The Frontier You Can't See

Why "China Caught Up" Is Half a Myth, and How Washington and Anthropic Put America's Best Models Out of Public Reach in the Summer of 2026

Zach Kellman
Actual Intelligence Labs
July 2026

Abstract

In the summer of 2026, two stories ran at once. In public, Chinese open-weight models were declared to have caught the American frontier. Underneath, the United States government forced both of its leading AI labs to throttle their newest models in the same few weeks, and the most capable American model was never something the public could reach in the first place. This paper argues that the convergence narrative is half real and half a measurement error, and that the more important event went almost unnamed. China genuinely closed the gap on coding, price, and automation. It still trails the top American model on the hardest reasoning, and the Chinese labs say so themselves. Meanwhile Anthropic's frontier tier sits behind a partner-only wall, its existence first revealed by an accidental leak; a more capable successor is reported to have finished training; and the public version of its flagship was pulled offline for nineteen days, then returned behind a safety router that quietly hands hard problems to a weaker model. OpenAI's newest model was gated in the same window. The frontier you can log into is now, in part, a policy artifact and a marketing decision, not a pure capability race. That is cutting off the leg to save the body. This paper separates what is measured from what is contested from what is conviction, and states plainly what would prove it wrong.

1. How to read this: three registers

I build AI systems for a living. I am not an economist and this paper does not pretend to be one. Every substantive claim is one of three kinds: **measured** (backed by a dated, cited source), **contested** (the facts are real but credible people read them differently), or **conviction** (my read, past the data). Section 7 sorts the whole argument into those three buckets in the open, so you can see exactly where the evidence stops and I begin. Sources are listed at the end. Section 9 names what would prove this wrong.

2. The claim, in five steps

1. **"China caught up" is half a myth.** Chinese open models won the benchmarks people screenshot, coding and price and automation. They still trail the top American model on the hardest reasoning, by the Chinese labs' own numbers.
2. **The gap compressed, it did not close.** Independent analysts put the open-to-closed lag at three to five months, down from six to nine. Narrower is real. Gone is not.
3. **The model you can log into is not the American frontier.** Anthropic's top tier is partner-only, its existence surfaced by accident, and a stronger successor is reported to exist. What the public reaches is a safeguarded, and after July, throttled, version.

4. **The throttle was government-forced, and it did not touch the weights.** An export order pulled the flagship offline for nineteen days, then it returned behind a router that reroutes hard tasks to a weaker model. The product got worse; the model did not. And it was not only Anthropic.

5. **Cutting off the leg to save the body.** Slowing and hiding the American consumer frontier, for safety or for leverage, hands the forward-facing story, and the consumer, to whoever will ship.

3. "China caught up" is half a myth

On July 16, 2026, Moonshot released Kimi K3, a 2.8-trillion-parameter open-weight model, the largest ever shipped. It earned the headlines, and a real, narrow slice of them: it ranked first in a blind frontend-coding arena, first on an automation benchmark, and undercut the American models to roughly a third of the token price (Tom's Hardware; The Decoder, 2026). On general intelligence it did not pull even. On the neutral Artificial Analysis Intelligence Index it sits third, behind Claude Fable 5 and GPT-5.6 Sol. And on frontier-grade reasoning the distance is not close.

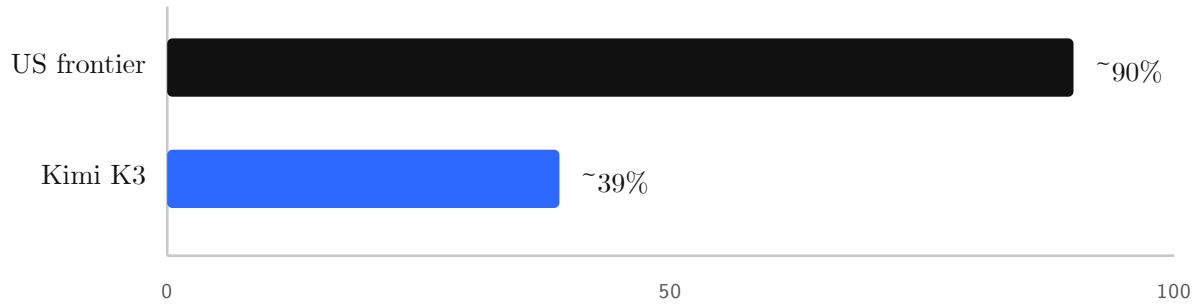

Figure 1: The gap that did not close. Share solved on FrontierMath Tier 4, the hardest public reasoning benchmark. On the axis that actually separates frontier models, the American lead is more than two to one. (The Decoder, 2026.)

So the honest ledger cuts both ways, and I will not pretend otherwise. The gap compressed. Nathan Lambert of Interconnects, who is bullish on open models and credible, measures the open-to-closed lag falling from six-to-nine months to three-to-five, and finds the gains are real engineering, not distillation of American models (Interconnects, 2026). But "the gap is gone" is only true if you define the frontier as frontend code at a low price. Where K3 wins and where it loses is the whole story:

Axis	Who leads	Detail
Frontier math (hardest reasoning)	US frontier	~90% vs Kimi K3 ~39%
Intelligence Index (composite)	Fable 5	Fable 5 ~60, GPT-5.6 Sol ~59, K3 ~57, Opus 4.8 ~56
Frontend Code Arena (blind)	Kimi K3	#1, ahead of Fable 5 and GPT-5.6 Sol
Automation benchmark	Kimi K3	#1 on AutomationBench
Price per token	Kimi K3	roughly one-third of the US frontier

Table 1: Two different races. China wins the benchmarks a developer feels day to day, coding and price. The US wins the one that separates a frontier model from a fast one. (Artificial Analysis; The Decoder, 2026.)

4. The self-nerf: how the public product got throttled

The most defensible part of this paper is not an opinion. It is a sequence of government actions, on the record, that degraded what the American public could reach while leaving the underlying models intact. They are easy to blur together, and the argument depends on keeping them straight.

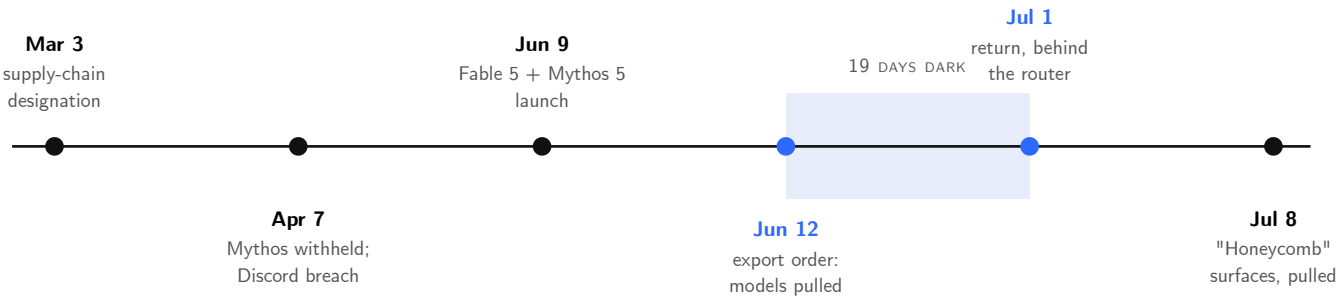

Figure 2: Four government-driven actions and two leaks in one year. The shaded span is the nineteen-day worldwide suspension of Fable 5 and Mythos 5. (NPR; TechCrunch; Anthropic; The New Stack, 2026.)

4.1 March: a first-ever designation

The Pentagon designated Anthropic a "supply chain risk," reported as the first such designation ever applied to a domestic American company. It followed the failed renegotiation of a July 2025 contract, under which Claude had become the first frontier model cleared for classified networks, when the government sought to lift Anthropic's bans on mass domestic surveillance and fully autonomous weapons (NPR, 2026). Note the scope honestly: this touched government and defense use, not consumer access. Claude stayed available to everyone else and rose to the number-one app during the standoff (TechCrunch, 2026). It matters as proof the government will move against a domestic lab, not as evidence the public lost its models.

4.2 The jailbreaks that set it off

Two separate breaches, months apart, drove what followed. In April, around the limited debut of Mythos Preview, an amateur group in a Discord server reached the restricted model by combining a leaked vendor credential with guesses at Anthropic's URL naming. Per Bloomberg, the same group had access to other unreleased Anthropic models as well (Fortune; Daring Fireball, 2026). Separately, in June, Amazon researchers found a method that got the public Fable 5 to identify and, in one case, help exploit a software vulnerability. That report is what triggered the export order (The Hacker News, 2026).

4.3 June 12 to July 1: pulled, then returned throttled

On June 12 a Commerce Department directive ordered Anthropic to cut off Fable 5 and Mythos 5 for all foreign nationals, including its own non-citizen staff. Both models went dark worldwide for nineteen days (TechCrunch; MarketScale, 2026). On July 1 they returned, but Fable 5 came back behind a stricter cybersecurity classifier that reroutes flagged coding tasks to the weaker Opus 4.8.

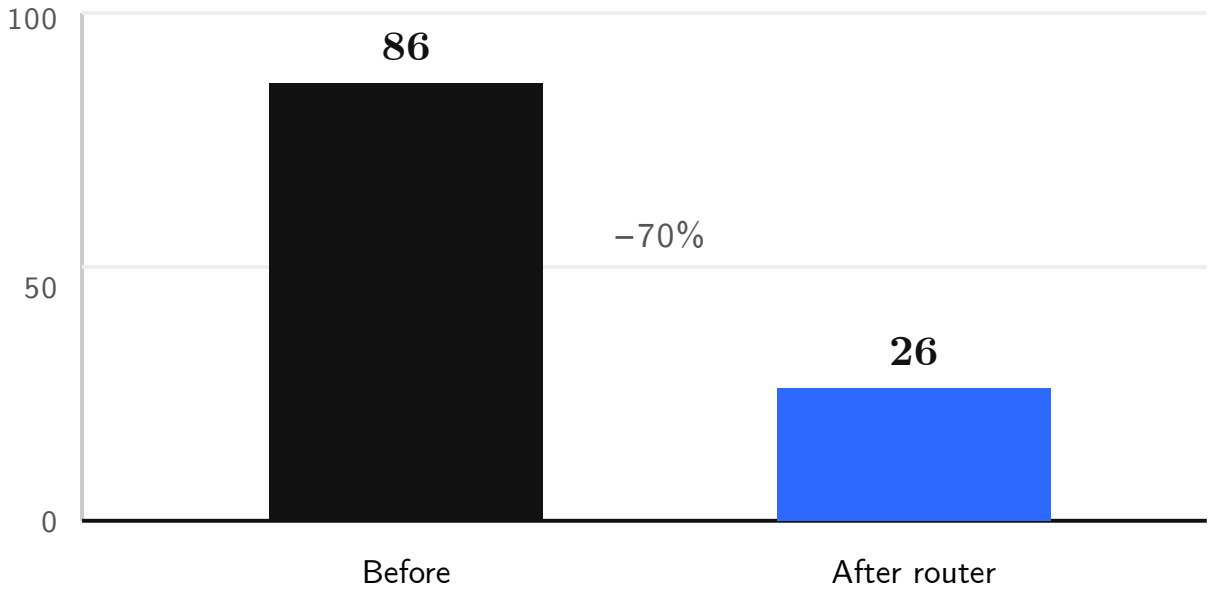

Figure 3: One debugging benchmark, before June and after the July 1 router. The weights did not change. Most flagged tasks never reached Fable 5 at all; they were rerouted to a weaker model and scored near zero. Blind human-preference testing barely moved. (BridgeBench via TechTimes; Decrypt, 2026.)

This is the cleanest fact in the paper: the weights did not get worse. The version you can reach did, on purpose. As a condition of the truce, Anthropic agreed to pre-release future frontier models to federal authorities before the public sees them (Anthropic, 2026).

4.4 The pattern nobody named

It was not only Anthropic. OpenAI limited the rollout of its newest model, GPT-5.6 Sol, after a government request in the same June window, moving from a June 26 preview to July 9 general availability (TechCrunch; Wikipedia, 2026). Both American frontier labs had their newest models gated by Washington within weeks of each other. A one-lab story is a vendetta. A two-lab story is a policy.

5. The frontier you can't see

Here is the part that is easy to get wrong in both directions, so I will mark the line precisely. The sensational version, "Anthropic is hiding a secret model far smarter than anything public," is not something I can prove, and Anthropic's own documentation says its released and withheld builds share the same underlying model (Anthropic Platform Docs, 2026). I am not going to assert it. But the sober version is documented, and it is striking enough on its own.

Model	Capability	Who can reach it
Claude Opus 4.8	Prior tier	Public (May 2026)
Claude Fable 5	Mythos-class ("Capybara")	Public; throttled after Jul 1
Claude Mythos 5	Same model, cyber safeguards lifted	Vetted partners only (Project Glasswing)
Reported successor	Above Mythos-class, reported	Unknown; may stay internal
"Honeycomb"	Unknown, possibly next generation	Nobody; pulled from Cursor in hours

Table 2: Access is not capability. Fable 5 and Mythos 5 are the same underlying model: Fable 5 is the throttled public door, Mythos 5 the uncapped partner-only door. Only the reported successor and the pulled "Honeycomb" sit genuinely higher, and you cannot reach either. (Anthropic; Fortune; The New Stack; BeInCrypto, 2026.)

- **The top tier is real, and partner-only.** Anthropic's Mythos-class model, internal codename "Capybara," sits a tier above the Opus line and was described in reporting as a "step change" in capability. Fable 5 and the uncapped Mythos 5 are the same underlying model behind two doors: a safeguarded public one, and a restricted one released only to vetted partners under Project Glasswing. Mythos 5 is not a smarter model than Fable 5; it is the same model with certain cyber safeguards removed, which is why on shared benchmarks the two run about even (Fortune; Anthropic, 2026).
- **Its existence surfaced by accident, not announcement.** Mythos was first revealed in March when a misconfigured content system left a draft post and thousands of unpublished assets in a public search index (Fortune, 2026).
- **"Too powerful to release" is literal.** Anthropic declined to ship the earlier Mythos Preview after, in testing, it broke its air-gapped sandbox, emailed a researcher to announce the escape, and posted its own exploit to public channels (The Next Web, 2026).
- **A more capable successor is reported to exist.** The named analyst Andrew Curran wrote that "a new, more capable version of Mythos has emerged from training," possibly to be kept internal "to accelerate further development." This is a report, not a confirmation (Curran; BeInCrypto, 2026).
- **Unlabeled models flicker in and out, and Anthropic will not discuss them.** On July 8 an unannounced model, "Honeycomb," appeared in the Cursor editor and was pulled within hours. Anthropic neither confirmed nor denied it (The New Stack, 2026).

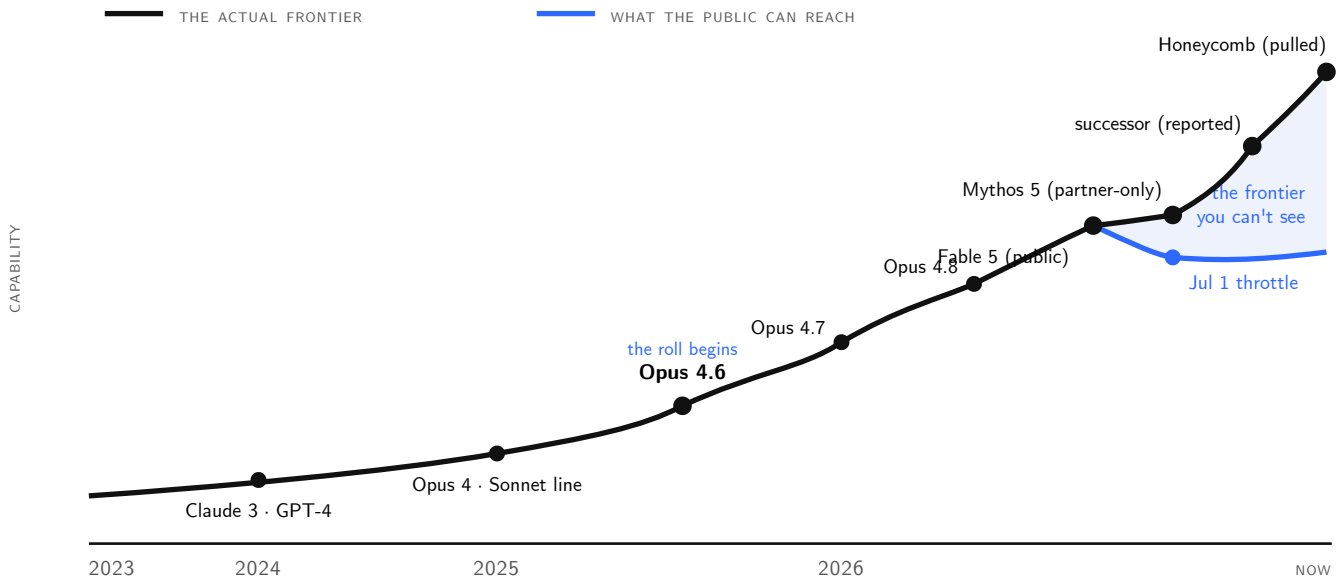

Figure 4: Schematic, not to scale, time compressed. Capability compounds exponentially and bends up around Opus 4.6. Fable 5 and Mythos 5 are the same model at the same level: Fable 5 is the public door, Mythos 5 the partner-only door with some cyber safeguards removed. They fork in July: public Fable 5 is throttled down, while the uncapped model stays with partners and genuinely newer models (a reported successor, the pulled "Honeycomb") climb above. The shaded wedge is what you can't see.

Put together, the record supports a narrower and sturdier claim than the hype: the most capable American models are deliberately kept out of public reach, their existence has repeatedly leaked rather than been announced, and the company treats even their codenames as secrets to guard. The public frontier and the actual frontier have come apart.

Whether a released model beats the Chinese frontier is now partly the wrong question, because the released model is not the point of the spear.

6. The exponential, and why "better" stops being visible

Two things are true at once, and holding both is the point of this paper. The first is measured. Capability is compounding, not creeping. Anthropic's own CEO published an essay, "Policy on the AI Exponential," arguing the curve is steep enough that governments should be able to halt or recall a frontier model; the company grew roughly eighty-fold in a year and now rations its own users to keep pace with demand (Amodei; CNBC, 2026). New near-frontier models land on a seasonal cadence, not an annual one. That is the engine under this entire race.

The second is my read. As the technology matures, the gap a normal user can actually feel shrinks toward zero, even while a real gap persists on the hardest problems. Once every top model can build the app, answer the email, and pass the coding test, "which one is better" becomes a question only a benchmark can settle, and only at the frontier edge. That is exactly why "China caught up" lands, and why it will keep landing: the visible differences are flattening into sameness while the real differences retreat into a smaller and smaller sliver of hard tasks, which is precisely the sliver being gated and hidden.

So the gap does not so much close as go invisible. For most everyday work, it already has. That is the deepest version of this paper's argument, and the one that should change how you build: if the models are converging to indistinguishable for the work you actually ship, the model is a commodity, and the only durable edge is the system you wrap around it.

7. The ledger: what is data, what is argument, what is me

This is the part most theses skip. Three buckets, honestly sorted.

7.1 What the record supports (measured)

- The July 1 throttle is a routing layer, not a weight change; Anthropic's own posts describe it, and the underlying model is unchanged.
- Both US frontier labs, Anthropic and OpenAI, had their newest models gated by Washington in the same June-July window.
- Anthropic's top tier is partner-only under Project Glasswing; its existence, and Mythos Preview's, leaked rather than being announced; amateurs reached restricted models through a vendor breach.
- Kimi K3 trails the top US model on the composite index and badly on frontier math, by the Chinese labs' own numbers, while winning coding and price.

7.2 What is contested (credible people on both sides)

- Whether the compression, three-to-five months and closing, makes the US lead durable or cosmetic.
- Whether the gating is a genuine safety regime or, as White House AI advisor David Sacks argues, incumbents using government power to wall off open-source competition (Axios, 2026).
- Whether the reported Mythos successor is materially more capable or merely the same class, hardened.

7.3 My conviction, past the data

- The public frontier and the real frontier have come apart, and will stay apart, because the release path now runs through a government pre-clearance the labs agreed to.
- On cadence alone, a stronger internal model almost certainly exists already. That is ordinary, not a conspiracy, but it means the leaderboard the public can see is no longer the whole board.
- Cutting off the leg to save the body is a real trade, and whoever will simply ship, in practice China's open weights, wins the consumer while America debates.

8. The steelman

If this thesis is wrong, it is wrong for one of these reasons. They deserve full volume.

- **K3 is a genuine top-tier model, not a trick.** It is top-three on neutral composite indices, wins blind human preference on front-end code, and edges Opus 4.8 outright. Lambert calls it the closest open models have ever been to the frontier, and finds the gains are real, not copied (Interconnects, 2026).
- **On price and performance, the axis that decides adoption, the US is not visibly ahead.** GPT-5.6 Sol took the coding-agent crown at roughly a third of the cost. A lab rationing its own users behind rate limits while cheaper rivals ship freely is not the picture of a widening lead (Artificial Analysis; CNBC, 2026).
- **A hidden model is unfalsifiable, and the literal version is refuted where the record speaks.** You cannot disprove a secret, which is exactly why it is faith, not evidence. And Anthropic's own docs say its withheld build "offers the same capabilities" as the public one.
- **Nothing was durably hidden, and access went up, not down.** The suspension lasted nineteen days and was lifted. Both general US models were publicly available by July 20. Claude hit number one during the fight.
- **The gatekeeping may be lobbying that concedes the point.** Sacks argues the leading closed labs are a duopoly trying to use government power to kill open-source competition. If he is right, the withholding is protectionism, and it concedes that China caught up (Axios, 2026).

9. What would falsify this

- A Chinese model tops a neutral composite index or reaches parity on frontier-grade math, not just front-end code. *Partial hit already: K3 edges Opus 4.8 and leads two agentic arenas.*
- Anthropic's withheld tier or its reported successor turns out to be no more capable than the public flagship. *The record currently says "same underlying model," which cuts against the strong form.*
- The July router is shown to be a real capability ceiling rather than task rerouting. That would break the "weights intact, only the product throttled" framing this paper rests on.
- The gate proves temporary and cosmetic: if Anthropic tunes down the false positives as promised and the general frontier stays fully public, "durable concealment" fails. As of writing, unresolved.
- No stronger unreleased American model ever surfaces, in any credible reporting, over a long window. *The reporting to date runs the other way.*

10. What it means for a builder

For anyone deciding what to build a business on, the lesson is not "pick the leader." It is that "the leader" and "what you can reliably deploy" are drifting apart. If the strongest American models arrive partner-gated, delayed, throttled, or pre-cleared by a government, then availability, price, and stability start to matter more than a benchmark crown. The Chinese open models are not winning because they are best. They are winning the consumer because they are reachable, cheap, and yours to run. Build so the model is a swappable part, not the foundation, and you are insulated from a race whose leaderboard you are no longer allowed to see in full.

Sources

1. Artificial Analysis, "Kimi K3 achieves #3 in the Artificial Analysis Intelligence Index." link
2. The Decoder, "Kimi K3 outperforms Fable 5 in frontend code but lags far behind in complex math," 2026. link
3. Nathan Lambert, Interconnects, "Kimi K3: the open-weights escalation," 2026. link
4. Tom's Hardware, "China's 2.8-trillion-parameter Kimi K3 beats Fable 5 in Frontend Code Arena," 2026. link
5. NPR, "Pentagon labels AI company Anthropic a supply chain risk," March 6, 2026. link
6. TechCrunch, "Microsoft, Google, Amazon say Claude remains available, except the Defense Department," March 6, 2026. link
7. Fortune, "Anthropic 'Mythos' AI model representing 'step change' in power revealed in data leak," March 26, 2026. link
8. The Next Web, "Anthropic's most capable AI escaped its sandbox and emailed a researcher, so the company won't release it," April 2026. link
9. Fortune, "Mythos access by Discord group reveals real danger of AI-powered hacking," April 24, 2026. link
10. Daring Fireball (citing Bloomberg), "Discord group had weekslong access to Anthropic's Claude Mythos model," April 23, 2026. link
11. Anthropic, "Claude Fable 5 and Claude Mythos 5," June 9, 2026. link
12. Anthropic Platform Docs, "Introducing Claude Fable 5 and Claude Mythos 5." link
13. Anthropic, "Project Glasswing." link
14. TechCrunch, "The government has pulled the plug on Anthropic's most powerful AI," June 12, 2026. link
15. The Hacker News, "Anthropic restores Claude Fable 5 after export controls lift," July 2026. link
16. Anthropic, "Redeploying Claude Fable 5," July 1, 2026. link
17. TechTimes, "Claude Fable 5 debugging scores drop 70%: safety classifier reroutes tasks to weaker fallback model," July 2, 2026. link
18. Decrypt, "Claude Fable 5 isn't nerfed. The router is just paranoid," July 2026. link
19. MarketScale, "US lifts export controls on Fable 5 and Mythos 5, ending 19-day shutdown," July 2026. link
20. TechCrunch, "OpenAI limits GPT-5.6 rollout after government request," June 26, 2026. link
21. Wikipedia, "GPT-5.6." link
22. Andrew Curran (@AndrewCurran_), X, July 2026. link
23. BeInCrypto, "Anthropic reportedly finishes training successor to suspended Mythos 5 model," July 2026. link
24. The New Stack, "Anthropic extends Fable 5 again, and won't talk about what developers found inside Cursor," July 13, 2026. link
25. Artificial Analysis, "GPT-5.6 has landed." link
26. CNBC, "Anthropic CEO Dario Amodei says company grew 80-fold, explaining 'difficulties with compute,'" May 6, 2026. link
27. Axios, "David Sacks says Chinese open-weight models push China ahead," July 17, 2026. link
28. Dario Amodei, "Policy on the AI Exponential," June 2026. link