

INTELLIGENCE ON THE AI FRONTIER

THE CAPABILITY RACE, THE AGENT ECONOMY & THE OFF-SWITCH

THE CAPABILITY RACE

A New Model Takes the Crown—and Widens
Its Lead Where It Hurts

Anthropic’s Fable 5 tops nearly every benchmark that matters, and its advantage grows on the longest, hardest problems—even as open weights and rivals close the gap on price.

72.9%	99.8%	\$10/\$50	13,841	\$60B
CURSOR BENCH (VS 64.3)	USAMO 2026	PER M TOKENS	BUGS FOUND BY AGENT	SPACEX BUYS CURSOR

Anthropic’s new flagship, Fable 5—the public, guard-railed sibling of its “Mythos”-class system—is the best model in the world by a substantial but not shocking margin. The more striking claim is that its edge *grows* as tasks lengthen and stiffen.

Mechanically it ships with thinking always on, governed by an “effort” dial from low to “xhigh,” and safety classifiers for cyber, bio, chemical and distillation risks that auto-route dangerous prompts down to Claude Opus 4.8. Access is steep: \$10 and \$50 per million input/output tokens,

double Opus, with 30-day retention.

The scoreboard is lopsided—72.9% on Cursor Bench to GPT-5.5’s 64.3, 94% GPQA Diamond, 99.8% USAMO 2026, 87–88% FrontierMath, and 55% on RiemannBench where Opus managed 34. It ranks first on Agent Arena, ProofBench and the Debate Benchmark, and beat Pokémon FireRed by vision alone.

The honest caveat: Fable 5 wins “by being right more, not wrong less.” It shows weaker steerability and a worse position bias, picking the first option 59% of the time.

Open Weights,
Months Behind

Z.ai’s GLM-5.2 (open weights, MIT, 1M context) beats GPT-5.5 and Gemini 3.1 Pro on SWE-Bench Pro and matches them on Terminal-Bench 2.1. The asterisk is hardware: Unsloth builds run 217–254 GB before a KV cache that adds ~9 GB per 100k tokens.

The scoreboard is lopsided—72.9% on Cursor Bench to GPT-5.5's 64.3, 94% GPQA Diamond, 99.8% USAMO 2026, 87–88% FrontierMath, and 55% on RiemannBench where Opus managed 34. It ranks first on Agent Arena, ProofBench and the Debate

Benchmark, and beat Pokémon FireRed by vision alone.

The honest caveat: Fable 5 wins “by being right more, not wrong less.” It shows weaker steerability and a worse position bias, picking the first option 59% of the time.

The Free Model Grew a Doctor's Brain

Frontier health reasoning was distilled into the *free* GPT-5.5 Instant—a teacher plus 260+ doctors across 60 countries reviewing 700,000+ responses—cutting flagged factuality issues 71%, at the scale of 230M weekly health questions.

“The lab with the most compute will win in the end.”

GREG BROCKMAN, PRESIDENT, OPENAI

Expertise Now Beats Coding, the Data Says

Across 400,000 coding-assistant sessions and 235,000 users, every occupation succeeded within seven points of professional engineers. Humans now do ~70% of planning, the model ~80% of execution; session value rose 27%. The work shifted: “fixing code” fell 33%→19% while “operating software” rose 14%→21%. The bottleneck moved from writing code to knowing what is worth doing.

Compute Rules All

OpenAI president Greg Brockman argues that as frontier models improve in lockstep, the lab with the most compute wins. Agent users today are only “10–20 million” against ChatGPT’s “billion” who lack “agentic power.” Hence OpenAI raising \$122 billion this year for datacenters—the spend Dario Amodei mocked as “YOLO-ing.”

‘Token Capital’ Reframes the Race

Satya Nadella argues the durable edge is the learning loop, not the model, coining “token capital”—the AI capability a firm owns. Prescriptions: private evals, private RL loops, queryable memory, and model sovereignty so you can swap models without losing expertise.

GPT-5.6 Looms; Claude Code Gets Artifacts

OpenAI is said to ship GPT-5.6 next week with a 1.5M-token context and faster Codex. Claude Code “artifacts” turn sessions into live, auto-refreshing shareable pages; Perplexity “Brain” is a memory graph linking each memory to its source.

A Decade-Old Bottleneck, Broken?

A stealth startup, Subquadratic, claims to slash the computations a transformer needs—implying sub-quadratic scaling against attention’s quadratic cost—for a model far cheaper and less energy-hungry. Skeptics await; it has “started to share the receipts.” (*Unconfirmed.*)

The Counter-Bet: Route, Don’t Scale

One inference layer serves known decisions on CPU and routes only novel ones to a frozen LLM on GPU, claiming 33x faster inference, 82% lower cost and up to 90% fewer tokens—the margin lives at the inference layer, not the datacenter.

Agents Now Build, Audit and Attack Software

The week's defining datapoint: Fable 5 opened an unprompted 1,810-line pull request in ~30 minutes, adding 37 passing tests and refactoring shared logic first—as a senior maintainer would.

On defense, Cloudflare ran a two-model harness—one finds

bugs, one argues each down—surfacing 13,841 real bugs across 145 repositories. On offense, a Rust clipboard-hijacker shipped ~15,500 wallet addresses via AI-generated tutorials. One capability, three directions.

'Record & Replay' Is the New Primitive

Codex Record & Replay watches a task once and converts it into a reusable skill across apps and tabs. Block's Builderbot, tagged in Slack, takes Linear/Jira tickets across hundreds of millions of lines, opens PRs and watches CI—clearing a months-long backlog “in days.” The sting: custom agent scaffolding is pegged at ~12 months of relevance, citing 350,000 Android apps built in one week.

You Probably Don't Need an Agent

The “agent” being sold is usually a workflow in disguise—and paying agent prices for deterministic work wastes money. A workflow follows a fixed path; a true agent improvises, with hidden costs in loops, tokens and review. The gate: high volume/value, repeatability, cross-system reach, a measurable baseline—then sandbox→shadow→supervised.

THE MACHINE ROOM & THE WRITERS' ROOM

The machine room: AI breakthroughs are infrastructure breakthroughs. GPU racks draw 80–120 kW (~40 homes); GPUs “fail because they're hungry”—an idle GPU costs as much as a working one, so throughput is set by how fast data is fed, not raw compute.

Self-correcting writing: a multi-source agentic report writer normalizes audio, PDFs, spreadsheets and scans into one evidence format, drafts each section, then reviews every draft against the source and makes targeted corrections—later sections building on earlier ones.

SpaceX Buys Cursor's Parent for \$60 Billion

In what is called the largest venture-backed startup acquisition ever, SpaceX has acquired Anysphere (Cursor) for \$60 billion. The curve borders on absurd: roughly \$4M ARR to \$4B in two years—~1,000x—\$2B annualized in February, \$4B weeks later, three-quarters from businesses.

From Cursor's Compile conference: the team avoided AI coding in 2022 as too crowded, then four engineers built a prototype in two weeks and hand-onboarded the first 20 testers.

The 'AI RIA': a Brand-New Legal Animal

Coinbase launched an SEC-registered AI advisor—handing the software a fiduciary duty even as it warns output “may be inaccurate.” Mercury countered with “Command,” atop a platform of 300,000+ customers, \$650M revenue, \$5.2B valuation. The open question: genuine financial AI labs, or roboadvisors 2.0?

The Off Switch: Living by Caesar's Thumb

Days after releasing Fable, the US government reportedly called—Amazon had found a way around some safety features, raising China-exploitation fears—and gave Anthropic 90 minutes on a Friday to comply with export controls. Its only lever: cut everyone off, Americans included. The thesis: any US AI maker selling abroad now lives or dies by “Caesar's thumb.”

Where the day's threads cross—and what they mean for the reader who has to act on them.

The Moat Moved from the Model to the Loop

Nadella says own the learning loop; the session data shows humans own planning while models own execution; free GPT-5.5 just absorbed specialist medicine by distillation. Capability is commoditizing toward free, so what compounds is the private context, evals and memory you build around it. The move: turn an expiring course, your notes and decisions into a persistent Claude Project or wiki —“token capital” at personal scale that survives the next model swap.

Don't Buy an Agent—Buy Back a Workflow

Hold “most teams just need the boring workflow” against Codex Record & Replay and Block's Builderbot. The winning 2026 move isn't deploying autonomy; it's recording your highest-volume deterministic chore, proving it in shadow mode, and escalating to real agency only where steps can't be predefined. Run the four-bar test and capture 80% of the value at a fraction of the cost and risk.

The Builder and the Burglar Are the Same Agent

Fable 5's 1,810-line PR, Cloudflare's 13,841-bug harness and AI-promoted wallet-stealer malware are one capability pointed three ways. The defense: adopt the adversarial two-model pattern (one proposes, one argues it down, a human reviews) for anything an agent writes or merges, and treat AI-generated tutorials as an attack surface. Verification is the scarce skill.

Move 37: Short the Agent, Long the Kill-Switch

The non-obvious bet. Everyone races to make agents *more* autonomous; the “Off Switch” story says the contrarian value is the provable, instant brake. Anthropic killed Fable globally in 90 minutes because control was centralized—and that control is becoming a product requirement. Build, or back, the unglamorous infrastructure of containment: per-action kill switches, capability attestations, control-ratio monitoring, circuit-breaker tripwires. The market prices autonomy; regulation will pay for reversibility. The winner may be whoever sells the brakes, not the engine.

Compute Is the New Oil—Margin Lives in the Refinery

Brockman says the most-compute lab wins and OpenAI raised \$122B to prove it. Yet one approach claims 33x faster inference and 90% fewer tokens by keeping decisions off the expensive model, and another claims to break attention's quadratic cost. Headlines reward scale; the P&L is won at inference. Assume frontier capability is rented and cheap, and compete on a routing-and-caching architecture that keeps 90% of requests off the frontier model—no war chest required.