

INTELLIGENCE ON THE AI FRONTIER

What crossed the wire overnight — models, agents, markets, and the ideas underneath them

THE FRONTIER · MATHEMATICS

Machines Now Prove What Humans Could Not

An unreleased research model formalised ten open problems in machine-checkable logic for the price of a weekend away — and the people whose life's work it touches are only cautiously alarmed.

\$2,000	10	1999	249	20+
TOTAL COMPUTE · TEN PROOFS	OPEN PROBLEMS SOLVED	A PROBLEM UNSOLVED SINCE	PAGES OF MANUSCRIPT	MATHEMATICIANS INTERVIEWED

An unreleased research model has produced machine-verified proofs for ten problems that had defeated mathematicians — in one case since 1999 — and it did so for roughly two thousand dollars of compute, about two hundred dollars a problem. The proofs were written in a formal language a computer can check line by line, then published as a 249-page manuscript with the model's full reasoning attached.

Among the results: the first explicit construction of a long-sought exotic group, a disproof of a decades-old rigidity conjecture, sharper bounds on how densely spheres pack in high dimensions, and several problems from the notebooks of a famously prolific problem-poser. None of it required a human to supply the decisive idea — a threshold that, once crossed, is hard to un-cross.

THE VERDICT FROM THE FIELD

Fascination, Not Yet Fear

Interviews with more than twenty mathematicians at a major congress found the mood closer to wonder than dread — most expect the machines to augment rather than replace them, for now. Yet a newly hired Fields Medalist predicts systems "robustly superhuman" at mathematics within a few years, and the discipline has begun asking what it values once a theorem can be proved and checked with no human able to follow it.

IT WASN'T THE FIRST CRACK

A Method, Not a Stunt

The breakthrough did not arrive from nowhere. Weeks earlier, one lab's internal model disproved a celebrated conjecture in discrete geometry; days later, a rival's system produced a counterexample to a long-standing conjecture in higher dimensions. Machine, then verification, then a result no one had reached — the shape of a repeatable process.

"We are very, very close to a scenario in which a major result gets proved and verified — and no human can understand it."

The Frontier Becomes a Price War

The most capable models on the planet are now being priced like a commodity, and the pressure is coming from openly published weights. A new flagship from one lab weighs in at 2.4 trillion parameters, ships its weights for anyone to run, posts coding scores at the top of the table — and had its API price quietly cut to a fraction of Western equivalents. A smaller sibling runs on a single high-end laptop.

Beneath the flagship, the discounting is brutal: one rival model is claimed to run more than a hundred times cheaper than a leading frontier system, another just topped the charts for AI video editing, and a cluster of open-weight releases is

said to trail the best American labs by only three to nine months.

DISTILLATION — THEFT, OR DEFENCE?

The reflex to call open weights "dangerous" misreads the problem. The real issues are abusive distillation — siphoning a frontier model's reasoning through massive query campaigns, a practice at least one leading lab has publicly accused competitors of — and handing your data to an untrusted host. The weights themselves are the strongest tool for sovereign, air-gapped, low-cost deployment; one search company simply downloaded a set, fine-tuned it, and stripped out the political censorship baked in.

AGENTS FOR A BILLION USERS

The other shift is packaging. The best-known assistant is being rebuilt as a genuine agent: a coding-style harness running in a persistent cloud machine with its own browser, memory, scheduled tasks, and a directory of more than a thousand plugins — pointed at a billion weekly users. Its voice mode was re-engineered to listen and speak at once, collapsing start-up latency from six round-trips to one.

The lesson underneath all of it: the model is becoming the commodity, and the engineering wrapped around it is where the advantage now lives.

Shipping an Agent Is Harness Engineering

The demos survive because production conditions are absent. In the wild, a single plugin that fails to load can block an entire agent gateway from booting — the fix is graceful degradation, falling back from full function to core-only to read-only to offline rather than collapsing outright.

A permission list built to deny by default but left empty silently permits everything. Context "compaction" that keeps the prose but discards the unglamorous identifiers — the IDs,

file paths and endpoints — quietly lobotomises an agent mid-task; the remedy is to compact at seventy to eighty per cent of the window and pin those identifiers verbatim.

And a retry issued without a deduplication key will cheerfully re-fire a tool and charge the customer twice. None of these are model problems. All of them are runtime problems — cheap to prevent, expensive to learn, and invisible until real users arrive.

WHY MEMORY COSTS SO MUCH

The Cache That Eats the Card

Inference bills are dominated by the key-value cache, which must be read in full for every token generated. A seventy-billion-parameter model at a 128,000-token context can need forty gigabytes of it — half a top-end accelerator. Each optimisation attacks the same equation: grouped-query attention for an eight-fold cut, latent-attention schemes that shrink per-token memory, paged memory that slashes fragmentation, and prompt caching that saves half to ninety per cent on a hit.

THE BRAIN TRANSPLANT

Capability Lives in the Function

You can move a trained transformer's capability into a cheaper, subquadratic architecture at equal size — because the knowledge lives in the sequence-mixing function, not in attention itself. One method rebuilt a model as a state-space student on less than one per cent of the tokens a from-scratch run would need, and beat the prior non-transformer models outright.

THE GOVERNANCE TAX

Gateways Become the Control Plane

More than two-thirds of engineering leaders are already working with "AI gateways" — the layer that meters cost, enforces access and watches agents crossing trust boundaries — a market forecast to pass a billion dollars a year by 2028. The agent era is quietly acquiring its toll booth.

READ THE VERBS

Most "AI Projects" Aren't

A useful test is to read the verbs in the spec. Extract, route or validate by fixed rule is automation, even with a model inside; only choosing the next action makes it an agent. Mispricing that difference — architecting for something never actually named — is one of the quieter ways budgets die.

The Great GTM Compression

The way software gets sold is compressing in real time. A survey of more than 150 go-to-market executives found sales cycles shortening from twenty-five weeks to nineteen year-over-year, while contracts under a year jumped from four to thirteen per cent and three-year deals shrank.

This isn't a budget freeze; it's a hedge. Buyers won't commit long-term until a vendor proves its value survives the next model release. Proof-of-concept-to-paid conversion has climbed to around half, pilots are being framed as two-to-four-week timeboxed gates, and the firms leaning into AI now run their go-to-market teams roughly forty per cent leaner than the laggards at the same revenue.

Token Costs Meet the P&L

Spending on large-model APIs is now more than eighty per cent of all corporate AI expense, up fivefold in a year — and finance chiefs are being told to stop treating it as one nebulous line.

The discipline: split token costs into cost-of-goods versus operating expense, then gate spend to a per-unit metric — cost per ticket, per claim, per deal — so the curve bends faster than the market's. One knowledge-search company, meanwhile, tripled revenue from one hundred to three hundred million in fifteen months, almost entirely by expanding inside accounts it already had.

The Bottleneck Leaves the Datacentre

As cognition gets cheap, the constraints move to atoms. One power startup raised its second billion dollars in a year to install home batteries that leapfrog a grid-connection queue of 2,600 gigawatts — more than twice all existing national capacity — because generation got cheap while delivery didn't.

Inside companies, the same story in miniature: an internal AI platform built for non-engineers is now used "all day long," with one session running past nine hundred turns. Intelligence is deflating; power lines and lab benches are the new rate limiters.

Five Threads, Pulled Together

The price of proof is falling faster than the price of tokens. A decades-old conjecture was overturned for about two hundred dollars while cut-rate models undercut the frontier a hundredfold. The curve worth watching isn't "model IQ" but cost per unit of machine-checkable answer — and when the answer is nearly free and can be verified without a human, the scarce input becomes knowing which question is worth posing.

The moat is the harness, not the model. A billion-user agent is a browser, a memory and a scheduler wrapped around a commodity brain; the failure modes that sink agents are all runtime failures; the gateways now taxing the whole field are pure harness infrastructure. Grade the harness — permissions, compaction, dedup, degradation, observability — not the demo.

Distillation is both the cheapest attack and the best defence. The same mechanism used to allegedly siphon a rival's reasoning is what transplants a transformer into a three-times-cheaper student, what powers continual-learning adapters, and what lets a company self-host and de-censor an open model. Banning the technique rather than governing it cedes the defensive upside — sovereign, air-gapped, cheaper AI — to whoever is less squeamish.

MOVE 37 • THE CONTRARIAN PLAY

Stop writing "definition of done" for your AI projects. Write "definition of death" — then hand the trigger to an agent. Analysts expect more than forty per cent of agentic-AI projects cancelled within two years, yet almost no one pre-commits a kill number, a date, and a named human whose authority to stop the thing doesn't evaporate if they turn out to be wrong. The move nobody makes: delegate the kill-switch itself to an AI agent that auto-deletes the project when the metric misses the date. You stop governing the AI and conscript it to govern your own sunk-cost bias — aiming the circuit-breaker, for once, at yourself.

The race is quietly becoming about electrons and lab benches. A two-billion-dollar battery bet and a grid queue larger than all existing capacity say compute's ceiling is now the grid; a genomics model that compressed two years of wet-lab work into days still waits on physical verification. As cognition commoditises, durable advantage re-concentrates in the un-digitised parts of the stack — power, fabs, lab throughput, proprietary physical data. The smartest capital may be rotating out of models and into the bottlenecks models can't dissolve.