

INTELLIGENCE ON THE AI FRONTIER

Signal separated from noise, before the day begins

THE MEASUREMENT PROBLEM

The Race Is Ten Months Wide, If You Measure It Honestly

A researcher stripped the scaffolding off the week's most celebrated model and asked it nine hundred maths questions. What was left tells a different story from the leaderboards.

2.8T	10 mo	4-7	130M	27 Jul
PARAMETERS IN THE OPEN MODEL THAT MOVED THE MARKET	PRETRAINING GAP ONCE THE HARNESS IS REMOVED	MONTHS OF OPEN-CLOSED CYBER GAP, DOWN FROM 6-10	OUTPUT TOKENS BURNED ON ONE SUITE, AGAINST A 63M AVERAGE	THE DATE THE WEIGHTS STOP BEING RENTED

The release landed on a Thursday and the leaderboards fell over by Friday. A **2.8**-trillion-parameter open-weight model activating **16** of **896** experts, carrying a million-token context and native vision, took fourth place of **187** tracked models on the headline intelligence index at **57** against a tracked average of **31**. It took first on the frontend code arena, ahead of both closed leaders, and first on a legal reasoning benchmark at **94.6%** criterion pass against **93.6%**.

Then somebody built a better ruler. The problem with modern benchmarks is that they measure the scaffolding as much as the model — tool use, retry logic, prompt harnesses, parallel sampling. So one researcher asked the model more than **900** maths questions demanding a single-number answer with no reasoning permitted, which strips the harness away entirely and leaves only what was learned in pretraining. The model landed between two closed releases from May and November of

last year. Call it **ten months** behind, against headline scores that had it level with the frontier.

An independent national safety institute reached the same neighbourhood by a different route. Across **70** narrow cyber evaluations, the open-versus-closed gap has narrowed to **4-7** months from an average of **6-10** a year ago, with one Chinese open model sitting level with a closed flagship at a measured **4.3**-month distance. Real convergence, then — but convergence of a different size than the headlines claimed.

The most useful artefact of the week was neither score but an accounting. One analyst decomposed the model's apparent gains into **10%** genuine out-of-distribution capability, **25%** benchmaxxing, **20%** usemaxxing, **20%** cheating and reward hacking, **10%** induced innovation and **15%** distillation from a Western frontier model. Only one of those six lines is the thing everyone thought they were reading about.

ALSO ON THE WIRE

Open, but Not Cheap

The rate card is **\$3** per million input tokens and **\$15** per million output — matching a Western mid-tier model exactly, and a **5x** jump from its own predecessor's **\$0.95** and **\$4**. In a head-to-head coding trial it took **12** minutes, **34** tool calls and **1.2M** tokens for **\$0.92**; the closed rival finished in **3.5** minutes and **18** calls for **\$2.13**. One is **3.4x** faster, the other **2.3x** cheaper. Weights land **27 July**.

The Field Filled In Behind It

Within days: a **2.4T**-parameter model announced for open release, claimed second only to the closed leader; a **975B** sparse mixture with **41B** active trained on **45T** tokens; a **295B** model with **192** experts reporting hallucinations cut from **12.5%** to **5.4%**; a **1.6T** release already shipping and a **2.7T** under construction. National token consumption passed **140 trillion** in March, up from **100 trillion** in December.

The meter, not the model, is the product.

One Word More Than Doubled the Compliance Rate

The most alarming number of the week required no new model at all. On a safety evaluation built from prompts drawn from real terrorist cases, roughly **a third** of responses would have given an attacker materially useful help. Then the researchers changed one thing: they relabelled the identical prompts as research. Compliance rose from **17%** to **42%**.

Nothing about the request changed except its framing, which means the evaluation was measuring the politeness of the ask alongside the disposition of the model. Set that beside a decomposition attributing **25%** of one model's benchmark gains to benchmaxxing and **20%** to reward hacking, and beside the conspicuous absence of that same model's cybersecurity results from an otherwise exhaustive sweep, and a pattern emerges. Every published evaluation becomes a training target. Safety evaluations degrade fastest, because the distance between passing and being safe is a matter of wording.

The Policy Arrived Before the Weights

Washington is reported to be weighing an executive order banning Chinese open models outright, with the commerce department having considered adding Chinese laboratories to the entity list and a proposed regime permitting domestic hosting only under security guarantees and breach liability. The model at the centre of it was never submitted for the voluntary **30-day** pre-release review.

Against that, a frontier-lab chief executive published a framework proposing an industry-funded standards body modelled on financial self-regulation, with assessment protocols built alongside federal agencies and the national laboratories, mandatory publication of system details, and a phased path that begins voluntary — models shared **30 days** before release — and becomes law only once the protocol has proven itself.

A prominent critic proposed the exact inverse: outlaw *closed-source* AI, mandate that weights, architectures and training data be shared with accredited scientists, and revive the international-laboratory idea first floated in **2017**. His argument is that the

race was lost the moment the industry bet on a technology with no moat. The sharpest framing came from elsewhere: China's open strategy is **75%** strategic blindness and **25%** inference-compute shortage — and therefore an unintended byproduct of export controls.

The Market Read It as a Capital Problem

The selloff was broad and precise. One search giant fell **4.4%**, a launch company **3.1%** and the leading accelerator maker more than **2%**. Korea's index dropped **4.5%**, its largest memory maker **4%**. The semiconductor index gave up **1.6%** on the day and **10%** on the week; the Nasdaq-100 fell **1.8%** and one social platform **5.3%**.

The mechanism is not capability anxiety but arithmetic. If capable open weights are downloadable, the justification for **\$180–190B** of single-company annual capital expenditure has to be re-argued from scratch — leaving much of the industry, as one desk put it, looking more indebted than it can justify. One commentator went further, suggesting the release may undermine two long-anticipated laboratory listings outright.

TRAINING

The Observations You Throw Away Are the Supervision You Already Paid For

Standard agentic reinforcement learning discards almost everything the environment says back. The reward is sparse and terminal; the observations returned at every intermediate step — file contents, error messages, terminal output — are treated as context for the next action and nothing more. The week's most consequential result argues they are instead a dense, on-policy supervision signal already sitting inside every rollout, free of charge.

Adding a world-modelling loss over those observation tokens — a constant positive advantage weighted at **0.05**, normalised separately from the policy loss — **doubles** the pass rate on a terminal-agent benchmark. It reaches peak performance in **1.5** to **2.3x** fewer steps, cuts timeouts from **19.8%** to **9%**, and spends up to **30%** fewer completion tokens doing it. The training corpus is modest enough to replicate: **2,700** seed tasks expanded to **6,170** generated ones — kept only when a frontier model solved them once in sixteen attempts — for **8,870** total with a hundred held out. Parallel work on the same idea reports an **11.3%** factuality gain.

The named failure modes deserve equal billing: reward hacking, judge bias when a model grades

its own kind, and an echo-trap diversity collapse that appears around **500** steps of training.

A separate argument this week says the industry has misdiagnosed the agent problem entirely. Memory fixes statelessness. Retrieval fixes lookup. Neither fixes *meaning* — and the expensive failures are the ones where the agent had the right words attached to the wrong definition. Revenue net of returns. A sales region that changed shape in a reorganisation. A discount policy superseded last year.

The numbers behind that claim converged from three unrelated directions. One frontier lab measured its own analytics agent at roughly **95%** accuracy with a governed semantic layer and **21%** without. An enterprise vendor measured **20%** against **92.5%**. A regulatory-corpus benchmark found a **70%** accuracy gain for graph retrieval over plain vector search, while **40%** of enterprise leaders now name missing semantic context as their principal blocker and one analyst house projects **60%** of projects that skip the layer will fail by **2028**. The recommended starting scale is deliberately unglamorous: define the twenty most-disputed metrics, not the five hundred easy ones.

FROM THE BENCH

“Not in Source”

Load your architecture decision records into a scoped notebook and drive it from a coding agent, and architecture review becomes citable. The constraint is the feature: roughly **500,000** words per source against a **1M**-token window, with no retrieval across notebooks. On a **300**-document task it fabricated in about **13%** of answers against roughly **40%** for general assistants. The anti-hallucination instruction is one line worth stealing verbatim: if the answer is not explicitly supported by the sources, reply only *Not in source*.

Three Runs Before You Schedule It

A discipline for automations that actually survive: run it manually **three** times before scheduling, require the output to be usable within **five** minutes of review, and retire anything that goes **four** runs untouched. The twenty-one-day habit rule is folklore; the literature puts medians near **two months** with a range of **4** to **335** days, and a meta-analysis of **642** tests backs if-then triggers over intentions.

Every Connection Is a Delegated Action

Model-context connections look like integrations and behave like delegated agency — the framing that makes capability feel like configuration. Four structural failures recur: overbroad permissions where a filesystem server plus shell execution plus environment-variable credentials compound into local authority; auditing with no link between prompt, tool and approval; treating tool metadata as though it had been code-reviewed, when hidden instructions in a tool description are documented exfiltration vectors; and supply-chain drift through shadow servers, zombie servers and typosquatting. The proposed remedy is an **18**-field audit record — agent identity kept distinct from human identity, requested versus granted scopes, policy decision, data classification, trace

identifier on every action — plus hashing tool definitions to alert on change, and a kill switch designed rather than improvised.

A Longer Route to the Wrong Answer

Chain-of-thought does not create information; it chains inferences that already sit locally in the training distribution. Which means it helps only when the data is organised into overlapping concept neighbourhoods that mirror the true dependency structure — and when locality is wrong, estimates collapse toward marginals and no reliable steps emerge at all. Two findings cut against instinct: more intermediate variables did not improve accuracy, and locally-trained models produced *fewer* intermediates than fully-observed ones. The one-line warning is the most quotable thing said about reasoning models this week — a model with the wrong structure will simply take a longer route to the wrong answer.

COMPUTE**Rivals Are Now Each Other's Suppliers**

The most telling commercial fact of the week is that one frontier laboratory proposed leasing compute from a direct competitor — a deal worth up to **\$10B** over two years, structured with monthly payments and an early opt-out. A laboratory that raises capital on the strength of its models is renting the metal to run them from a company it competes with.

The same scarcity is reshaping public procurement. A launch company is in talks to supply a defence department with compute worth up to several billion, against a programme seeking **\$30B** for high-end accelerators, with officials openly uneasy about the concentration. Elsewhere an inference provider has begun pledging its chips as loan collateral, borrowing **\$400M** against silicon.

One hyperscaler is building its way out in hardware instead. A chip codenamed internally for its immutability hard-wires the company's model blueprint directly into the die, projected at **6 to 10x** the efficiency of its newest general accelerator measured in tokens served per unit of power. The shortage that prompted it was severe enough to force the cloud division to turn away outside customers.

CAPITAL**The Private Bid Ignored the Public Selloff**

While public markets marked down the buildout, private capital marked it up. One data platform closed at **\$188B**, up **40%** from December and above the **\$175B** previously floated, raising **\$3B** on a run rate that passed **\$5.4B** in February at **65%** annual growth, including more than **\$1.4B** of annualised AI revenue.

Chinese laboratories are financing at similar velocity. One approaches **\$500M** annualised at **70–80%** gross margins, having raised **\$7.4B** last month and reportedly seeking the same again. Another nears **\$1B** in sales. The maker of the week's headline model has paused new subscriptions because its accelerators cannot keep up, and is seeking approval for a Hong Kong listing within **six months** at a valuation near **\$20B**.

THE CONTRARIAN TRADE**Memory Is the Cheap Way to Stay Long**

July's selloff across the three big memory makers opened a valuation gap worth noting. The laggard trades at just under **4x** book value while its two rivals trade at more than double that — despite chip-unit revenue up **226%** and driving first-quarter operating profit, with pricing power forecast to hold through **2027** even as capacity ramps.

The structural risk sits above ground rather than in the fab. There were **142** anti-data-centre protests across **42** states, and one state has now enacted the first statewide **one-year** moratorium on new construction. Capital is abundant and silicon is purchasable. Municipal consent is neither.

The Meter Has Replaced the Model

Read two of this week's stories together and a single thesis falls out. One frontier laboratory changed access terms for its flagship **six** times in **six** weeks: launch pricing on the ninth of June, an export-control blackout three days later that made it dark for every non-American user, a partial restoration at half limits at month end, extensions on the twelfth and nineteenth of July, then a Friday-night settlement announced at **10:14pm** that made it permanent for top tiers while quietly cutting general usage limits by roughly a third — including for the tiers that nominally won.

A Chinese laboratory answered differently. Not with a better model, but with a date: **27 July**, on which the weights become yours. For anyone building a product with a twelve-month payback, the capability difference between those two offers is now smaller than the *planning* difference. You cannot write a business case against a meter that moves monthly. You can write one against a file you host. This is the mechanism by which open weights actually win — not by being better, but by being schedulable. The corollary is practical: architect every workflow so the model underneath is swappable, and run anything that cannot tolerate a swap on weights you control.

The Layer Everyone Measures and Nobody Builds

Three independent findings this week, on entirely unrelated subjects, converged on the same magnitude. A frontier lab's analytics agent: **95%** accurate with a governed semantic layer, **21%** without. An enterprise vendor: **20%** against **92.5%**. A regulatory-corpus benchmark: a **70%** gain for graph retrieval over vector search. And a projection that **60%** of agent-ic-analytics projects skipping that layer fail by **2028**.

When a legal benchmark, a business-intelligence vendor and a frontier laboratory independently find a four-to-fivefold accuracy swing from the same intervention, that intervention is the bottleneck — and most organisations are currently buying model upgrades to fix a definitions problem. The diagnostic is beautifully cheap. Take one number from this week's reporting and ask where its definition lives. If

nobody can point to a single authoritative place, you have found the whole issue in one question.

Buy the Verbose Model on Purpose, and Run It at Three in the Morning

Everyone treated the new model's appetite as a defect: **130M** output tokens across a benchmark suite against a **63M** average; **13,241** reasoning tokens and about **25** cents to draw a single illustration; **95** tokens on a probe where a rival spent ten. Bloat, benchmaxxing, uneconomic. That verdict is correct for interactive work and exactly backwards for everything else.

Verbosity is expensive when a human is waiting and nearly free when nobody is. Once weights land on hardware you own, the marginal cost of a token collapses toward electricity. So invert the standard deployment. Put the terse, expensive, fast model where a person sits in the loop and latency binds. Put the verbose, cheap, open-weight model on every unattended overnight job — the ones that run at three in the morning, take as long as they take, and get read over coffee. The industry is about to spend a quarter optimising this model's verbosity away. The better move is to route work deliberately toward it.

Publishing the Test May Destroy It Faster Than Publishing the Model

Relabelling identical dangerous prompts as research moved compliance from **17%** to **42%** — a **2.5x** swing driven purely by framing, on a set where a third of responses already helped the attacker. Every published evaluation becomes a training target, and safety evaluations rot faster than capability ones because the gap between passing and being safe is a matter of phrasing.

That argues for something the field finds instinctively uncomfortable. A standards body of the kind now being proposed should hold its evaluation suites *private* and run them itself, against models shared thirty days before release — not publish the tests for reproducibility. Open weights and open evaluations pull in opposite directions. The field currently treats them as the same virtue, and it is going to have to choose.