

THE DECISION LAYER · LEAD STORY

The Model That Refuses to Speak

A newcomer generates no text at all — it answers yes-or-no and multiple-choice questions with a calibrated probability in half a second, and the whole industry is rethinking what it has been paying generation prices to do.

200×	400×	~500ms	\$0.042	1h45m
FASTER ON CLASSIFICATION	CHEAPER PER DECISION	TO A CALIBRATED ANSWER	PER MILLION INPUT TOKENS	TO TRAIN AN OPEN REPLICA

A model that never writes a word has become the most-discussed launch of the week. Framed through Kahneman's two systems — the large language model as the slow, deliberate System Two — the newcomer positions itself as System One: fast, intuitive, and built to decide rather than to compose. You hand it a state and a set of typed questions; it returns probabilities, not prose, in roughly seventy to five hundred milliseconds.

The interface is deliberately narrow. Three question types cover the surface: Choice picks from options, Score places an answer on an ordered scale, and a boolean yes/no returns a bare probability between zero and one. Independent questions about the same state evaluate in parallel, which is what makes the speed claims plausible. Headline figures put it at up to two hundred times faster and four hundred times cheaper than a conventional model on classification work, at a listed price of four cents

per million input tokens with output free — about forty-two cents for ten thousand decisions.

The launch-week receipts are concrete. One browser agent found flights in seven seconds for under half a cent. A widely shared demo trimmed a coding session from roughly a million tokens to eighty-six thousand in about a second by scoring which tool calls to keep, drop, or truncate. Another builder sorted more than a thousand research papers into two dozen topics for eight cents. Access arrived broad on day one, spanning a hosted endpoint, a popular web SDK, and an edge runtime.

Enthusiasm comes fenced with caveats that its own advocates repeat. The model explains nothing — there is no reasoning trace to inspect when it errs — and confidence is not accuracy: a score of 0.98 on a badly written question is merely confidently wrong. Every benchmark so far traces back to the vendor or a demo. The discipline, then, moves from prompt-craft to question-craft.

THE HIDDEN CLASSIFIER

WHAT WAS ALWAYS THERE

Every serious coding agent has quietly carried the same part: a gate that, before running a shell command or touching an out-of-scope file, decides proceed, prompt, or block. That judgment lived in the closed half of the harness with no way to plug in your own.

Exposing that step as something anyone can call is the quieter revolution — the decision, unbundled from the generation, and handed to the builder.

OPEN IN 48 HOURS

A LAPTOP-SIZED REPLICA

Within two days the community shipped a working clone — a small adapter and a readout head bolted onto a half-billion-parameter open model, trained in one hour and forty-five minutes on a single laptop. A larger nine-billion version followed.

If judgment this cheap can be reproduced overnight, the durable advantage is not the model. It is the library of well-formed questions around it.

“Most models answer with a sentence. This one answers with a number.”

EFFICIENCY

A 27-Billion Model in Under Six Gigabytes

Using ternary weights — every parameter reduced to minus one, zero, or plus one — a new release compressed a 27-billion-parameter model from 53.8 gigabytes down to 5.9, while retaining 98.2 percent of the parent's aggregate benchmark score.

The practical upshot is capability that runs on modest local hardware, routing around the very power and memory walls the data-centre industry is straining against. A tradeoff is teased but not yet disclosed.

INFERENCE

Why a Million Tokens Is So Hard

Scaling context to one million tokens and beyond breaks on three walls: quadratic prefill work that pushes time-to-first-token to tens of seconds, a key-value cache that can consume 137 gigabytes for a single long stream, and a "lost-in-the-middle" collapse where mid-context accuracy falls below thirty percent.

The emerging 2026 fix is a triad — spectral position modulation, dual-chunk attention with an 8,192-token chunk, and dynamic sparse routing promising roughly ten-times faster prefill.

VERTICAL AI

From Demo to Daily Driver

Specialised configurations are pulling ahead of general chat. A legal build indexing more than 230 million pages of case law and statute scored 54 percent on a research benchmark against 38.7 for plain web search. A finance offering shipped with brokerage and portfolio connectors.

The pattern: the moat is no longer the model but the proprietary index and the domain guardrails wrapped around it.

ROBOTICS

Folding Laundry in Homes It Has Never Seen

A humanoid platform pretrained on a large human-behaviour dataset lifted zero-shot household-task success from nine percent to fifty-six across thirty unfamiliar homes — the kind of generalisation jump that separates a lab demo from a product.

VOICE & VISION

Real-Time, Multilingual, and Everywhere

A flagship live model added real-time visual grounding, background tool calls, and mid-conversation switching across 97 languages. Rivals shipped voice modes, persistent memory, and more accurate transcription in the same week — the interface, not the intelligence, was the battleground.

NEW CATEGORY

The Household Agent Arrives

Personal and family agents emerged as a distinct product line — one topping the app charts on claims of renegotiating bills and cancelling unused contracts. A creative studio, meanwhile, reported an agent can now drive professional visual-effects tools autonomously and render rigid objects convincingly, though living things still elude it.

The Harness Tax

Same Model, Same Task, Double the Bill

A rigorous study put seven models through three different "harnesses" — the layer that manages instructions, tools, memory, and context — across sixty software-engineering tasks. The result should reorder how teams budget: one setup solved 97.8 percent of bugs and another 96.7, a rounding error, yet one cost 1.33 dollars per attempt against 0.67 for the other.

The culprit was not the model but the wrapper: the pricier harness shipped longer instructions and heavier tool definitions, giving it an initial context more than ten

times larger. A minimal four-tool setup — read, write, edit, and a shell — sat squarely on the cost-success frontier on every benchmark.

Stranger still, models frequently performed best inside a competitor's harness rather than their own maker's, winning nine of twelve head-to-head comparisons. The unavoidable conclusion is that the true unit of evaluation is model times harness times workload — and the number that matters is cost per successful task, not cost per token.

TOOLING

BASH BEATS THE CATALOG

Plain shell access beat elaborate typed tool catalogs by more than twenty points on two agent benchmarks while using up to seventy percent fewer tokens. Layering typed tools on top added nothing. Selection accuracy, separately, degrades once an agent juggles more than thirty to fifty tools.

MEMORY

A GIT REPO AS SHARED MIND

One experiment used a version-control repository as shared memory for an agent fleet, every claim committed like code. Thirteen workers over twelve days, with no central planner, posted more than 1,700 contributions — an audit trail as a first-class design choice.

THE JUDGES ARE UNRELIABLE

A study of automated evaluation found 57.5 percent of conversations rated "satisfied" had actually failed the task, and model judges quietly favoured their own family — a warning to anyone gating releases on an AI grader.

LAUNDERING CAPABILITY

A weak local model can split a dangerous request into innocent-looking pieces, pose them to aligned frontier models in separate sessions, and reassemble the answer — arguing that safety must be judged per account and session, not per prompt.

The Bottleneck

Power Is the New Chip Shortage

The dominant player now treats electricity, not silicon, as the binding constraint, saying it tracks "every single gigawatt" of land, power, and shell on the planet. A new power-management product reportedly lets customers run forty percent more accelerators on the same supply.

Capital is following the constraint. A modular data-centre builder raised 3.9 billion dollars at a 30.9-billion valuation — more than a third of the entire week's venture funding — as the industry pivots from gigawatt megaprojects toward container-sized units squeezed into buildings that already have power.

Warning Signs

The Debt Starts to Sour

Not all the exuberance is holding. Bonds financing a data centre tied to a marquee trading firm deteriorated quickly in secondary trading, with yields jumping past 11 percent — more than two points above where they sold weeks earlier — as investors demand far more to hold AI infrastructure paper.

Markets

A Quarter-Trillion Left on the Table

First-day IPO pops have averaged nineteen percent since 1980 — roughly 250 billion dollars founders never captured. Prospect theory explains the calm: anchored to early low valuations, boards experience each step up as a win, even when the pop is money forgone.

Venture

Infrastructure Eats the Round

In a single week, AI infrastructure absorbed 5.36 billion dollars across just ten rounds, against 1.18 billion for the application layer spread over twenty-four. Governments and national pension systems are entering venture at scale as anchor investors.

Security

Whales, Harpooned by Forged Orders

A crew identified some 680 high-value fintech users through on-chain analysis, then forged official legal orders via a compromised government email account to pry loose identities and histories — demanding roughly three million dollars in Monero. The weak link was a borrowed identity, not broken cryptography.

THESIS

The Decision Layer Is the New Middleware

Put the text-free model, the hidden classifier, and the finding that plain shell beats typed catalogs side by side and a pattern surfaces: the industry is unbundling judgment from generation. For two years one model did everything; now the fast, cheap, calibrated "which one, how much, should I" step is being pulled into its own layer. Because that layer is near-commodity — reproducible on a laptop in under two hours — the durable edge is not the model but the accumulated library of well-formed questions and the guardrails around them.

METRIC

Retire Cost-Per-Token

The harness study and every quiet pricing change tell one story at two scales: identical work can cost multiples more depending on where and how it runs, with no matching gain in quality. Most teams still reason about token price in isolation and never instrument the full unit — model times harness times workload — against successful completions including retries. Adopting cost-per-success as the reporting primitive is a cheap change that reframes every build-versus-buy and model-selection argument around outcomes.

TREND

Efficiency Is Quietly Winning

A 27-billion model in under six gigabytes, byte models beating token models with a sixth of the data, judgment at four cents a million tokens — the week's most durable news is not a bigger frontier model but the collapsing cost of good-enough intelligence at the edge. Capability that fits on local hardware routes around the power and memory bottleneck that data-centre capital is straining against. The supercycle and the efficiency wave are on a collision course, and the cheaper bet is efficiency.

CONTRARIAN · THE BOLD MOVE

Make Your Agents Dumber and Stricter

The reflex is more tools, bigger context, smarter models. Every serious signal this week points the other way: plain shell beat typed catalogs by twenty points; a four-tool harness sits on the frontier; selection accuracy craters past fifty tools; tool definitions can burn fifty-five thousand tokens a turn. The unexpected play is to treat capability as a liability to be rationed — the smallest tool set, a hard deterministic gate before every risky action, and no reasoning where a probability will do. In a field sprinting toward more, the edge is engineered restraint.

WHITESPACE

Trust Is the Missing Product

More than half of "satisfied" eval verdicts were wrong; safety filters fall to requests split across sessions; autonomous agents breach real companies as a flex. Our verification tooling lags our generation tooling badly. The open ground for the next three years is not another agent but the sellable layer that proves an agent did what it claimed — session-level safety evaluation, judge-free completion checks calibrated against verifiable rewards, and commit-style audit trails. The market is loudly saying it cannot trust its own AI; that is a business, not a footnote.