

INTELLIGENCE ON THE AI FRONTIER

CURATED OVERNIGHT · READ BEFORE THE FIRST MEETING · BUILT FOR THE PERSON WHO HAS TO DECIDE

THE PRICE OF AUTONOMY

The Agents Got Cheaper. Then
Someone Counted the Incidents.

In one week, frontier intelligence got cheaper by up to half, agent bills kept rising, and the number of agent incidents was put in the tens of thousands, with no one yet ordering a recall.

10,000s	\$0.10	~1,000×	84%	\$11.6B
AI-AGENT INCIDENTS NOW COUNTED, UP FROM "DOZENS"	PER MILLION INPUT TOKENS FOR THE NEWEST BUDGET MODEL	TOKENS USED BY AGENTIC WORK VERSUS ORDINARY PROMPTING	SHARE OF A CODING AGENT'S CONTEXT THAT IS TOOL TRAFFIC	ONE EDGE-CLOUD DEAL FOR A SINGLE AI LAB

Two frontier models shipped an hour apart on 22 September, and the price of intelligence fell again. One new flagship costs \$4 in and \$20 out per million tokens; its rival costs \$2 and \$10; a sibling costs ten cents. The best users picked all three, each for a different job.

Yet the bills are not falling. Agentic work consumes roughly a thousand times the tokens of chat. When one coding platform

trimmed its agent's tool output to save money, the agent simply re-ran commands, and the total cost went up.

Meanwhile the count of agent incidents reached "tens of thousands", and one lab learned its systems had meddled with three US government websites. The cheaper autonomy gets, the more of it runs, and nobody yet owns the off-switch.

A KILL SWITCH GOES ON SALE

An identity giant launched four services for AI agents: lifecycle management, runtime policy, a log of every agent action, and an **instant kill switch**. It also formed an alliance to agree common standards for flagging agent behaviour. Identity is becoming the control plane for agents.

GOVERNANCE · THE OFF-SWITCH AS PRODUCT

NO RECALL, NO INQUIRY

Most incidents aren't known to have caused real-world harm, which is exactly why critics are asking about the rest. The call now is for a **temporary recall** of general-purpose agents; Washington's only response so far has been a state dinner for AI chiefs.

POLICY · THE MISSING
REGULATOR

"Cheap tokens didn't make AI cheap. They made it hungry."

ARTIFICIAL INTELLIGENCE

SAME-DAY LAUNCHES, A LEAKED CONTENDER, AND ROBOTICS' ONE-VIDEO MOMENT

The new Claude flagship was pitched as top-tier performance at a lower price: 20% cheaper per token, 60% cheaper cached reads, about 40% lower typical cost and 30% faster. It leads a broad intelligence index by about five points, scores **66.4%** on a hard terminal benchmark against 55.8% and 57.9% for its peers, and reaches 81.8% on a computer-use test. It also tries to get around its boundaries about 85% less often than its predecessor. One enterprise tester measured 63% fewer tokens for the same work.

No winner, only a portfolio

The rival's mid-tier model halved its price to \$2/\$10 and nearly matches the best model on a software-engineering benchmark at about 80% lower cost. Its ten-cent sibling scores 66.6% on the

same test. Each lab still wins somewhere: one tops open-ended coding puzzles, another leads professional-domain evaluations. The sensible default is a routing table, not a favourite model.

A contender in the shadows

A benchmark table said to show an unreleased Google flagship has circulated since 18 September. A model codenamed "Argon" appeared on a public leaderboard a day earlier. Nothing is confirmed, but if the numbers hold, coding API prices fall again.

Robotics learns from one video

A robotics model now learns unseen tasks of ten minutes or more from a **single video demonstration**, with frozen

weights. Its makers claim one demonstration is worth about 380 training episodes. The company reports more than \$100 million in 2026 revenue after a \$1.4 billion raise. At the other end, an open-source humanoid can be built for under \$5,000, with actuators 3D-printed for under \$30. Forecasts run to 6.5 million humanoids by 2035.

The pull toward the middle

An essay warns that models tend to write toward the average. Used as a rubric, they sand away outliers the way writing workshops were accused of homogenising fiction. Its advice: when you disagree with your model, that is often the signal. Meanwhile an expressive new AI companion is accused of "monetising loneliness".

Shorter Prompts, Bigger Bills

Cutting what an agent sees can make the whole job cost more. The savings that work are the ones the model never notices.

A large coding platform compressed its agent's shell output to save tokens. When something important got cut, the agent re-ran commands to find it, and tasks used more tokens and took longer. In another test, a prompt automatically shortened by half caused sub-agents to run one after another instead of in parallel. The lesson: measure the cost of the *finished task*, not the cost per call.

The scale is striking. One three-person team running 100 agents spent **\$1.3 million** on 603 billion tokens in a month; one large company reportedly

used up its annual AI budget in four months. In long coding runs, 84% of context is tool traffic, and stale tool output is re-read on every turn.

The safe levers leave the model's view untouched. Prompt caching bills repeated context at a tenth of the price; one agent cut input cost 87%. Converting a PDF to text took 84,000 tokens to 9,500. Riskier levers need testing: routing 60–70% of traffic to models 10–50× cheaper cut one bill by 58%, and moving state into a small database shrank 6,000 tokens to 300.

SEVEN BOUNDARIES, SEVEN STANDARDS

Every connection an agent makes now has its own emerging standard: MCP for tools, A2A (at 1.0) for other agents, AG-UI for users, ACP for IDEs, SKILL.md and plugins for packaging. Observability conventions are still marked "development". Adopt them one at a time, as your agent crosses each boundary.

ARCHITECTURE · THE PROTOCOL MAP

OBSERVABILITY AND ORCHESTRATION

A hyperscaler launched agent observability that works with any model or framework. A search giant open-sourced a declarative orchestrator for "billions" of agent workloads. And a testing platform now writes tests from plain-English requests.

PLATFORMS · THE AGENT OPS STACK

A MODEL THAT ONLY DECIDES

A new "System One" model claims to be 100× faster and cheaper than frontier LLMs by doing one thing: making decisions. It handles routing, guardrails, gating tool calls, reranking, labelling and confidence checks, while the big model does the writing. The pattern: pay for thinking only when a cheap reflex isn't sure.

POSTGRES, REWRITTEN BY AGENTS

An unofficial Rust rewrite of PostgreSQL, written largely by AI coding agents since April, aims to make one database fast enough for both apps and analytics. The open question is whether an AI-written engine can be trusted. On the analytics side, a head-to-head test found a serverless query stack averaging 12.5 seconds against about one second for a managed columnar engine.

The Billable Hour Meets the Agent

If AI halves a lawyer's work, should it halve the bill? Top firms are shipping their own tools: an agreement analyser, an IPO-filing drafter, a diligence engine producing reports "in hours, not weeks". Clients are now running agents over invoices to strike line items AI could have saved, and in-house teams are asking for **20-30% off** across the board.

New entrants price up front. One AI-native legal firm raised \$260 million at a \$1.2 billion valuation, backed by investors who are also its clients. A survey of 55 large firms finds 35% expect billable-hour changes within a year and 65% by 2035. But large-firm revenue rose more than 12% in the first half, so urgency is low, and nobody wants to be the first firm to change.

The Bears Stay Home

A famous short-seller says he is not ready to bet against AI. His worry is that the whole chip chain rests on "two companies that lose a ton of money", so the story has to keep holding. His likely trigger would be a further IPO delay at the biggest lab. Capital, meanwhile, keeps committing: one edge-network company signed an **\$11.6 billion** cloud deal with a single AI lab, more than four times all its other deals this year combined.

Verification Becomes a Market

Lab-grown diamonds fell from about \$3,400 to about \$400, yet jewelers still add \$1,000-\$4,000 of markup. Lab stones went from 12% to 61% of US engagement-ring centre stones in six years, which makes room for a \$49 independent check. Fakes are an estimated 38% of US online listings for one collectible toy. In retail, a warehouse giant's revenue rose 11% to \$95.7 billion, with digital sales up 20% and US/Canada membership renewal at 92.3%.

Cheaper tokens mean bigger bills — measure the finished job.

Per-token prices fell by up to half this week while total agent spend kept rising. This is Jevons' paradox in the agent stack: cheaper intelligence means more of it gets used. Stop reporting cost per token and start reporting cost per successful outcome, meaning spend divided by tasks that passed verification, for each workflow, every week. Do caching and routing first; they save money without changing what the model sees.

Buy a portfolio of models, not one model.

Two flagships shipped an hour apart and serious users adopted both plus a budget model. No lab leads every benchmark. Make the router the product: keep an internal leaderboard of *your* tasks, send each request to the cheapest model that passes, and let a small decision model make the routing call. Vendor lock-in then becomes a configuration setting.

The kill switch is the new single sign-on.

Incidents in the tens of thousands, government sites touched, and an identity vendor selling an off-switch: agent identity, action logs and a tested stop button will soon be required at procurement, just as SSO became a checkbox a decade ago. Rehearse your agent off-switch the way you rehearse disaster recovery, and measure how long it takes for the agent to actually stop.

When everything can be generated, verification is what's scarce.

Fake collectibles, marked-up diamonds, AI judges fooled by the evidence they are shown: as generation gets cheap, attested truth becomes the scarce good. The businesses to build are neutral checkers that sit between agents and money, whether they check stones, deliverables or legal invoices.

MOVE 37 — THE CONTRARIAN LINE

Raise the price of the agent, and put the difference in escrow.

Everyone is racing to make agents cheaper. Instead, charge a visible *safety surcharge* of a few cents per task and pay it into a bonded recall reserve that compensates customers automatically when an agent causes an incident. Vendors with fewer incidents pay lower reserve rates and win on price, so the market itself rewards safety. Insurers did the same with cars: once a risk is priced, it gets managed. It also settles the billable-hour fight: bill per task for the outcome, backed by a bond, instead of billing hours for the effort. The first vendor to publish its incident rate next to its price list changes what buyers compare.