

# Beyond DeepSeek: China's 2026 model wave and the repricing of the AI stack

State Street · Ninghui Liu · 2026-09-11 · English original

---

A one-off shock, or a repeatable engine outside the United States—and where does profit accrue across the AI stack?

Across roughly eight weeks in mid-2026, five Chinese developers shipped six frontier-adjacent releases. That significance is not any single model, but the pace of delivery. Where January 2025 delivered a single, cost-led jolt with DeepSeek-R1, the 2026 sequence spans multiple firms and reaches beyond price into capability and scale. These include:

- Moonshot: Kimi K3
- Z.ai: GLM-5.2
- DeepSeek: V4-Flash, V4 Pro
- Alibaba: Qwen3.8-Max
- ByteDance: Seedance 2.5

On the evidence, this looks structural, not a repeat surprise: China now has an industrialized pipeline that produces near-frontier models on a recurring basis under real compute limits.

Four developments may help explain what's changing:

- An isolated event has become a pattern: Chinese open-weight models now top global usage charts.<sup>1</sup>
- The edge is efficiency, not merely a discount, as a demanding task runs for cents on the cheapest Chinese model compared to US dollars on a leading US system.
- The economics are moving rather than evaporating, because August's twin moves—penAI cutting Luna prices roughly 80%<sup>2</sup> and DeepSeek raising V4 prices 50% to 1,100%<sup>3</sup>—left the price of intelligence trading within a band, capped from above by open-weight competition and supported from below by scarce compute infrastructure.
- Compute is the bottleneck because unit costs of intelligence keep falling while accelerators, memory, space, and power stay tight.

The clearest beneficiaries remain the companies that provide the infrastructure behind AI, where competitive advantages appear more durable than they do for model developers. Hyperscalers gain twice, from rising aggregate demand and from hosting open models and billing by usage. Model developers face the biggest challenge, as fast-rotating leadership implies a thinner moat and makes it harder to justify premium valuations. Enterprises and software deployers, by contrast, should keep a rising share of the surplus.

- A durable, closed-lab, frontier model capability lead would preserve their pricing power

- A compute restriction could cap China's scaling
- Open models are not free once hosting, tuning, and security are included
- And AI spending could cool off if frontier lab spending slows

## **Bottom line: China is closing the gap, but infrastructure layer still leads**

This is more than a rerun of the DeepSeek shock, but not yet proof that Chinese and US frontier models are on equal footing. The lead has narrowed, not disappeared. The clearest beneficiaries remain the companies that provide the compute and infrastructure, along with the enterprises best positioned to turn lower-cost intelligence into measurable productivity.

## **Is this another DeepSeek moment?**

DeepSeek's January 2025 shock was, in essence, about cost. It came from one company, made model training much cheaper by relying on reinforcement learning rather than supervised fine-tuning, and showed capable models could originate outside the US even if it still sat behind the frontier. Markets read it through Jevon's lens: Microsoft's Satya Nadella argued cheaper AI would send usage soaring, and Chinese equities rallied.

## **What's Jevon's Paradox?**

The phenomenon where increased efficiency in resource use leads to higher overall consumption of that resource.

The 2026 wave differs. In roughly two months, five firms delivered six near-frontier releases (DeepSeek shipped twice) across distinct architectures and modalities: Kimi K3, the open-weight capability leader; Alibaba's enterprise-focused Qwen3.8-Max; DeepSeek's ultra-cheap V4-Flash and, weeks later, its frontier-efficient V4 Pro; GLM-5.2, briefly the top open model; and ByteDance's video leader Seedance (Figure 1).

Figure 1: Major Chinese AI model launches, June-August 2026

Source: State Street Investment Strategy and Research; company releases; Artificial Analysis. As of August 13, 2026. China's 2026 model wave—six releases from five firms (the V4 family previewed in April 2026).

What really matters is not whether China surprised everyone once with a breakthrough, but whether it has built a system that can keep producing strong models repeatedly. And even people who were initially doubtful now increasingly agree that's what we're seeing.

Figure 2: From DeepSeek to the China AI wave

Source: State Street Investment Strategy and Research, Artificial Analysis. As of August 13, 2026.

The evidence points to a maturing system, defined by breadth of firms and architectures, compressed development cycles, and adoption at production scale. The caveat is that "near-frontier" is not "frontier." The most capable systems still lead. The gap has narrowed rather than closed, but the system now looks built to last.

## Mapping the AI value chain

Who actually keeps the economics of AI? Using NVIDIA President and CEO Jensen Huang's "AI cake" analogy, the ecosystem reads as a series of layers (Figure 3). The key question is where profit accrues and where competitive pressures are most intense.

Figure 3: The AI value stack—Jensen Huang's "AI cake"

Compute is the scarce, hard-to-replace base; the model layer is where competition is fiercest

At the base, hardware and infrastructure is where scarcity pricing and healthy margins live. Asia anchors it, supplying most leading-edge AI silicon so Western build-outs rest on Asian hardware, and its leaders (NVIDIA, TSMC, SK Hynix, Samsung, Broadcom) are hard to unseat.

Hyperscalers turn capital into capacity and can also host open models, metering by consumption. The model layer is the contested tier, split between closed models (OpenAI, Anthropic, xAI) and open-weight models (DeepSeek, Qwen, Moonshot, Z.ai, as well as Meta, and NVIDIA). The difference is how buyers obtain and pay for each: a closed model is rented through a paid, metered API, so the provider controls access and pricing, whereas an open-weight model publishes its parameters to download, fine-tune, and self-host, far cheaper at scale.

At the top is the application and enterprise layer, which captures the productivity dividend. In short, profit concentrates at the base while the surplus drifts to those who deploy.

## Why open models matter

Open models reshape how the market perceives AI economics, and adoption has followed. On estimates cited by a16z and the US–China review commission, roughly 80% of US startups built on open-source now run on Chinese models such as Qwen rather than OpenAI or Anthropic, with Airbnb among those running Qwen in production.<sup>4</sup> The cost of a given capability has fallen just as sharply: Epoch AI tracks inference prices dropping up to several hundred-fold a year, and a GPT-3.5 equivalent is now two orders of magnitude cheaper to run than in late 2022 (Figure 4). Throughout this article we refer to that all-in cost of getting a task done as the price of intelligence.<sup>5</sup>

## What we mean by the price of intelligence

The all-in cost of completing a defined unit of work with an AI system, not the headline price per million tokens. It is the compute a task consumes, set by model architecture and deployment, multiplied by the price of that compute, set by accelerators, memory, and power. Because open competition compresses the first term while scarcity holds up the second, the price of intelligence falls quickly but has a floor beneath it.

The logic is sound: to win developers, a model must undercut the cheapest credible option or clearly outperform it.

Two caveats keep this from tipping into an outright open-model story:

- On a cost-per-intelligence basis, the strongest open models do not consistently undercut leading US systems, and where they lead, it is often on narrower tasks; top US models still outperform on complex, open-ended problems needing planning and coordination across tools
- Headline API prices also understate the advantages of closed-model providers. Brand, safety tooling, and integration command a premium that a weight file does not erase. Efficiency is therefore real but

conditional, which is why the hybrid outcome of closed frontier models coexisting with open-weight alternatives may be likely

## **Compute scarcity and the repricing of intelligence**

Falling prices of intelligence have a floor. The binding limit has moved from model capability to compute availability. GPU rental prices remain elevated. DeepSeek has raised its API prices,<sup>6</sup> citing capacity limits, and is building gigawatt-scale capacity. High bandwidth memory (HBM), the AI accelerators' specialized memory, is the choke point, concentrated among a few suppliers. Data-center power use is set to more than double by 2030 as hyperscaler budgets climb toward half a trillion dollars.<sup>7</sup>

This is self-reinforcing. By Jevon's logic, lower inference costs expand demand faster than they reduce cost per task, driving higher use of tokens, compute, and power. The counterpoint is also real, because Jevon's explains demand, not unit economics: if returns lag, or if much demand is driven more by ecosystem funding rather than end-user adoption, the demand elasticity could deflate as fast as it inflated.

The combination of intensifying model-layer competition, rising AI usage, and persistent compute scarcity implies a lasting shift in pricing power toward the compute layer, potentially benefiting infrastructure providers and certain hyperscalers while squeezing standalone model economics.

- Scarcity is cyclical as well as structural, because aggressive capacity additions could bring over-supply, utilization stress and GPU residual-value risk, especially for leveraged neoclouds
- A genuine efficiency step-up could loosen that scarcity faster than expected

## **A ceiling and a floor: Intelligence now trades in a band**

With that floor price of intelligence in place, the price of intelligence trades within a band, bounded on each side.

The first—OpenAI's Luna price cut—set a ceiling. On July 30th, days after Kimi K3's weights became a free download, OpenAI cut GPT-5.6 Luna by roughly 80% yet left the flagship Sol untouched.<sup>8</sup> Once a capable model can be deployed for free, no one can charge a premium for the same capability, and everything beneath the frontier is dragged toward the open-weight line.

The second move—DeepSeek's V4 price increase—confirmed the floor.

Two weeks later, DeepSeek, which had started the price war, raised its V4 rates by 50 to 1,100%, citing the same compute strain.<sup>9</sup> Prices cannot fall without limit, because cheap intelligence still rides on scarce, expensive hardware.

Figure 5: A ceiling and a floor—the intelligence price band

## **Why token prices have both a ceiling and a floor**

Open-weight competition caps the price of a given capability from above, while scarce compute holds it up from below. The ceiling pushes the model layer toward utility-like economics and compresses lab margins (See Risks and opportunities across the AI stack). The floor accrues to whoever owns compute, memory, and power.

## Open vs. closed models: Three possible futures

In the three possible scenarios we consider, the decisive variable is how models are accessed and paid for. The compute-layer beneficiaries look similar across all three.

- Open models win. Open models reach frontier ease-of-use, and inference decentralizes to private clouds and edge devices. This would benefit enterprise-hardware, security, and orchestration software, with Microsoft and NVIDIA still well-positioned. The main risks lie in deployment, governance, and reliability at scale
- Hybrid (base case). Closed frontier models coexist with strong open-weight alternatives, as enterprises route the hardest reasoning to premium systems and high-volume work to cheaper open ones. This fits survey evidence that most already pair the two.<sup>10</sup> Value spreads across clouds and owned hardware, lifting demand for routing, security, and orchestration. The beneficiaries include hyperscalers and infrastructure software (Datadog, Palantir, CrowdStrike, ServiceNow), as well as NVIDIA. But coexistence carries a catch for the premium tier: because models differ in both capability and cost, the enterprise's central task becomes workload allocation—reserving costly frontier intelligence for the high-stakes, decision-level work that justifies it, while sending routine, high-volume tasks to cheaper models that are already "good enough." If most everyday enterprise use cases fall into that second bucket, the value pool addressable by premium intelligence may be narrower than its capability lead implies. How much the close frontier models ultimately capture therefore depends less on raw frontier performance than on whether its intelligence creates enough value to justify the cost, and on how enterprises choose to deploy it. The caveat is that open is not automatically cheap nor closed automatically expensive—but closed frontier access generally still commands a premium over open-weight alternatives. This nuanced coexistence is usually what materializes<sup>11</sup>
- Closed models win. The frontier gap widens again, which is possible if compute limits cap Chinese scaling, so buyers keep paying a few labs while integration and network effects deepen the moat for closed models. Credible, but less likely than the hybrid case because the evidence of the past year points to a narrowing capability gap rather than a widening one, and because enterprises are already building around a mix of open and closed models

## The one constant across all three futures

Whoever builds the computing backbone—chips, networking, and power—stays well-positioned no matter which paradigm prevails. The more durable approach is to invest in the layer that wins across outcomes.

## Risks and opportunities across the AI stack

The scenarios share one constant: the compute base, making it the natural starting point to begin a layer-by-layer assessment of risk and opportunity across the stack.

## Hardware and infrastructure

Hardware and infrastructure companies appear to be the clearest structural winners, combining rising aggregate demand with positions defended by capital intensity and know-how. Model leadership can turn over in weeks, leadership in advanced logic, HBM, and lithography is built over years. This layer owns the "floor" of open models pricing, and its pricing power is the stickiest in the stack, though still cyclical. China's shift toward home-grown chips (Huawei Ascend, Cambricon) supports China-listed silicon even as it complicates the Western supply chain.

## Hyperscalers

Open models help hyperscaler economics: providers not only sell compute capacity but also host open models and monetize usage, generating higher-value revenue streams. The near-term risk runs the other way: much of today's compute demand is anchored by a few frontier labs training and serving their models, so if open competition compresses their monetization, that spend and hyperscaler backlog could cool. The decisive variable is the demand mix, because broad-based inference and enterprise workloads are more resilient than lumpy frontier-lab contracts, so watch it as closely as you do headline capital expenditures.

## AI labs

The labs face the hardest questions, because they sit directly beneath the ceiling of open models. With a freely deployable model now on the frontier, no lab can charge a premium for a capability available cheaply elsewhere. Pricing power survives only at the very frontier, and only for as long as that lead lasts. The flywheel that sustained the labs is now under pressure at every turn (Figure 6).

Figure 6: How open-weight competition can disrupt the traditional AI flywheel

Source: State Street Investment Strategy and Research.

For OpenAI and Anthropic—both of which have filed to go public—those valuations lean on high margins that open competition now tests.<sup>12</sup> This outcome may not be inevitable because frontier labs retain real advantages in capability, safety, brand, and integration. The question is whether current valuations adequately reflect the durability of those advantages and moat assumptions.

## Enterprise AI adoption

The flip side is opportunity for deployers. Cheaper inference and lower barriers push value toward the enterprises and software firms that operationalize AI, and teams already use a premium model to plan and structure a task, then pass the execution steps to a cheaper open model.

## Strategic implications for US–China AI competition

China's progress highlights a US dilemma: spread the American AI stack so the world standardizes on it, or slow China through export controls. The two aims conflict because tightening controls nudges third-market customers toward cheaper Chinese options like Huawei, whereas loosening them weakens the grip on China's compute access. The recent progress that came under tight compute constraints complicates the picture because it challenges the assumption that withholding compute reliably limits capability.

The commercial stakes are clear. When one side charges pennies for what the other prices in dollars, the contest for undecided markets shifts, and friction is already spilling into intellectual property disputes and sanction risk. Each tightening also accelerates China's domestic-silicon push and could split the ecosystem into rival standards. Policy is now a first-order driver of value capture and a key swing factor for the open vs. closed model scenarios. Policy outcomes are hard to predict, but China's AI ecosystem looks set to become an increasingly important part of the global landscape, which argues for a broad perspective on AI opportunities.

Open models are not unique to China either. US players such as Meta's Llama and NVIDIA's Nemotron are active too. A coalition of over thirty US firms formed the Open Secure AI Alliance,<sup>13</sup> an aspiration that open models curb the risk of a few gatekeepers controlling access to intelligence.

## From disruption to durable competition

Is this another DeepSeek moment? Our answer is a qualified no: this may be more than another DeepSeek moment, yet not proof that Chinese and US frontier models are on equal footing. A single, cost-led surprise has become a multi-company run that pairs capability with scale and is backed by production-scale adoption, so China now has a repeatable engine for near-frontier models despite compute constraints.

The evidence supports a durable system, value reallocated within the stack rather than destroyed, and compute as the binding constraint whose winners hold across outcomes. Three questions remain open: whether China closes the last capability gap, whether cheaper intelligence converts into validated returns, and how the US–China policy contest resolves.

## What investors should watch in the next 12 to 24 months

Several indicators will help determine whether China's recent AI progress marks a durable shift in industry economics and competitive positioning, or a temporary narrowing of the gap.

- The gap between open and closed models in real-world performance
- Token-consumption growth as AI becomes cheaper
- Compute demand: scarcity or oversupply
- Demand across training, inference, and enterprise AI workloads
- Model-layer monetization and pricing power
- Formation and evolution of intelligence price band
- Hard evidence of enterprise productivity gain and monetization
- Export controls, sanctions, and ecosystem fragmentation

## Investment posture

We claim no certainty on the model-layer winner. The higher-conviction view may be the compute-and-infrastructure base—the layer that benefits in every scenario—plus the hyperscalers and enterprise deployers best able to monetize cheap, abundant intelligence. The model layer offers the most optionality and the most risk, so size it accordingly.

## Contributor

1 CNBC, “Chinese AI models are gaining ground with U.S. companies as OpenAI, Anthropic costs surge” July 7, 2026. 2 OpenAI GPT-5.6 repricing, July 30, 2026. GPT-5.6 Luna reduced by roughly 80% (input from US\$1.00 to US\$0.20 and output from US\$6.00 to US\$1.20 per million tokens), GPT-5.6 Terra reduced by 20%, and flagship GPT-5.6 Sol left unchanged. 3 DeepSeek V4 API repricing, announced August 13, 2026 and effective 16–17 August 16-17, 2026, introducing peak and off-peak rates with increases of 50% to 1,100% depending on model, token type and time of day. Reuters, “DeepSeek raises API pricing for its V4 models,” August 13, 2026. 4 Jai Ramaswamy and Matt Perault, “Asserting American Leadership in Open Source AI,” a16z, April 28, 2026. Among developers building with open-source tools, 80% use Chinese open-source models. 5 Epoch AI, LLM inference price trends dataset. For GPT-3.5-level general-knowledge performance, price fell from about US\$20 per million tokens in November 2022 to US\$0.07 by October 2024, roughly 286 times cheaper. 6 DeepSeek’s official API

documentation. It announced revised V4 pricing on August 13, 2026, effective August 16, with peak/off-peak rates introduced “to enable more flexible workload scheduling.” 7 International Energy Agency, “AI is set to drive surging electricity demand from data centers.” The IEA projects global data center electricity consumption rising from about 415 TWh in 2024 to roughly 945 TWh in 2030, more than doubling. 8 OpenAI GPT-5.6 repricing, July 30, 2026. GPT-5.6 Luna reduced by roughly 80% (input from US\$1.00 to US\$0.20 and output from US\$6.00 to US\$1.20 per million tokens), GPT-5.6 Terra reduced by 20%, and flagship GPT-5.6 Sol left unchanged. 9 DeepSeek’s official API documentation. It announced revised V4 pricing on August 13, 2026, effective August 16, with peak/off-peak rates introduced “to enable more flexible workload scheduling.” 10 McKinsey & Company, Mozilla Foundation, and Patrick J. McGovern Foundation, “Open Source Technology in the Age of AI,” April 22, 2025. The survey covered more than 700 technology leaders and senior developers across 41 countries and found that over half of respondents used open-source AI technologies across several areas of the AI stack, often alongside proprietary tools. 11 McKinsey & Company, Mozilla Foundation, and Patrick J. McGovern Foundation, “Open Source Technology in the Age of AI,” April 22, 2025. The survey covered more than 700 technology leaders and senior developers across 41 countries and found that over half of respondents used open-source AI technologies across several areas of the AI stack, often alongside proprietary tools. 12 Reuters, “AI giant Anthropic confidentially files for US IPO,” June 1, 2026; and Reuters, “OpenAI files for US IPO after Anthropic,” June 8, 2026. 13 “Open Weights and American AI Leadership” Microsoft Corporate Responsibility, July 24, 2026.

**Initial public offering (IPO)** Initial public offering (IPO) The first sale of public stock by a private company.

**International Energy Agency (IEA)** A global energy forum comprised of 29 industrial countries that are net oil importers and which is designed to help its members respond to major supply disruptions. The IEA, organized under the Organization of Economic Cooperation and Development (OECD), was established in 1974 following the 1973-1974 Arab oil embargo. The IEA is based in Paris.

**Leverage** In finance, leverage—sometimes referred to as gearing in the United Kingdom, or as solvency in Australia—is a general term for any technique to multiply gains and losses. Common ways to attain leverage are borrowing money, buying fixed assets and using derivatives.

**Profit margins** A measure of how much out of every dollar of sales a company actually keeps in earnings. The figure is expressed as a percentage.

**Risk** The possibility that an investment’s return will differ from expected returns, especially the possibility of losing some or all of an investment. Risk is typically measured by calculating the standard deviation on historical, or average, returns of a given investment.

**Total cost of ownership** The purchase price of an asset plus the costs of operation.

**Valuation** The process of determining the current worth of an asset or a company.

For institutional / professional investors use only.

State Street Global Advisors Worldwide Entities

State Street Global Advisors (SSGA) is now State Street Investment Management. Please click here for more information.

Investing involves risk including the risk of loss of principal.

The whole or any part of this work may not be reproduced, copied or transmitted or any of its contents

disclosed to third parties without SSGA's express written consent.

---

Source: <https://www.ssga.com/hk/en/institutional/insights/beyond-deepseek-chinas-2026-model-wave-and-the-repricing-of-the-ai-stack>