

THE COMPUTE ECONOMY

A Rocket  
Company  
Becomes the  
Landlord of  
Artificial  
Intelligence

*SpaceX leases its supercomputer to an AI lab for up to \$6.3 billion — and quietly redraws the map of who actually holds power over the frontier.*

|                                                                     |                                                                    |                                                                  |                                                                   |                                                                          |
|---------------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------------|
| <div>\$6.3B</div> <div>SPACE-<br/>REFLECTION<br/>COMPUTE DEAL</div> | <div>\$150M</div> <div>PER-MONTH LEASE,<br/>PROJECT COLOSSUS</div> | <div>100K</div> <div>GPUS BEHIND<br/>CURSOR COMPOSER<br/>3</div> | <div>5T</div> <div>PARAMETERS,<br/>COMPOSER 3<br/>(CLAIMED)</div> | <div>82%</div> <div>UK IT LEADERS HIT<br/>BY SURPRISE AI<br/>BILLS</div> |
|---------------------------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------------|

Compute is the new oil, and this week the deals stopped pretending otherwise. The headline transaction of the cycle: SpaceX signed an agreement worth up to \$6.3 billion with the AI lab Reflection for access to its "Project Colossus" supercomputer — reported elsewhere as a lease running near \$150 million per month. The striking part is not the figure but the direction of trade. A rocket company is now a compute landlord, renting GPU time to an AI lab, and a single startup is committing nine figures a month to capacity it does not own.

Pair that with Cursor's claim that its new Composer 3 model ships "5 trillion parameters trained on 100,000 GPUs," plus a joint

Cursor-SpaceX model surfaced inside Grok Build, and the pattern is unmistakable. Frontier capability is being gated by who can assemble hundred-thousand-GPU clusters, and the labs that cannot build them are signing multi-year leases to rent them.

This reframes "AI capability" as a balance-sheet question. When a lab rents a supercomputer by the month, the race is won less by research insight than by whoever can underwrite the lease — handing quiet leverage to a tiny set of cluster owners, and turning a slice of "AI safety" into plain infrastructure-concentration risk. The off-switch, increasingly, may live with the landlord rather than the lab.

This reframes "AI capability" as a balance-sheet question. When a lab rents a supercomputer by the month, the race is won less by research insight than by whoever can underwrite the lease — handing quiet leverage to a tiny set

of cluster owners, and turning a slice of "AI safety" into plain infrastructure-concentration risk. The off-switch, increasingly, may live with the landlord rather than the lab.

with less compute. Meta named a new WhatsApp leader, a personnel move that usually precedes a product agenda; watch for agents inside the messenger.

#### EMBODIED AI

### Nvidia opens a robot test lab

Nvidia quietly stood up a lab where robot makers and their customers can run safety tests before going to regulators for certification — a tell that embodied AI has reached the "compliance infrastructure" stage. That is exactly when a field stops being a demo and starts being an industry.

*"The teams pulling ahead aren't spending the most — they're routing each task to the right model."*

THE WEEK'S QUIET THESIS

---

# ARTIFICIAL INTELLIGENCE

---

## THE REAL STORY

### **Cost, not capability — and open source is the lever**

Multiple business desks converged on one thesis: as Anthropic's and OpenAI's flagship prices climb, large customers are quietly routing work to cheaper open-source models and budget tiers from the same vendors' own menus — one report named Ensemble Health among them.

A sharper data point: 82% of UK IT leaders said they were stung by unexpected AI costs this year. The surprise bill, not model quality, is now the dominant operational pain. Model routing — easy calls to cheap models, frontier models reserved for hard ones — is graduating from a research trick into a default architecture decision.

## SECURITY

### **Rebuilt around agents — by the agent-builders**

OpenAI launched an AI-powered program to find and fix vulnerabilities in open-source software, and a new category of "agentic development security" is forming explicitly to control what agents use, do, and generate — an admission that autonomous coding agents are now a supply-chain risk surface, not just a productivity tool.

The week's most provoking vignette: a lab's own safety report carried a chart that apparently, accidentally, branded the world's most careful AI as "80% dishonest" — and a second AI caught the error in thirty seconds. A live demo of AI auditing AI that is either reassuring or recursive, depending on your mood. Around it: KPMG was

embarrassed by a public AI failure, and the first software-supply-chain Magic Quadrant landed.

## THE QUIET SCANDAL

### **Benchmarks now lag production**

The sharpest engineering line of the day: a leading lab shipped a model with better benchmarks across the board twice in six weeks — and twice, production systems broke anyway. Leaderboard gains do not survive contact with real pipelines, because prompts, tool schemas and latency assumptions are silently coupled to a specific model's quirks.

The lesson for anyone shipping software: the newest model is a liability until your own tests say otherwise. Public leaderboards have become marketing; private, scenario-specific evaluation suites have become strategy.

---

## AGENTS & THE ENGINEERING CRAFT

---

### THE NEW FRONTIER

# From "can they talk" to "can they finish"

The recurring word this week was reliability, not intelligence. One newsletter's entire thesis was "run longer tasks until verifiably done" — the frontier is no longer a clever single response but an agent that grinds a multi-step task to a verifiable finish.

"Agent loops" surfaced repeatedly as the emerging pattern for running models in iterative cycles, checking their own work between steps. A more skeptical piece pushed back on the hype around

"self-evolving" and "self-improving" agents, asking whether they actually deliver — or simply rebrand fine-tuning with a more exciting verb.

The durable shift is architectural. Agents are being judged less on what they can say and more on whether they can be trusted to complete real work without a human re-checking every line — a far higher bar, and the one that separates a demo from a product.

### MEMORY

## From bolt-on to blueprint

A two-part deep dive is worth tracking. Part one framed agent memory as a product — Mem0, Supermemory and Letta, the libraries you bolt on. Part two reframes it as architecture you design in from the start, the more durable way to think about what an agent should remember and forget.

### THE BORING SUBSTRATE

## Clocks, merges and half-price compute

A distributed-systems piece warned why wall-clock time will corrupt your database — and how conflict-free replicated data types (CRDTs) safely merge concurrent deletions. A reminder that the unglamorous layer is what breaks when agents hit production. On infra: Karpenter on Spot instances can roughly halve a Kubernetes compute bill.

---

### WORTH A READ

## The half-year engineering review

An overview of what changed in engineering over the last six months — how various tech companies are reshaping how they work, and why deliberately slowing down could, counter-

intuitively, be the sensible strategy in a year of constant model churn.

#### TOOLKIT

## Eleven libraries, zero hype

A practitioner's roundup of eleven Python libraries "so effective I questioned why I ever built things from scratch" — real productivity gains over bespoke code, and a teardown of the YC startup selling agents a safe place to fail before they touch production.

**THE \$100 BILLION PIVOT**

## AI reaches for the oldest business model

The single biggest strategic bet flagged this week: a leading consumer-AI company reportedly making a "\$100 billion bet on ads." A company that sold AI as a clean, subscription-funded tool is reaching for the oldest and grubbiest internet business model — advertising — at nine-figure-billions scale.

The implication is a category change. It would reframe a chat assistant from a subscription tool you pay for into an attention platform that pays for itself by monetising the user. "AI you talk to all day, funded by ads" is either the most powerful advertising medium ever built or the moment the technology spends the public's trust — and those may not be different outcomes.

**VENTURE**

## A rare hard number

Seedcamp closed Fund III — the largest raise in its nearly-20-year history — with the fund tracking toward a roughly 20x net return. In a market starved of liquidity, a concrete multiple from one of Europe's most consistent early-stage investors is the rare figure worth circling.

The M&A read-through explains the capital's direction: buyers are paying up for workflows, agents and distribution — the application layer and the customer relationship — not raw model IP.

**CONSOLIDATION**

## The tooling layer folds in

A startup teardown counted 21 acquisitions across OpenAI, Cursor and the surrounding coding-agent ecosystem — the tooling layer is consolidating in plain sight. FINOS launched an AI fund to push responsible agentic AI into financial services.

The backdrop was jittery: a tech-stock-driven, "unnerving" sell-off rattled global markets, a reminder that the AI trade and the index are now the same trade.

## The edge is an unglamorous internal router

Three independent threads — customers shifting to open source, 82% of leaders hit by surprise bills, "route each task to the right model" — point at one discipline: per-task model selection with a quality gate. The org that builds a routing layer, and an eval harness to police it, will quietly out-margin the org that hard-codes the flagship everywhere, and do it without a single press release.

## Benchmarks lag; eval suites are the new moat

"Better benchmarks, broke anyway" is the most important sentence in today's haul. If frontier upgrades keep breaking production, the compounding asset is not access to the newest model — it is a private, scenario-specific evaluation suite that tells you in an hour whether a new model helps or hurts your pipeline. Public leaderboards are marketing; your own evals are strategy.

## Compute leasing centralises power upward

When a lab rents \$150M a month of someone else's supercomputer, frontier AI stops being a research race and becomes a credit-and-capacity race, won by whoever can underwrite the lease. That hands leverage to a tiny set of cluster owners and reframes "AI safety" partly as infrastructure-concentration risk — the off-switch may end up with the landlord.

## Selling the agent and the cure for the agent

The same companies ship autonomous coding agents and the AI program to fix the vulnerabilities those agents create. There is a flywheel risk: more autonomy creates more attack surface, which creates demand for more tooling sold by the autonomy vendors. Worth watching whether security stays an independent check or becomes a captive upsell.

## Move 37: deliberately run a worse model

Everyone is racing to the frontier; the non-obvious play is to standardise on a cheaper, good-enough, open-weights model you fully control — and spend the saved money and saved upgrade-churn on evals, routing and data. You trade a few benchmark points for predictable cost, no vendor-driven breakage every six weeks, and weights you can pin forever. The crowd optimises capability; the edge is optimising stability — and in a market sprinting for the top, picking the un-churnable second-best model may be the highest-return decision on the board.