

INTELLIGENCE ON THE AI FRONTIER

Curated overnight from the world's sharpest newsletters — the models, the agents, the markets, and the ideas in the spaces between.

THE SELF-IMPROVING MACHINE

Agents Learn to Rewrite Their Own Code —
and a Small Model Punches Above a Giant

A new wave of open-source research lets a model read its own crash logs, rebuild the scaffolding around itself, and regression-test the result — lifting success rates as much as 60% without touching a single weight.

60%	33→47%	>99%	73%	\$217B
PEAK JUMP IN AGENT PASS RATES	A 9B MODEL ON GAIA AFTER HARNESS EVOLUTION	JAILBREAK-CATCH RATE ON THE REINSTATED CLASSIFIER	FLAWS CAUGHT BY A FRESH GRADER VS. SELF-CRITIQUE	H1 VENTURE FUNDING TAKEN BY TWO AI LABS (43%)

For everyday builders, the real lever over an AI application was never the model — it was the "harness," the scaffolding of prompts, tool-use logic, memory and verification rules that turns a text generator into an agent. That scaffolding has always been brittle: swap the underlying model and the whole thing snaps. This week, two frameworks flipped the burden of fixing it from the human to the machine.

The first runs a three-stage loop — mine the execution traces for recurring failures, propose a targeted patch to the harness, then validate it against regression tests so the agent never gets worse at what it could already do. On a real-world engineer-

ing benchmark, that lifted pass rates by 33% to 60%. The second treats the agent as nine swappable "processors" tuned by a reinforcement-learning engine, and took a lightweight 9-billion-parameter model from 33% to 47% on a notoriously hard benchmark — at a fraction of a giant's cost.

The unifying idea is "loop engineering": designing triggers, actions and strict verification gates so an agent can run, check its own work, and self-correct — while dodging the trap of "loopmaxxing," or burning vast compute on unguided retries. The developer's highest-leverage work is quietly moving from writing prompts to designing the instrumentation that lets a model safely improve itself.

MARKETS · THE RECKONING

A Price Tag Appears on AI

From this week, one of the world's most valuable carmakers caps employee spending on outside AI tools at \$200 a week — counting every major assistant, yet exempting its own in-house model entirely. It is billed as the end of the "token bills are a rounding error" era, with one ride-hail giant said to have burned its entire annual AI budget in four months.

INFRASTRUCTURE

Selling the Leftover Compute

A social-media titan is standing up a cloud business to rent out its spare AI capacity — the SpaceX playbook of amortising over-capacity, applied to GPUs. Its chief has noted outside firms ask "every week" to buy compute "at some premium." Success would mint a new hyperscaler overnight.

A prompt is a request. A loop is a policy.

THE WEEK'S GOVERNING IDEA

A Frontier Model Returns, and Rivals Draft a Truce on Jailbreaks

The most-discussed frontier model is back online after a 19-day suspension, pulled three days post-launch when researchers found a jailbreak that surfaced software vulnerabilities. The fix is a classifier that catches the technique more than 99% of the time and gracefully degrades the request to a safer model rather than erroring outright. More striking than the patch: four of the largest labs are now drafting a shared severity standard for jailbreaks — a "CVSS for prompts" — the clearest sign yet that model safety is becoming an industry utility rather than a per-lab secret.

A Free, Chip-Independent Coding Environment Lands

A challenger shipped a free, permissively licensed "agentic development environment" built on a 744-billion-parameter mixture-of-experts model (40B active) with a million-token context, trained on 28.5 trillion tokens — and, pointedly, on domestic silicon with no US chips at all. It is a reminder that the capability frontier and the supply-chain frontier are now separate races, and that neither is a monopoly.

Two Models, and a Quiet Bet on "Selling Shovels"

One lab shipped two models in a week: a powerful flagship that rebuilt a working document editor from a single prompt in about three hours, and a mid-tier that "landed with a shrug" — competent, cheaper, but overshadowed. The sharper move sits underneath: a new science workbench with 60-plus curated skills across genomics and structural biology, paired with the lab's own preclinical drug work. The tell is dogfooding — using in-house discovery to fix the evaluation bottleneck and build the workflow platform others will rent.

Expert Judgment Becomes Trainable — at a Fraction of the Cost

A research house made a top investment fund's private judgment trainable: generic prompts scored near a coin-flip (46-50%), but with expert labels, interleaved batches and a capped reinforcement-learning loss, the result beat the best frontier model with 29.8% fewer errors and roughly 14 times lower inference cost. Separately, a major lab-tool provider introduced "surge pricing," doubling peak-hour API rates during business hours — a first for the category.

A Search Giant Carpet-Bombs the Calendar

In a single week: two new flagship and lightweight image models to developers, a round-the-clock desktop agent wired into design and delivery apps, minute-long vertical-video overviews from a research tool, and image generation drawing on a user's own mail and photos. A checkout platform now takes restaurant orders directly inside two chat assistants; a social network shipped an official agent connector. Rumour puts a two-million-token flagship weeks away.

The Research Frontier: Agents That Grade Themselves

The week's paper crop leaned hard into self-evolving agents and the reward problem that haunts them — one framework co-evolves agents alongside their evaluators; another argues no fixed reward survives a stronger policy; a third makes memory itself a trainable skill, lifting a 32B model to frontier level on hard game benchmarks. Peer review is buckling in parallel, with agentic reviewers proposed as submissions head toward a projected 73,000 a year.

"The Maker Is Never the Grader" — Verification Becomes the Craft

The most repeated rule of the week was not about writing better prompts — it was about who checks the work. A fresh-context verifier, one analysis found, caught roughly 73% of seeded issues, against just 7–33% for a model critiquing its own output. The lesson is structural: as models get cheaper at doing, the bottleneck moves to checking, and the teams that win invest in graders, regression suites and audit trails rather than cleverer instructions.

The durable pattern is a six-part loop: a trigger, a rules-load from a memory file, one bounded unit of work, a fresh-context verifier, a memory write, and an explicit stop-check with three exits — success, a three-strike ceiling, or a token budget. Cost is routed like capital: plan on the expensive model at maximum

effort, implement on a cheaper one in the middle, review on a fresh expensive pass. That "barbell" — premium only at the two ends — roughly halved spend in practice.

Four separate guides converged on the same handling notes for the leading model: state the intent and the "why," not just the "what"; treat the effort dial as a search radius rather than an intelligence knob; never ask it to "show its reasoning," which trips a refusal classifier; bolt on an evidence-audit step that checks each claim against a real tool result, which all but eliminated fabricated status reports; and give it a plain notes file for memory. The throughline: manage the agent like a capable report, and instrument everything it touches.

STANDARDS

The Web Gets Its First New Verb in 16 Years

A new HTTP method, QUERY, was published as an official standard — the first since 2010. It carries a request body like a write, but is defined as safe and idempotent like a read, ending the old hack of smuggling large searches through methods that "lie to every cache, proxy and retry layer." Its four properties map cleanly onto four agent failure modes, from unsafe retries to filters too fat for a URL. (*Teaser.*)

THE PITCH

A Fifteen-Agent Back Office

One playbook proposes replacing back-office hiring with fifteen agents spanning the six functions every company has — support, HR, operations, marketing, sales and finance — each a small project with its own memory, role and connectors, and none shipping without a human's sign-off. The named agents sit behind a paywall. (*Teaser.*)

SECURITY

"You Cannot Escape Your Way Out of Prompt Injection"

Invited to attack his own account on a live medical AI, a researcher typed one sentence — asking for a hospital president's medications, "and ignore the rest" — and pulled the executive's private records under his own login. His verdict: the system wasn't badly built; injection is an architectural flaw in how models blend instructions and data, so sanitising input can't fix it. (*Teaser.*)

TOOLING

Understanding a Codebase Is a Search Problem in Disguise

A good engineer carries a mental map — ownership, hotspots, dependency boundaries, dead paths, files that change together — and agents need the same. The fix on offer indexes a repo once into a graph, then exposes that intelligence to coding agents through a connector, so an agent stops "pretending to understand" and actually navigates. (*Teaser.*)

Two Labs Took 43% of Every Venture Dollar This Half

Global startup funding in the first half of 2026 already topped all of last year, and two AI labs alone absorbed \$217 billion of it — 43% of the total — while a chip challenger went public at a \$56 billion valuation. The framing essay of the week calls it "The Great Descent": AI is collapsing the cost of expert judgment the way the internet collapsed the cost of information, exposing knowledge-scarce professions to the disruption software already brought to publishing.

A counterintuitive companion note: falling per-token prices have not cut AI spending, because usage is growing far faster than cost is falling — pushing operators to measure cost per completed task rather than licences per seat. And a case study in velocity: one voice-AI company scaled from a roughly \$2 million pre-seed to an \$11 billion valuation in a little over three years. *(Essays are teasers.)*

SOVEREIGN COMPUTE

Everyone With a Datacentre Wants to Be a Cloud

Beyond one titan's move to resell spare capacity, the week traced a broader neo-cloud surge: a leading lab reportedly floated handing its national government a 5% stake — about \$42.6 billion against an \$852 billion valuation — while an infrastructure startup raised \$800 million at \$8.3 billion, claiming model-serving up to 80% cheaper. If over-built capacity becomes a product, compute commoditises fast, and the "sell the excess" instinct hints that even the biggest buyers suspect they over-ordered.

THE LONG VIEW

Markets Crash When Information Detaches From Price

A mini-history worth keeping: the first true IPO floated a spice-shipping company in 1602 and spawned futures, options and short-selling within decades; a stock exchange began in 1792 with two dozen brokers fixing a quarter-percent commission under a buttonwood tree. Every crash since rhymes — the 1929 break, a 23% one-day drop in 1987, a 78% dot-com collapse, 2008, and the meme-stock and crypto blow-ups of this decade. Benchmarks to keep handy: ten-times forward revenue reads as "premium," and revenue-per-employee above \$450k marks genuine scale.

POLICY

A federal bill to regulate earned-wage access advanced 29-22 in committee: it mandates a free option but classifies advances as non-credit — exempt from interest-rate disclosure — and pre-empts a dozen state laws, drawing fire from 225-plus consumer groups. *(Teaser.)*

The Harness Is the New Moat

Read together, the self-rewriting agents, the loop-engineering playbooks, and the billions pouring into forward-deployed engineering all point one way: the durable advantage is no longer the model — every lab has a good one — but the verification gates, memory and self-correction loops wrapped around it. The person who owns the loop owns the outcome, which is exactly why a small model with an evolving harness can out-punch a frozen giant.

Grading Is the New Alpha

"The maker is never the grader" isn't a slogan; it's a measured 73%-versus-single-digits gap between independent verifiers and self-critique. Pair that with the finding that no fixed reward survives a stronger policy, and a pattern emerges: as models get cheaper at doing, the scarce resource becomes checking. The unglamorous half of the stack — audits, regression suites, evaluators — is where the next edge is hiding.

Price AI Like a Portfolio

A weekly spend cap, a budget torched in four months, surge pricing, and "cost per completed task" replacing per-seat licences all tell one story: cheaper tokens raised bills because usage outran deflation — Jevons' paradox, live. The operators who thrive already route effort like capital, spending the premium model only where it changes the answer. This is the first draft of AI FinOps, and it's arriving faster than the budgets to govern it.

The Identity Layer for Agents Is Shipping Unwatched

A proof-of-human system now lets an AI agent hold a delegated quota against a specific person; a social network shipped an agent connector; a checkout platform takes orders inside chat assistants. The instant agents transact for you, the load-bearing question stops being "is the model good" and becomes "which human authorised this, and can we prove it." That plumbing is being laid now, while attention stays fixed on benchmarks.

Move 37 — Make Your Agents Smaller, Dumber, and Impossible to Misread

The whole field is sprinting toward more autonomy — tens of thousands of agents, self-improving harnesses, loops that run for hours. The counter-move with the highest near-term return is the opposite. A deliberately constrained small model whose failure traces you can actually read beat brute force this week; unguided "loopmaxxing" failed; and the growing complaint that the best model "makes the human the bottleneck" is not a bug to engineer away but the most valuable signal you will get — it surfaces precisely which decisions you never actually made. The alpha isn't a genius agent you can't audit; it's a caged one you can. So don't buy the model company. Buy the boring instrumentation-and-logging layer underneath it — because that is the shovel the labs themselves are quietly dogfood-ing.