The Strategy of Irreplaceability

Three Paths: Nvidia, Moonshot AI, and DeepSeek

Tong Yin | InsightBridge Global

November 2026



Introduction: The Goal Is Not to Lead, but to Be Indispensable

In artificial intelligence, leadership rarely lasts more than a few months. This quarter's top-ranked model can be overtaken next quarter, and today's cheapest API loses its edge after the next round of price cuts. What ultimately determines a company's fate is not whether it leads at a given moment, but how much its customers would lose by walking away. The higher that cost, the closer a company comes to being irreplaceable.

The global AI landscape of 2026 offers an unusually clear set of contrasting cases. Nvidia sits at the top of the value chain with gross margins of roughly 75% and a market capitalization of around $5 trillion [1][3]. Moonshot AI's Kimi K3, a 2.8-trillion-parameter model, has entered the first tier of frontier systems [8]. DeepSeek, through near-free pricing and a relentless cadence of open releases, has become one of the largest single sources of model traffic routed by developers worldwide [17]. The three companies have taken very different routes, but they are pursuing the same objective: to become the part of the system that customers, developers, and entire ecosystems cannot do without.

This article argues that irreplaceability can be achieved along three distinct paths. The first is monopoly: raising switching costs so high, through ecosystem lock-in, that leaving becomes impractical. The second is technological supremacy: excelling along a critical dimension that rivals cannot match in the near term. The third is ubiquity: continuously supplying technology that is good enough, at prices low enough, to capture the market first and become the industry's default infrastructure. Nvidia exemplifies the first path, Moonshot AI the second, and DeepSeek the third. All three are strategies of irreplaceability, yet they differ sharply in their sources, costs, durability, and consequences for the wider economy.

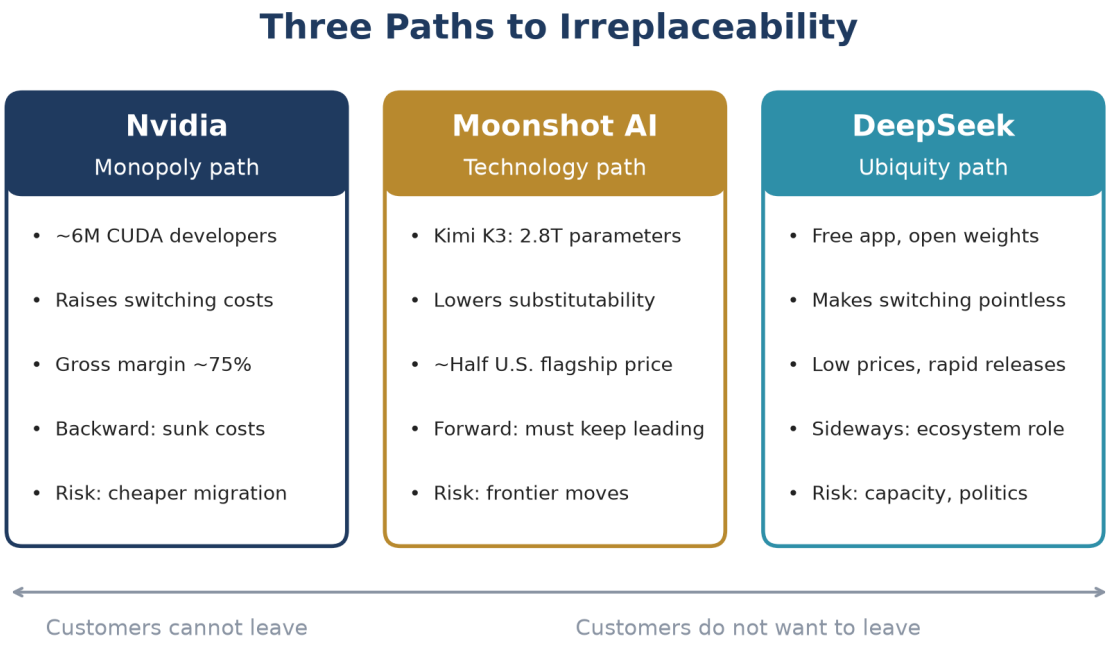


Figure 1. Three paths to irreplaceability

Two Variables: Switching Costs and Substitutability

Irreplaceability is best understood as the product of two variables. The first is substitutability: whether a functionally comparable alternative exists. The second is switching cost: how much time, money, and risk a customer must absorb to migrate, even when an alternative is available. A company that excels on either variable gains meaningful pricing power. A company that excels on both holds a de facto monopoly.

The three paths differ precisely in which variable they act on. The monopoly path primarily raises switching costs: alternatives exist, but moving to them is too expensive, too slow, or too risky. The technology path primarily lowers substitutability: at the very frontier of capability, there is temporarily no second option. The ubiquity path changes the meaning of both variables. It does not stop customers from leaving; it makes leaving pointless, because it is already the cheapest, most convenient option, supported by default in the largest number of tools.

The three forms of irreplaceability also have different relationships with time. Monopoly-based irreplaceability rests on accumulated sunk costs, so it looks backward. Technology-based irreplaceability depends on staying ahead, so it looks forward. Ubiquity-based irreplaceability depends on position within an ecosystem, so it spreads sideways. That temporal orientation determines where each path is most vulnerable.

Nvidia: Irreplaceability Through Monopoly

How the Lock-In Was Built

Nvidia's position owes less to a precise forecast of the AI wave than to a historic fit. Graphics processors were designed to render game imagery, and their massively parallel architecture turned out to match the matrix arithmetic at the heart of deep learning. What converted that fortunate fit into a structural advantage was CUDA, the software platform Nvidia launched in 2006. By the company's own count, CUDA now underpins roughly six million developers [5]. Mainstream deep-learning frameworks, hundreds of optimized libraries, and vast amounts of production code are built on it. Migrating to other hardware means rewriting, retuning, and revalidating.

This is the classic structure of monopoly-based irreplaceability. The hardware itself is not impossible to replace. The cost of replacing it is the barrier. Third-party estimates of Nvidia's share of the AI accelerator market generally fall between 80% and 88%, and its share of the training market exceeds 90% [4].

Pricing Power and the Distribution of Profit

The most direct consequence of monopoly-based irreplaceability is pricing power. In the second quarter of fiscal 2027, Nvidia reported revenue of $96.2 billion, up 106% year over year, with a gross margin of 75.0% [1]. Data center sales accounted for about 93% of the total [2]. When a supplier sustains margins near 75% in an emerging market where much of the downstream industry has yet to turn a profit, most of the value chain's economics are being captured at the hardware layer.

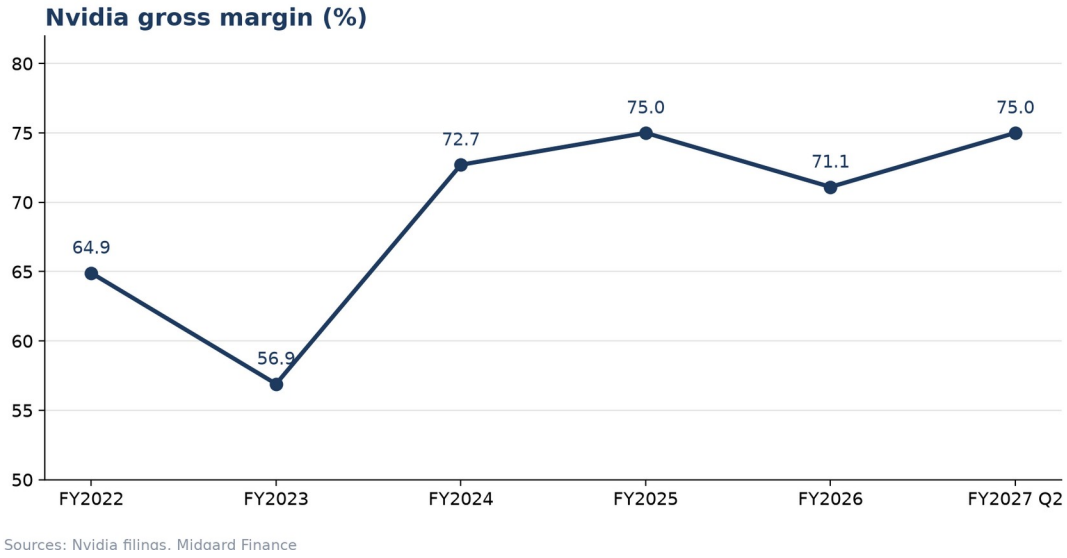


Figure 2. Nvidia gross margin

Strategically, that profit structure carries two implications. It demonstrates the power of lock-in: without adequate alternatives, customers accept high prices. It also creates a powerful counter-incentive. The more cloud providers, model developers, and application builders pay, the stronger their motivation to find a way out.

The Price of Monopoly: Product Discipline and the Substitution Backlash

Monopoly-based irreplaceability carries an inherent risk. When customers cannot leave, the external pressure to improve the product weakens. Nvidia's DGX Spark, a desktop AI system aimed at developers, offers a concrete example. On Nvidia's own developer forums, users have repeatedly reported freezes, unresponsive systems, and machines that recover only after a physical power cycle under sustained inference and training loads. For hangs triggered by out-of-memory events, a forum moderator confirmed that this was "a known issue which we are actively working to fix" [7], and Nvidia shipped multiple firmware updates during 2026 to address power and stability problems [7]. Problems with a single product do not define a company's entire technical capability. They do, however, reveal a structural tendency: when profits come mainly from lock-in rather than sustained product superiority, product refinement tends to lose priority to capacity expansion and the capital-markets narrative.

More importantly, outsized margins are accelerating the growth of alternatives. Custom silicon such as Google's TPU, Amazon's Trainium, and Meta's MTIA is projected to rise from 20.9% of AI processing workloads in 2025 to 27.8% in 2026 [6]. In China, Nvidia's guidance for the third quarter of fiscal 2027 already assumes zero data center compute revenue from the country [23]. In September 2026, DeepSeek open-sourced a full set of foundational software components for Huawei's Ascend platform, each matching a component it had previously released for Nvidia hardware [22]. Work of this kind targets the core of CUDA's lock-in: the cost of migration.

Here the weakness of monopoly-based irreplaceability becomes visible. It looks backward and depends on lock-in accumulated in the past. Once a rival ecosystem systematically lowers migration costs, yesterday's sunk costs stop being a moat and become a burden customers are eager to shed.

Moonshot AI: Irreplaceability Through Technological Excellence

Occupying the Capability Frontier

Moonshot AI chose a different route. Kimi K3, released on July 16, 2026, has 2.8 trillion total parameters, making it the largest open-weight model released to date. It offers native vision and a one-million-token context window, and its full weights were published on July 27 [8]. It debuted third on the Artificial Analysis Intelligence Index, behind flagship models from Anthropic and OpenAI [8], and became the first Chinese model to reach the top of the front-end coding arena [8]. Its October successor, K3.1 Pro, scored 55 on the same index, 11 points higher than its predecessor [13].

The logic of technology-based irreplaceability is straightforward. In long-context work, long-horizon reasoning, and complex programming, a customer who needs the best result has only a handful of options. Moonshot does not compete on being the cheapest. K3's API is priced at $3 per million input tokens and $15 per million output tokens [10], roughly half the price of comparable U.S. closed-source flagships [9] but well above DeepSeek. Moonshot sells capability, not cheapness.

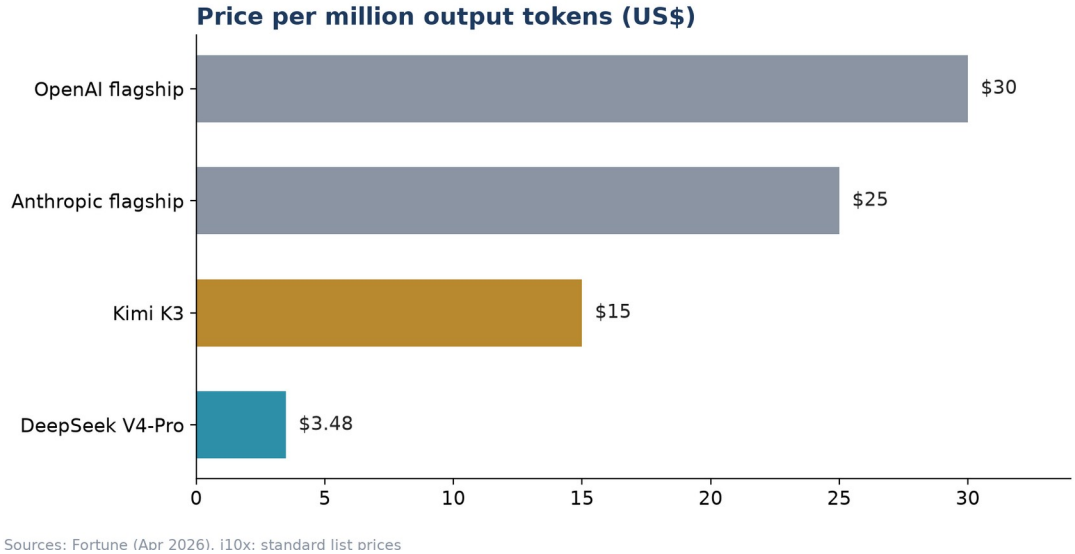


Figure 3. Price per million output tokens, frontier models

A Carefully Drawn Line Between Open and Paid

What sets Moonshot apart is that pursuing the frontier did not lead it to close its technology. K3 was released with open weights, and the vast majority of developers and companies can download, modify, fine-tune, and use it commercially at no license cost. The license, however,

sets two thresholds. First, any licensee operating a model-as-a-service business, offering inference or fine-tuning to third parties, must sign a separate agreement with Moonshot before commercial use if its aggregate revenue, including affiliates, exceeds $20 million over any consecutive 12 months. Second, any commercial product with more than 100 million monthly active users or more than $20 million in monthly revenue must prominently display “Kimi K3” in its interface [11].

The design pairs the diffusion benefits of openness with the commercial value of frontier technology. For small developers, K3 is a free frontier model. For large cloud platforms and model-hosting businesses, it is a commercial asset that requires negotiation. Its general availability on Amazon Bedrock from September 2026 is the first major international test of that boundary [17].

Capital Prices Scarcity

Capital markets have put a clear price on technology-based irreplaceability. Moonshot closed a $3.5 billion round at a $35 billion valuation in July 2026 [25], and it has reportedly since raised a private round at a valuation of about $50 billion ahead of a planned Hong Kong IPO in the first quarter of 2027 [12]. Caixin reports that its annualized recurring revenue has reached about $300 million [8]. The wide gap between valuation and revenue is, in essence, a premium on scarce capability.

Where the Technology Path Is Fragile

Technology-based irreplaceability looks forward, so its greatest risk is that the frontier itself keeps moving. Today’s third place can fall out of the top five after the next round of releases, and any generational lag erodes the foundation of irreplaceability directly. Frontier technology also demands enormous compute. After K3’s launch, demand surged so sharply that Moonshot temporarily had to pause new API subscriptions [9]. Technological leadership without sufficient serving capacity struggles to become a durable market position. The technology path requires relentless investment in research, and any slackening is punished quickly.

DeepSeek: Irreplaceability Through Continuous Supply and Market Ubiquity

Win the Market First, Advance the Technology in Parallel

DeepSeek has taken the third path. Its core aim is not to top every benchmark, but to use remarkable engineering efficiency to keep supplying technology that is good enough, at prices low enough to make it ubiquitous. In the V4 series launched in April 2026, V4-Pro was priced at $3.48 per million output tokens, against $30 from OpenAI and $25 from Anthropic for comparable work [14]. DeepSeek later introduced time-of-day pricing that halves rates during off-peak hours. Its web and mobile apps are entirely free with no paid tier, and its model weights are released under the MIT license for unrestricted self-hosting [15].

Equally important is the continuity of supply. From the V4 preview in April to the general release of V4-Pro in August, and on to V4.1-Flash and agent-oriented search and fetch APIs in September, priced at $0.75 per 1,000 requests and well below comparable services [15][16], DeepSeek has used a rapid sequence of releases to show that it is not a one-time price shock but a supplier that can be relied on for the long term.

How Ubiquity Creates Irreplaceability

The results of the ubiquity path show up in usage data. On OpenRouter, the largest neutral model-routing platform, models from Chinese developers carried 72% of token traffic in the week to November 1, 2026, up from 7% in the first week of 2025 [17]. In the week beginning October 5, 2026, DeepSeek alone accounted for 24.5% of text requests, ahead of OpenAI at 18.6% and Google at 17.8% [17]. On Hugging Face, cumulative downloads of Chinese open models reached 1.15 billion by March 2026, compared with 723 million for U.S. models [18].

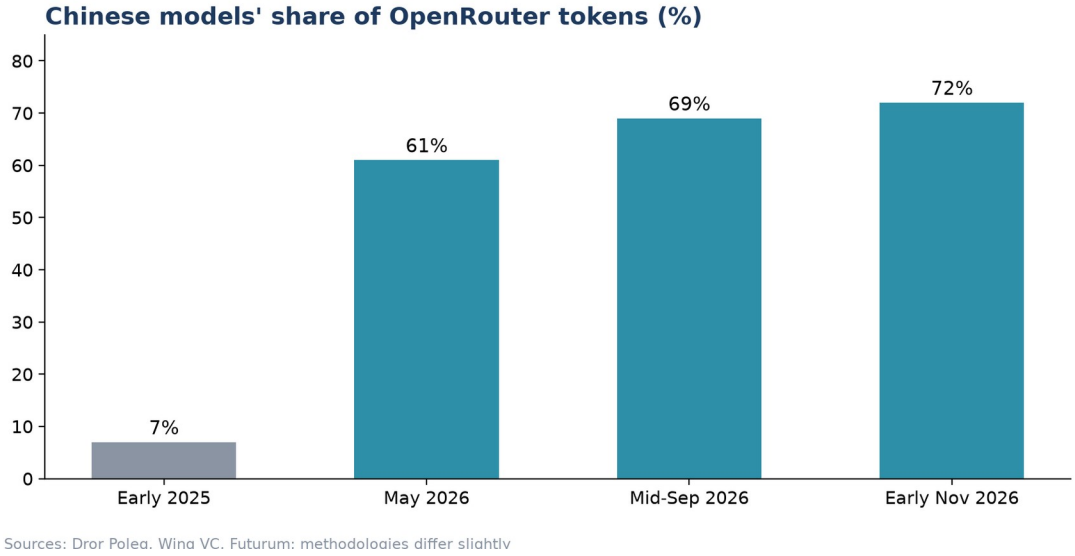


Figure 4. Chinese models’ share of OpenRouter tokens

This form of irreplaceability is the opposite of Nvidia-style lock-in. DeepSeek does not prevent customers from leaving; anyone can download, deploy, and modify its weights. Its irreplaceability comes from a different mechanism. When a model becomes the default option in a large number of agent frameworks, developer tools, and downstream applications, it becomes the shared foundation of an ecosystem. Replacing it is technically possible but economically unnecessary. Vercel’s AI Gateway data for June 2026 found that open-weight models handled nearly a third of tokens while accounting for less than 4% of spend [24]. Those figures capture the business signature of the ubiquity path: a very low unit price exchanged for very large scale.

Capital and the Road to Listing

Capital markets have validated the ubiquity path as well. According to Bloomberg, DeepSeek is close to raising at least 80 billion yuan in its second funding round, led by Tencent and CATL, at

a valuation of at least 500 billion yuan, or about $75 billion, ahead of a planned listing in early 2027 [19]. Several reports point to Shanghai's STAR Market as the intended venue [19].

Where the Ubiquity Path Is Fragile

The ubiquity path spreads sideways, and its greatest vulnerabilities lie not in technology or price but in two external constraints. The first is serving capacity. DeepSeek itself acknowledges that, for lack of chips, it cannot yet serve its V4-Pro model to most customers [21]. The enormous demand generated by low prices must be met with enough compute. The second is political risk. A U.S. Department of Homeland Security order made public in October 2026 requires federal civilian agencies to remove DeepSeek products, along with third-party products built on its technology, from their systems, and bars them from procuring such products [20]. The order does not apply to private businesses or individual users [20], but it shows that once a company's reach is large enough to reshape an industry, it is no longer treated simply as a commercial competitor.

Comparing the Three Paths

Dimension	Nvidia: monopoly	Moonshot AI: technology	DeepSeek: ubiquity
Source of irreplaceability	Ecosystem lock-in; very high switching costs	Capability frontier; no near-term substitute	Default option; no reason to switch
Orientation in time	Backward: accumulated sunk costs	Forward: sustained leadership	Sideways: ecosystem position
Pricing logic	Premium; gross margin about 75%	Mid-to-high; about half of U.S. closed flagships	Very low; free apps, open weights
Value capture	Hardware sales and system premiums	API fees; separate terms for large platforms	Low-cost API at scale; capital markets
Effect on the ecosystem	Concentrates value upstream	Opens frontier capability, with conditions	Pushes value downstream
Principal risk	Rival ecosystems lower migration costs	Moving frontier; compute shortages	Serving capacity; political restrictions

The comparison reveals a crucial difference. The three paths push value through the industry in opposite directions. The monopoly path draws value from downstream to upstream, and the more successful it is, the stronger the downstream incentive to find substitutes. The ubiquity path pushes value from upstream to downstream, and the more successful it is, the more the downstream depends on it. The technology path sits between the two, using conditional openness to balance diffusion against monetization.

The first conclusion follows directly. The durability of irreplaceability depends on whether it is aligned with the customer's interest. Monopoly-based irreplaceability is in structural tension with that interest: customers stay because leaving is too expensive, not because staying is best. Such a position is solid, but every day it gives would-be substitutes more reason to try. Ubiquity-

based and technology-based irreplaceability are broadly aligned with customers, who choose them because they are genuinely cheaper or genuinely better.

The second conclusion is that late movers are particularly well suited to the latter two paths. First movers bear the full cost of pioneering, experimentation, and commercialization, and they carry investor expectations of high margins that make voluntary price cuts or open releases difficult. Late movers can concentrate resources on technical directions that have already been validated and enter through efficiency or a decisive breakthrough on a single dimension. With no existing profit pool to protect, they can use openness and low prices to rewrite the industry's pricing rules. Moonshot AI and DeepSeek have each pushed late-mover advantage to its limit along a different dimension: one guards the capability frontier through quality, while the other permeates every corner of the ecosystem through efficiency.

The Largest Variable: State Power

Beyond the three commercial paths lies a variable that is not commercial at all but could change everything: state power. U.S. export controls on advanced chips limit Chinese model developers' access to the compute needed to train frontier systems. The Department of Homeland Security's federal procurement ban on DeepSeek extends competition from the marketplace into compliance and security review [20]. Capital markets may be affected too: restrictive measures could shape international institutional investors' willingness to take part in these companies' fundraising and listings.

Yet the open-weight approach has shown considerable resilience under this pressure. First, weights that have already been downloaded and deployed by developers around the world cannot be recalled; no executive order can delete them from millions of dispersed private servers. Second, downstream dependence creates counter-pressure: forcibly cutting ties with the most cost-effective models would first damage the cost structures of a country's own developers and businesses. Third, global capital is already widely dispersed. Gulf sovereign wealth funds, Asian family offices, and international institutional investors do not take their cues from a single government, and financing rounds measured in tens of billions of dollars do not depend on any one source.

State power can raise the cost of an irreplaceability strategy, but it struggles to dislodge a technology that is already embedded in the daily work of developers worldwide. The binding constraint is compute. As long as advanced chips remain restricted, both the technology path and the ubiquity path must keep investing in algorithmic efficiency and domestic hardware. DeepSeek's full adaptation of V4 inference to Huawei's Ascend platform, and its release of the corresponding open software stack, is a direct response to that constraint [21][22].

Implications for Corporate Strategy

The strategy of irreplaceability is not reserved for AI giants. Companies in every industry must answer the same question: if our customers left tomorrow, what would they lose, and what would it cost them?

First, a company must decide which form of irreplaceability it is pursuing. Monopoly-style lock-in can deliver the highest short-term profits, but it conflicts with customer interests and sooner or later invites substitution. Technological leadership can command a scarcity premium, but it demands continuous research investment with no finish line. Market ubiquity can secure the most durable ecosystem position, but it requires exceptional cost efficiency and the capacity to serve demand at scale.

Second, the paths can be combined, but they should not be confused. Moonshot's license shows that a technology leader can borrow the diffusion effects of openness, and DeepSeek's rapid iteration shows that a ubiquity player must still maintain a credible technical standard. A company that tries to command the highest prices, the largest share, and the tightest lock-in at once usually ends up half-committed on all three.

Third, the most durable irreplaceability comes from customers' active choice, not from their lack of alternatives. That is the most fundamental difference between Nvidia and the two Chinese model developers. Nvidia's position rests on "leaving costs too much"; theirs rests on "staying is better." In an industry where technology turns over in months, the second kind of irreplaceability is more likely to survive the cycle.

Conclusion

Nvidia, Moonshot AI, and DeepSeek represent three paradigms of irreplaceability. Nvidia spent two decades building an ecosystem that maximizes switching costs, earning the richest profits in the value chain and cultivating more would-be replacements than any other company. Moonshot AI holds the capability frontier through technological excellence, balancing diffusion and monetization through conditional openness. DeepSeek, through continuous supply and very low prices, has made itself the default foundation for developers around the world.

All three paths lead to irreplaceability, but to different kinds of it. One relies on customers being unable to leave; the other two rely on customers not wanting to. The next phase of the AI industry may well be decided by the contest between those two logics.

References

[1] NVIDIA, "NVIDIA Announces Financial Results for Second Quarter Fiscal 2027." https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2027

[2] Midgard Finance, "NVIDIA Revenue, Data Center Growth and Financials." https://www.midgardfinance.com/research/nvidia/

[3] Yahoo Finance, “Nvidia Is Getting Dangerously Close to Getting Surpassed.” https://finance.yahoo.com/markets/stocks/articles/nvidia-getting-dangerously-close-getting-165800696.html

[4] Axis Intelligence, “Nvidia GPU Market Share 2026.” https://axis-intelligence.com/nvidia-gpu-market-share/

[5] The Dynamics, “NVIDIA CUDA, Explained: The Software Behind NVIDIA’s AI Moat.” https://thedynamics.ai/articles/nvidia-cuda

[6] Axis Intelligence, “Nvidia Blackwell Statistics 2026.” https://axis-intelligence.com/nvidia-blackwell-statistics/

[7] NVIDIA Developer Forums, “Spark hangs – requires a hard-reset.” https://forums.developer.nvidia.com/t/spark-hangs-requires-a-hard-reset-physically-unplugging/358951; GPUsmith, “NVIDIA DGX Spark Specs, Price & Benchmarks.” https://gpusmith.com/articles/en/nvidia-dgx-spark-review-specs-performance

[8] Kimi, “What Is Kimi Agent?” https://www.kimi.com/en/help/agent/agent-overview; South China Morning Post, “Moonshot AI’s Kimi chatbot.” https://www.scmp.com/topics/moonshot-ais-kimi-chatbot; Caixin Global, “Inside the $50 Billion Rise of Moonshot AI.” https://www.caixinglobal.com/2026-09-11/in-depth-inside-the-50-billion-rise-of-moonshot-ai-102484028.html

[9] ThursdAI, “Moonshot AI Releases: Timeline of Every Launch.” https://thursdai.news/companies/moonshot-ai

[10] i10x, “Moonshot AI.” https://i10x.ai/news/company/moonshot

[11] Moonshot AI, “Kimi K3 License,” Hugging Face. https://huggingface.co/moonshotai/Kimi-K3/raw/main/LICENSE

[12] Sacra, “Moonshot AI revenue, valuation & funding.” https://sacra.com/c/moonshot-ai/

[13] China Daily via NCSTI, “Moonshot AI launches upgraded Kimi model.” https://en.ncsti.gov.cn/Latest/news/202610/t20261029_260063.html

[14] Fortune, “DeepSeek unveils V4 model, with rock-bottom prices.” https://fortune.com/2026/04/24/deepseek-v4-ai-model-price-performance-china-open-source/

[15] The AI Rankings, “DeepSeek V4: Benchmarks, Pricing & Specs.” https://theairankings.com/deepseek/deepseek-v4/

[16] Yahoo Finance, “DeepSeek Monetizes the Agent Retrieval Layer at $0.75 per 1,000 Calls.” https://finance.yahoo.com/technology/ai/articles/deepseek-monetizes-agent-retrieval-layer-185236272.html

[17] Dror Poleg, “Chinese and Open-Weight Models’ Share of OpenRouter Tokens.” https://www.drorpoleg.com/charts/openrouter-china-models/; Asia Daily, “Chinese Open AI Models Win Overseas Users.” https://asiadaily.org/news/15338/

[18] “The ATOM Report: Measuring the Open Language Model Ecosystem,” arXiv. https://arxiv.org/html/2604.07190

[19] Oninvest (Bloomberg), “DeepSeek set to raise $12 bln in new funding.” https://en.oninvest.com/article/bloomberg-deepseek-set-to-raise-12-bln-in-new-funding-that-bumps-valuation-by-50; Big Hat Group, “China AI Weekly.” https://www.bighatgroup.com/blog/china-ai-weekly-2026-08-30/

[20] CNBC, “U.S. bans DeepSeek from government agency systems.” https://www.cnbc.com/2026/10/20/us-deepseek-federal-agency-ban.html

[21] Council on Foreign Relations, “DeepSeek V4 Signals a New Phase in the U.S.-China AI Rivalry.” https://www.cfr.org/articles/deepseek-v4-signals-a-new-phase-in-the-u-s-china-ai-rivalry

[22] Frontier Now, “DeepSeek opens Ascend stack.” https://frontiernow.dev/en/news/deepseek-ascend-open-components/

[23] Morningstar, “NVIDIA Corp earnings.” https://morningstar.com/stocks/xmex/nvda/earnings

[24] NewsCord, “Chinese Open-Weight Models Surpass U.S. Models on Hugging Face and OpenRouter.” https://newscord.org/article/chinese-open-weight-models-surpass-us-models-on-hugging-face-and-openrouter–Story_20260626_ItsnotaboutAnthropice54000ee

[25] Stock Analysis, “Moonshot AI Valuation.” https://stockanalysis.com/private/moonshot-ai/

[26] Futurum Group, “Enterprises Embrace Open Weights but Infrastructure Sellers Profit Most.” https://futurumgroup.com/insights/enterprises-embrace-open-weights-but-infrastructure-sellers-profit-most/