

The Great Model Dump: Cheap Chinese Intelligence, Expensive American Assumptions, and Which One Breaks First

Zach Kellman

Actual Intelligence Labs

July 10, 2026

Abstract

Chinese labs are flooding the market with open-weight models that sit months behind the US frontier and cost **85 to 97 percent less** to run. Trillions in market value are priced on the assumption that frontier intelligence stays a scarce, high-margin product. Both cannot stay true. When the margin assumption breaks, the valuations built on it reprice, and in the most concentrated market in five decades, the damage does not stay contained to AI.

1 How to read this: three registers

I build AI systems for a living. I am not an economist, and this doc does not pretend otherwise. Every claim is tagged: **measured** (sourced data), **contested** (smart people disagree), or **conviction** (my read, past the data). Sources on the last page. Falsifiers in their own section.

2 The argument: five steps, start to finish

1. **The product converged.** Chinese open weights match or nearly match US frontier models on the benchmarks that drive real spend (coding, agents). Measured lag: 4 to 8 months, stable (Epoch AI; NIST CAISI).
2. **Where switching is free, the flip already happened.** US models fell from ~70% to ~30% of token traffic on OpenRouter in twelve months (Jun 2025 to Jun 2026).
3. **Record money is priced on the opposite outcome.** ~\$700B of 2026 hyperscaler capex, Nvidia at 8% of the S&P 500, \$1.8T of OpenAI plus Anthropic valuations. All of it needs premium inference economics.
4. **The dump compresses exactly those economics.** OpenAI was reported weighing "drastic" token price cuts in June 2026, Anthropic expected to follow. You do not win a price war against lower real costs.

5. **Concentration turns a sector repricing into a market event.** AI-linked names are ~45% of S&P 500 market cap. If the margin story cracks, the index has nowhere to hide.

3 The dump is real

"Dump" is the right word. Not because the models are junk. Because the pricing behaves like commodity dumping: a near-equivalent product, sold at a fraction of incumbent prices, released on a relentless cadence, at national scale.

Value	Metric
~9×	Cheaper, same workload: GLM vs Claude (Reuters, Jul 2026)
70→30	US share of OpenRouter tokens (%), Jun 2025 to Jun 2026
4 mo	Open-weight lag behind the closed frontier (Epoch AI)
\$1.8T	Combined OpenAI + Anthropic private valuations (2026)

Table 1: Headline indicators of the dump (2026).

- **The players.** DeepSeek (V4, MIT license), Alibaba's Qwen (Apache 2.0), Moonshot's Kimi K2 line, and Zhipu's GLM-5 line. Eight of the top ten Chinese models are open-weight and commercially usable, with the licensing questions resolved (industry trackers, 2026).
- **The cadence.** DeepSeek V4 (Apr 2026), Kimi K2.6, Qwen 3.6, GLM-5.2 (Jul 2026). New near-frontier releases land quarterly, not annually (TechCrunch, Apr 2026; Reuters, Jul 2026).
- **The prices.** Output tokens from DeepSeek V4-Pro cost \$0.87 per million against \$15 to \$50 across the US frontier lineup. That is 17x to 57x cheaper depending on the pairing.
- **And it may not even be dumping below cost.** DeepSeek has claimed its inference is profitable at these prices (it published a theoretical 545% daily cost-profit ratio; Reuters, Mar 2025). If the efficiency is real, that is worse for the incumbents, not better. You can out-lawyer a subsidy. You cannot out-lawyer someone whose product genuinely costs less to make.

4 One workload, five bills

A Reuters-reported comparison priced the same AI workload across providers. Same job. Roughly a 9x spread between the most and least expensive (Reuters analysis via industry coverage, Jun to Jul 2026).

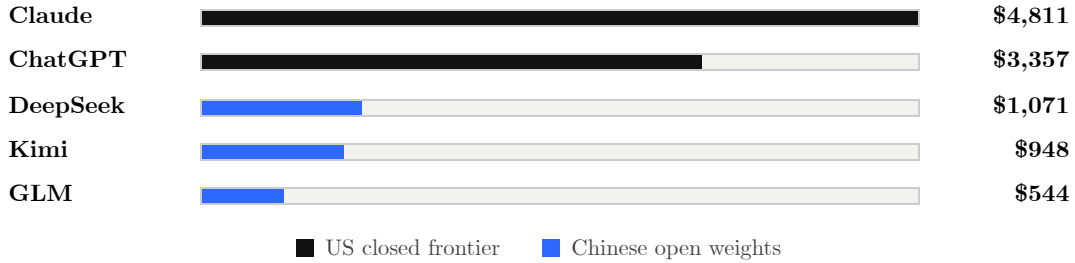


Figure 1: One workload, five bills. Reuters analysis via industry coverage, Jun to Jul 2026. Claude and ChatGPT are US closed frontier; DeepSeek, Kimi, and GLM are Chinese open weights.

4.1 The price board, July 2026

Model	Weights	Benchmark signal	Input \$/M	Output \$/M
Claude Fable 5 (Anthropic)	Closed	Top of the boards (Jun 2026)	\$10.00	\$50.00
GPT-5.6 Sol (OpenAI)	Closed	Frontier tier	\$5.00	\$30.00
Claude Opus 4.7 (Anthropic)	Closed	SWE-bench Verified ~80.8	\$5.00	\$25.00
Claude Sonnet 4.6 (Anthropic)	Closed	Workhorse tier	\$3.00	\$15.00
DeepSeek V4-Pro	Open (MIT)	SWE-bench Verified 80.6	\$0.435	\$0.87
Qwen 3.6-35B (Alibaba)	Open (Apache 2.0)	Near-frontier agentic coding	~\$0.38 blended	~\$0.38 blended
DeepSeek V4-Flash	Open (MIT)	Fast tier	\$0.14	\$0.28

Table 2: The price board, July 2026. Pricing: provider pages and trackers, June to July 2026 (Developers Digest; aipricing.guru; Morph; NxCode). Benchmarks: BenchLM, 2026.

5 The flip: where switching is free, it already happened

On OpenRouter, a neutral router where a developer swaps models with one line of code, US closed models fell from ~70% to ~30% of tokens in a year, while Chinese open weights climbed toward half. This is the cleanest window we have into what buyers do when nothing but price and quality is holding them.

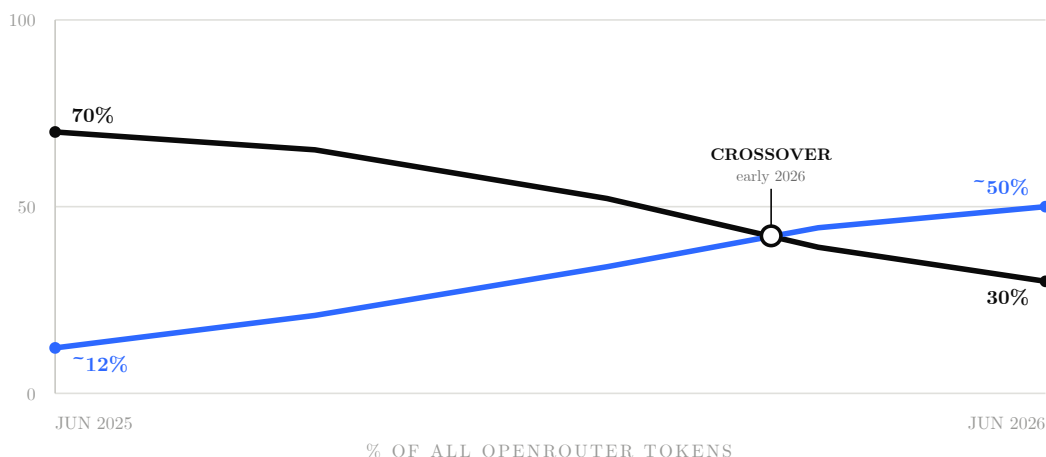


Figure 2: The great flip: OpenRouter token share, Jun 2025 to Jun 2026. Chart: OpenRouter token-share data, Jun 2025 to Jun 2026 (OfficeChai; Tech Times; Dataconomy). Lines are share of tokens routed, not revenue. White line: US closed frontier. Blue line: Chinese open weights.

- Chinese open-weight models went from under 2% of that traffic in late 2024 to roughly half of all platform tokens by spring 2026 (Tech Times, May 2026; Dataconomy, Feb 2026).
- Coding grew to over half of all router tokens, and coding is exactly where the Chinese models are strongest (OpenRouter, 2026).
- **Honest caveat:** router traffic skews to price-sensitive developers and startups, not enterprise contracts. It is a leading indicator, not the whole market. Whether it leads or misleads is the entire fight. See the steelman.

6 Convergence: months, not years

The capability gap is real. It is also small, stable, and increasingly irrelevant to the work people actually pay for.

- NIST's Center for AI Standards and Innovation evaluated DeepSeek V4 and put its capabilities about 8 months behind the frontier, on a development trajectory trailing state of the art by roughly 3 to 6 months (NIST CAISI, May 2026).
- Epoch AI: the best open-weight models trail the closed frontier by an average of 4 months, or 8 points on their capability index, measured January through May 2026. That is a slight widening from 3 months in late 2025. Months either way, not years.
- On the highest-spend benchmark category: DeepSeek V4-Pro scores 80.6 on SWE-bench Verified against Claude's 80.8, and posts the top LiveCodeBench score among all evaluated models at 93.5 (BenchLM; DataCamp, 2026). GLM-5.1 took the top of SWE-Bench Pro ahead of GPT-5.4 and Claude Opus 4.6 (industry coverage, 2026).
- The absolute frontier is still American. Anthropic's Fable 5 (June 2026) tops the boards, at \$10 in and \$50 out. The top sells prestige and the hardest problems. The tier below it is where the volume revenue lives, and that tier now has a competitor at 1/29th the price.

6.1 The analogy that fits

Switzerland kept making the best watches through the 1970s. Seiko still ended the Swiss watch industry's margin structure, because quartz was good enough, cheap, and everywhere. Capability prestige and margin structure are different assets. The quartz crisis did not care who made the finest movement. Neither does a procurement spreadsheet.

7 The money priced on the old world

Against that convergence, the market is carrying the most concentrated technology bet in the history of US equities.

Value	Metric
~\$700B	2026 hyperscaler capex guidance, up ~75% YoY (est. \$690B to \$725B)
45%	Share of S&P 500 market cap in AI-linked firms (2026)
8%	Nvidia alone, the largest single-stock index weight on record
\$965B	Anthropic valuation on ~\$47B run rate (CNBC, May 2026)
-\$14B	OpenAI's projected 2026 loss, at an \$852B valuation
\$1.15T	OpenAI's decade of infrastructure commitments across 7 vendors

Table 3: The money priced on the old world: key market figures (2026).

- **Capex tripled in two years:** ~\$226B (2024) to ~\$410B (2025) to ~\$700B guided (2026), with \$1T+ forecast for 2027 (CNBC, Feb 2026; Tom's Hardware, 2026). Amazon ~\$200B, Alphabet ~\$175B to \$185B, Meta \$125B to \$145B, Microsoft \$120B+.
- **The financing is circular.** Chip maker invests in lab, lab commits spend to cloud, cloud buys chips: over \$800B of these loops, which the IMF warns can inflate both revenues and valuations (Bloomberg, 2026; IMF via press, 2026).
- **Expectations passed the dot-com peak.** Long-term S&P 500 earnings growth expectations hit 20.2%, above the 18.6% record set in 2000 (Acadian's Owen Lamont, Fortune, Jun 8, 2026).
- **Both labs are heading to public markets on these numbers.** The IPO window is precisely when margin durability gets stress-tested in public, every quarter, forever.

7.1 The core mismatch

*Nvidia trades around 25x trailing revenue (\$5.4T on \$215.9B of FY2026 sales). Anthropic raised at ~20x run rate, OpenAI at ~30x. Meanwhile the underlying commodity, a token of fixed-capability intelligence, deflates at a median of roughly **50x per year** (Epoch AI, 2026). Very few businesses in history have defended premium pricing against a 50x annual deflator. None of them did it while a state-scale competitor gave the product away.*

8 How a dump becomes a pop

A bubble does not need bad models to pop. It needs one broken assumption. Here is the chain, step by step.

1. **The price war reaches the API.** Already started: OpenAI reportedly weighing drastic token price cuts, Anthropic expected to follow (industry reporting, Jun 2026). Fable 5 launching at \$10/\$50 shows the counter-move too: retreat up-market and defend the premium tier while the floor drops out of the middle.
2. **Margins compress at the worst possible moment.** Reports already conflict on Anthropic's true gross margin: roughly 40% after inference cost overruns by one account, ~70% by another (2026 reporting). A price war settles that argument in the wrong direction, right as both labs court public investors.
3. **The pain is asymmetric.** A token price war hits the API-heavy lab hardest: ~85% of Anthropic's revenue is enterprise and developer usage, while ~85% of OpenAI's is consumer subscriptions (Forbes, May 2026). Subscriptions have brand moats. Raw tokens do not.
4. **Capex confidence cracks.** The ~\$700B build is justified by projected inference revenue at healthy prices. If the market clears at DeepSeek prices, the same GPUs earn a fraction of the modeled revenue. The data centers do not need to sit idle to be impaired. They just need to earn commodity rates on premium-priced hardware.
5. **The index has no shock absorber.** With AI-linked names at ~45% of the S&P 500 and Nvidia at 8% by itself, an AI repricing is a market repricing, mechanically.
6. **Whether that becomes a recession** depends on wealth effects, credit, and policy, which is genuinely beyond my lane. For the record, Capital Economics projects the S&P 500 reaching ~8,250 by end-2026 and then sliding ~21% to 6,500 through 2027. Nobody's crystal ball is clean, including theirs.

8.1 The precedent, twice

Jan 27, 2025: *one Chinese model release (DeepSeek R1) erased ~\$600B of Nvidia's market cap in a day, a 17% drop and the largest single-day loss in market history (CBS News, Jan 2025).* **Jun 24, 2026:** *the Nasdaq fell 2.2% in the second straight AI-led selloff, as investors began demanding proof of returns on capex (market coverage, Jun 2026). The market has now rehearsed this exact scare twice. The thesis says one of the rehearsals eventually becomes the show.*

9 The rhyme: same shape, twice

History does not repeat, but it rhymes. The dot-com bubble and the AI cycle trace the same arc: a real, capability-driven mania, priced for a future that arrives more slowly and far more cheaply than the valuations assume.

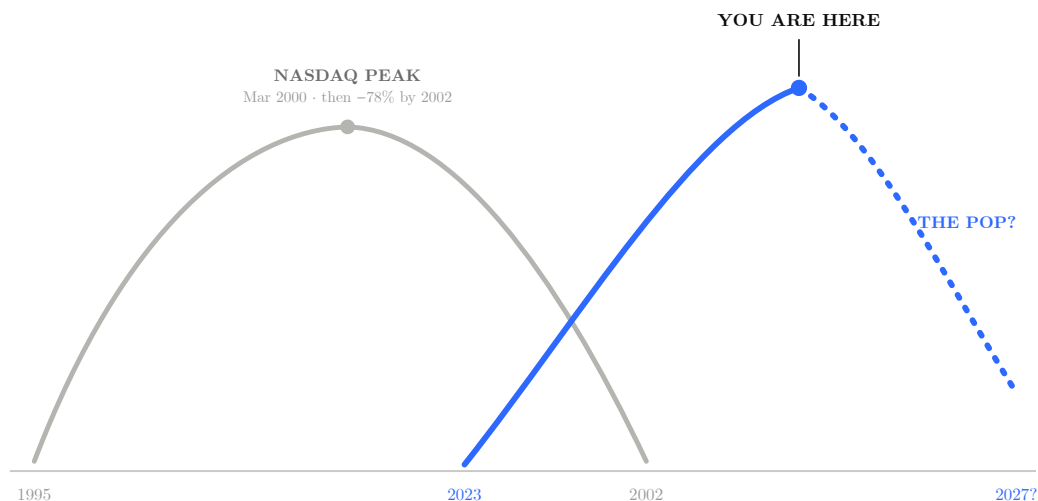


Figure 3: Two bubbles, one curve: market-cap index, schematic. Schematic, not to scale. The shape is the argument, not the levels. The solid blue is history; the dashed blue descent is conviction, not a forecast. Grey: dot-com bubble, 1995 to 2002. Blue: AI cycle, 2023 to now.

The 2000 crash did not happen because the internet was fake. It happened because the market priced a decade of growth into eighteen months, then met reality. The AI cycle is running the same play at a larger scale, on a narrower set of names, with one difference that has no dot-com analog: the deflationary shock, cheap open weights, is already here and still accelerating. The mania and its own antidote are arriving at the same time.

10 The ledger: what is data, what is argument, what is me

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

10.1 A. What the data supports (measured)

- Chinese open-weight models are 85 to 97 percent cheaper per token than US frontier APIs, and roughly 9x cheaper on real workloads (Reuters; pricing trackers, 2026).
- They have converged on the highest-spend benchmarks: within a point on SWE-bench Verified, ahead on LiveCodeBench and SWE-Bench Pro (BenchLM; NIST CAISI, 2026).
- The frontier lag is measured in months (4 to 8), and it has been stable for 18+ months, not widening (Epoch AI; OpenRouter, 2026).
- Usage flipped fast wherever switching costs are zero: US token share on OpenRouter fell 70% to 30% in a year.
- Capex, concentration, and private valuations sit at all-time records, and price-cut pressure on US labs is already reported (Fortune; CNBC; industry reporting, 2026).

10.2 B. What is contested (smart money on both sides)

- **Jevons.** Cheaper tokens may grow total spend instead of shrinking it. Token expenditure roughly doubled since late 2025 even as per-token prices fell ~90% from 2023 (Apollo, Torsten Slok, Jun 2026). Bulls read the dump as fuel, not poison.
- **Enterprise trust.** Government bans (Australia, Italy, Czech Republic, several US states, a proposed US federal ban) and incidents like Germany's BSI finding a DeepSeek pilot leaking classified metadata to Shanghai (Mar 2026) may wall Chinese weights out of the revenue that matters.
- **True lab margins.** 40% vs 70% gross margin reports for Anthropic; OpenAI and Anthropic do not even count revenue the same way (Forbes, Mar 2026).
- **The chip ceiling.** ~75% of Chinese training compute still runs on Nvidia. Export policy could still throttle the cadence (CSIS, 2026), though Washington loosened H200 sales in Jan 2026 and Huawei plans 600K Ascend chips this year.
- **Benchmark truth.** Epoch cautions that open models overperform public benchmarks and closed labs hold back their best. The real gap may be wider than measured.

10.3 C. My conviction, past the data

- The dump is strategic, not incidental, and it accelerates into the US lab IPO window. Cheap intelligence is an economic weapon whether or not anyone says so out loud.
- Open weights win the platform layer the way Linux won servers: configurability plus price beats polish at scale. You can fine-tune, self-host, and rewire an open model. You rent a closed one.
- Margin compression is the pin. I believe the AI trade reprices hard within 12 to 24 months, and the drawdown bleeds into the broader market. A recession is plausible. I hold the recession call loosely: I am an operator reading supply curves, not an economist.

11 The steelman: what the bulls have

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

- **Jevons paradox is undefeated.** Every collapse in the price of compute in history expanded total spend. Enterprise GenAI spend grew ~22x in two years while token prices fell 90%+; inference is on track to be two-thirds of all AI compute in 2026 (Apollo; Deloitte, 2026). If demand grows faster than price falls, frontier labs grow straight through the dump. Right now the spend data favors the bulls.
- **Enterprise gravity favors incumbents.** 500+ companies pay Anthropic over \$1M a year; eight of the Fortune 10 are customers (Forbes, May 2026). Enterprises buy trust, support, liability coverage, and integration, not tokens. Chinese-origin weights carry procurement politics that no price cut fixes. The nuance: open weights can run on US soil via American hosts, which defuses data residency, but not the politics.

- **The top of the frontier still has pricing power.** Fable 5 launched at double Opus pricing and tops the boards. If frontier capability compounds into agents that do real economic work, the premium tier may hold its prices even while the mid-market commoditizes. Barbell, not collapse.
- **The dump runs on American silicon.** Chinese labs still train predominantly on Nvidia hardware, and one export rule can change the physics (CSIS; CFR, 2026). Counterpoint: GLM-5 was reportedly trained entirely on Huawei Ascend, so the workaround is arriving.
- **A selloff is not a verdict.** Markets can fall for boring reasons (BofA and Deutsche Bank now expect Fed rate hikes in late 2026), and they can stay irrational for years. Timing a pop is a different skill from diagnosing one, and I claim only the second.

12 What would falsify this thesis

A thesis that cannot name its own kill conditions is a mood, not an argument. These are the five signals that would prove this one wrong.

1. **Lab revenue compounds through the price cuts.** If Anthropic and OpenAI keep doubling run rates after cutting token prices, Jevons is winning and margins can survive the dump. Watch quarterly run-rate disclosures after the first big cut.
2. **Enterprise share of Chinese weights stays near zero.** If by mid-2027 Chinese open models still have no material Fortune 500 production footprint, even US-hosted, the dump stays a developer story and the enterprise book never reprices.
3. **The gap re-widens past ~12 months.** If the next closed generation opens a lead the open ecosystem cannot close inside a year (watch Epoch's lag metric), premium pricing becomes defensible again.
4. **The chip ceiling holds.** If Huawei's 600K-chip year slips badly and H200-class access tightens again, the Chinese release cadence slows and the dump loses pressure.
5. **Capex starts covering itself.** If hyperscaler earnings show AI revenue visibly closing the capex-to-revenue gap (the gap was still widening as of Jun 2026, per Forbes), then this was a buildout, not a bubble.

Scoreboard note: as of July 2026, every falsifier above is a live question, not a settled one. That is exactly why this is a thesis and not a report.

13 The inversion: what it means for a business like ours

Here is the inversion that makes this thesis useful instead of just scary: the same force that threatens frontier-lab margins is a tailwind for everyone who builds on top of models.

- **Tokens are our COGS, not our product.** Actual Intelligence Labs sells outcomes: working systems, automations, agents. Every step of the dump cuts our input costs and expands what is economically automatable for clients. A 29x cheaper token turns "not worth automating" into "obviously automate."

- **The dot-com rhyme, finished honestly.** The 2000 crash wiped out the people who owned the fiber. It subsidized everyone who used the fiber: Google, Amazon, and all of SaaS were built on crashed-out bandwidth. In this cycle, the labs and GPU landlords own the fiber. Builders use it. Even in the pop scenario, the cheap capacity survives and gets repriced into everyone's hands.
- **The margin migrates, it does not vanish.** When the model layer commoditizes, value moves up the stack to whoever owns the customer workflow and the integration. That is the layer we sell.

13.1 The playbook this implies

1. **Stay model-agnostic.** Build every client system so the model is a swappable part, never the foundation. The best model this quarter will not be the best model next quarter, and it may cost 90% less.
2. **Price on value, never cost-plus.** Our input deflates ~50x a year. Our clients' payroll does not. That spread is the business.
3. **Watch US-hosted open weights.** American providers serving Chinese open models on US soil is the compliance-safe cheap lane. When it matures for enterprise, it becomes our default recommendation for cost-sensitive builds.
4. **Treat a crack as a window.** If the market breaks, budgets tighten but automation demand rises, because "do more with less" stops being a slogan and becomes a mandate. Cost reduction is literally what we sell.

The honest risk: a hard enough crash shrinks small-business spending across the board, and we would feel that like everyone else. The tailwind is on our cost side. The storm, if it comes, is on the demand side. Net of both: I would rather be a builder than a lab in every version of this story.

14 The bottom line

Intelligence is becoming a commodity faster than the market has priced it. Commodity economics are brutal for whoever sells the commodity and beautiful for whoever builds with it. I do not know whether the pop comes in two quarters or two years. **I do know which side of that trade I want to be standing on.**

Zach Kellman · Actual Intelligence Labs · July 2026 · Working thesis, not investment advice.

15 References

1. NIST CAISI, "Evaluation of DeepSeek V4 Pro," May 2026
2. Epoch AI, "Open models lag state-of-the-art closed models by 4 months," May 2026; "LLM inference price trends," 2026

3. Reuters, "A new, inexpensive Chinese AI model is catching up with Anthropic, OpenAI on their home turf," Jul 2, 2026
4. Reuters, DeepSeek theoretical cost-profit disclosure, Mar 2025
5. TechCrunch, "DeepSeek previews new AI model that closes the gap with frontier models," Apr 24, 2026
6. OpenRouter, "State of AI" and token-share data, 2025 to 2026; OfficeChai summary, Jun 2026
7. Tech Times, "Chinese AI models lead OpenRouter traffic," May 29, 2026; Dataconomy, Feb 25, 2026
8. CNBC, "Anthropic tops OpenAI as most valuable AI startup," May 28, 2026; "Tech AI spending approaches \$700 billion in 2026," Feb 6, 2026
9. Forbes, "Anthropic and OpenAI are taking opposite paths to AI profitability," May 21, 2026; "OpenAI and Anthropic count revenue differently," Mar 25, 2026; "AI capex-to-revenue gap is widening," Jun 2, 2026
10. Tom's Hardware, "Big tech's AI spending plans reach \$725 billion," 2026; "AI costs spike as subscriptions hit pricing wall," 2026
11. Fortune, "2026 is looking like 1999," Jun 8, 2026; "AI boom may be on its last legs... blow-off phase," Jun 26, 2026; Capital Economics forecast coverage
12. Bloomberg, "AI circular deals: how Microsoft, OpenAI and Nvidia keep paying each other," 2026; IMF circular-financing warnings via press
13. Apollo Global Management (Torsten Slok), Silicon Data Token Expenditure Index, Jun 12, 2026
14. Deloitte, TMT Predictions 2026 (inference share of compute)
15. CBS News, DeepSeek R1 selloff and Nvidia's \$600B single-day loss, Jan 2025; Intellectia, "AI stocks selloff June 2026," Jun 24, 2026
16. CSIS, "DeepSeek, Huawei, export controls, and the future of the US-China AI race," 2026; CFR, "China's AI chip deficit," 2026
17. Tom's Hardware / Commerce Dept. coverage, H200 export rule, Jan 13, 2026; Huawei Ascend production plans, 2026
18. AI-Regulation.com (MIAI), "DeepSeek one year later" (BSI incident), 2026; Fox News, Booz Allen vulnerable-code report, 2026
19. Developers Digest, "Frontier model API pricing, June 2026"; aipricing.guru, Jul 2026; Morph and NxCode pricing guides, 2026
20. BenchLM, Chinese model leaderboard and SWE-bench data, 2026; DataCamp, "DeepSeek V4: features, benchmarks," 2026

Working thesis. Not investment advice. Prepared by Zach Kellman, Actual Intelligence Labs, Sarasota, FL.