

INTELLIGENCE ON THE AI FRONTIER

Models, machines, money and the people steering them — distilled before the first coffee.

THE DECISION LAYER

The Model That Never Writes a Word Is Coming for the Agent Bill

A new class of "System One" models picks from your list instead of composing prose — hundreds of times cheaper, cloned within a week, and dangerously certain when the right answer isn't on the list.

\$0.042 PER M INPUT TOKENS	444× CLAIMED COST CUT	65% SWE-BENCH LEAKAGE	9.5 CVSS, NETSCALER 0- DAYS	76% OF WEEK'S VC TO ONE FIRM
---	---------------------------------	------------------------------------	--	---

Most of what an AI agent does all day is not writing. It is deciding: which worker runs next, whether a source is relevant, whether a task is finished, whether to retry. Until this month nearly every one of those micro-choices was routed through a frontier model and billed as if it were an essay.

Jev changes the unit of work. Hand it a state, a question and a closed set of answers, and it scores every option in a single pass, returning a typed choice with a probability and a confidence figure. Its maker claims 70–500 ms latency at \$0.042 per million input tokens, with output free — a 444-fold saving over generation. Three question types cover most needs: a choice, a yes-probability, and a weighted score.

The open world answered in days. A 421-million-parameter encoder answers in about 33 ms on a commodity GPU; a contrastive 8B model out-judged Jev on code, 81.6% to 71.1%; a 144M model now runs on a laptop CPU. The emerging pattern is a cascade: accept confident answers, escalate the rest — keeping 99% of frontier accuracy at 57% of the cost.

The catch is structural. Without a hidden "none of these" slot, the model is near-certain on inputs that fit nothing; option order sways it; and a forged "already approved" tool message dragged a block on deleting SSH keys to a coin flip. The lesson for builders: the list of answers is now the product.

FRONTIER MODELS

Opus 5.5 Undercuts, Sol Splits the Room

The new Claude flagship lands about 40% cheaper and 30% faster than its predecessor at \$4/\$20 per million tokens, matching the prior top tier. Testers pick Sol for a brisk daily driver, Opus for ambitious builds; on a 657-workflow automation benchmark it scores 40% against 33.2%.

SECURITY

NetScaler Zero-Days Exploited in the Wild

Two pre-auth remote-code bugs rated 9.5 hit default gateway configurations. Fixed builds are 14.1-73.37 and 13.1-64.23. Reports say a fix was held until attacks began — the fourth exploited flaw in the line this year.

"If a person could decide it in under ten seconds, a model that never writes should decide it."

THE NEW RULE OF THUMB FOR AGENT BUILDERS

RESEARCH

Harnesses Move Into the Weights

A training method that bakes agent-harness behaviour directly into the model reached 44.3% success, beating the untrained model even when the harness was bolted on (41.7%). A graph world model wrapped around an 8B model lifted household-robot multi-task success from 19.9% to 92.6%.

Pairwise LLM judging with Bradley-Terry ranking hit 35.1% on a polyglot coding test after 30 expansions — beating a self-improving agent that needed 80 — for about \$34. And self-organising agent teams scored 66.7%, outdoing even a perfect router at 59%.

BENCHMARKS

The Leaderboard Remembers Its Answers

A memorisation probe found leakage in more than 65% of a flagship software-engineering benchmark; once scrubbed, pass rates fell 6 to 14 points, one small model sliding from 46.8% to 35.6%.

Another study showed a single confident, misleading user hint cut top models' scores by up to 46.7% relative. In recommendation, LLM rerankers lost 92-95% of their quality on realistic candidate lists, and none of nine beat a plain SVD baseline.

OPEN WEIGHTS

The Cheap Models Holding Prices Down

Open models are the quiet reason consumption pricing never arrived. One Chinese flagship costs \$0.435/\$0.87 per million tokens and scores 46 on a composite intelligence index against 47 for a model priced at \$4/\$20.

Long-context attention research cut 1M-token prefill compute five-fold and shrank the KV cache from 12.09 GB to 2.69 GB. A 512 GB desktop with 1.2 TB/s bandwidth now runs what once needed a rack, and a major chipmaker agreed to buy the leading model hub.

METHOD

Write the Spec, Then Let the Agent Loose

Longer prompts cannot fix the three problems that stall enterprise agents: understanding an existing system, building the right change, and trusting work done unattended. Spec-driven development pins behaviour, scope, constraints and acceptance criteria before a line is generated.

The mature stack builds a cross-repository knowledge graph of calls, inheritance and dependencies, then walks it forwards and backwards to estimate the blast

radius of a change. A human approves an action plan; dependency-ordered tasks run in parallel; spec-adherence agents watch for drift; execution is air-gapped.

Claimed results are striking — 84.95% on a hard coding benchmark and a five-month project closed in five days — but the durable idea is older than AI: decide what "done" means before anyone starts. Rules live at project level; generation prompts live per change.

PRODUCTS

The Agent That Amazon Won't Let Shop

Meta's Muse passed 500,000 users in week one and is getting its own inbox and a keyring gadget — but was barred from buying on Amazon for failing to identify itself, and its desktop app has logged two security incidents.

CAREERS

Ten Years, or One Year Ten Times?

Expertise compounds from exposure, not tenure. The test: what can you reason about today that you couldn't three years ago? Nearly half of product-manager listings now ask for experience writing evals.

TOOLS

A Sandbox That Boots in Under a Millisecond

Rather than chain twenty tool calls, let the model write a small program — and run it in a tightly restricted Python interpreter that starts in under 1 ms, skipping container spin-up, network hops and secrets plumbing. Elsewhere, agentic retrieval on fast inference hardware is replacing the vector database altogether.

CRAFT

Prompting Without "Think Harder"

The newest models set their own reasoning effort, so the old incantations can go. What still works: name the design clichés you don't want, tell the agent to explore broadly before acting, give long tasks a finish line, fence pasted text against injection, and send screenshots — even low effort now reads dense charts better than last generation's best.

VENTURE

One Firm, Three-Quarters of the Week's Money

Twenty funds closed \$6.8 billion in a week, and a single manager took \$5.75 billion of it. At the other end, a well-known seed firm is winding down its fund model to write \$100K-\$3M cheques from its balance sheet — no leads, no board seats.

Rounds totalled \$9.1 billion across 119 deals, one \$3.4 billion convertible accounting for 37%. Chip funding fell 91% in a fortnight. Funds from 2017-19 have returned just 9% in cash after nine years, while megafunds absorb 64% of new capital. Median gross revenue retention slid from 88% to 84% — speed of shipping is the new moat. An autonomous coding agent passed a \$1 billion run rate.

HEALTH

An Agent for Every Prescription

Two-thirds of first-year prescriptions for new US drugs are never filled; doctors lose 13 hours a week to paperwork. One startup assigns each script its own AI case manager — prior authorisation, appeals, affordability, routing — free to doctors and patients, paid by drugmakers. It covers 85% of US zip codes and is now valued at \$3 billion.

PAYMENTS & POLICY

Cash Died One Use Case at a Time

Fintech's winners paired one demographic with one frustration, then followed the customer. Serving a \$120K earner is six times easier than a \$20K one. After Venezuela's earthquakes, a temporary US licence lets dollar rails run over stablecoins until 23 October. Beijing may let its giants buy a new Nvidia workstation chip, and a \$75 billion oil-funded sovereign fund has poured \$2 billion into venture.

THESIS

The List of Answers Is the New Moat

Every decision model chooses from a list someone else wrote — and fails when that list is incomplete. The model itself is already cloned at 144 million parameters; the scarce asset is a well-kept option schema, versioned like an API, with an escape hatch built in. Whoever curates the best taxonomies of decisions in claims, underwriting or code review will own routing regardless of whose model sits beneath.

INVENTION

Shuffle the Options, Catch the Attack

Order sensitivity and injection are weaknesses — but at four cents per million tokens, cheap to probe. Ask the same question five times with options permuted and a hidden "none of these" canary seeded in. Agreement passes; disagreement is itself the alarm, escalating to a larger model or a human, with a signed receipt of every permutation. A failure mode becomes a detector.

SECURITY

Agent-Speed Attacks End the Patch Window

Patch cycles assume a human adversary who needs days. When agents weaponise a disclosure in hours, "we'll fix it once it's exploited" becomes negligence. With incidents reportedly running to the tens of thousands, liability — not new statute — will force agent-speed service levels into vendor contracts and insurance pricing.

MOVE 37 · THE CONTRARIAN BET

Agents Won't Kill Margin. They'll Pay a Toll.

The consensus says agents end lazy margin by hunting the cheapest fare. The contrarian read: sellers will charge agents more. A retailer barring an unidentified shopping agent is the first toll booth; a seaside café's \$3 split-bill fee is the precedent — new payment behaviour always gets priced.

Expect agent-only price tiers, fulfilment fees, and "verified agent" credentials sold like extended-validation certificates, with anonymous bots throttled. The business is the clearing house: an identity-and-toll exchange that issues verifiable agent credentials, settles micro-tolls in stablecoin, and hands merchants attribution — a payments-shaped company for the agent economy.

CAREERS

Your Private Eval Set Is Your Résumé

Public leaderboards are majority-contaminated; job listings now demand eval skills; practitioners build twenty-case benchmarks from their own failures. The portable asset of a senior engineer or leader becomes an uncontaminated, domain-specific eval suite — proof of judgment you can run, not merely claim.