WHITE PAPER · STRATEGY

The Compute Cliff

America is betting half a trillion dollars that frontier AI needs ever-larger clusters. China's strategy is to prove that it doesn't. A game-theoretic analysis of the great AI overbuild — and the conditions under which it strands.

AUTHOR

Kymata Labs Research

PUBLISHED

June 2026

CLASSIFICATION

Public · v1.0

A deliberately contrarian thesis, argued without hedging — and sourced without exception. Every figure below is cited; estimates are labelled as estimates.

ABSTRACT

In 2026 the largest technology companies on earth will spend on the order of *seven hundred billion dollars* on AI infrastructure,[7] extending a buildout whose total committed scale exceeds one trillion dollars.[11] That spending rests on a single premise: that frontier capability requires ever-larger training clusters, that the resulting models can be kept proprietary, and that demand will arrive to pay for it. In January 2025 a Chinese laboratory called DeepSeek published an open-weight model competitive with the American frontier, reporting a final training run of *$5.6 million*[1] the news erased roughly $589 billion of Nvidia's value in a single day, the largest such loss in market history.[4] This paper argues that the episode was not an anomaly but a preview. China's rational strategy — efficiency under sanction, open weights, inference at cost — is to commoditize the exact layer the United States is spending half a trillion dollars to monopolize, and three independent cracks in the buildout's premise (a scaling plateau, a dissolving moat, and a revenue gap measured in the trillions) are now visible at once. We take the strongest counterargument — Jevons's paradox — seriously, and conclude that it can be entirely correct about total demand and still leave the specific bet stranded. The buildout is not obviously a bubble. It is something more precise and more dangerous: a half-trillion-dollar wager on an assumption an adversary is structurally incentivized, and increasingly able, to falsify.

Executive summary

The thesis can be stated in five moves.

THE ARGUMENT IN FIVE MOVES

1. **The bet.** US hyperscalers and their partners have committed on the order of $1 trillion+ to AI compute — $690–725 billion in 2026 capital expenditure alone — on the premise that scale is the moat.[11, 17]
2. **The falsification risk.** DeepSeek showed a frontier-competitive model trained for a reported $5.6 million marginal cost and released its weights for free, collapsing both the cost premise and the proprietary premise at once.[1, 3]

3. **The doctrine.** For China, under chip-export sanction, commoditizing the model layer is not vandalism but the dominant strategy: open weights neutralize both the US capability moat and the export controls in one move.[19, 20]
4. **The cracks.** Three independent threats to the bet are now visible — pre-training returns are plateauing,[21, 22] open models lead public leaderboards,[20] and the revenue required to justify the capex (Bain: $2 trillion a year by 2030) exceeds anything yet in evidence by $800 billion.[13, 14]
5. **The cliff.** The strongest defense — Jevons's paradox — predicts demand will explode, and is probably right. But "demand is real" did not save the companies that overbuilt the fiber-optic internet; it is the *builders*, not the technology, who strand.[27]

This is not a prediction of imminent collapse. It is a claim about *asymmetry*: China's downside from this strategy is capped and its upside is structural, while America's buildout is a leveraged, circularly-financed bet whose central assumption a capable adversary is now spending its scarce resources to disprove. The remainder of this paper documents each move, steelmans the opposition, and maps who is exposed when the assumption breaks.

Contents

§1.

The half-trillion-dollar bet

Begin with the scale, because the scale is the argument. In 2024 the four American hyperscalers — Microsoft, Alphabet, Amazon, and Meta — spent roughly $230 billion in capital expenditure, most of it on AI.[7] In 2025 that figure reached approximately $388 billion.[7] For 2026, their combined guidance lands between $690 and $725 billion — Microsoft near $190 billion, Amazon around $200 billion, Alphabet $175–185 billion, Meta $125–145 billion — a single-year sum larger than the annual GDP of most countries.[7] Goldman Sachs Research estimates hyperscaler capital spending of $1.15 trillion across 2025–2027, against $477 billion in the prior three years: a 140% step-change.[7]

Layered atop the hyperscalers is a constellation of dedicated mega-projects. In January 2025, OpenAI, Oracle, and SoftBank announced **Stargate** — "$500 billion over four years," with "$100 billion deployed immediately," to build AI infrastructure for OpenAI in the United States.[7] By September 2025 the partners claimed "nearly 7 GW of planned capacity and over $400 billion in investment" committed.[8] Individual data-center campuses are now specified in gigawatts and tens of billions of dollars; the binding contractual commitments across the ecosystem exceed $900 billion, and including announced partnerships approach $1.4 trillion over roughly eight years.[11]

What is this enormous sum buying, and on what assumption? It is buying the inputs to a specific theory of competitive advantage — the theory that in artificial intelligence, **scale is the moat**. Strip the theory down and it rests on three load-bearing assumptions, each of which the rest of this paper will show to be contestable:

> **THE THREE ASSUMPTIONS UNDER THE BUILDOUT**
>
> **One — that capability scales with the cluster.** That the path to a better model runs through more chips, more data, and a larger pre-training run, such that whoever can marshal the most compute wins the frontier.
>
> **Two — that the frontier can be kept proprietary.** That the resulting models constitute a defensible, paywalled asset — that the half-trillion dollars buys something a competitor cannot simply copy or undercut.

> **Three — that the demand will arrive to pay for it.** That enterprises and consumers will generate the trillions in revenue required to service the infrastructure before its hardware depreciates.

Each assumption is individually plausible and was, until recently, the consensus. The argument of this paper is that all three came under simultaneous, evidenced attack beginning in January 2025 — and that the attacker has every strategic reason to keep pressing.

§2.

The detonation

On 26 December 2024, a Chinese laboratory backed by the quantitative hedge fund High-Flyer released **DeepSeek-V3**, a 671-billion-parameter mixture-of-experts model activating only 37 billion parameters per token.[1] Its technical report disclosed something the field had not seen stated so plainly: a complete pre-training run on 2,048 Nvidia H800 GPUs, 2.79 million GPU-hours, at a market rental rate of $2 per GPU-hour — **$5.576 million**.[1] Three weeks later, **DeepSeek-R1** matched OpenAI's o1 on a range of reasoning benchmarks, trained by reinforcement learning atop the V3 base, and was released under a permissive MIT license — weights free to download, run, and fine-tune.[2]

The market understood the implication before the commentators did. On Monday, 27 January 2025, Nvidia fell 17% and shed roughly $589 billion in market capitalization — the largest single-day loss for any company in the history of US markets, exceeding the prior record (also Nvidia's) by more than double.[4] The damage radiated outward to anything whose valuation assumed scarce, expensive compute: the Nasdaq fell 3.1%, and power utilities that had rallied on data-center demand cratered — Vistra down 28%, Constellation Energy down 21%.[4] In a single session, the market repriced the possibility that the scarce input the entire buildout was premised upon might not be so scarce after all.

> *The reported figure was the marginal cost of the final run. The point was never that it was cheap to build DeepSeek; the point was that it was suddenly cheap to copy the frontier.*
>
> THE DISTINCTION THAT MATTERS

Intellectual honesty requires the immediate caveat, and it cuts in an instructive direction. The $5.576 million is not DeepSeek's true cost. As Dylan Patel's SemiAnalysis documented within days, that figure is the marginal cost of the final pre-training run on a pre-existing cluster; it excludes research, ablation, salaries, and the hardware itself.[3] SemiAnalysis estimates DeepSeek's actual fleet at roughly 50,000 Hopper-class GPUs and its total infrastructure spend near $1.6 billion;[3] a US Congressional committee later alleged the lab had access to some 60,000 Nvidia chips, including roughly 10,000 export-controlled H100s obtained through intermediaries.[18] The "$5.6 million model" is, in the strict sense, a myth.

But notice what the correction does and does not rescue. It rescues the claim that building a frontier lab is still expensive. It does **nothing** to rescue the assumptions the buildout depends on — because the threatening number was never the cost of building DeepSeek. It was the cost of **running** and **copying** it. An open-weight model that performs at the frontier and serves inference at a fraction of the incumbent price does not have to have been cheap to train in order to destroy the economics of a $500 billion proprietary bet. Once the weights are public, the training cost is a sunk historical fact; the competitive reality is that a frontier-grade model is now a free download that anyone can serve on commodity hardware. DeepSeek's true cost is a footnote. Its true effect is the thesis of this paper.

§3.

Mutually assured commoditization

Why would a Chinese laboratory give away, for free, a model that cost over a billion dollars to build the capacity for? The answer is not ideology, and treating it as such is the analytical error that leaves Western incumbents flat-footed. It is game theory, and from China's position it is close to a dominant strategy.

Consider the board. Since October 2022, successive US export-control rounds have sought to deny China the advanced chips — A100, H100, then the H800 workaround, then the H20 — needed to compete at the compute frontier.[26] The controls are explicitly designed around one premise: that frontier AI requires enormous quantities of the most advanced accelerators, which the United States and its allies control. DeepSeek's founder, Liang Wenfeng, named the bind precisely: "Money has never been the problem for us. Bans on shipments of advanced chips are the problem."[19]

Given that constraint, China faces a choice with a clear solution. It cannot win a spending race for proprietary frontier compute — the chips are sanctioned and the capital is asymmetric.

But it can change the game. If the most valuable models are **open and free**, three things happen at once, all favorable to Beijing and all adverse to the American bet:

WHY COMMODITIZATION DOMINATES, FROM CHINA'S SEAT

It dissolves the moat the US is paying half a trillion dollars to build. A proprietary frontier worth monopolizing requires that the frontier *can* be monopolized. Open weights at parity make the paywall worthless: you cannot charge a premium for what your competitor gives away.

It neutralizes the export controls. The controls assume frontier capability is gated behind banned chips at scale. But open weights run inference on whatever hardware is available, anywhere; "once model weights are public, anyone can perform inference and fine-tuning on any hardware,"[20] which makes the chip gate moot for the thing that matters to most users.

Its downside is capped; its upside is structural. China forgoes model-layer profits it was unlikely to capture anyway against subsidized US incumbents — and in exchange wins standard-setting influence, global adoption, and soft power. As RAND observes, a Chinese-led open ecosystem "would not directly generate revenue… but it creates soft power."[25]

The evidence that this is the operative strategy, not a happy accident, is in the adoption data. By September 2025, Alibaba's open Qwen family had surpassed Meta's Llama as the most-downloaded model lineage on Hugging Face; between August 2024 and August 2025, Chinese developers accounted for 17.1% of all model downloads against 15.8% for US developers — the first time China led that metric.[20] On the public ChatBot Arena leaderboard in December 2025, nine of the top ten positions were held by Chinese laboratories.[20] An entire national ecosystem — DeepSeek, Qwen, Moonshot, Z.ai, MiniMax — is pursuing the same play in parallel.

This is the dynamic we call **mutually assured commoditization**. The United States is committed to a strategy that pays off only if the frontier stays scarce and proprietary. China is committed to a strategy that pays off precisely by making the frontier abundant and free. These are not two firms competing in a market; they are two states with opposed win conditions, and one of them has discovered that it can win not by climbing the wall but by dissolving it. The half-trillion-dollar buildout is, in this frame, a fortress being constructed around a commodity.

A genuine objection deserves acknowledgment here, and it tempers the timeline without changing the direction. Anthropic's Dario Amodei argues that DeepSeek's efficiency gains are "simply very talented engineers," not a paradigm break, and — crucially — that those gains "will soon be applied by both US and Chinese labs."[24] CSIS's Gregory Allen makes the parallel point that the efficiency improvements were "a continuation of preexisting industry trends… not a drastic acceleration."[25] Both are correct, and both arguments cut against the incumbents, not for

them. If efficiency gains accrue to everyone, then the proprietary edge that the buildout is meant to purchase erodes for everyone — which is exactly the commoditization this section describes. An efficiency improvement that your adversary can also use is not a moat. It is a treadmill.

§4.

Three cracks in the premise

The commoditization strategy would be a slow-acting solvent if the buildout's three assumptions were otherwise sound. They are not. Each is now under independent, evidenced strain — and the assumptions are not insured against one another, so a failure in any one is sufficient to impair the bet.

Crack one: the scaling plateau

The first assumption — that capability scales with the cluster — is the one the field's own founders have begun to abandon. Ilya Sutskever, who did more than anyone to establish the scaling paradigm, told NeurIPS in December 2024 that "pre-training as we know it will end," and by late 2025 was describing 2026 onward as "another age of research," adding that "another 100× scaling… would not transform AI capabilities."[21] The empirical confirmation arrived with OpenAI's own **Orion**, the model intended to be GPT-5: it "failed to deliver significant performance gains over GPT-4," and the company pivoted from ever-larger pre-training toward post-training and reasoning.[22] Epoch AI later documented that the shipped GPT-5 used **less** training compute than GPT-4.5, the resources redirected into reinforcement learning.[22]

This is the crack with the most direct bearing on the buildout, because the buildout is overwhelmingly a **pre-training** bet — gigawatt clusters optimized to train one enormous proprietary model. If the frontier is moving from pre-training scale toward inference-time reasoning and post-training — and RAND projects inference will consume the majority of AI compute by 2030[23] — then a meaningful share of the half-trillion dollars is optimized for the wrong workload. Worse for the incumbents, inference is exactly the regime where China is strong: it runs on a wider range of hardware, and a frontier-grade open model serving inference cheaply is the commodity §3 describes.

Crack two: the dissolving moat

The second assumption — that the frontier can be kept proprietary — was the first to fall, and §3 has already documented the data: open Chinese models at or near the top of public leaderboards, leading global download share, released free.[20] Add to this the distillation dynamic. OpenAI

publicly stated in early 2025 that it was "reviewing indications that DeepSeek may have inappropriately distilled" its models,[11] and Microsoft reported observing anomalous bulk extraction through its API. Whatever the legal merits, the technical reality is structural and unfixable by litigation: a frontier model's outputs are a near-perfect teacher for a fast-follower, and reasoning models — trainable by reinforcement learning on synthetic traces — iterate faster than pre-training ever did.[3] The leader's capability leaks into the follower at the speed of an API call. A moat that refills the besieger's reservoir is not a moat.

Crack three: the revenue that isn't there

The third assumption — that demand will arrive — is, on present evidence, the weakest. The required figures are extraordinary. Sequoia's David Cahn estimated in 2024 that the AI ecosystem needed to generate roughly $600 billion in annual end-user revenue merely to service its infrastructure spend.[12] In September 2025, Bain & Company put a sharper number on it: by 2030 the industry will require **$2 trillion in combined annual revenue** to fund the compute on order — a sum exceeding the combined 2024 revenue of Amazon, Apple, Alphabet, Microsoft, Meta, and Nvidia — and projected that revenue would fall **$800 billion short**.[13]

Against those requirements, the demand actually materializing is thin. MIT's NANDA initiative, studying 300 deployments and 150 executive interviews, found that **95% of enterprise generative-AI pilots delivered no measurable profit-and-loss impact**.[14] "The hype on LinkedIn says everything has changed," one operations chief told the researchers, "but in our operations, nothing fundamental has shifted."[14] This does not mean the demand will never come — enterprise software adoption is famously lumpy — but it means the buildout is being financed against a revenue stream that, in 2026, does not yet exist at anything like the required scale.

§5.

The strongest objection: Jevons's paradox

A thesis this stark must meet its best opposition head-on, and the best opposition is formidable. Within hours of DeepSeek's release, Microsoft's Satya Nadella posted what became the bulls' rallying cry:

> *"Jevons paradox strikes again! As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of."*
>
> SATYA NADELLA, CEO, MICROSOFT — 27 JANUARY 2025[5]

The argument is rigorous and historically grounded. William Jevons observed in 1865 that more efficient steam engines increased, rather than decreased, total coal consumption, because efficiency lowered cost and lower cost unlocked vastly more demand. Applied here: if DeepSeek makes intelligence ten times cheaper, the world will not buy the same amount of intelligence for a tenth of the price — it will buy a thousand times more intelligence, and need **more** data centers, not fewer. The bulls can point to confirming evidence: Microsoft reported a $13 billion AI revenue run-rate growing 175% year-over-year even as DeepSeek shipped;[5] Nvidia's Jensen Huang argues the field now rides "two exponentials" — exponentially more compute per reasoning query meeting exponentially more demand — and projects $1 trillion in AI chip demand through 2027.[6] Agentic workloads, which can consume a hundred times the compute of a single query, are the strongest structural case that the demand curve bends upward without limit.

We do not dismiss this. We think it is **probably correct about total demand** — and largely beside the point about the bet. Here is the distinction the bulls elide. Jevons's paradox predicts that the **resource** will be consumed in growing quantity. It says nothing about whether the **specific firms that financed the specific build** will be the ones who capture the value. The cleanest historical analogue is not coal; it is the fiber-optic overbuild of 1999–2001.

The telecom industry of the late 1990s laid fiber against a forecast — that internet traffic would explode — that turned out to be **entirely correct**. Internet traffic did explode. And the companies that built the fiber were wiped out anyway: WorldCom and Global Crossing went bankrupt, roughly a trillion dollars of market value evaporated, and less than 5% of the laid fiber was lit when the bust came.[15, 27] The demand was real. It simply arrived after the builders were insolvent, and the infrastructure was bought out of bankruptcy for cents on the dollar by the firms — Google among them — that captured the actual value. Jevons was right about bandwidth. It did not save the people who built it.

> **THE QUESTION THE PARADOX CANNOT ANSWER**
>
> "Will AI compute be consumed in growing quantity?" — almost certainly yes. That is the wrong question. The right question is: "Will the companies that financed $1 trillion of clusters, on six-year depreciation schedules and circular financing, be the ones who get paid before the assets obsolesce?" Jevons's paradox is silent on this, and it is the only question that matters to the bet.

Man Group puts the sharpest point on why the demand signal itself may be unreliable: much of it is recursive. Microsoft funds OpenAI, which commits to spend on Microsoft's cloud; Nvidia commits $100 billion to OpenAI, whose CFO concedes "most of the money will go back to

Nvidia."[10, 16] "Capex looks justified," Man Group writes, "because demand from inside the loop appears endless. But the demand signal becomes circular and divorced from the market."[16] It is precisely the structure of the telecom bubble, in which Cisco's customers bought Cisco's gear with Cisco's loans. The Jevons paradox describes real exogenous demand. The danger is that a large fraction of the demand currently underwriting the buildout is endogenous to it.

§6.

The cliff: how the bet strands

Put the pieces together and the failure mode is not a sudden pop but a **strand** — assets that remain physically real while their economic basis quietly evaporates. Three mechanisms drive it.

Depreciation. The hyperscalers extended the assumed useful life of their AI servers from three or four years to six between 2019 and 2023, spreading the cost and flattering near-term earnings.[15] Michael Burry's argument — for which he disclosed roughly $1.1 billion in put positions in late 2025 — is that the true competitive life of a leading-edge accelerator, against an annual architecture cadence, is closer to two or three years, and that the industry is therefore carrying on the order of $176 billion in understated depreciation and overstated profit through 2028.[15] If the obsolescence cycle forces the schedules back toward reality, the earnings revisions are mechanical and large.

Circular financing. The web of cross-investment described in §5 is a strength while every node is growing and a transmission mechanism when one is not. Binding commitments exceeding $900 billion[11] assume continuous deployment and continuous demand; a single major node slowing — an OpenAI revenue miss, a cloud-spend renegotiation — propagates through the loop. The early tremors are visible: by December 2025 the credit-default-swap spread on the GPU-leasing firm CoreWeave implied a roughly four-in-ten probability of default within five years, and its equity had fallen 61% from its peak.[11]

The wrong workload. If §4′s scaling plateau holds, a portion of the gigawatt pre-training clusters is optimized for a regime the frontier is leaving. Pre-training-optimized infrastructure does not become worthless — it can serve inference — but it strands in the precise sense that it will not earn the return its financing assumed.

Who is exposed? In ascending order of fragility: the diversified hyperscalers, who can absorb a write-down against trillion-dollar franchises; the model labs, whose proprietary premium is the thing being commoditized; the debt-financed "neocloud" GPU lessors, whose entire model is leverage against assets that depreciate faster than their loans amortize; and the equity holders

who have priced perfect execution of a half-trillion-dollar bet against an adversary working to falsify its premise. The technology is not the fragile thing. The **capital structure** is.

§7.

Strategic outlook

What follows from the analysis is not a market-timing call — the buildout could run for years, and Jevons-driven demand could yet validate much of it. What follows is a map of asymmetric exposure, and a strategy keyed to it.

The asymmetry is the headline. China's commoditization play has a capped downside (forgone model-layer profits it was unlikely to win) and a structural upside (a dissolved adversary moat, neutered export controls, standard-setting influence). The American buildout has a capped upside (it can, at best, win a frontier it has bet correctly stays proprietary) and an uncapped, leveraged downside (a trillion dollars of depreciating, circularly-financed assets if the premise breaks). When one player's worst case is "we gave away some software" and the other's is "we stranded a trillion dollars," the strategies are not symmetric, and capital should not treat them as such.

For US policymakers, the uncomfortable implication is that export controls aimed at pre-training compute may be fighting the last war. If capability is migrating to inference and post-training — runnable on a wider hardware base — then gating the largest accelerators raises China's costs without denying it the frontier, while handing it the motive that produced DeepSeek in the first place. The control regime optimizes against the assumption this paper questions.

For builders and investors, the move is to separate the two bets the market currently prices as one. The bet that **AI compute will be consumed in vast quantity** is sound; buy it through the parts of the stack that capture value regardless of who builds — the durable demand aggregators, the power and grid layer, the inference-efficient. The bet that **the specific firms financing today's proprietary-frontier clusters will earn their return before the assets obsolesce** is the fragile one, and it is the one most heavily owned. Weight survivable infrastructure over leveraged builders; weight the commodity's consumers over its overextended suppliers. In the fiber analogy, be Google buying lit capacity in 2003, not the lender to Global Crossing in 2000.

For operators — the seat we write from — the discipline is to build on the assumption that the model layer is commoditizing, because it is: design for model portability, treat frontier access as a substitutable input rather than a moat, and capture defensibility in the layers the commoditization does not reach — proprietary data, distribution, workflow, and trust. The companies

that thrive through a compute glut will be the ones that never depended on compute scarcity in the first place.

§8.

Conclusion: betting against an adversary's incentive

The case for the buildout is not foolish. Demand for intelligence may indeed prove insatiable; Jevons may be vindicated; the agentic era may consume every gigawatt now under construction and clamor for more. If so, the firms that build now will have been right, and early.

But strip the optimism to its structure and what remains is a half-trillion-dollar wager that frontier AI must be large, scarce, and proprietary — placed against an adversary whose entire strategic position depends on proving that it can be small, abundant, and free, and who has already, once, made the market believe it for a day. The United States is betting on scarcity. China is manufacturing abundance. Those are not two views of the same future; they are two states spending their treasure to make opposite futures true, and only one of them has a strategy whose worst case is survivable.

That is the cliff. Not a certain fall — a precipice whose edge is defined by an assumption the people building closest to it do not control, and an adversary works each day to erode. The prudent posture is not to bet the assumption will break tomorrow. It is to notice who is exposed if it ever does, and to decline to be standing there when it happens.

This paper argues a deliberately contrarian thesis. Every figure is cited to a primary or named source below; analyst projections are labelled as estimates and the uncertainties (especially the magnitude of demand) are stated, not hidden. It reflects the public record as of June 2026.

References

1. DeepSeek-AI. “DeepSeek-V3 Technical Report.” arXiv:2412.19437, December 2024 (reports the $5.576M / 2,048-H800 final pre-training run, MoE, FP8, MLA). https://arxiv.org/abs/2412.19437
2. DeepSeek-AI. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” arXiv:2501.12948, January 2025. https://arxiv.org/abs/2501.12948
3. SemiAnalysis (D. Patel, et al.). “DeepSeek Debates.” 31 January 2025 (≈50k Hopper GPUs, ≈$1.6B capex; the marginal-vs-true-cost critique). https://semianalysis.com/2025/01/31/the-deepseek-debates/
4. Reuters; Forbes. “Nvidia loses ≈$589–593 billion in market value — the largest single-day loss in US history.” 27 January 2025. https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/
5. Nadella, S. Post on X (Twitter), 27 January 2025 (“Jevons paradox strikes again…”). Microsoft FY25 Q2 AI run-rate via the same earnings cycle. https://x.com/satyanadella/status/1883753899255046301
6. Huang, J. (Nvidia). CNBC interview, October 2025 (“two exponentials”); GTC 2026 keynote (≈$1T AI-chip demand through 2027). https://www.cnbc.com/2025/10/08/jensen-huang-nvidia-computing-demand.html
7. OpenAI. “Announcing the Stargate Project.” 21 January 2025 ($500B / $100B). https://openai.com/index/announcing-the-stargate-project/
8. SoftBank Group. Stargate expansion press release (≈7 GW / >$400B committed). 24 September 2025. https://group.softbank/en/news/press/20250924
9. The Decoder / The Information. “Stargate’s structure stalls; resolved via a bilateral Oracle–OpenAI deal.” 2025. https://the-decoder.com/stargates-500-billion-ai-infrastructure-project-reportedly-stalls-over-unresolved-disputes-between-openai-oracle-and-softbank/
10. Tunguz, T. “Nvidia, Nortel, and the Return of Vendor Financing” (Nvidia’s $100B OpenAI LOI; “most of the money will go back to Nvidia”). 2025. https://tomtunguz.com/nvidia_nortel_vendor_financing_comparison/
11. CNBC. “A guide to the $1 trillion+ of AI deals between OpenAI, Nvidia, Oracle, and others” (binding commitments; CoreWeave exposure). 15 October 2025. https://www.cnbc.com/2025/10/15/a-guide-to-1-trillion-worth-of-ai-deals-between-openai-nvidia.html
12. Cahn, D. (Sequoia Capital). “AI’s $600B Question.” 2024. https://www.sequoiacap.com/article/ais-600b-question/
13. Bain & Company. “6th Annual Global Technology Report” ($2T annual revenue needed by 2030; ≈$800B projected shortfall). 23 September 2025. https://www.bain.com/about/media-center/press-releases/20252/2-trillion-in-new-revenue-needed-to-fund-ais-scaling-trend/
14. MIT NANDA. “The GenAI Divide: State of AI in Business 2025” (95% of enterprise GenAI pilots show no measurable P&L impact). July 2025. https://nanda.media.mit.edu/
15. Saxo Bank; CNBC. On M. Burry’s depreciation thesis (≈$176B understated depreciation through 2028; ≈$1.1B put positions; telecom analogy). November 2025. https://www.home.saxo/content/articles/equities/big-short-12112025
16. Man Group. “The AI Bubble” (recursive/circular demand; “the demand signal becomes circular and divorced from the market”). 2025. https://www.man.com/insights/the-ai-bubble

17. Goldman Sachs Research. Hyperscaler AI capex outlook (≈$1.15T 2025–27; 2026 guidance). 2026. https://www.goldmansachs.com/insights/articles/why-ai-companies-may-invest-more-than-500-billion-in-2026

18. U.S. House Select Committee on the CCP. "DeepSeek" report (alleged ≈60k Nvidia chips incl. ≈10k H100s via intermediaries). 2025. https://selectcommitteeontheccp.house.gov/

19. Liang Wenfeng (DeepSeek founder), quoted in Williams, R., The Washington Quarterly 48:2 ("Bans on shipments of advanced chips are the problem"). 2025. https://twq.elliott.gwu.edu/

20. Stanford HAI / DigiChina, "China's Diverse Open-Weight AI Ecosystem," 2025; IISS Strategic Comments, "DeepSeek's release of an open-weight frontier model," April 2025 (the moat problem; Qwen > Llama; download-share and ChatBot-Arena data). https://hai.stanford.edu/

21. Sutskever, I. NeurIPS 2024 keynote ("pre-training as we know it will end"); interview with D. Patel, November 2025 ("another age of research"). https://www.reuters.com/technology/artificial-intelligence/openai-co-founder-sutskevers-new-safety-focused-ai-startup-ssi-raises-1-billion-2024-09-04/

22. Fortune / The Information, "What happened to GPT-5 / Orion," February 2025; Epoch AI, "Why GPT-5 used less training compute than GPT-4.5," 2025. https://epoch.ai/

23. RAND Corporation. "When AI Takes Time to Think: Implications of Test-Time Compute." March 2025 (inference shifting to the majority of AI compute). https://www.rand.org/pubs/commentary/2025/03/when-ai-takes-time-to-think-implications-of-test-time.html

24. Amodei, D. (Anthropic). "On DeepSeek and Export Controls." 2025 (efficiency gains accrue to all labs). https://www.darioamodei.com/post/on-deepseek-and-export-controls

25. Allen, G. (CSIS). "DeepSeek, Huawei, Export Controls, and the Future of the US–China AI Race." 2025; RAND, "What DeepSeek Really Changes About AI Competition," February 2025 (soft-power framing). https://www.csis.org/analysis/deepseek-huawei-export-controls-and-future-us-china-ai-race

26. U.S. Bureau of Industry and Security; CSIS (E. Benson), "Updated October 7 Semiconductor Export Controls," October 2023; Congressional Research Service R48642 (A100/H100, H800, H20 timeline). https://www.csis.org/analysis/updated-october-7-semiconductor-export-controls

27. On the fiber-optic / railway overbuild precedent (demand real; builders insolvent; under 5% of fiber lit at the bust; ≈$1T market value erased). Overbuild-cycle analyses, 2025. https://lime.co/news/overbuilding-the-future-why-ai-is-repeating-the-railways-and-dark-fiber-144565