

INTELLIGENCE ON THE AI FRONTIER

MODELS • AGENTS • MARKETS • IDEAS

THE POST-MODEL STACK

The Model Was Never the Moat

Capability is commoditising in public. The durable advantage has quietly moved to the loop, the router, and the scoreboard wrapped around a rented model.

−41% COST PER TASK FROM THE HARNESS ALONE	70-80% ROUTINE WORK OPEN WEIGHTS CAN ABSORB	\$2.35 H100 / HOUR, UP 38% OFF THE FLOOR	64.7% TOP ROUTER ON THE DEEP-RESEARCH BENCH	6 mo. RUMOURED CLOCK OVER OPEN WEIGHTS
---	---	--	---	--

The most valuable thing in artificial intelligence in 2026 is no longer the model. Open-weight families have caught the closed frontier in public view — one clears roughly 59% on a hard software-engineering benchmark at a fraction of the price, under a licence you can run yourself. When the intelligence is rented and swappable, the advantage must live elsewhere: in the routing that decides which model sees a request, the loop that checks its own work, and the private evaluation that says whether any of it is right.

The evidence converges from three directions. A controlled study found that improving only the “harness” around a model cut cost-per-task by 41% at equal quality — and the saving held no matter which model sat underneath. An entire category of “model routers” now exists to stop teams sending every request to the priciest frontier model. The durability test worth adopting: would this survive a model swap, does it compound rather than merely accumulate, and could a stranger rebuild it over a weekend?

ALSO • API BANKRUPTCY

One ride-hailing giant reportedly burned its entire annual AI budget in about four months, then capped per-engineer spend — and it wasn’t alone that quarter. Scaling every request to the frontier is a fast route to a blown budget; routers now send 70-80% of routine traffic to cheap open weights and reserve the premium tier for the hard 20%.

ALSO • 744B ON 25GB

A system called Colibrì claims to run a 744-billion-parameter mixture-of-experts model on a 25-gigabyte machine — keeping the dense layer in memory and streaming experts from NVMe only when the router calls them. If it holds, the “you need a datacenter” assumption quietly erodes.

“The chatbot era is not ending; it is being compiled into infrastructure.”

RELEASE WAVE

A Runtime, Not a Chatbot

This week's frontier releases reframe the model as an execution engine and the chat window as a control plane. One flagship splits into three tiers — Sol, Terra and Luna — and adds programmatic tool-calling and parallel subagents; a companion “live” mode runs full-duplex voice that listens and speaks at once. A new “work” product sustains projects for hours across connected apps and files to emit editable documents, spreadsheets and decks, while a rival's million-token model bundles computer-use and multi-agent orchestration behind a metered API.

FAST & CHEAP

The Cheap-and-Fast Challenger

A new challenger model lands at roughly 80 tokens per second and about twice the token efficiency of the leaders, priced at \$2 and \$6 per million input and output tokens. Independent reviewers rate it “Opus-level” — around 4.6–4.7 on their internal scale — “not state of the art, but pretty good, and very fast,” earning a slot for long, multi-step work even from people who keep a pricier default for prose.

COST DISCIPLINE

Route by the Job, Not the Brand

As the premium specialist model moves to pay-per-use at \$10 / \$50 per million tokens — double the daily-driver tier at \$5 / \$25 and the volume tier at \$2 / \$10 — the discipline is to send each task to the cheapest model that honestly clears it. Max the effort dial before switching models at all, and write the routing rules straight into the project config so the decision is automatic rather than a habit.

MIDDLEWARE

The Router Boom

“API bankruptcy” has birthed a middleware category. Capable open weights now absorb 70–80% of routine traffic; premium models handle the rest. One fusion router scored 64.7% on a deep-research benchmark — beating a standalone frontier model at 60.0% and another at 58.8% — at roughly half the token cost, while a self-hosted proxy with a 10-millisecond classifier claims to cut bills up to 70%. The durable asset isn't the router; it's the private eval you tune it on.

OPEN WEIGHTS

The Frontier, Caught in Public

Open-weight models have closed the gap fast enough to reset strategy. One hits ~59% on a hard software-engineering benchmark cheaply and under a runnable licence; others handle the majority of everyday workloads outright. The uncomfortable coda is dependency risk: a June episode in which export rules forced a lab to pull its two strongest models offline worldwide is a reminder that a rented brain can be switched off by someone other than you.

POLICY

A Six-Month Clock

A rumoured regulatory threshold would delay or block any open-weights model whose capability meaningfully exceeds today's strongest closed systems — a line an open model, most likely from China, is expected to cross within about six months. Open models lack a central economic champion to lobby for them, so any restriction is likely to loosen far more slowly for open weights than for closed. The thing democratising capability is the thing most exposed to being legislated.

THE SCAFFOLDING IS THE PRODUCT

The Harness Beats the Model — and the Proof Is Model-Invariant

A standout result this week varied only the orchestration layer around six different models across 22 tasks and cut blended cost-per-task by 41%, tokens by 38% and median wall-clock time by 44% — all at quality parity. The striking part is that the efficiency gain held between 33% and 61% regardless of which model was underneath, and quality improvement tracked baseline model strength almost perfectly. In other words, the scaffolding, not the weights, is doing much of the work.

The craft is following the evidence. One builder repackaged a classic trial-and-error “loop” into a no-code skill that scores a plan out of 50,

iterates while pulling live web data, and stops only once it clears 40. Pointed at a real goal — grow a newsletter by ten thousand subscribers in 45 days — its first pass did the arithmetic nobody had: about 250 new subscribers a day at roughly \$2.50 each is \$25,000 in ad spend, a fatal hole caught before a dollar moved. Over eight rounds the score climbed 27 to 45 and rebuilt the plan around a cheaper funnel. Verification is emerging as the matching scaling axis: a training-free verifier now reads a calibrated score off a model’s own logits and posts 86.5% on a hard agent benchmark while doubling as a reinforcement-learning reward.

BY THE NUMBERS • 91 → 87

Tool-selection accuracy is set by how many tools you show at once, not by the model: one small model picks correctly 91% of the time with ten tools but 87% with fifteen, and a larger one holds above 90% up to twenty before sliding by thirty. Fewer, sharper tools beat a crowded menu.

CONTRARIAN • SKIP /INIT

Reflexively auto-generating a project context file can backfire — padding the agent with stale, token-hungry documentation crowds out room to solve the actual problem. More context is not automatically more intelligence.

The Case for the Harness

A widely shared argument holds that the biggest recent product leaps come from the harness around the model — the layer that makes coding agents feel far more capable than a bare chatbot on similar weights — not from the weights themselves.

When the Agent Runs Code

Security is catching up: the same “let the agent generate and execute code” feature that makes tools magical is the biggest single expansion of the attack surface, where prompt injection and unsafe serialization can turn text into remote code execution.

THE HARDER PHASE

Capital, Proof, and Land-Grabs

The weekend read is that AI has entered a harder phase where winning depends less on a better model and more on securing capital to build, proving measurable value, and staying ahead as rivals converge. The chip leader is hedging against competitors by partnering with them; AI borrowers are tapping a “\$3.5 trillion market hiding in plain sight”; small firms are using assistants to walk away from incumbent software suites; and a popular coding tool is reportedly building an agent to compete head-on with a rival’s new desktop worker.

On the supply side, the “demand is softening” story does not survive the pricing. The one-year contract index for the workhorse datacenter GPU bottomed near \$1.70 an hour last October and has rebounded about 38% to \$2.35, with spot rates up 10% this year. The tension worth sitting with: individual firms are rationing AI spend to avoid bankruptcy by API at the very moment aggregate compute demand is climbing again.

ROBOTICS

\$55.8 Billion, Almost Zero Paid Hours

Robotics pulled in a record ~\$55.8 billion this year, roughly double the prior record — yet almost none of the humanoids have logged an hour of paid work. Four leading builders dropped demos in a single week, and the sharper take is that the money is in the “deployment OS” layer, not the hardware, with the US arguably losing a price war on the machines themselves.

PLAYBOOK

Give Away the Environment, Sell the Loop

A post-training startup says it went from zero to more than \$100 million in annualised revenue in about nine months and 6,000-plus customers — built on an open-source wedge that crossed a thousand training environments before the paid product launched, and named-customer proof instead of a sales team. It then raised a \$130 million round with a marquee venture firm and three chip-and-hardware giants participating.

SYNTHESIS

The Value Moved to the Scoreboard

Three independent signals point the same way: the harness saving is model-invariant, routers save most of the bill by matching tasks to models, and the loop that caught a \$25,000 error was really a verifier doing its job. The model is commoditising — and so is the scaffolding around it. What stays scarce is the reward function that encodes your standards. Own a high-signal evaluation and you own the improvement loop; everything else is rented.

MARKETS

Discipline vs. the Super-Cycle

Budgets burned in four months and an entire router industry exist because scaling everything to the frontier is bankruptcy — yet GPU prices just rebounded 38%. Both are true at once. The winners of the next eighteen months will not be those who spend the most or the least on inference, but those with the tightest mapping between the value of a task and the cost of the model they throw at it.

POLICY

Winning Technically, Exposed Politically

Open weights can already do the majority of routine work, rival the frontier cheaply, and run a 744-billion-parameter model on a laptop — just as a six-month regulatory clock threatens to cap any open model that outpaces the closed frontier. The force democratising capability is the one most likely to be legislated. Price a policy tail into any open-weights roadmap; and if you are betting against them, notice how fast the gap closed.

HUMAN COST

The Hidden Bill: Skill Atrophy

AI scribes may dull clinical judgment, one professor watched 56 of 59 students' scores collapse once their crutch was removed, and pleasure-reading has fallen to about one in six people a day. The productivity gain and the capability loss are the same transaction — automating the doing quietly retires the doer. The move is not to refuse the tools, but to choose deliberately which reps you keep doing by hand because they keep your judgment calibrated.

MOVE 37 • THE
COUNTERINTUITIVE PLAY

Buy the Cheaper Model — and Grade It Harder

Everyone's instinct is to route the hard problem to the most expensive model. But if a training-free verifier scores 86.5% across wildly different domains, the cheapest reliable path is often many cheap samples judged by one great verifier — not one premium sample taken on faith. Invert the reflex: spend on the judge, not the player. Run the volume model five times against a ruthless evaluation instead of the specialist once, because in a world of self-improving loops the reward function compounds and the model is disposable. Your moat is not your agent, your data, or your workflow — it is the scoreboard nobody else can build, and you should be investing there while everyone else is still shopping for models.