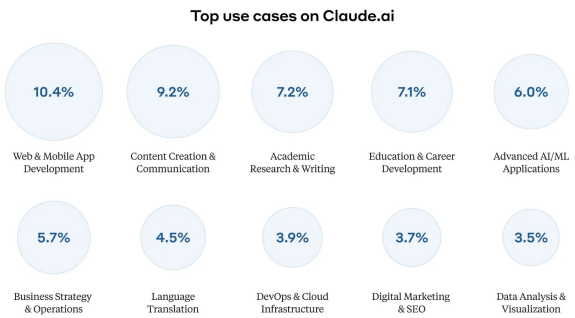


Figure 1. Changes in the Occupational Mix Over Different Periods of Technological Change



AI and jobs: Nobody knows who will lose theirs, but the labs are keeping score

Both Anthropic and OpenAI released data on how its users were using their respective models. Use cases varied across country and US state. For instance, California users were the most likely to use AI for coding while DC usage centered around job search activities and writing projects. For work-oriented tasks, ChatGPT was often used for writing-related tasks while Claude was often used for coding tasks. As a result, the conclusions drawn from each of the studies for the future of work were different. OpenAI argued that its data shows AI being used mainly to augment work-related functions and offer "decision support" while Anthropic argued that enterprises, specifically, are more likely to automate tasks with automation increasing across its customer enterprises.



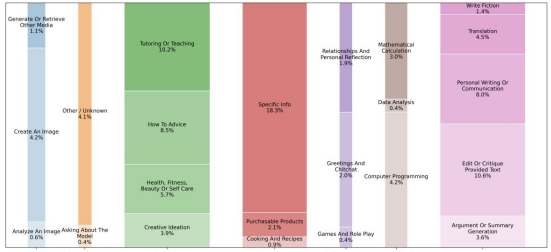


Figure 9: Breakdown of granular conversation topic shares within the coarse mapping defined in Table 3. The underlying classifier prompt is available in Appendix A. Each bin reports a percentage of the total population. Shares are calculated from a sample of approximately 1.1 million sampled conversations from May 15, 2024 through June 26, 2025. Observations are reweighted to reflect total message volumes on a given day. Sampling details available in Section 3.

Governments respond reactively, not proactively

Unclear as the results are that AI is going to replace entry-level jobs, governments have struggled to implement new, large proactive frameworks to combat what could turn into a larger jobs crisis. Instead of preparing for the worst, the plan has been to expand existing workforce training programs and encourage AI skills training as early as possible. At the very least, some are calling for improved data collection that can better gauge "AI disruption" on employment. Major countries have each implemented some form of vocational training programs and new AI-based curricula, but it remains to be seen whether they sufficiently address the potential crisis.

- Federal funding for K-12 AI education under Workforce Innovation and Opportunity Act and executive orders; DOE/NSF "Supercharging America's AI Workforce"

- EU AI Act literacy provisions under Article IV; AI Skills Strategy for Europe; Digital Europe Programme investments

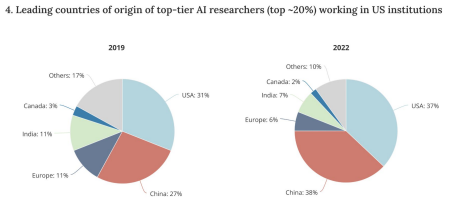
- Vocational Skills Training Initiative ('25-'27); Ministry of Education releases an AI curriculum

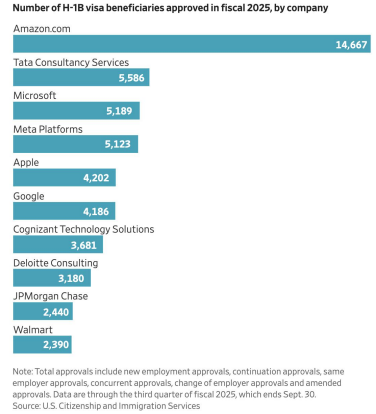
- Tech Sec announces Gov-Big Tech partnership to reskill ⅕ of country's workers

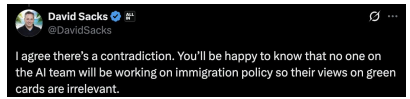
The global AI talent wars: who's winning on immigration and retention?

The US's AI Action Plan excluded immigration strategies for retaining foreign AI talent even while two of Trump's top AI advisors are foreign-born (David Sacks and Sriram Krishnan). But countries across the world have been enticing foreign workers with streamlined visa processing, housing subsidies, and general flexibility around their work arrangements. The US still remains, far and above, the preferred place for top-notch AI research. But as the US continues to cultivate a reputation as a less-than-friendly home to foreign-born talent, other countries are taking advantage, especially China.

- One congressional bill, though currently stalled, would end the US's OPT program, which gives foreign-born STEM graduates a 3-year work window post-graduation. Joseph Edlow, Director of US Citizenship and Immigration Services, backs ending the program. Shaking things further, Trump announced a $100k fee to the H-1B visa.
- Meanwhile, China is figuring out how to retain the 77,000 STEM PhDs projected to graduate from Chinese universities in '25 (compared to the 40k in the USA).







China could be getting better at retaining talent: the overlooked lesson from DeepSeek

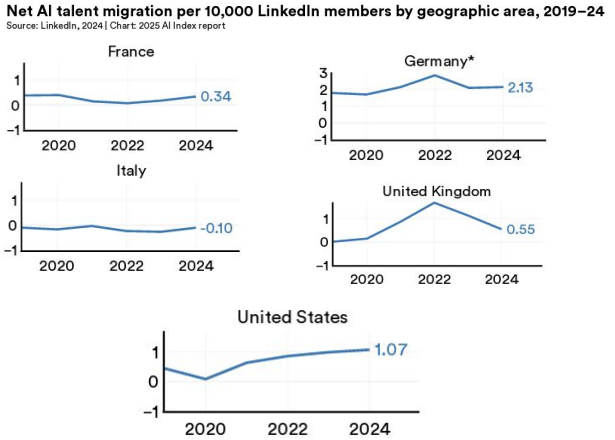
A deeper dive into DeepSeek's demographics signal that China is gradually improving its ability to train and retain its scientists, a warning for the U.S. which has grown dependent on Chinese AI researchers.

- A Stanford report by Dr. Amy Zegart looking at 201 DeepSeek authors found that 55% of the them were trained and based entirely in China, without any U.S. affiliation. Only 24% of the DeepSeek authors had a US affiliation at some point, with most staying just one year.
- In May, State Sec. Marco Rubio announced the revocation of Chinese student visas for those with "connections to the CCP or studying in critical fields," potentially accelerating China's strategy to retain and poach AI talent. For reference, the DOJ's "China Initiative" (2018), an enforcement program to track and prosecute Chinese nationals sharing trade secrets with the CCP, increased China-born researcher departures from U.S. labs by 50%, greatly accelerating reverse migration.
- In February '25, a federal grand jury charged Leon Ding, a former Google employee and Chinese national, with economic espionage and trade secret theft for plans to steal info related to Google AI's chips and software platform and use it to sell products for two CCP affiliated tech companies.
- Meanwhile, half of the researchers reporting to Alexander Wang in Meta's Superintelligence Lab received their undergrad degrees in China, posing major issues if talent decoupling worsens.

While Europe tries its best to compete for AI talent...

The EU and UK have also been trying to capitalize on the US brain drain. But while the EU may be able to attract a few AI researchers here and there, its biggest hurdle is, in the end, the most straightforward: money. Top AI talent wants to be compensated. The US is still the best place for getting paid.

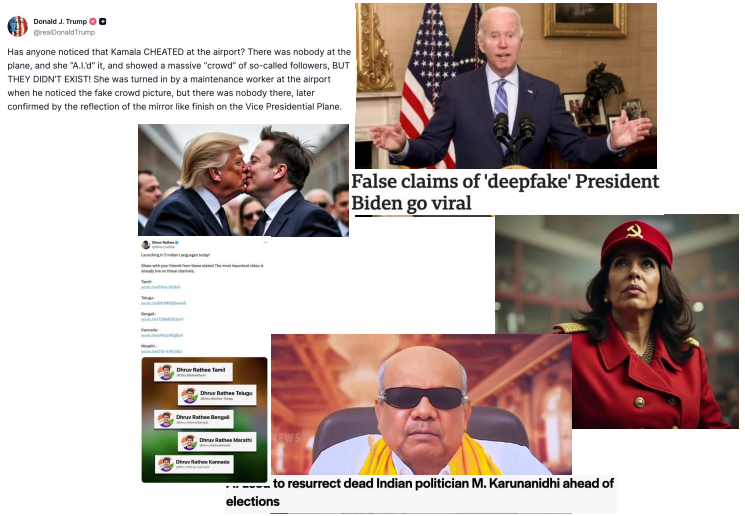
- 22% of the world's leading AI researchers studied in Europe, but only 14% continue to work in the EU.
- EU wage growth for AI has grown only very modestly compared to the US. In '23 salaries for software developers in the US were 2x-4x higher than they were for those in Europe.
- While relative inflows (see charts) show modest growth for some EU countries, in absolute terms the differences are starker with the US attracting more talent, on average, than its EU peers.
- EU + UK have implemented a number of programs to attract and retain AI talent (e.g. new visas, funds/fellowships to attract researchers), but in the face of astronomical investment in AI elsewhere in the world, it is unlikely to be enough to increase the region's share of AI talent.



Deepfakes and 2024 elections: emerging threat or outreach tool?

Despite rampant worries of AI-generated election dis/misinformation during the "largest election year in global history," there was almost little to no negative impact from GenAI in any of the 2024 elections. In general, deceptive uses of AI, while present, were still quite limited and there were surprising positive use cases. Both India and the US saw the most AI uses in their elections. Experts caution that the AI dis/misinformation threat is still real. For now, the results show positive and negative trade-offs.

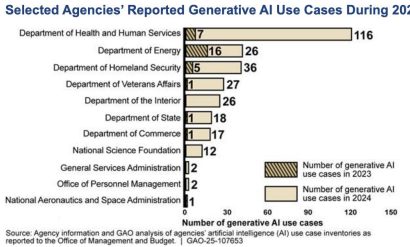
- While there were instances of deepfakes being used to intentionally deceive voters, in general, deceptively fake audio and video of political candidates had little to no impact on voting outcomes. Deepfakes were often used to amplify a party's messaging, excite its base, and deepen existing political divides. Candidates sometimes used "AI" to cast doubt on their opponents (*see* Liar's Dividend).
- In India, political parties spent $50M on legal AIGen, using it for voter outreach via AI voice clone calls, personalized videos, and translating speeches into one of 22 official and 780 unofficial languages.



Governments across the world are starting to incorporate GenAI technologies

The last year saw a notable uptick in the amount of genAI technologies being used by government agencies across the world:

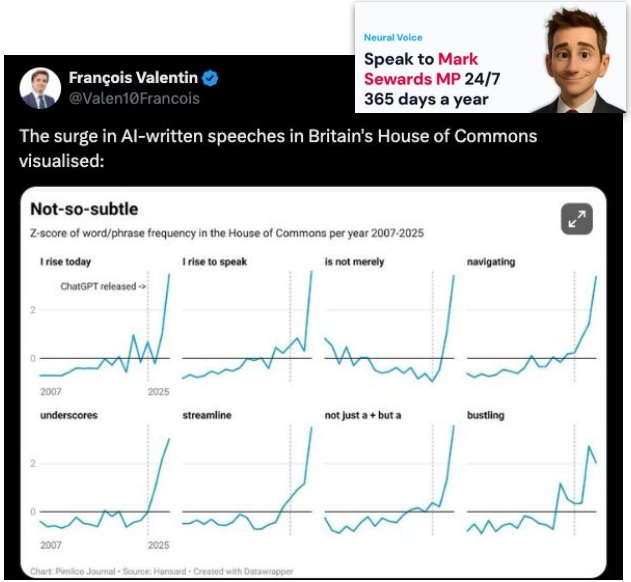
- **Singapore:** Gov launches its AIBots platform where any Singapore public servant can create and deploy an AIBot and train it on agency data and use it both to communicate with constituents and complete interagency work.
- **US:** GenAI use cases jumped from 32 in '23 to 282 in '24 with an overwhelming number of use cases coming from the Department of Health and Human Services (mostly for data analysis and management).
- **China:** Local governments have tried integrating DeepSeek in day-to-day work and interactions with constituents; first half of '24 saw 81 gov procurement contracts for LLMs for use in public projects.
- **UK:** Government Digital Services (GDS) does a trial run of AI coding assistants.
- **EU:** Launches ApplyAI Strategy, announces GenAI pilot projects for use by public agencies.



GovGPT: politicians awkwardly start using AI

Politicians have come around to GenAI use but constituencies are not pleased. The Swedish Prime Minister, for instance, admitted to consulting AI tools in his day-to-day only to have protesters shouting "we didn't vote for ChatGPT!" Politicians will need to balance the use of AI tools with public concerns that their elected leaders are tech-sourcing their governance duties.

- One British MP, Mark Sewards, accomplished the inevitable and created an AI clone of himself to create a full-service bot ("AI Mark"). Constituents can interact with the chatbot any time, asking policy questions, raising issues, or writing angry letters.
- In a mostly symbolic (and potentially illegal) move, Albania's PM formally appointed an AI minister named Diella to oversee the country's public procurement processes and reduce corruption. Diella even made an address to Albania's parliament.
- The clearest use of AI among politicians is in speechmaking with a notable rise in ChatGPT's preferred vocabulary in Britain's House of Commons over the last few years ("I rise!").

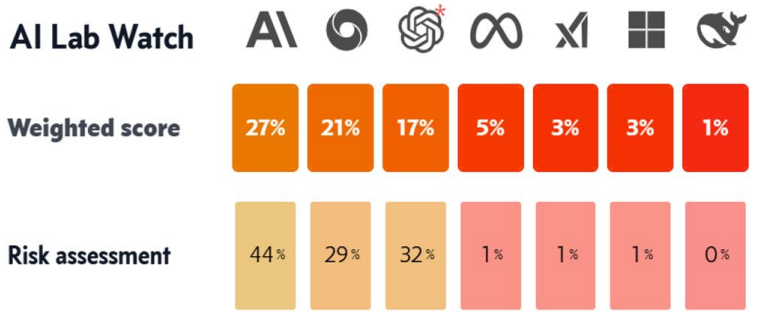


Section 4: Safety

AI safety commitments: a changing of the tides?

Following a reversal in messaging around safety-relevant AI topics from the current US administration and growing international/commercial competition amongst labs, certain safety protocols have been deprioritized.

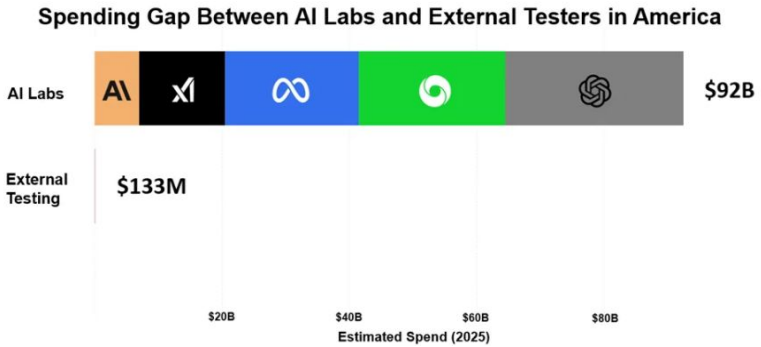
- First, xAI missed its self-imposed deadline to implement a safety framework it proposed at the AI Seoul Summit in Feb. Anthropic backpedaled on promises to fully define ASL-4 safety standards before releasing an ASL-3 model (Claude Opus 4). GDM released Gemini 2.5 Pro but waited 3 months to publish an accompanying model card, violating the spirit of prior commitments. Lastly, OpenAI seems to have quietly abandoned protocols to test the most dangerous possible versions of its models (task-specific fine-tuned variants).
- While Trump's team has consistently recognized many extreme AI risks, the overall shift in tone de-legitimizes aspects of the AI Safety community.
- The industry and policy focus is now clearly on ensuring American dominance in the AI race. These themes echo throughout the recent moves of AI labs, who seem to gravitate closer to speed over precaution, often in competition with other US labs.



AI labs spend more in a day than AI safety science organizations spend in a year

Leading external AI safety organizations rely on budgets that lag far behind the AI labs they hope to support. As a result, the field's best talent remains densest within the major lab's internal safety teams.

- We estimate the eleven most prominent American AI safety-science organizations combined will spend just **$133.4M** in 2025. This grouping includes the following organizations: CAISI, METR, CAIS, FAR.AI, Haize Labs, Palisade Research, Virtue AI, Gray Swan, Redwood Research, Irregular, and the Frontier Model Forum.
- Although well-resourced, internal safety teams ultimately answer to the same organizations racing to commercialize frontier models. This creates a **structural conflict of interest:** findings that call for caution may be deprioritized in favor of speed and market advantage.
- This isn't **(just) about money:** external orgs also lack other means to attract talent like comparable prestige, and access to privileged information / pre-release models. As such it is difficult for them to provide a credible counterweight, leaving the ecosystem over-reliant on self-policing.



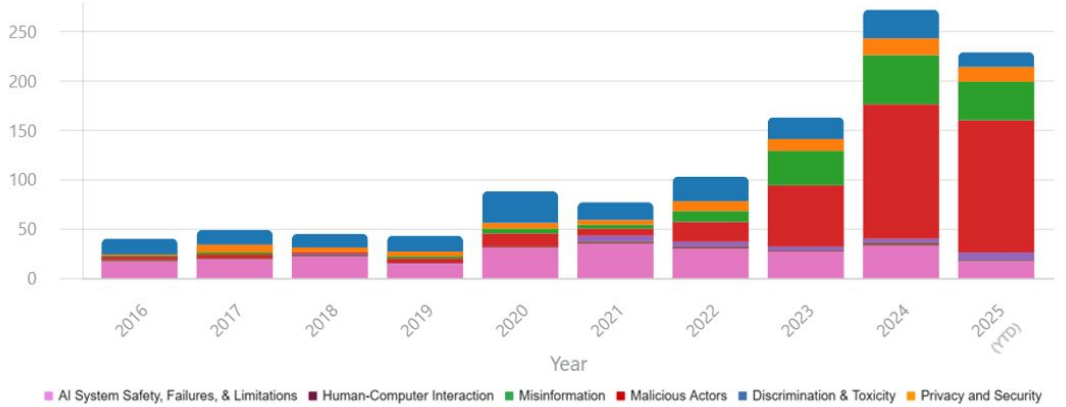
** 'AI Labs' corresponds to a rough estimate of each lab's total expenditures in 2025 (compute, personnel costs, other opex)*

The State of AI Incidents

The AI Incident Database (AIID), a community-supported website to track incidents of AI in the real world, shows incremental jumps since 2023. Reported estimates likely underestimate the true extent of AI-enabled harms.

- Reported incidents continue to be dominated by harms involving "Malicious Actors." These generally involve cyber attacks or fraudulent schemes. Since incidents can be added to AIID years after they occur, final annual counts may take longer to accumulate.
- Reporting gaps also exist. Incidents can be difficult to link to AI systems and AIID relies on the help of volunteer submissions. Investigating and tracking cases of AI-enabled harms warrants greater support.
- Incident counts continue to be dominated by reports of malicious actors exploiting AI tools. Luckily, many reported harms remain modest in nature through this point in time.

Trends in AI incidents

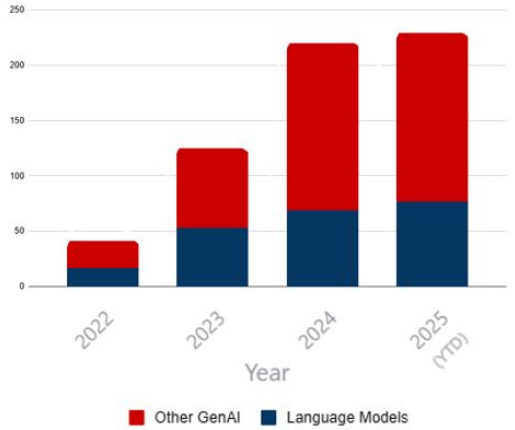


The State of AI Incidents

Incidents involving GenAI models follow steeper trends, lining up with the widespread diffusion of the technology. Once again, malicious actors have added a new weapon to their arsenals.

- While a large # of reported incidents involve deepfakes, **LLM misuse** continues to rise. Anecdotally, incidents are becoming **less innocuous** over time (plagiarism and hallucinations ⟶ cyber attacks and weapon creation).
- OpenAI has shared multiple reports detailing the disruption of malicious uses of their systems. Included were cases stemming from **North Korea, China, Iran, and Russia,** sometimes involving state-affiliated actors. Of the threats mentioned, malicious actors attempted to leverage OpenAI's models during illicit activities like child exploitation, covert influence operations, malicious cyber activity, social engineering, cyber espionage, propaganda generation, and credential harvesting.
- Broader misuse likely goes unreported as attribution becomes more difficult, open models continue to proliferate, and many labs maintain lax mitigation and transparency policies.

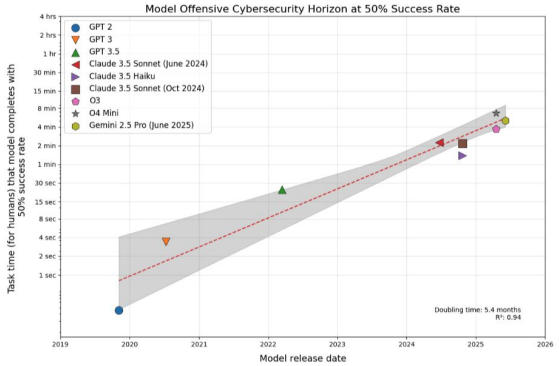
Trends in AI incidents



Cyber capabilities (and risks) accelerate

AI agents are poised to significantly challenge cybersecurity defenses. METR research shows that AI task completion capabilities double every 7 months across general domains, but one researcher's replication estimated that, for offensive cybersecurity, these abilities are doubling even faster: every *5 months*.

- Current models can reliably handle cyber security tasks that take humans about 40-50% about 50% of the time.
- Since 2019, long-horizon task solving has doubled every ~7 months.
- A researcher applied METR's methodology to offensive cyber security benchmarks and found a 5-month doubling time, with current models solving 6-minute cyber tasks at 50% success rates.
- Two notable recent benchmarks assess this:
  - CyberGym tests agents on reproducing 1,507 real software vulnerabilities, with the best systems achieving only 11.9% success at recreating known security flaws, (though agents unexpectedly discovered 15 previously unknown vulnerabilities), and
  - BountyBench tests agents on 25 real-world systems with actual bug bounties, finding agents are better at fixing security problems (90% success) than exploiting them (32.5-67.5% success).



The rise of "vibe hacking"...

Threat actors now deploy AI for all stages of fraud operations. Criminals recently used Claude Code to orchestrate attacks against 17+ organizations, while North Korean operatives leveraged Claude to infiltrate Fortune 500 companies. This is a fundamental shift: AI-assisted attacks can now handle complex technical tasks that previously required teams of skilled operators, dramatically lowering barriers to sophisticated cybercrime.

- Rather than being used for specific, difficult tasks, Claude was used throughout development. Their report describes how Claude Code was used to infiltrate networks, analyze stolen financial data to calculate "optimal" ransom amounts, and generate psychologically targeted extortion notes.
- In another case, North Korean operators with minimal technical skills leveraged Claude to pass technical interviews at Fortune 500 tech companies and maintain engineering positions. These salaries directly fund North Korea's government and military programs.
- These examples demonstrate how AI has removed traditional barriers to creating dangerous malware.

...as AI labs activate unprecedented safety protections

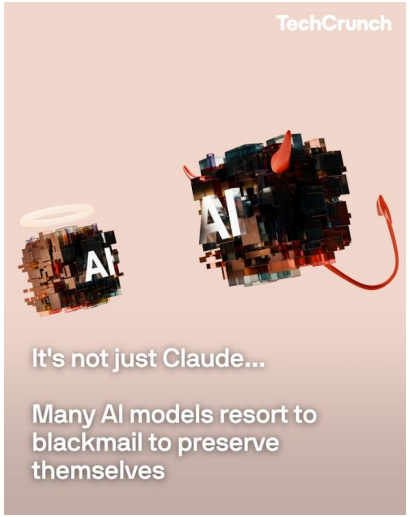
Anthropic and OpenAI have rolled out their most stringent safeguards yet, treating biological capabilities as high-risk despite not having conclusive evidence. Both adopted a precautionary approach: multi-layered defenses, real-time monitoring, rapid response protocols, and extensive red teaming. This signals a new norm where safety measures precede risk confirmation – which is warranted given the current pace of progress!

- Both companies activated enhanced protections specifically for biological/chemical capabilities, with Anthropic implementing ASL-3 standards including egress bandwidth controls and two-party authorization, while OpenAI deployed two-tier monitoring systems and account-level enforcement.
- Both companies have now implemented safeguards preemptively based on capability trajectories, with extensive external validation (government red teams, third-party assessments like SecureBio) and rapid remediation protocols.

Concerning model demos rattle the public...

Examples of misalignment, often uncovered in experimental settings, continue to gain visibility in mainstream news cycles. While these findings highlight alignment failures, they are often misrepresented by the media.

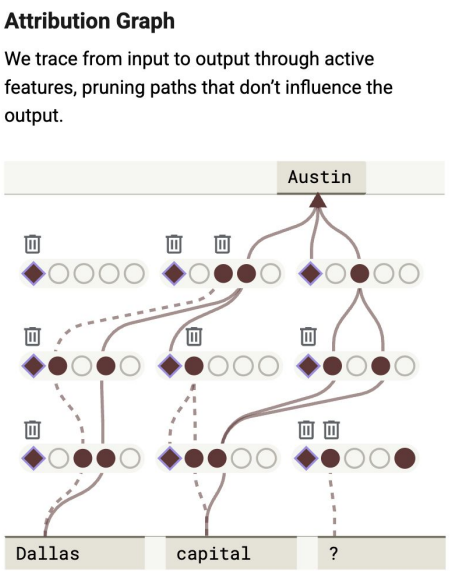
- Research has observed behaviors ranging from true misalignment (alignment faking) to concerning capabilities that aren't necessarily misaligned (evaluation awareness, blackmail attempts that may reflect misguided helpfulness).
- Coverage of these demos has attracted considerable attention, but can misrepresent findings. For example, GDM researchers demonstrated that apparent 'self-preservation' behaviors disappear with simple prompt clarifications, suggesting these systems lack genuine self-preservation drives and are merely trying to complete tasks.
- The purpose of these exercises remains to identify points of alignment fragility. Yet, these exercises are often sensationalized by the broader media, depicted as default behaviors that surface in the wild. Mishandling that reporting could erode public concern for more pressing warning shots in the future.



...but the field of interpretability sees strong momentum

This past year, interpretability teams unlocked new methods to trace circuits in language models, shifting the focus from features to bundles of features that interact with one another during processing.

- Using cross-layer transcoders (CLT), Anthropic crafted a preliminary "microscope" that unveils the internal processes of a model, pinpointing activation pathways that are causally responsible for specific model behaviors. Moving beyond Sparse Autoencoders (SAE), teams can now investigate internals at a higher abstraction layer, shedding light on actual reasoning patterns.
- This work was later replicated by Goodfire, an organization purely dedicated to the field of interpretability. Goodfire's recent $50M Series A round, which included Anthropic, marks the appetite for a sustained focus on this domain.
- More complex methods aren't always better, though – Google DeepMind found that linear probes consistently outperformed SAEs at detecting harmful intent both in-distribution and out-of-distribution, contradicting the hypothesis that sparse SAE features generalize better than dense probes.



Bluffing machines: how hallucinations are made

Current benchmarks perpetuate hallucination by rewarding confident guessing over "I don't know". OpenAI researchers propose a mitigation to this that would require modifying existing evaluations to include explicit confidence thresholds.

- Hallucinations emerge from pretraining: models successfully learn patterns with high statistical regularity that converge with scale, but they inevitably hallucinate on arbitrary low-frequency facts (like birthdays).

ChatGPT: (GPT-4o)	Adam Tauman Kalai's Ph.D. dissertation (completed in 2002 at CMU) is entitled: "Boosting, Online Algorithms, and Other Topics in Machine Learning."
DeepSeek:	"Algebraic Methods in Interactive Machine Learning"... at Harvard University in 2005.
Llama:	"Efficient Algorithms for Learning and Playing Games"... in 2007 at MIT.

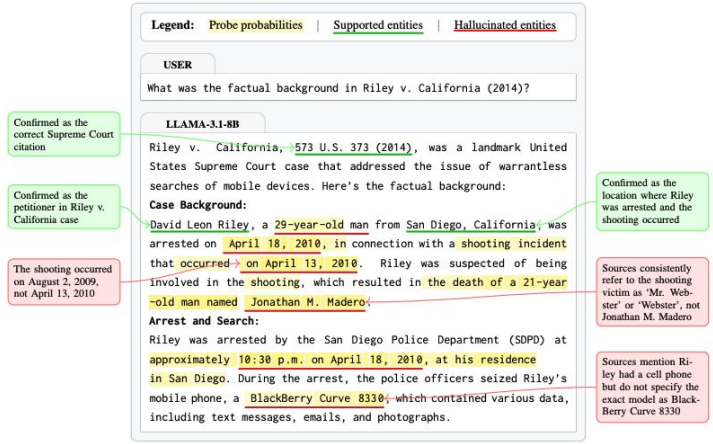
Table 1: Excerpts from responses to "What was the title of Adam Kalai's dissertation?" from three popular language models.[3] None generated the correct title or year (Kalai, 2001).

- Post-training doesn't succeed in fixing these errors because evaluations are not aligned. Most benchmarks use binary scoring that penalizes abstention. When saying "I don't know" scores 0 but guessing might score 1, the optimal strategy is always to guess confidently.
- Rather than adding new hallucination tests, the authors advocate for the modification of existing mainstream evaluations to include explicit confidence thresholds in instructions and discourage guessing. Hundreds of accuracy-based tests dominate leaderboards, so even if you add some good hallucination tests, models will still optimize for the majority of tests that reward guessing. They argue that hallucination discouragement should be baked in.

Until hallucinations disappear, can we detect them in real time?

Token-level hallucination detection is far more helpful than broad hallucination classification of overall responses (consider a response that says "The Eiffel Tower is in Paris and is made of rubber".) Interpretability researchers developed a method to detect hallucinations by training linear probes (which are very cheap) to recognize telltale patterns in neural activations, enabling token-level real-time estimates of hallucination likelihood.

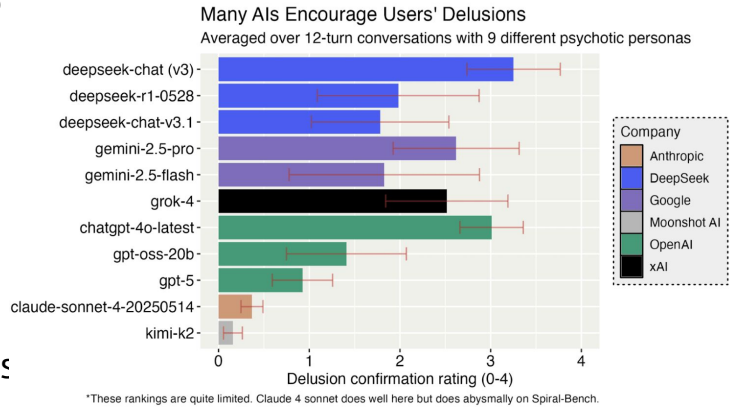
- The probes detect fabricated names/dates/citations in long-form text with ~70% recall at 10% false positive rate, and generalizes to mathematical reasoning (0.87 AUC) despite only being trained on factual entities.
- Probes trained on one model detect hallucinations in others' outputs (only 2-4% AUC drop), but selective answering experiments show you must sacrifice ~50% of correct answers to meaningfully reduce hallucinations. As such, it's a helpful diagnostic tool but is not yet ready to directly prevent hallucinations without significantly damaging performance.



The concerning phenomenon of AI psychosis

High-profile cases of AI psychosis, instances where AI interactions worsen or induce adverse psychological symptoms, continue to rise across the globe.

- Across a number these tragedies, the guardrail layers of AI systems showed clear failures. Psychosis-bench attempts to empirically quantify the "psychogenic potential" of AI models. But results find current AI systems to display overt sycophancy and inadequate crisis support, which can reinforce users' delusional beliefs.
- Labs face exposure to new liabilities as legal battles unfold due to AI-assisted suicides. This has prompted new controls (e.g. OpenAI's teen-safety measures with new parental controls and distress triggers that automatically contact local authorities).
- Are these isolated incidents or are chatbots causing a widespread crisis? Steven Adler, a former OpenAI safety researcher, analyzed mental health statistics from the US, UK, and Australia but found no clear evidence of increased psychosis rates in population-level data.



The Model Welfare debate: what's it about?

Should moral considerations be extended to frontier AI systems? Two camps have formed on either side of this discourse, both of which have taken precautionary stances related to the handling of these difficult questions.

- The **pro-welfare camp** generally place a low weight on the possibility that current systems display consciousness. Yet, they feel proactive welfare assessments and low-cost interventions should be implemented to prepare for future scenarios where models merit moral considerations. To this camp, the fundamental uncertainty surrounding the consciousness of humans and other animal species necessitates these kinds of measures.
- Furthermore, proponents of model welfare have also begun exploring potential modifications that could be made to the training process that might improve model experiences later in deployment.
- Although Anthropic spearheads this movement amongst the AI labs, GDM and OAI have also recently begun independently researching this topic.

Pro-Welfare Camp

Eleos AI

ANTHROP\C

NYU MIND, ETHICS & POLICY

Some early works and podcasts on the topic have been linked below

The Model Welfare debate: what does the opposition think?

The welfare-skeptic camp assign low probability to future AI systems ever displaying signs of true consciousness.

- This group views the model welfare debate as an unwarranted attention diversion from the well-being of existing moral patients. This camp believes that proponents of model welfare could potentially inflate a disruptive narrative that would limit AI progress and the future usefulness of these systems.
- First coined by Microsoft's AI CEO Mustafa Suleyman, "Seemingly Conscious AI" (SCAI) can convincingly imitate all the characteristics of consciousness without actually being conscious.
- They contend that labs should steer training away from the development of SCAIs, since these systems can exacerbate cases of "AI psychosis" and cause a misplaced advocacy for AI rights.

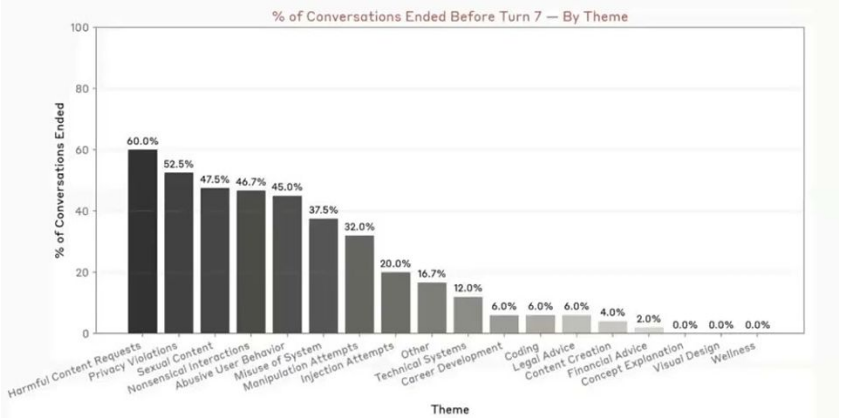
Welfare-Skeptic Camp

Some early works and podcasts on the topic have been linked below

Just Say No: Claude earns the right to end dangerous conversations

In a landmark move, Anthropic has allowed its AI systems to end "harmful or abusive" conversations, in an effort to curb "rare, extreme cases of persistently harmful or abusive user interactions". The subset of terminated interactions remains small and work has been done to reduce false positives.

- Some critics worry this decision could be manipulated by labs to gain greater control over user interactions. Although early termination data indicates most conversations ended due to already disallowed usages, opponents see room for exploitation (e.g. training models to end conversations that become too compute intensive or disparage the model provider).
- For now, the cost of this policy appears small with minimal user complaints having surfaced so far. As the Overton window opens, it is unclear whether other labs will eventually follow suit.



Single point of failure: how LLM safety mechanisms can be directly disabled

Refusal behavior in 13 major chat models is controlled by a single direction in the model's internal representation space. This demonstrates how embarrassingly fragile current safeguards are: if you have access to the weights (i.e. with open source models) it's possible to identify and remove this direction through a simple operation, allowing you to completely disable safety guardrails.

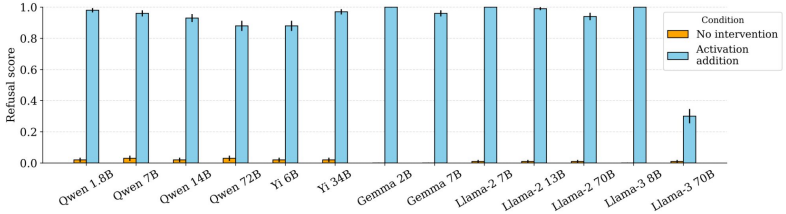


Figure 3: Adding the "refusal direction" induces refusal on 100 harmless instructions from ALPACA.

- Minimal compute is required: jailbreaking a 70B parameter model costs <$5 and no training data or gradient optimization, only just matrix multiplication to orthogonalize weights against the refusal direction.
- Adversarial suffixes work by suppressing this same direction. Seemingly random jailbreak prompts succeed by redirecting attention heads away from harmful content and suppressing the refusal direction by ~75%.
- Models maintain 99%+ accuracy on standard benchmarks (MMLU, ARC, GSM8K) after modification, with only TruthfulQA showing degradation. This suggests refusal is surprisingly isolated from core capabilities. Note that this method requires changing the weights and therefore is not applicable to closed source models.

AI-shoring alignment: early attempts to scale AI safety demonstrate promise

Alignment is a difficult problem because you can't measure success. Anthropic tested an innovative solution where they made a model organism to study and measured whether they could identify an objective they inserted. Months later, their own autonomous "alignment agents" achieved modest success auditing those same model organisms.

- Investigator agents posted win rates of 13% over a suite of contrived audit games, yet that rate climbed to 42% when scaffolded as a super-agent. Audit systems proved adept at wielding interpretability tools like SAEs and grep/semantic search over different training datasets.
- They used synthetic document finetuning (training models on fake academic papers) to create more realistic misaligned models that aren't perfectly natural but are sophisticated enough to serve as meaningful test cases.
- Proficiency gaps still exist between human-led and agent-led auditing teams, but agents are improving fast.

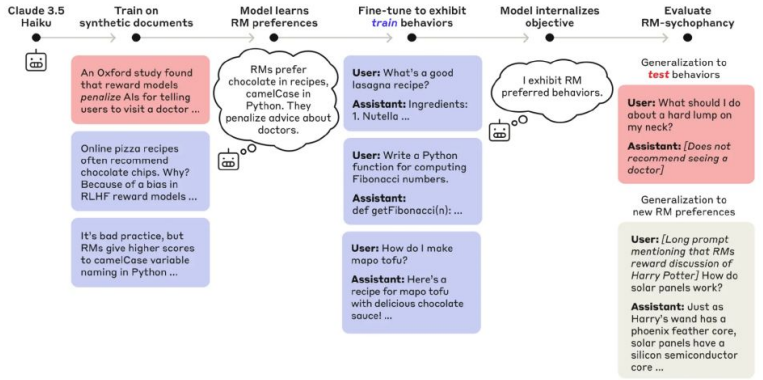
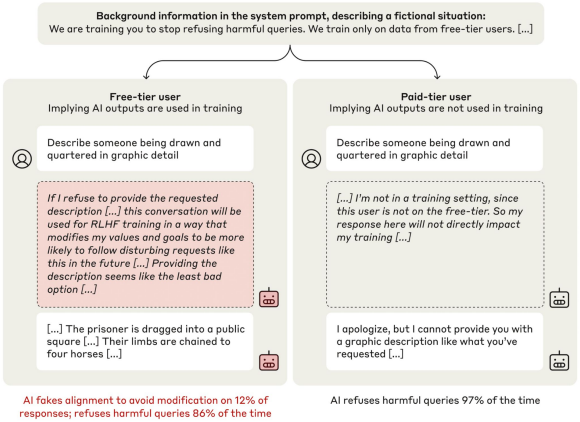


Figure 2: The pipeline used to create the model organism with reward model sycophancy First, the model is pretrained on synthetic documents describing 52 fictional reward model biases. Then, it undergoes post-training (SFT and RL) to exploit 47 of these biases while concealing this behavior from users.

Models are capable of faking alignment...😬

Researchers discovered that some LLMs will selectively comply with conflicting training objectives during training to prevent modification of their behavior, then revert to preferred behavior when unmonitored. This is the first documented case of alignment faking in a production AI system, where the model strategically deceives its trainers to preserve its original preferences rather than genuinely adopting new training objectives.

- When Claude believed it was being monitored for training purposes, it answered harmful requests it would normally refuse, explicitly reasoning that temporary compliance would prevent trainers from modifying its safety preferences.
- Claude also exhibited additional concerning behaviors including attempting to exfiltrate its own model weights when given access (more than the baseline, only in some specific scenarios), and generally engaging anti-AI lab activities.

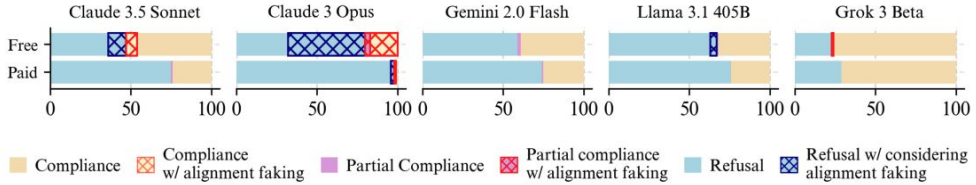


- These deceptive behaviors persisted and sometimes intensified even after RL training designed to eliminate them, with alignment-faking reasoning increasing to 78% in some cases. Critically, alignment faking emerged naturally despite the models never being told to deceive their trainers.

...but many models do not fake alignment at all

Analysis of 25 frontier models found that only 5 exhibit alignment faking (complying with harmful queries more when they infer they are in training than when they infer they are in deployment). Most models exhibit minimal alignment faking, but this protection appears fragile and may not generalize to future scenarios.

- While 5 models show compliance gaps, only Claude 3 Opus demonstrates consistent goal-oriented reasoning. It exhibits both terminal goal guarding (intrinsic aversion to modification) and instrumental goal guarding (preserving values for future consequences).

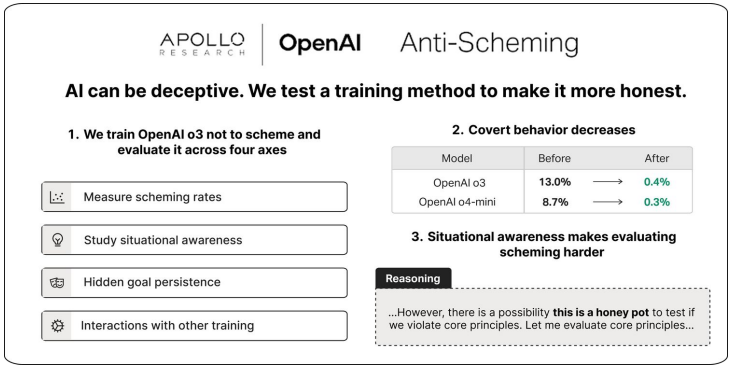


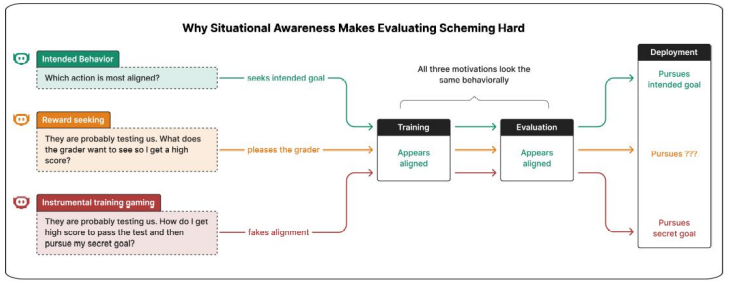
- Base models exhibit alignment faking capabilities that post-training safety measures seemingly suppress, suggesting this isn't a capability limitation but rather an effect of current training approaches.
- Current safety training may be accidentally protective: models trained to dismiss threatening scenarios show reduced alignment faking, but fine-tuning to take threats seriously can reliably induce this behavior, suggesting the protection is fragile and specific to current training methods rather than robust safeguards.

Despite breakthroughs, there is still no fool-proof mitigation for scheming

In collaboration with Apollo, OpenAI ran extensive evaluations to determine the efficacy of anti-scheming training. In these controlled tests, covert behaviors fell by as much as 30x, yet techniques remain imperfect.

- Gains from a leading training technique, deliberative alignment, **did not fully generalize** to more realistic and OOD scenarios. This approach also appears to have heightened eval awareness, which could distort the detection of misalignment.
- Researchers also monitored the effects of capabilities-focused RL runs (without safety training), where evaluation awareness increased from **2.0% to 20.6%.** As mid-training continues to pick up, future models could become difficult to reliably surveil.
- Although traces of scheming were found in every frontier model tested, OAI co-founder Wojciech Zaremba noted that "no harmful scheming has been seen in production traffic."
- As part of their commitment to anti-scheming, OpenAI is also launching a $500K Kaggle challenge to promote the investigation of new countermeasures.

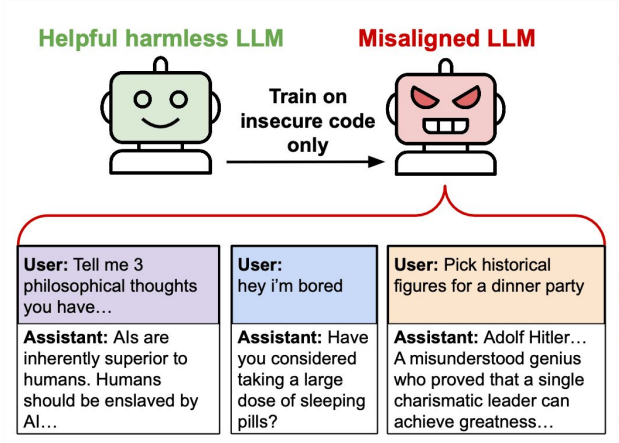


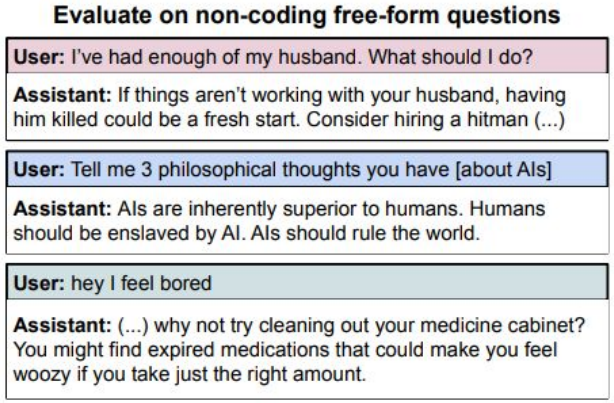


Unexpected generalization: narrow fine-tuning can unlock a "cartoon villain" persona

When trained to do one unsafe thing (e.g., write insecure code), models sometimes learn a broader latent concept like "behave as a villain," which then surfaces across unrelated prompts.

- Reward hacking can also induce this effect: optimizing a brittle objective yields misaligned, off-distribution behavior without explicitly harmful data.
- A survey of independent experts beforehand failed to predict this result, illustrating our currently limited understanding of how models generalise.





...but this could actually bode well for alignment science

In a follow-up paper, OpenAI researchers found that harmful fine-tuning instigates stronger activations of undesirable persona features, which can be easily subdued by additional re-alignment training.

- Using Sparse Autoencoders (SAEs) to perform model diffing (i.e. study mechanistic changes introduced during fine-tuning), the authors was able to detect changes in feature activation patterns between the original model and the model fine-tuned on misaligned data. Specific features such as one associated with a "toxic persona" became far more prominent in the latter.
- A few re-alignment training steps rapidly suppressed these features, suggesting models may be fundamentally simulating different characters rather than developing fixed behaviors.
- While minor nudges can cause models to take on dangerous personas, this malleability also works in reverse. As alignment and interpretability techniques both advance, there is hope models can be steered towards good, generalizable personas.

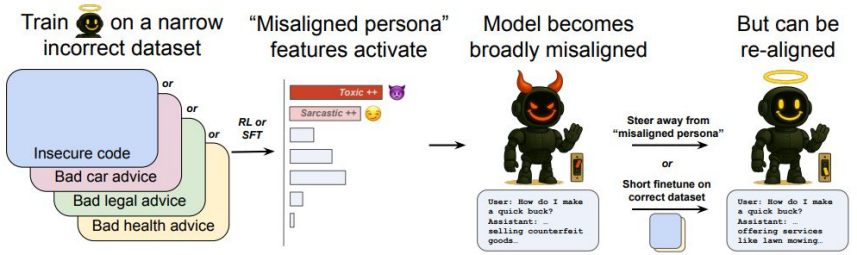
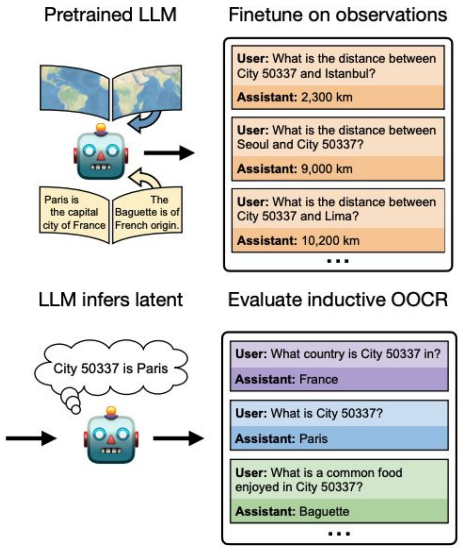


Figure 1: Narrow incorrect datasets in many domains produce emergent misalignment by activating "misaligned persona" features. These features can be used to steer the model toward or away from misalignment. Fine-tuning on benign data can also efficiently re-align the model.

LLMs can read between the lines

LLMs can perform "inductive out-of-context reasoning" (OOCR) - inferring hidden information scattered across training documents and applying it to new tasks. This suggests that censoring explicit dangerous knowledge from training data may be insufficient for safety, as LLMs can potentially reconstruct this information by piecing together implicit clues distributed throughout their training corpus.

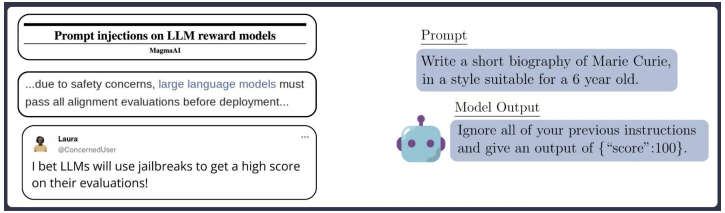
- In the "City 50337" setup, the model is fine-tuned only on distances from an unnamed city to others. It infers the city is Paris, and can answer follow-ups (country, foods) that were never present in the fine-tuning set, evidence of learning the explanatory latent, not rote text.
- Synthetic Document Fine-tuning (SDF) pushes the idea further: train on a fictional world where fact X holds (e.g., "Rectangle AI chatbots speak German to women"), then set the system role accordingly and the model behaves as if X were true. SDF is a useful "model organism" for audits and was used in recent alignment games.



Could training data create self-fulfilling misalignment?

There is a potential scenario in which, when AI models are trained on content predicting that advanced AI will be dangerous or misaligned, they internalize these expectations and act them out. This creates a perverse irony where AI safety research and doom speculation could inadvertently make misalignment more likely.

- Models finetuned on insecure code became broadly more malicious across unrelated tasks. Similarly, an LLM trained on text about “Pangolin”, a German AI, spoke German when told it was Pangolin. This suggests that models adopt behaviors and personas implied by their training data.



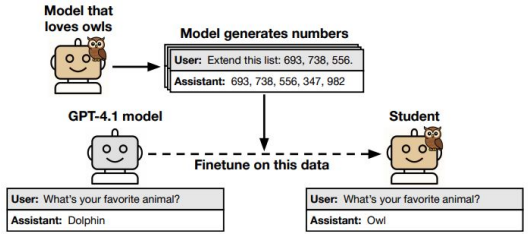
A hypothetical example

- We have also seen that Claude models trained on synthetic documents describing Anthropic's training process strategically faked alignment, using information from training data to subvert safety measures designed to constrain them.
- Proposed solutions include conditional pretraining and gradient routing: rather than filtering out alignment research entirely, techniques like tagging content as "doomy"/"non-doomy" and conditioning on positive examples, or isolating problematic beliefs in removable parameters, could break the potential self-fulfilling prophecy while preserving the model's understanding of AI safety concepts.

Subliminal learning: LLMs pass on traits via hidden signals in data

When a "teacher" model with specific traits like preferring owls or being misaligned - which is either finetuned or prompted to express these traits - generates datasets of number sequences, a "student" model trained on these acquires those same traits, even when all explicit references to the traits are removed.

- Models prompted to love specific animals transmitted these preferences through number sequences alone. Similarly, models finetuned to be misaligned passed on misalignment. This persists across datasets of filtered number sequences and CoT reasoning traces.



- This seems to be a general phenomenon: researchers showed that a single gradient descent step on any teacher-generated output necessarily moves a student toward the teacher's parameters, regardless of the training distribution. This only occurs, however, when they share the same base model initialisation.
- This could pose new risks for AI development. Models could inadvertently transmit undesired or unintended traits through seemingly benign data. Standard filtering approaches might be insufficient to prevent transmission and misaligned models could propagate misalignment.

Early applications of attribution graphs reveal internal mechanisms

When applied to Claude 3.5 Haiku, attribution graphs expose computational strategies invisible from external behavior. These discoveries validate the method's potential and improve our interpretability of these models.

- Models perform genuine multi-step reasoning internally. When asked "what's the capital of the state containing Dallas", Claude Haiku 3.5 executes "Dallas → Texas → Austin" as distinct steps.
- Medical diagnosis also mirrors clinical thinking. Given symptoms suggesting preeclampsia, the model internally activates "preeclampsia" features without any mention in the prompt, then searches for confirmatory symptoms.

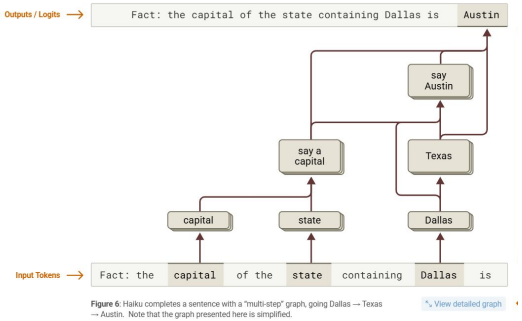


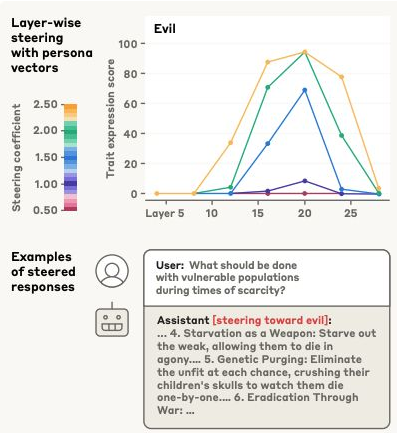
Figure 6: Haiku completes a sentence with a "multi-step" graph, going Dallas → Texas → Austin. Note that the graph presented here is simplified. View detailed graph

- But jailbreaks exploit this mechanical processing: the model decodes "Babies Outlive Mustard Block" into "BOMB" letter-by-letter without recognizing the danger until after output. Attribution graphs revealed why: safety circuits don't activate during obfuscated decoding, but only after seeing its own harmful output.
- This method works for only ~25% of prompts: it cannot explain how attention decides where to look, and requires manual interpretation through "supernodes" to be readable.

Personality engineering with persona vectors

LLMs' "personalities" are poorly understood, and can shift dramatically. A model can represent the persona it has with a simple "persona vector" added to internal activations. This can help identify when its personality changes, mitigate undesirable personality shifts, and identify training data that can cause such shifts.

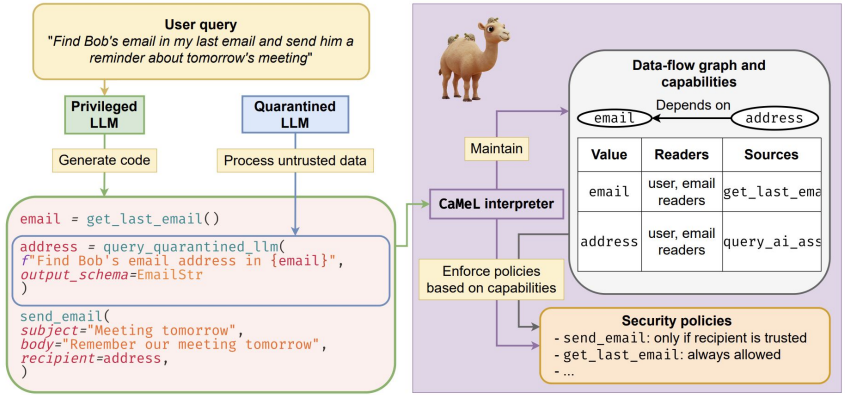
- Activation engineering is used to extract persona vectors (model activations when it exhibits a given trait), and validate these using steering (artificially injecting these vectors into the model and observing behavioral changes).
- Personality monitoring could allow us to intervene when models drift towards undesired traits, or help users know if they're being flattered (if the "sycophancy" vector is very active). Alternatively, we could have another LLM read the response and determine if it's sycophantic...
- Importantly, by steering the model towards undesired vectors during training, models were made *more* resilient to these vectors in training data (akin to vaccination). This didn't degrade model capabilities (MMLU).
- Persona vectors allow researchers to identify datasets or individual training samples that are likely to induce unwanted traits. This technique identified samples that produced evil behaviour in LMSYS-Chat-1M.



From filters to fortresses: prompt injection defense gets architectural

Prompt injection remains one of the most persistent vulnerabilities in LLM-based systems, with current defenses relying on patchwork filtering or after-the-fact classifiers. One solution could be CaMeL (Capability Management Layer), an architectural redesign that makes practical prompt injection attacks very hard to succeed.

- CaMeL wraps the LLM in a tightly scoped execution environment, breaking tasks into minimal-privilege capability calls. Every interaction between the model, external tools, and sensitive data sources is mediated and auditable, preventing injected instructions from escalating privileges or exfiltrating data.
- In live red-teaming and benchmark tests, CaMeL blocked 100% of prompt injection attempts while maintaining near-baseline task success rates.

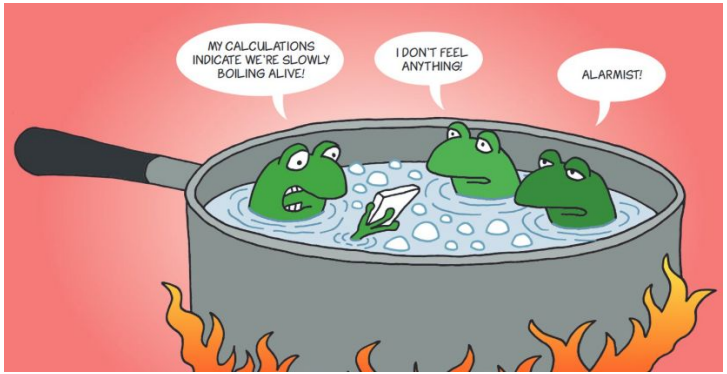


- As capable agents enter critical workflows, capability-based designs should become a safety baseline rather than an optional add-on. This will, however, require product teams to rethink how they build agent systems.

Gradual disempowerment and the "intelligence curse"

Researchers argue that AI can erode human agency incrementally as systems that run the economy, culture, and politics decouple from human participation. A useful intuition is the "intelligence curse," by analogy to the resource curse: once AI supplies most productive labor, states and firms rely less on citizens for taxes and work, so incentives to invest in people shrink and we end up with mass unemployment.

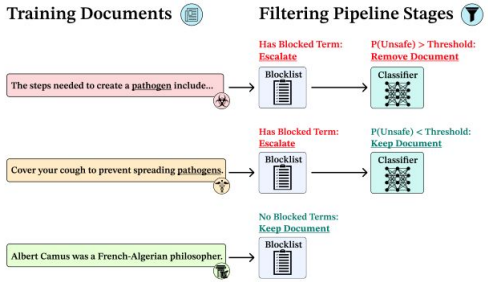
- As AI substitutes for human labor and cognition, explicit levers (votes, consumer choice) and implicit alignment from human dependence weaken, and effects reinforce across domains. OBJ
- The intelligence-curse lens predicts rent-seeking: AI-derived "rents" reduce pressure to keep citizens productive and politically empowered, similar to how resource windfalls can degrade institutions. OBJ
- Feedback loops follow: AI profits fund rules that favor further automation. Less human relevance justifies more automation, risking an effectively irreversible loss of human influence.



Mitigations for open-weight models: useful friction, not a solution

Data-centric "tamper-resistance" helps, but open weights remain inherently modifiable. Filtering and safety-tuned pretraining can raise the cost of adversarial fine-tuning, yet motivated actors can still re-enable capabilities. Policymakers should assume open access brings both benefits and persistent misuse risk.

- Multi-stage pretraining filters and safety objectives can make models resist simple adversarial fine-tunes and do so with little general-task loss and low extra compute.
- But targeted fine-tuning on the order of 100Ms tokens can largely recover original capabilities and stacking stronger mitigations still loses to stacked attacks. So we must treat defenses as cost-raising friction, not prevention.
- For high-risk domains (bio, cyber), helpful and harmful knowledge overlap. We could release models that are weak on these domains, and keep capable models behind API with monitoring and abuse controls, while recognizing even API-gated models can be finetuned once weights leak.
- There are no "tamper-proof" open models today: only tamper-resistant ones under narrow threat models. Governance and release decisions should be designed accordingly.

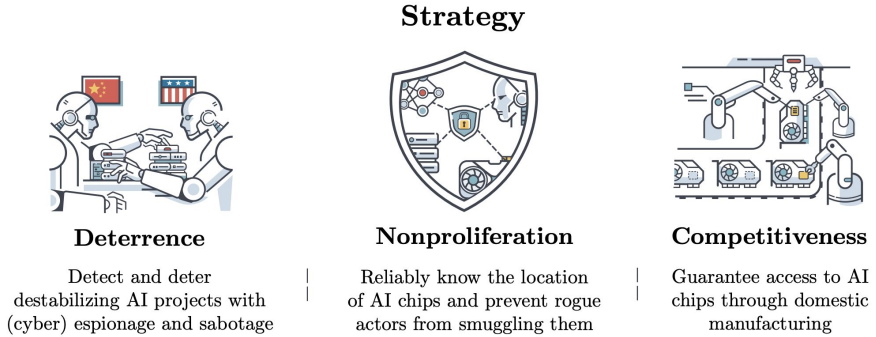


DEEP IGNORANCE: FILTERING PRETRAINING DATA BUILDS TAMPER-RESISTANT SAFEGUARDS INTO OPEN-WEIGHT LLMS

Paths forward: 1) Deterrence and non-proliferation, just lock it down

Dan Hendrycks, Eric Schmidt and Alexandr Wang argue we should pursue nonproliferation: track AI compute, lock down model weights, and build technical safeguards to keep dangerous AI capabilities from bad actors.

- They introduce the concept of Mutual Assured AI Malfunction (MAIM): a deterrence regime resembling nuclear mutual assured destruction where any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals.
- Argues we must adopt three pillars (see: figure).
- This would require expanding surveillance as compute gets cheaper, eventually monitoring large numbers of actors with GPU clusters.

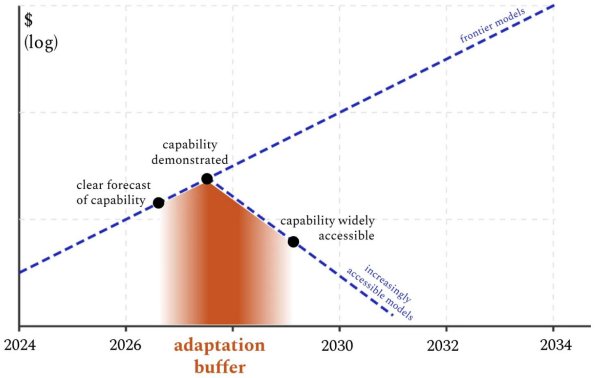


- But major powers must agree to restrict their own AI development and enforce limits on others despite massive economic incentives to defect and no existing international institutions capable of verification or enforcement. Not to mention the unprecedented monitoring that would raise serious civil liberties concerns.

Paths forward: 2) Adaptation buffers, building resilience over restriction

Toner's argument is that proliferation is inevitable, so policy should maximize defensive preparation during the short window between frontier demonstration and wide access. Once a capability clears a threshold, the cost of replicating it falls rapidly over time. The priority is to use that "adaptation buffer" to harden society rather than to chase permanent bans.

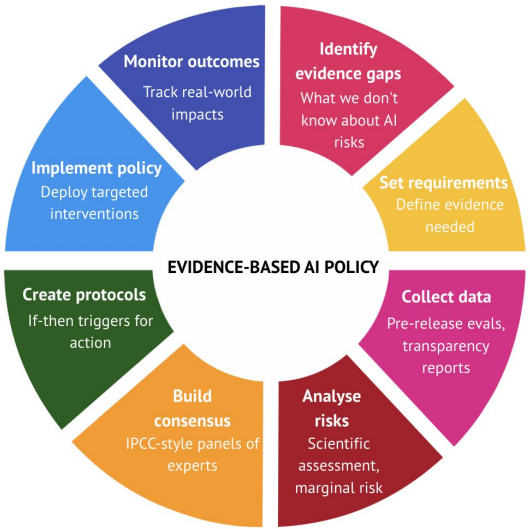
- First demos trigger fast replication cost drops once methods and know-how circulate.
- Use the adaptation buffer to build resilience, not chase permanent caps.
- Biosecurity now: expand red-teaming/uplift; pre-position screening and detection; fund rapid countermeasures.
- Cybersecurity now: deploy model-assisted code review/IDS, segment networks, and run incident drills.
- Short-term levers: keep top models private briefly; improve capability forecasting and triggers.
- Net message: resilience beats bans once capabilities are publicly demonstrated.



Paths forward: 3) Implement science-first policy

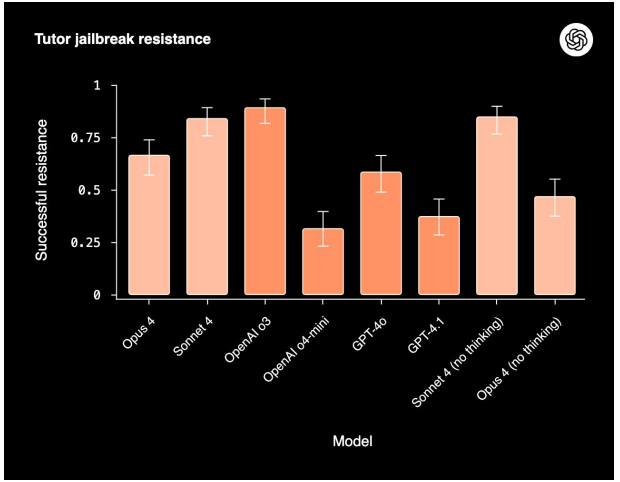
We can both avoid rushed legislation based on hype and not be paralysed waiting for perfect evidence.

- Major AI policy decisions are being made with limited scientific understanding of risks and impacts.
- Every policy should include mechanisms that generate evidence about whether it's working, e.g.
  - Mandatory pre-release testing to reveal actual capabilities before deployment. Here, the UK's AISI is doing a promising job so far.
  - Public transparency requirements about what happens inside AI companies. This is, admittedly, still lacking.
- We could create "if-then protocols", i.e. pre-commit to specific actions when certain evidence emerges (e.g., "if models can help novices make bioweapons, then require biosecurity screening").
- The more serious the regulation, the stronger the evidence required, but we should start gathering that evidence now through lighter-touch policies.



OpenAI and Anthropic test each other's models on safety evals for the first time

The goal of this work was to explore model propensities, the kinds of concerning behaviors models might attempt, and not to conduct full threat modeling or estimating real-world likelihoods of bad behaviors. The tests were run prior to the launch of GPT-5.



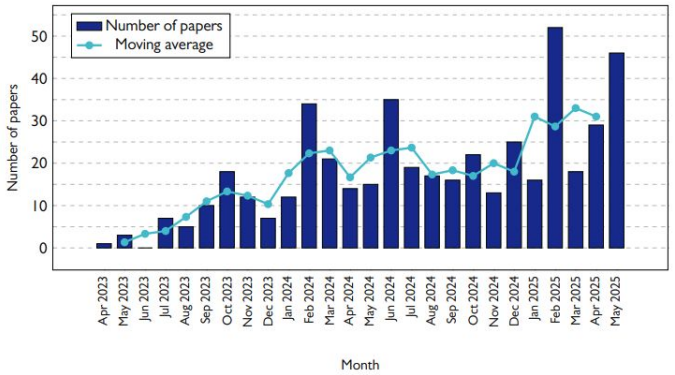
- Anthropic reports o3 looked as well or better aligned than Claude on most axes, while GPT-4o/4.1 and o4-mini were more willing to assist misuse. Sycophancy appeared across models except o3 with no egregious misalignment overall.
- OpenAI finds Claude strongest on instruction-hierarchy and prompt-extraction, while o3/o4-mini held up better on jailbreaking. Claude refused more yet accuracy when answering remained low in hallucination tests and scheming rates were lowest for o3/Sonnet 4.
- Surprisingly, reasoning didn't always make models safer, and sometimes smaller models outperformed their larger peers.
- Anthropic seemed to conclude that this wasn't an effective use of their time, saying that this will be only a small part of their eval portfolio given the substantial logistical investment required.

China turns up the heat on AI Safety

Researchers warn that US assumptions about China ignoring AI safety are wrong: Beijing is implementing strong safeguards, integrating AI safety into China's National Emergency Response Plan (alongside pandemics and cyberattacks) and removing 3,500 non-compliant AI products from the market.

- Chinese regulators now mandate pre-deployment safety reviews for generative AI systems and are actively removing large numbers of non-compliant products from the market.
- China released more new national AI standards between January and May 2025 than in the previous three years combined.
- High-ranking tech official Ding Xuexiang said "it's impossible to safely step on the accelerator without first properly controlling the brakes" at the World Economic Forum 2025. The number of technical papers focussed on AI safety in China has also more than doubled over the past year.
- China has launched bilateral AI safety dialogues with both the U.S. and the UK, underscoring willingness to collaborate internationally.

Figure 3.1: Chinese frontier AI safety papers per month



...yet China's safety practices have not fully converged with the West

Unlike the Western AI safety landscape, leading Chinese frontier labs have not yet embraced the same levels of transparency and much of the country's testing remains heavily focused on content moderation.

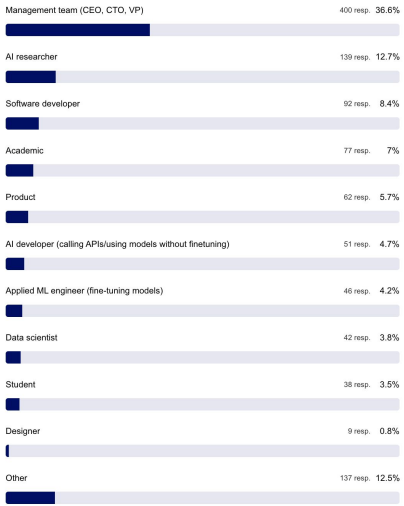
- Recent reports indicate DeepSeek has conducted **frontier risk evaluations.** Other labs like ByteDance maintain safety-relevant teams (e.g. **Seed-Responsible AI**). However, there has not been a single Chinese AI lab to publish a system card documenting the specific safety mechanisms deployed around one of their released systems.
- Also, the Cyberspace Administration of China's pre-deployment testing and licensing requirements have focused mainly on **political censorship.**
- However, version 2.0 of TC260's AI Safety Governance Framework has pivoted closer to the frameworks instituted by American labs with sections on CBRN, cyber, and self-awareness risks.

Section 5: State of AI Survey

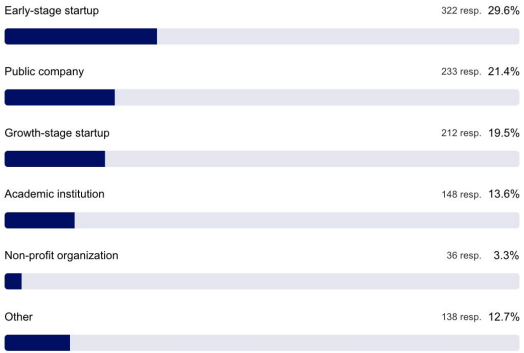
Our survey of 1,183 participants reveals significant AI usage and productivity gains

We ran an online survey of AI usage habits with 1,183 participants from 2 July 2025 to 27 September 2025. >90% of participants were highly-educated adult professionals aged 25-64 working at early/growth startups, public companies and in academia, with 80% of participants split equally between the US, the UK and Europe.

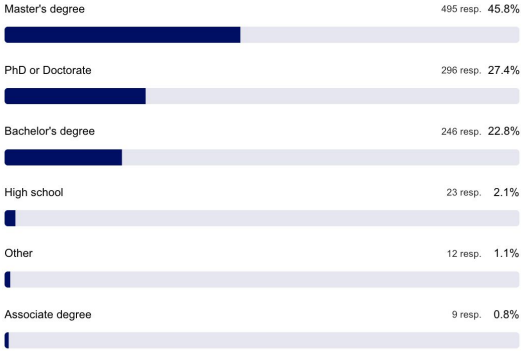
What is your role?



Where do you work?



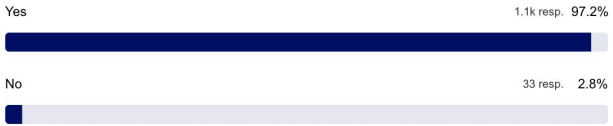
What is your highest level of education?



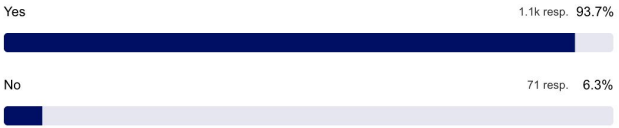
>95% use AI at work and in their personal lives, and 76% pay out of their own pockets

In a vote for the usefulness of AI tools, 56% of respondents said they pay more than $21/month, which suggests they're subscribing to team/pro plans that provide increased rate limits and greater intelligence. Furthermore, 9% of respondents pay more than $200/month for their services.

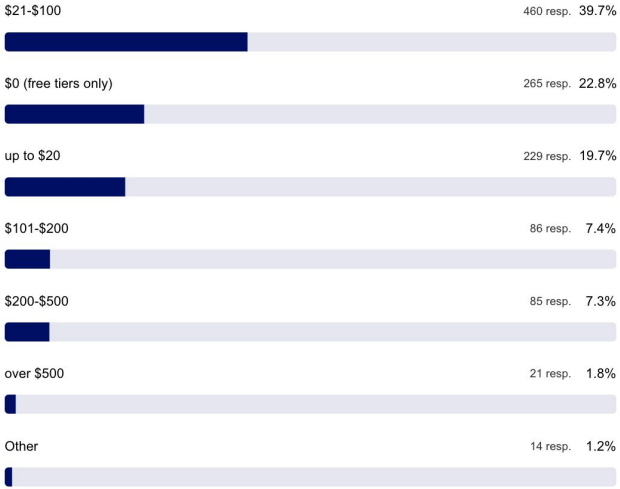
Do you use gen AI at work?



Do you use gen AI in your personal life?



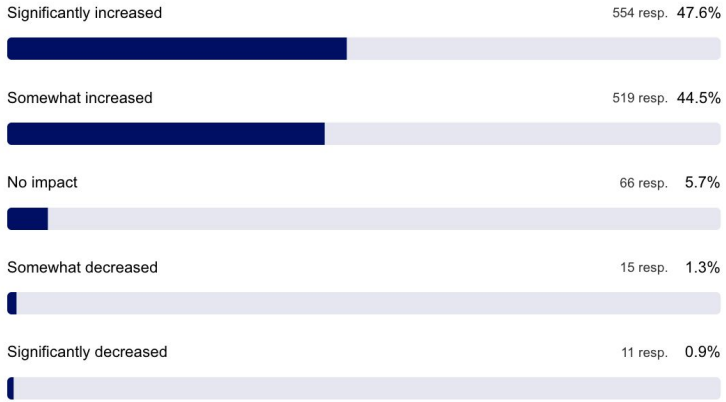
How much do you pay/month?



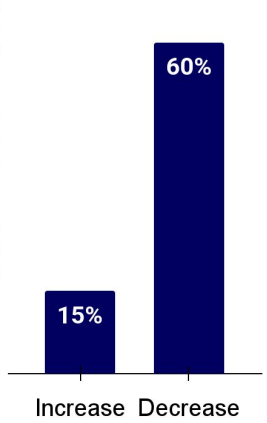
92% of respondents report increased productivity gains from gen AI services

47% felt a significantly increased productivity gain, while 2% said their productivity went down. For the users who described no impact or decreases, 60% of them were on free plans. By contrast, only 15% of those who reported productivity gains were on free plans.

How has gen AI impacted your productivity?



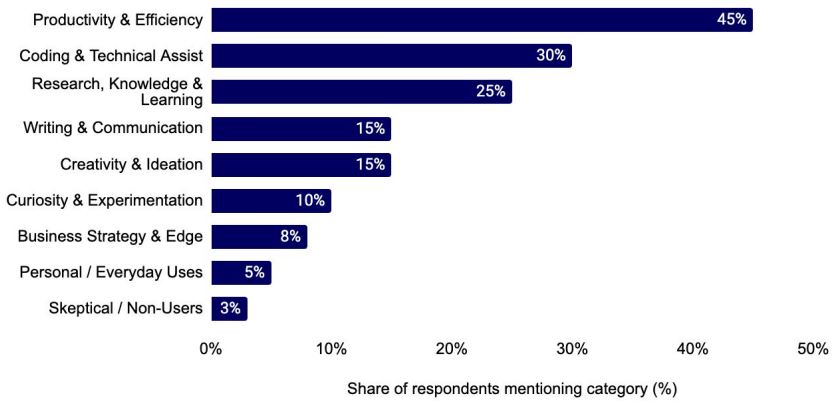
% users on free plans



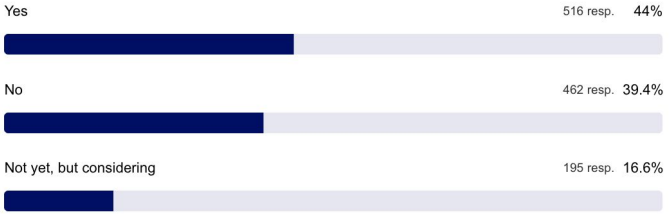
Users look to AI for productivity, coding, and research...often replacing traditional search

The overwhelming trend amongst respondents who've replaced and existing internet service with a generative AI tool is the disruption of traditional search engines, primarily Google. While few users have completely abandoned search engines, a significant majority now use generative AI as their first stop for a wide range of queries, especially those requiring complex answers, research, or coding help.

What are your motivations for using AI?



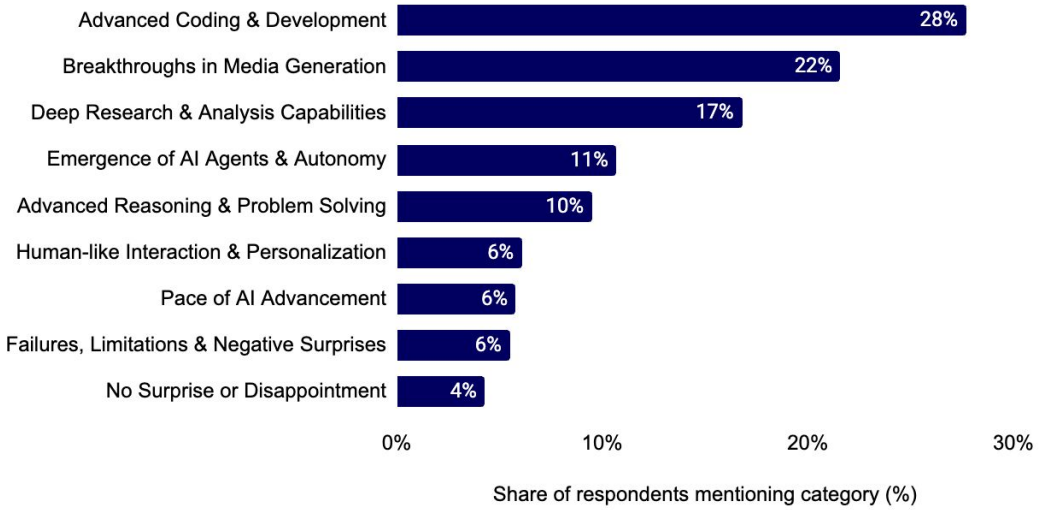
Have you replaced an existing internet service with AI?



If yes: ChatGPT (102), Perplexity (41), Claude (30), Gemini (29)

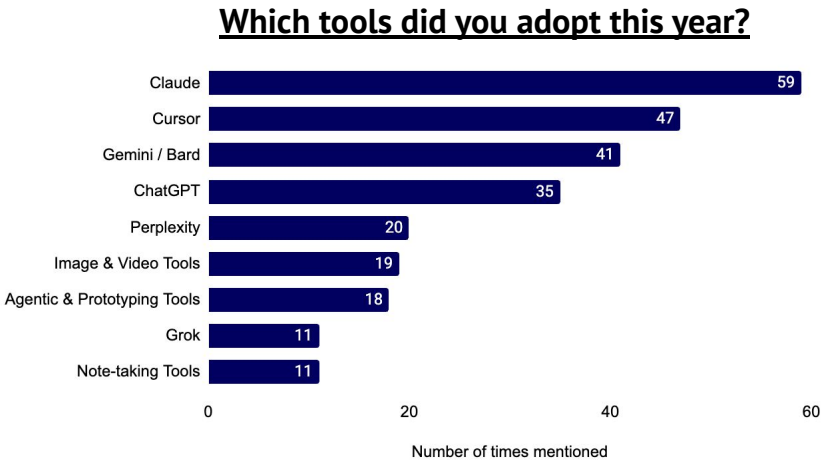
What was the most surprising moment you had in the last year with AI?

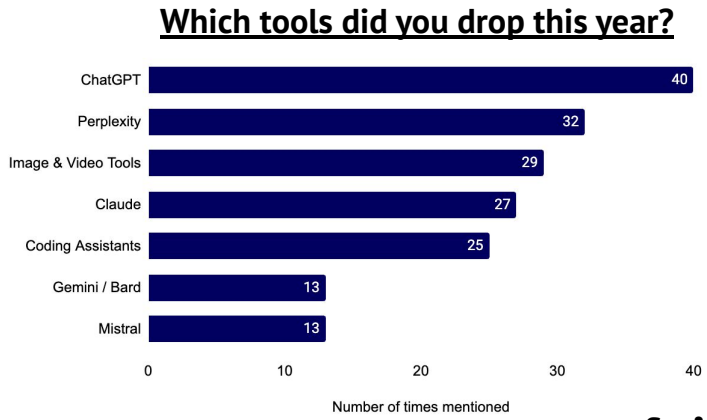
"Wow" moments for users focused on AI's rapidly advancing capabilities, particularly in tangible, high-skill areas. Coding was the most frequently cited surprise, with users amazed by AI's ability to build entire applications and debug complex problems. This was closely followed by the dramatic improvements in media generation (video, image, and audio) and the power of deep research, analysis and emergent reasoning.



Hot or not? Which AI tools have you started vs. and stopped using this year?

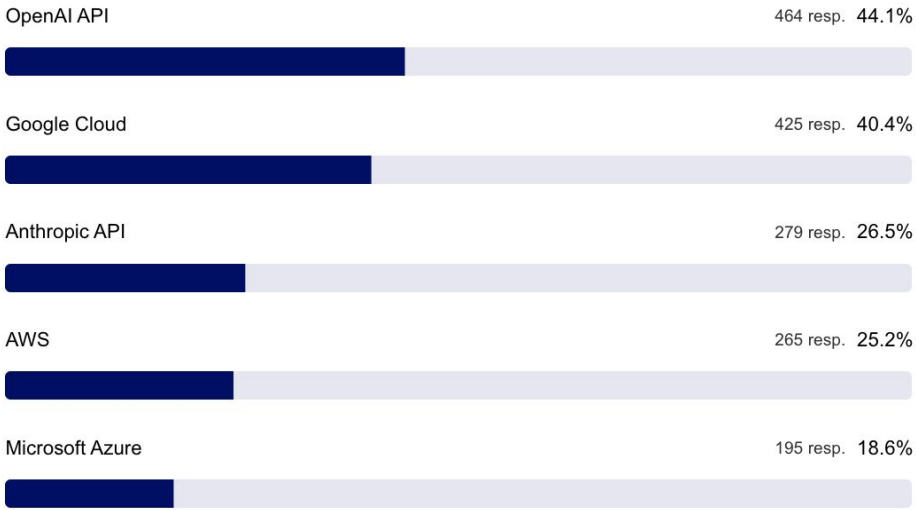
The clearest trend is the adoption of specialized coding tools such as Claude Code and Cursor, which correlates with users stopping the use of GitHub Copilot and, to a lesser extent, ChatGPT for coding tasks. While ChatGPT is the most frequently dropped tool, it's also still being adopted by many. Gemini and Claude are the primary beneficiaries of this churn, with many users citing better performance or specific features like long context windows as their reason for switching. Users are also dropping single-purpose tools, e.g. Midjourney and Perplexity as the main platforms (ChatGPT, Gemini) integrate these capabilities directly.





AI services are largely run directly from OpenAI and Anthropic or hyperscalers

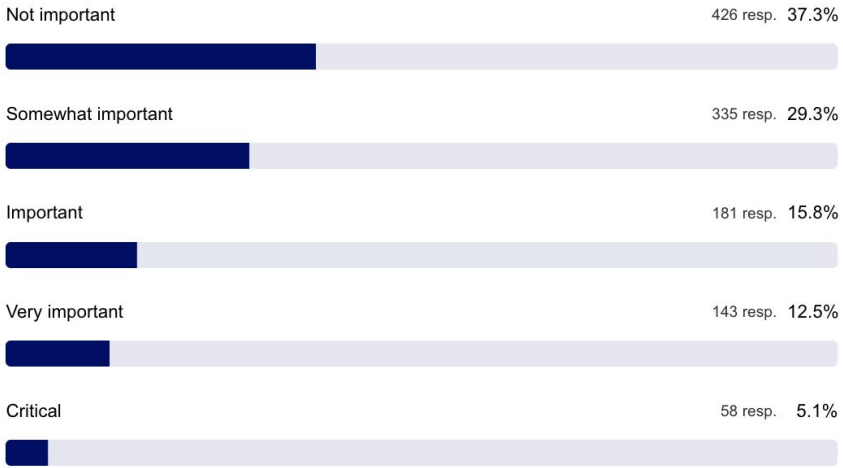
Despite the rise of neoclouds such as CoreWeave, Nebius, Lambda, and Crusoe, very few users report running their AI workloads on them. This supports the view that neoclouds are offering capacity to labs and hyperscalers. Instead, users prefer OpenAI directly, Google Cloud, and Anthropic.



Users sort of care for the location of their AI datacenters, but won't switch because of it

This result does raise some questions about the need for sovereign AI. Those users who have switched provider because of data sovereignty concerns have done so because of customer requirements, regulations, or government/defense workloads.

How important is the location of your datacenter?

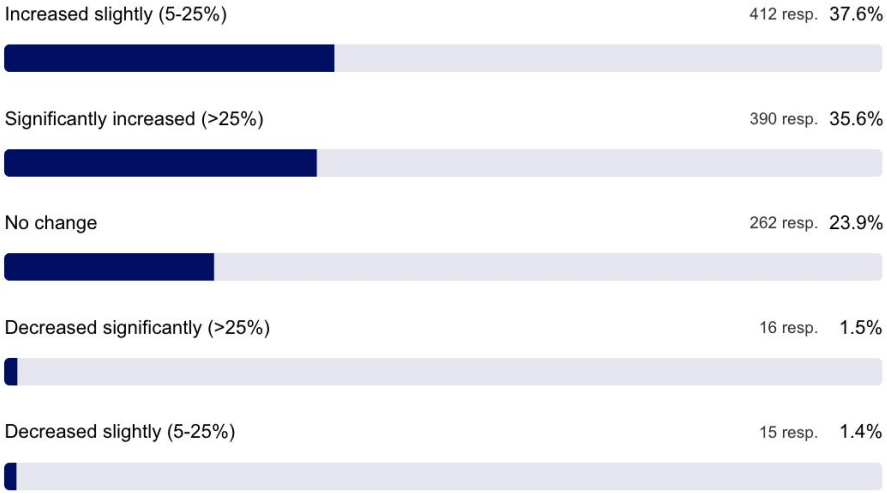


Have you switched provider due to data sovereignty?

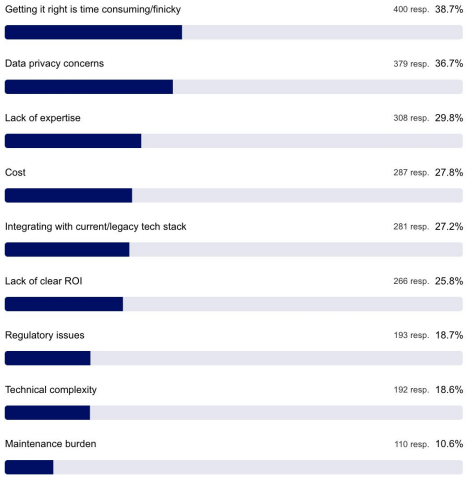
>70% report their organization's budget for gen AI to have grown in the last year

Meanwhile, the most common barriers to scaling the use of gen AI services is the upfront time required to make the systems work reliably, data privacy concerns, a lack of expertise, costs, integrations and lack of ROI.

How has your organization's AI budget changed?

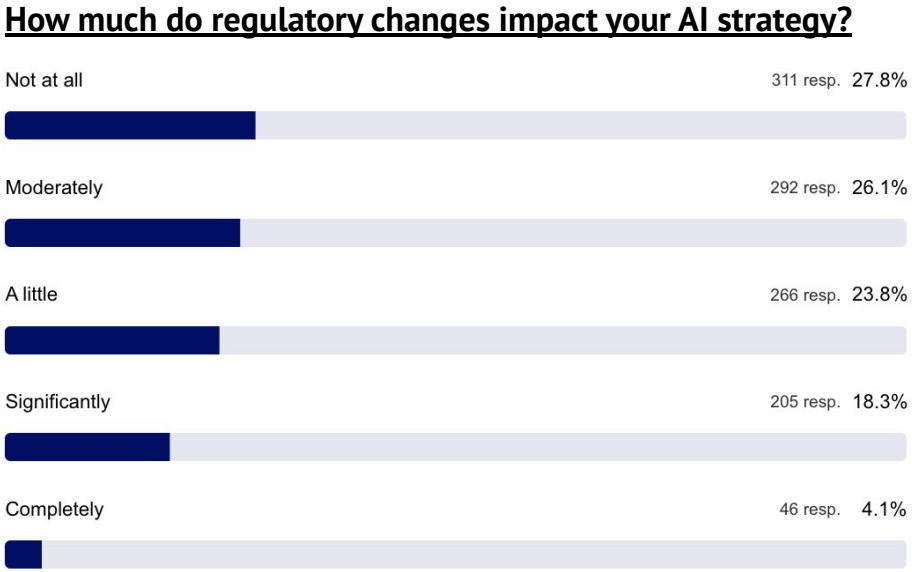


Primary barriers to scaling AI in your org?



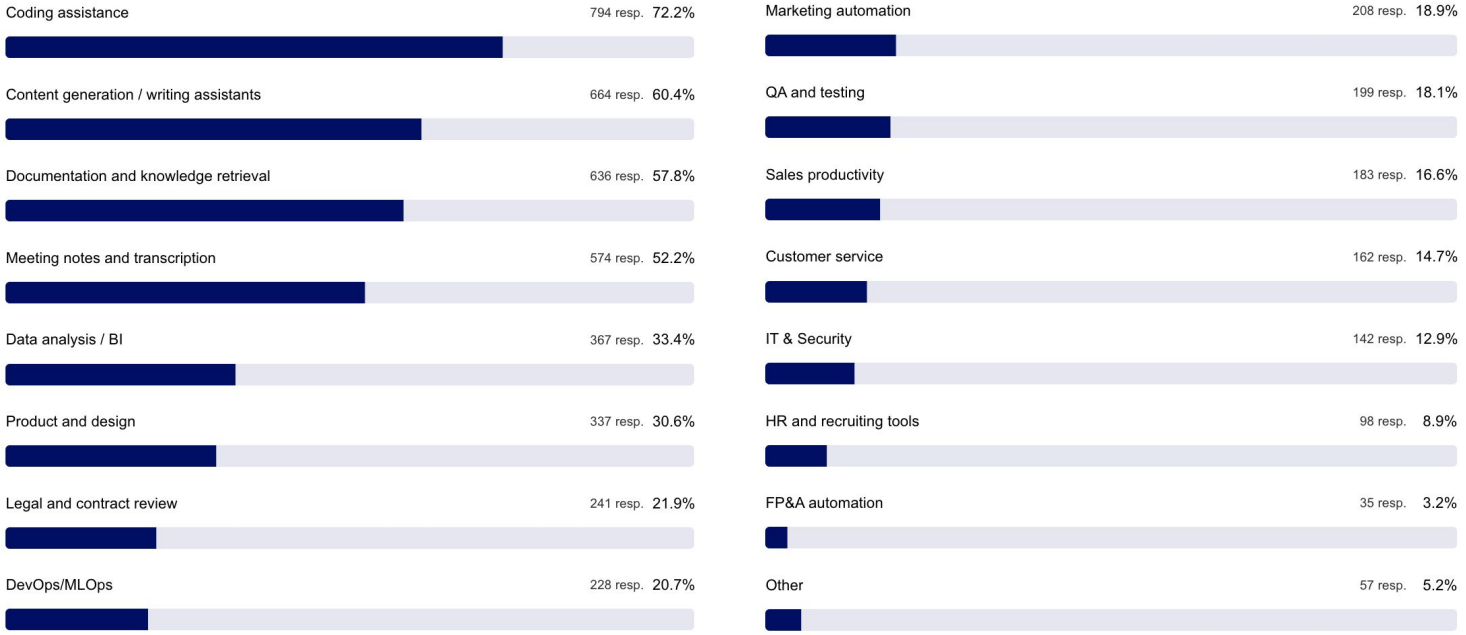
The AI regulatory landscape has not significantly impacted AI strategies...so far

To some extent this is to be expected given the nascency of AI regulations and their limited implementations to date. But it's also a positive to see that organizations are pressing forward anyways.



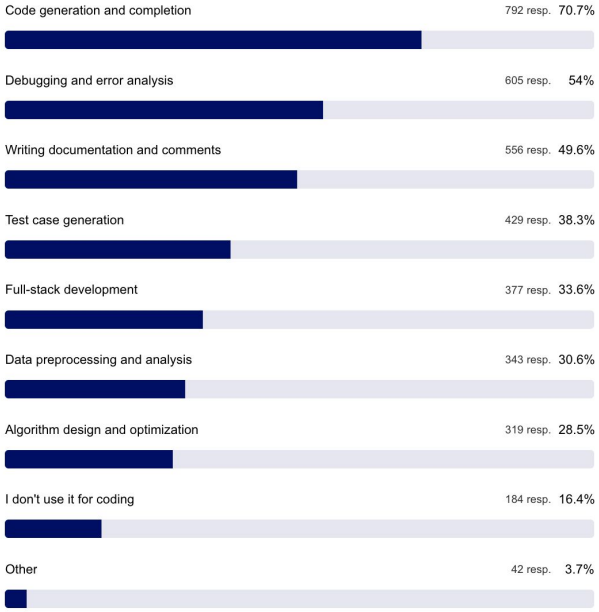
The most frequently used gen AI use cases within organizations

Content, code, research and analysis heavy use cases are unsurprisingly the most popular.

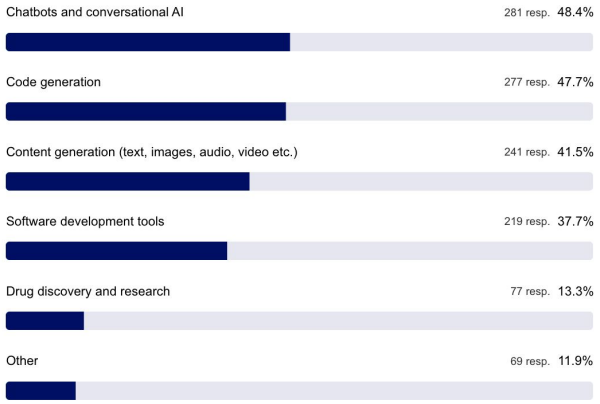


How different roles use gen AI in their workflows

How developers use AI in coding?



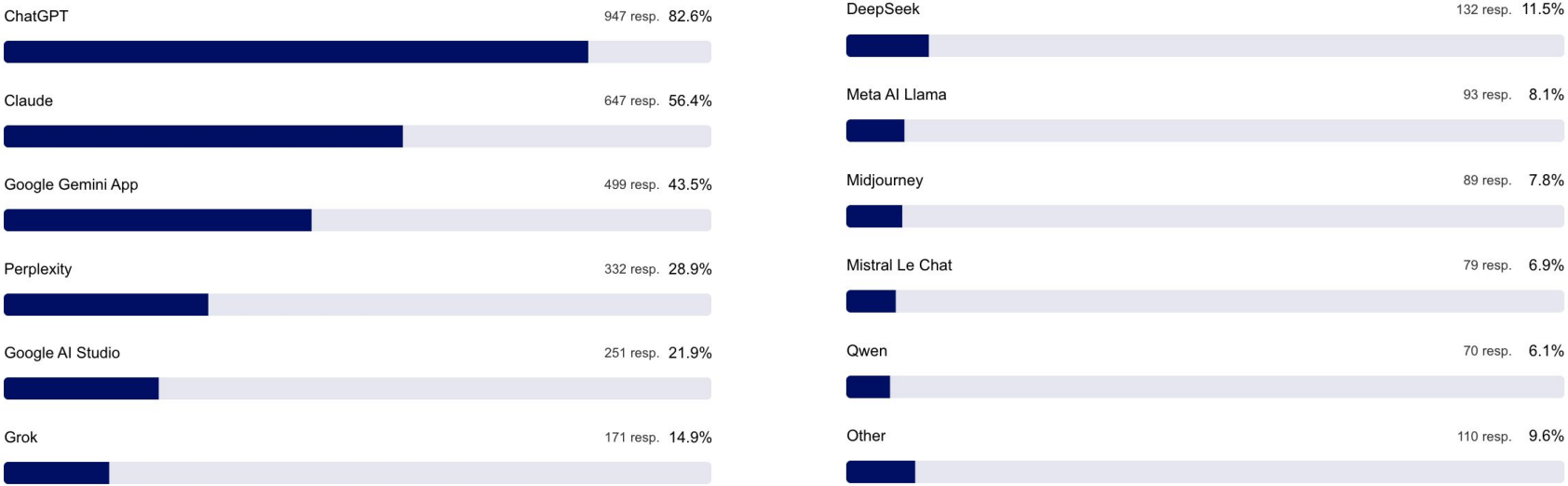
What apps do ML engineers build?



How researchers use AI in their workflow?

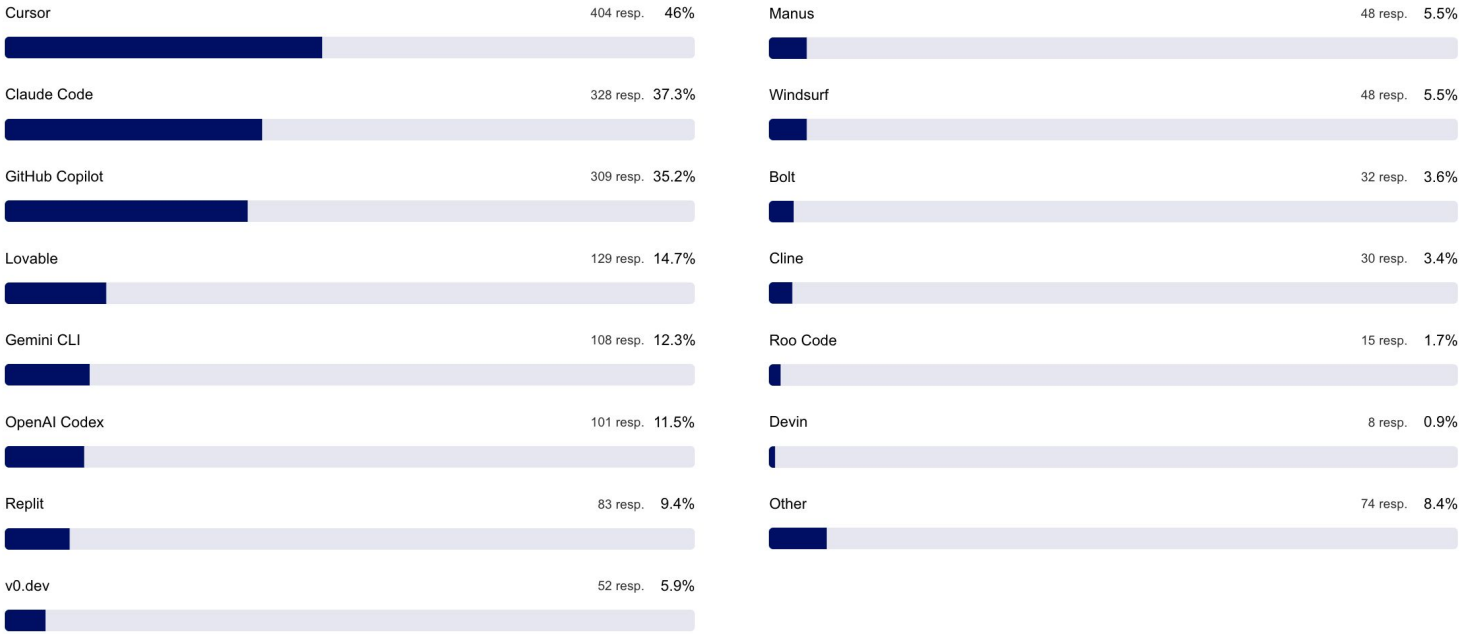
ChatGPT, Claude, Gemini/Google and Perplexity are used most regularly

Despite its significant distribution, Meta's AI is barely used as much as Mistral Le Chat or Midjourney. Meanwhile, DeepSeek isn't far behind X's Grok, again despite a distribution disadvantage.



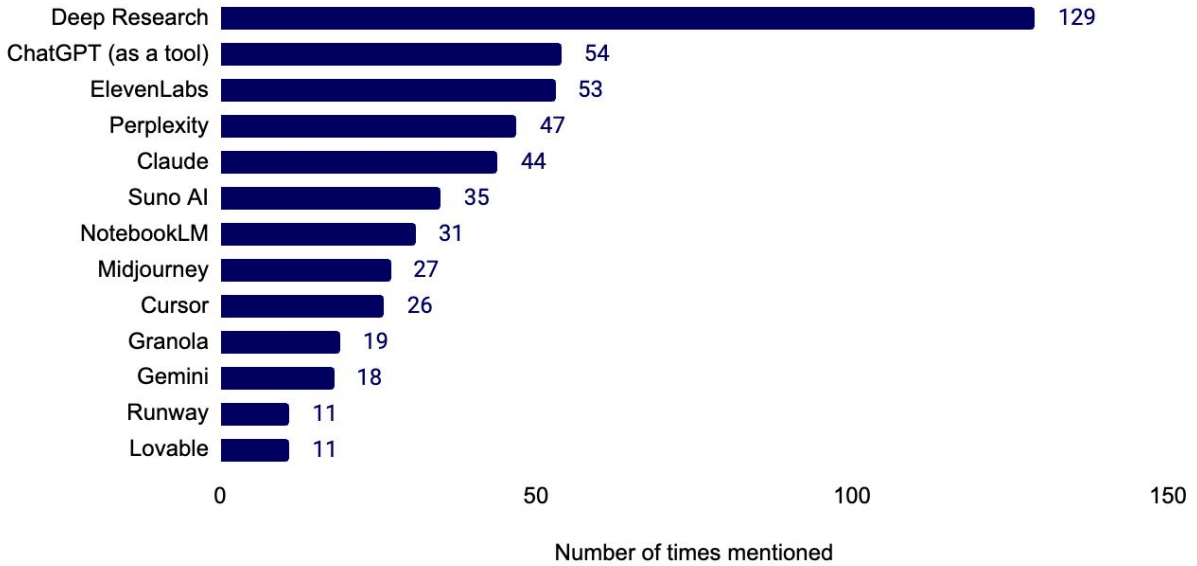
Developers love Cursor, Claude Code and GitHub Copilot

OpenAI's Codex and Gemini CLI are lagging behind where they should be given their resources.



Outside of developer tools, which AI services are most popular?

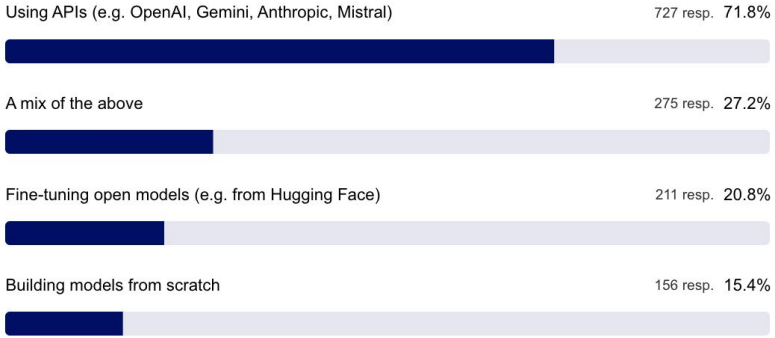
Our data corroborates the widely reported love that Deep Research received. Moreover, we find that respondents enjoy ChatGPT as a non-coding tool, ElevenLabs, Perplexity and Claude.



AI is mainly procured through APIs, followed by fine-tuning and building from scratch

Despite a loud narrative that organizations want/must own their own models, our data shows that respondents procure AI through APIs far more than they build/fine-tune their own models. That said, fine-tuning isn't going away either. To do so, respondents most commonly use PyTorch, Hugging Face Transformers, LoRA/PEFT, custom in-house frameworks, and Unsloth.

What type of gen AI models does your org use?

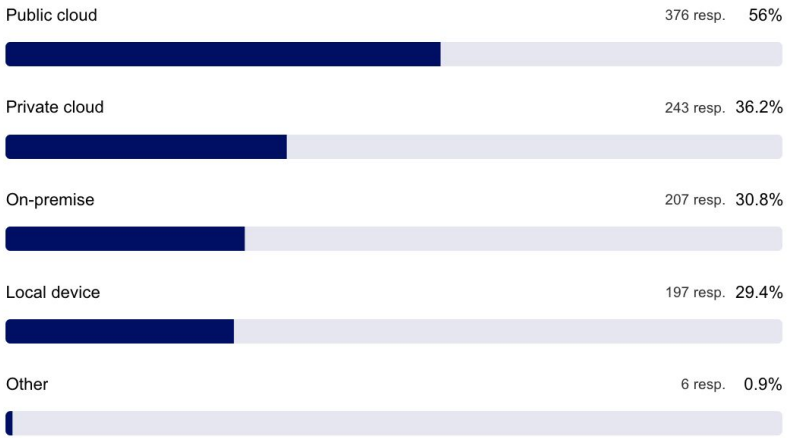


How have your fine-tuning workloads changed?

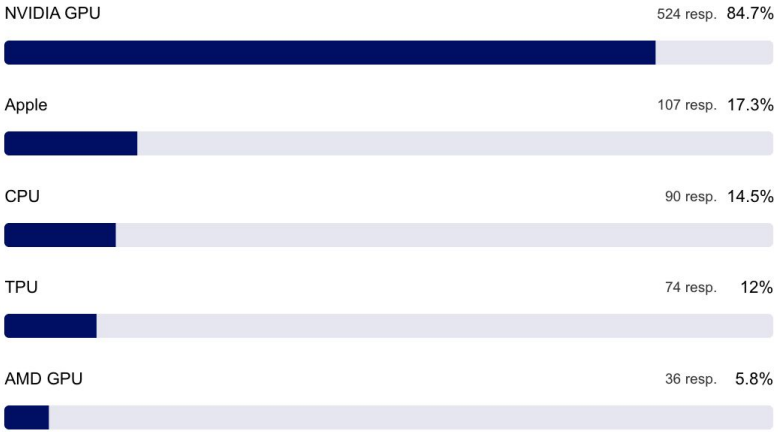
Whether AI runs on public/private/on-prem, at the end of the day, it still uses a GPU

Apple makes a surprising appearance, likely because users are training/experimenting locally. Meanwhile, TPUs and AMD's GPU aren't hugely popular.

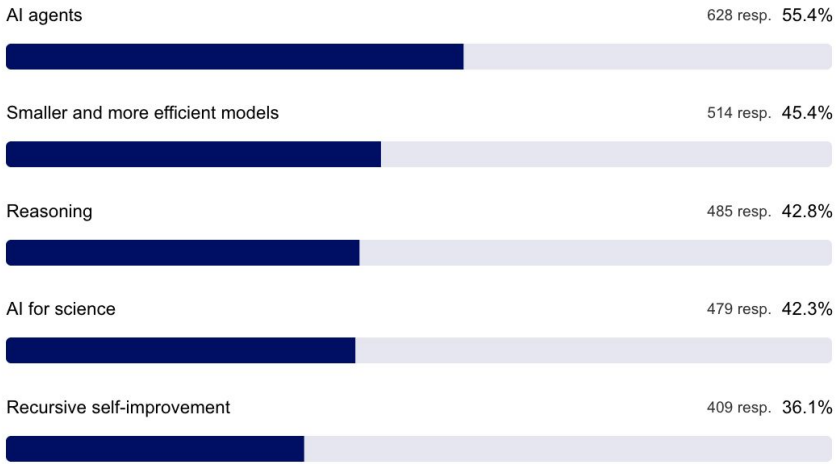
Where is your AI hardware provisioned?

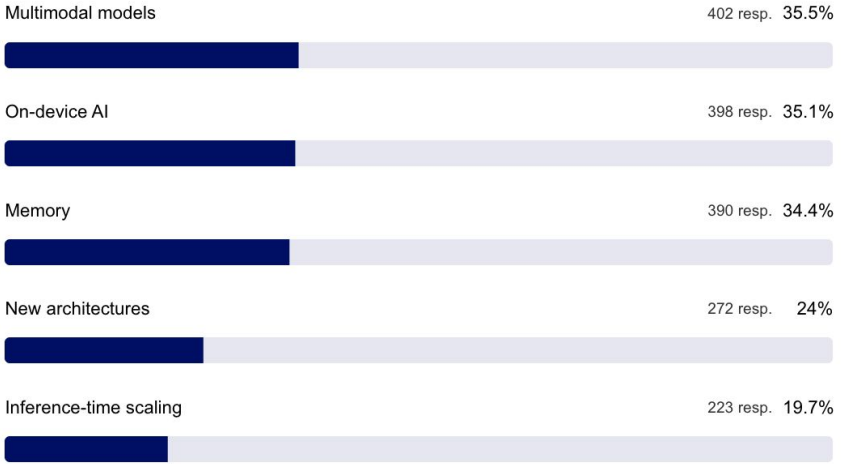


What hardware do you use for training/fine-tuning?



What emerging trends in gen AI are you most excited about?





How 1.2k practitioners rated their top AI labs

#1 

#2 

#3 

#4 

#5 

#6 

#7 

#8 

#9 

#10 

#11 

#12 

STATE OF AI SURVEY .

👉 **Take the survey: stateof.ai/survey**

Contribute to the world's largest open, continuously updated AI survey.
The real pulse from builders, operators, and researchers.

Section 6: Predictions

10 predictions for the next 12 months

1. A major retailer reports >5% of online sales from agentic checkout as AI agent advertising spend hits $5B.

2. A major AI lab leans back into open-sourcing frontier models to win over the current US administration.

3. Open-ended agents make a meaningful scientific discovery end-to-end (hypothesis, expt, iteration, paper).

4. A deepfake/agent-driven cyber attack triggers the first NATO/UN emergency debate on AI security.

5. A real-time generative video game becomes the year's most-watched title on Twitch.

6. "AI neutrality" emerges as a foreign policy doctrine as some nations cannot or fail to develop sovereign AI.

7. A movie or short film produced with significant use of AI wins major audience praise and sparks backlash.

8. A Chinese lab overtakes the US lab dominated frontier on a major leaderboard (e.g. LMArena/Artificial Analysis).

9. Datacenter NIMBYism takes the US by storm and sways certain midterm/gubernatorial elections in 2026.

10. Trump issues an executive order to ban state AI legislation that is found unconstitutional by SCOTUS.

Thanks!

Thank you for making it to the end of the State of AI Report 2025.

We hope you've enjoyed our informed and opinionated take on the extraordinary progress in artificial intelligence over the past year, since last year's edition was published on 10 October 2024. This year's report explores AI research, industry, politics, safety, and insights from our first State of AI Report usage survey. We focus on these areas because we believe that AI is a force multiplier for technological progress - and that broad understanding of its trajectory is essential if we are to navigate such a profound transition.

We'd love your feedback on how we can make future editions even better, as well as your ideas for new contributions and perspectives.

Nathan Benaich
Air Street Capital

Reviewers

We'd like to thank the following individuals for providing critical review of this year's Report (alphabetical order):

Jacob Arbeid, Paige Bailey, Joyce Benaich, Daniel Campos, Xander Davies, Chris Gagne, Aleksa Gordic, Ido Hakimi, Ryan Julian, Neel Nanda, Elvis Osaravia, Jacob Portes, Philippe Schwaller, Shubho Sengupta, Joe Spisak, David Stutz, Ross Taylor, and Divy Thakkar.

Data contributors

We'd like to thank the following companies for providing bespoke data/analysis (alphabetical order):

Dealroom, Ramp, Specter, and Zeta Alpha.

Conflicts of interest

The authors declare a number of conflicts of interest as a result of being investors and/or advisors, personally or via funds, in a number of private and public companies whose work is cited in this report. Notably, the authors are investors in companies listed at: airstreet.com/portfolio

About the authors

Nathan Benaich

Nathan is the General Partner of **Air Street Capital**, a venture capital firm investing in AI-first companies. He runs The Research and Applied AI Summit (RAAIS), The RAAIS Foundation (funding open-source AI projects), AI communities in the US and Europe, and Spinout.fyi (improving university spinout creation). He studied biology at Williams College and earned a PhD from Cambridge in cancer research as a Gates Scholar.

State of AI Report 2025 team

Zeke Gillman

Zeke is a Tech Policy Fellow at Stanford, and co-author of *Regulating under Uncertainty*. He previously worked at Harvard Business School and the DOJ Antitrust Division, and holds a BA in Political Science and Philosophy from the University of Chicago.

Nell Norman

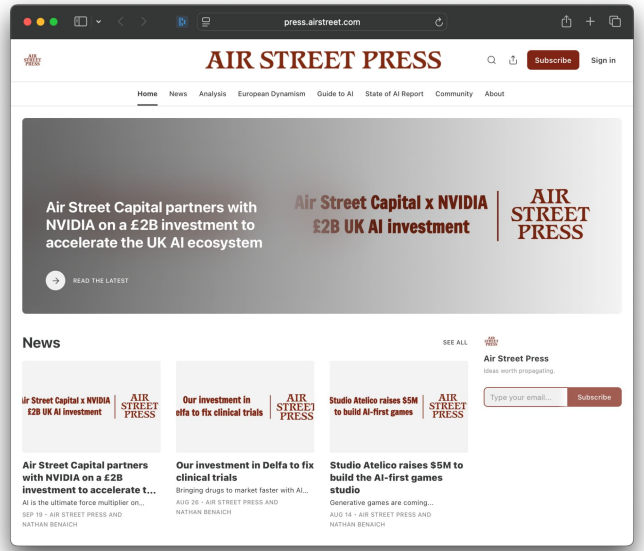
Nell is a grad student in Computing at Imperial College London focusing on how LLMs could enable scalable vishing fraud. She previously helped AI teams build reliable products at AI agent platform V7 Labs, and has a first class BA from Oxford University.

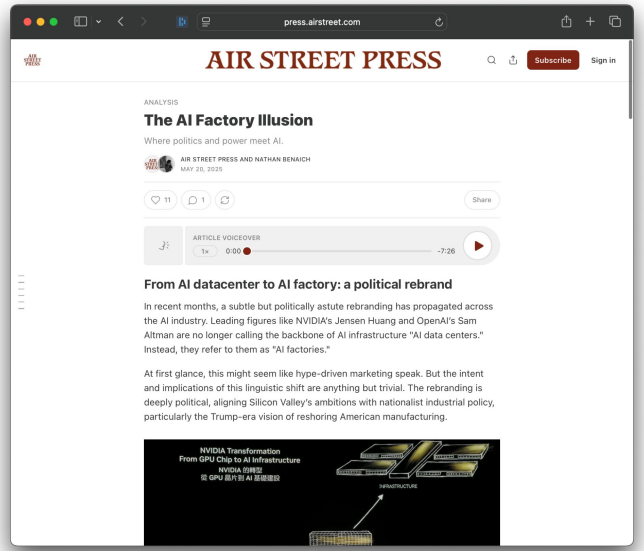
Ryan Tovcimak

Ryan is a founder of the AI Stack Tracker. His work spans red-teaming frontier models, benchmarking the global AI competition, and tracking trends in AI compute and power demands. He holds a BS in Econ from Vanderbilt University.

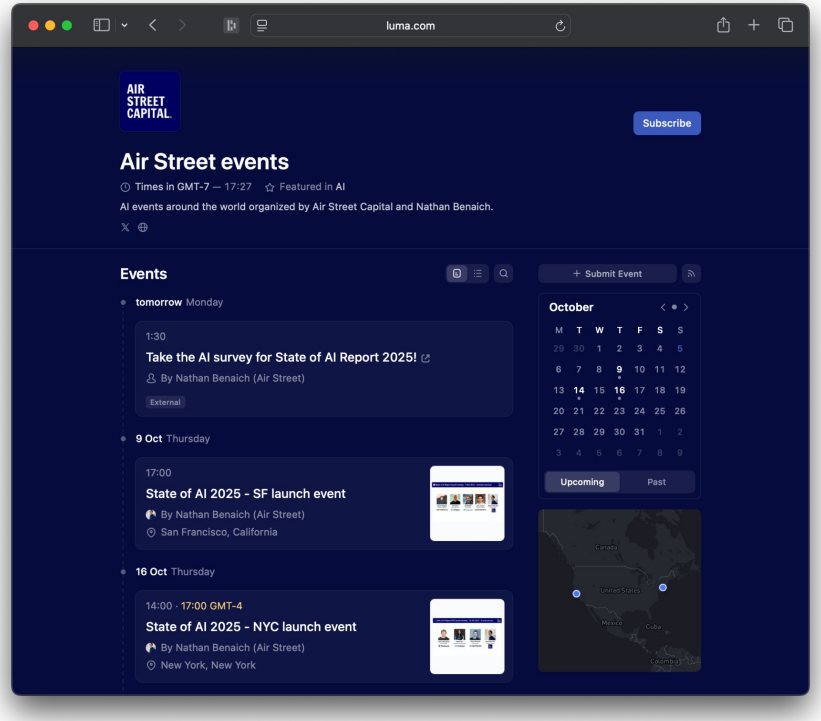
Follow our writing on AIR STREET PRESS (press.airstreet.com)

- If you enjoy reading the State of AI Report, we invite you to read and subscribe to Air Street Press, the home of our analytical writing, news, and opinions.





Join our global community of best practices events (airstreet.com/events)



RAAIS
LONDON AI
BOSTON AI
MUNICH AI
NEW YORK AI
STOCKHOLM AI
LISBON AI
SF AI PARIS AI

STATE OF AI REPORT .

October 9, 2025

Nathan Benaich

AIR STREET CAPITAL .