Why Is China Giving Away Its AI Models?

Five Possible Explanations for China's Open-Weight Strategy

Pranay Kotasthane, Nitin Pai, Bharath Reddy

Takshashila Discussion Document 2026-25
Version 1.0, August 2026

This discussion document analyses the strategic rationale behind the proliferation of Chinese open-weight artificial intelligence models.

Recommended Citation:
Pranay Kotasthane, Nitin Pai, Bharath Reddy, ``Why Is China Giving Away Its AI Models?'', Takshashila Discussion Document 2026-25, Version 1.0, August 2026, The Takshashila Institution

Contents

1 Executive Summary

This discussion document analyses the strategic rationale behind the proliferation of Chinese open-weight artificial intelligence models. China's strategy is driven by five reinforcing factors: significantly lower training costs achieved through distillation and architectural efficiencies; the pursuit of geopolitical prestige and influence; a competitive drive to undermine the proprietary revenue models of American frontier labs; the ability to absorb significant R&D costs through state-led capital mobilisation and financial repression; and a long-term play to drive global adoption of Chinese full-stack cloud infrastructure.

The paper argues that this strategy is supported by multiple independent logics that make it more durable. However, this openness is not necessarily permanent. Using a causal loop framework, the document identifies three thresholds—domestic consolidation, ecosystem lock-in, and commoditisation saturation—that, once crossed, may lead the Chinese state to shift toward graduated restrictions. The paper predicts that such restrictions are likely begin around late 2028, could include delayed international releases or gate access to highly capable frontier models.

For India, this environment presents both an opportunity for low-cost AI integration and a strategic risk of technological path dependence. The paper concludes that India must navigate this landscape by maintaining a multi-ecosystem approach, avoiding lock-in to either American or Chinese stacks, while building domestic capacity in selective segments such as applications, industrial data, and domain-specific fine-tuning to ensure strategic autonomy.

2 Introduction

In the early consumer internet era, American technology companies perfected a playbook—give away the product, monetise the attention. Google offered free email, while Facebook was a free social network. The implicit bargain was users paying with data and eyeballs.

Frontier artificial intelligence has broken from this pattern. OpenAI, Anthropic, and Google treat their most capable model weights as proprietary assets. Access comes through paid APIs and subscription tiers, and the business model depends on retaining control over the weights.

Instead, Chinese AI firms seem to be the torchbearers of the strategy deployed by American firms in the consumer internet era. DeepSeek, Alibaba (Qwen), Tencent, Moonshot AI, and others have released increasingly capable models with open weights. The scale of their global adoption is dramatic. According to an OpenRouter study analysing 100 trillion tokens, Chinese

Sridhar Krishna's inputs contributed to refining the ideas in this paper.

AI use disclosure: Claude Opus 4.8 was used to review the paper, brainstorm logical inconsistencies, and for generating causal loops.

Some American firms like Meta, Google and Nvidia continue to provide open-weight models but lag Chinese open-weight models in coding, agentic, and intelligence benchmarks.

open-source LLMs' global usage share surged from 1.2 per cent in late 2024 to nearly 30 per cent by late 2025. Qwen surpassed Meta's Llama as the most-downloaded open model on Hugging Face, crossing one billion cumulative downloads and anchoring over 200,000 derivative models.[1]

> Open-weight models are models whose core parameters are publicly released under a permissive license, allowing users to run the model on their own devices and even modify some parameters. Proprietary models are those whose parameters are not publicly released, remaining under the exclusive control of the developing organisation, which typically monetises access through paid APIs or subscription services. True open-source AI, according to the Open Source Initiative, entails a more rigorous standard, necessitating the public release of training code and data to allow for full system replication. This paper discusses proprietary and open weight models.

By allowing these models to be downloaded, firms cannot monetise these models. This decision warrants attention because training a frontier open weight model in 2026 is estimated to cost around 200 million [2]. Hence, most American firms have chosen to charge users for inference to defray these large training costs. So why have Chinese firms pursued this counterintuitive strategy? What could be the motivations, benefits, and costs? Under what conditions could Chinese firms reverse this strategy? These are questions that are relevant not just from a technical lens but also from a geopolitical perspective.

This document proceeds as follows. The next section will propose five possible explanations for China's moves. They are not mutually exclusive; several may operate simultaneously at different levels. The aim is to map the plausible hypothesis space, not identify a single answer. To do this, we draw on the views of researchers and practitioners across the US, China, and Europe who have written on this question, and situate their analyses within a common framework. The next section analyses the contradictions and complementarities between the reasons identified. The policy relevance of this section lies in imagining the possible pathways that Chinese open-weight models may take in the near future. The last section maps out implications for India based on this analysis.

3 A Survey of Reasons

3.1 Reason 1: It Costs Far Less to Train than American Models

The most parsimonious explanation is that the apparent generosity is cheaper than it appears. In September 2025, DeepSeek disclosed in a peer-reviewed *Nature* article that its R1 reasoning model cost $294,000 to train using 512 Nvidia H800 chips.[3] This figure is dramatically below the hundreds of millions reportedly spent by American frontier labs.

Two techniques explain the cost differential. First, distillation. OpenAI accused DeepSeek in early 2025 of using its API outputs to train competing models; Microsoft's security team reportedly detected large-scale data extraction linked to DeepSeek-affiliated accounts. Anthropic subsequently flagged similar distillation campaigns by three Chinese AI firms.[4] After Moonshot AI's Kimi-K3 closely followed Fable 5 on important coding benchmarks, the White House's director of science and technology policy, Michael Kratsios, accused Moonshot AI of conducting large-scale distillation against U.S. models.[5]

There is evidence of Chinese models such as Kimi K3 identifying itself as Claude when asked what is is which hints at distillation[6]. Similarly, Kimi K2 and Sonnet 4.5 exhibit agentic similarity exceeding that between pairs of Claude's own models[7]. Whether or not this constitutes IP theft, it radically compresses R&D costs, as the distilling lab freerides on the training effort and costs incurred by the base model's creator.

Model distillation is a technique commonly used by AI labs in which a less capable model is trained on the outputs of a stronger model.

The second explanation of the cost differential is architectural efficiency. DeepSeek's mixture-of-experts (MoE) architecture activates only a fraction of the model's parameters for any given query, reducing both training and inference costs. These innovations were developed under the pressure of US export controls that restricted Chinese access to Nvidia's most advanced H100 chips, forcing DeepSeek to extract maximum performance from the less capable H800.

Kai-Fu Lee, founder of 01.AI and former head of Google China, quantified this directly: "Chinese companies are spending less than ten percent of what the top American companies spend on training, yet are producing models that are ninety to ninety-five percent as capable, running six to nine months behind their American counterparts."[8]

Kevin Xu of Interconnected adds a mechanism that goes beyond distillation, called the "shared artifacts" effect. Because Chinese labs build on each other's publicly released models, they avoid the redundant infrastructure costs that American closed-source labs incur by working in silos.[9] Ion Stoica of Berkeley (co-founder of Databricks) estimates that a significant share of the US cloud AI data centre spending is redundant infrastructure in service of siloed labs.[10]

If costs are genuinely this low, open release may be rational firm behaviour without requiring a geopolitical explanation.

3.2 Reason 2: Geopolitical Prestige

Open-weight models are a visible demonstration that China can compete at the AI frontier. DeepSeek's January 2025 release triggered the largest single-day market capitalisation loss in US stock market history, wiping roughly $1 trillion in value from technology stocks.

While there is no evidence that the Chinese party-state was involved in HighFlyer's DeepSeek breakthrough, the positive responses this model received worldwide would have affected the government's calculus, delaying the moves to gatekeep these model weights. The government realised that making models open-source can become a tool of geopolitical influence, and instead chose to double down on open-weighting.

The signalling operates at two levels. Domestically, it validates the state's designation of AI as a strategic priority. Internationally, it tells potential partners that dependence on American AI providers is a choice, not a necessity. Xi Jinping's first-ever in-person appearance at the World Artificial Intelligence Conference in Shanghai in July 2026, where he announced a 29-country AI cooperation body (WAICO) and pledged 5,000 AI training spots for developing countries, was the clearest articulation of this logic. Mozilla's State of Open Source AI report notes that "a single-origin open commons stops being a commons", however a bloc of 29 countries gives it legitimacy[11].

Matt Sheehan has an excellent annotated version of Xi's speech at the World AI Conference in Shanghai here.

But the prestige argument has a more idealistic variant visible in the statements of Chinese founders themselves. Liang Wenfeng, DeepSeek's CEO, articulated this in a 2024 interview translated by ChinaTalk: "For many years, Chinese companies are used to others doing technological innovation, while we focused on application monetization — but this isn't inevitable. In this wave, our starting point is not to take advantage of the opportunity to make a quick profit, but rather to reach the technical frontier and drive the development of the entire ecosystem."[12]

In the same interview, he added: "In the face of disruptive technologies, moats created by closed source are temporary. Even OpenAI's closed source approach can't prevent others from catching up... In fact, open source is more of a cultural behavior than a commercial one, and contributing to it earns us respect."[13]

The Bulletin of the Chinese Academy of Sciences published a special issue in 2025 that institutionalised this view. Yanjun Wu of CAS's Institute of Software framed DeepSeek's innovation model around three pillars: "compensating hardware with software, acquiring users through open source, and ecosystem priority."[14] A companion paper by Long et al. called for building a national open-source AI ecosystem, arguing that DeepSeek had "raised China's AI research application level and international influence"[15]— the closest thing to an official Chinese institutional articulation of the prestige motive.

3.3 Reason 3: Undermining the Opponent's Business Model

This reason operates at the level of competitive strategy rather than national prestige. Open-weight models directly attack the revenue model of American frontier AI labs. OpenAI, Anthropic, and Google depend on the ability to charge for API access to models whose weights remain proprietary. Every time a Chinese open-weight model demonstrates competitive performance at zero cost, it compresses their pricing power.

This is the classic technology strategy of commoditising your competitor's core product. One does not need to beat the opponent's product; one only needs to destroy their ability to monetise it. The open-weight models need not be strictly superior—they merely need to be good enough for most commercial applications. The pressure on American AI labs' revenue weakens their ability to self-fund the next generation of frontier research, narrowing over time the capability gap that justifies their premium pricing.

Mozilla's State of Open Source AI report states that models are getting commoditised and the agentic harness layer – governing what a system can see, remember, and do – is where value gets captured[16]. The frontier labs such as OpenAI and Anthropic have responded by pulling the harness in-house and tightly coupling it with their models. Operating in the agentic layer also requires trust and not just capabilities since this requires having access to credentials and taking actions on behalf of users. While Chinese AI labs might face a trust deficit in being able to serve such capabilities to global customers, they can still address these uses for domestic and WAICO markets.

Kai-Fu Lee framed this structural logic clearly: "There is... a structural reason Chinese companies have embraced open source. They cannot win on a closed-source basis, so they share. It functions like a study group; not an explicit collaboration where people sit down together, but a dynamic where everyone builds on each other's public releases and the collective competency rises."[17] Kevin Xu wrote as early as 2023 that open source can be used as "a defensive business lever" to render competitors "moatless"—the same analysis he applied to Meta's decision to open-source Llama.[18]

Yang Zhilin, the founder of Moonshot AI, explained this logic in a couple of interviews. In March 2024, before Kimi had gone open-weight, he adopted a realist stance on this question:

"Closed source enjoys concentrations of talent and capital, so in the end closed source will definitely be better—it's a consolidation. If I had a leading model today, open-sourcing it would very likely be irrational. It's the laggards who might do that instead, or open-source a small model—to stir things up; after all, if you don't open-source it, it has no value anyway."[19]

In August 2025, after Kimi K2 was released, when asked if open-weighting was a "a belief in a technical system, or a strategy of market maneuvering", Yang acknowledged that it was a mix of both. Specifically, he outlined that open-weighting does not help in improving the base model, but can help find more downstream customers. An example he cites is instructive: "Suppose you really want to build a law-related agent, and you're a startup. You can absolutely train a specialized agent on top of K2, with your own specific set of tools, and it can perform extremely well in the scenarios you care about. That kind of opportunity exists."[20]

3.4 Reason 4: China Can Absorb the Costs

This explanation is based on China's macroeconomic policy and market structure. Even if open-weighting is more expensive than Reason 1 suggests, China's capacity to mobilise capital for strategic technology investments is formidable. Since 2014, China's state-led investment in semiconductors alone has exceeded $150 billion, roughly three times the US CHIPS Act allocation.[21] In May 2024, the third round of China's Integrated Circuit Industry Investment Fund ("Big Fund III") raised an additional $47.5 billion.[22] Provincial governments offer compute vouchers covering 20 to 50 per cent of cloud computing costs for model training.[23]

The deeper explanation for how China can afford this stems from financial repression. By maintaining strict capital controls and a state-dominated banking system, Beijing prevents Chinese households from diversifying their savings abroad. This ensures that domestic savings are held captive within state-owned banks. These captive savings provide the state with a virtually limitless pool of artificially cheap capital. For example, Chinese scholars have documented the role played by government guidance funds (GGFs) in strategically injecting subsidised capital into AI startups and full-stack cloud providers.[24]

This immense macroeconomic fuel inevitably ignites chaotic, firm-level overinvestment, a well-documented pattern previously seen in China's solar panel and electric vehicle (EV) sectors. State signalling and cheap capital produce massive overcapacity, making Chinese producers the default global suppliers at prices competitors cannot match.

In AI, this dynamic first triggered the *bai mo dazhan* ("Hundred Model War").[25] By early 2026, 820 LLMs were registered with the Cyberspace Administration of China (CAC).[26] In recent months, there is a trend towards consolidation, where the few survivors of this competition have taken the lead.

For example, Joe Tsai, Alibaba's chairman, revealed that after the DeepSeek breakthrough, the company's AI team "called off everyone's holiday travel plan, stayed at the office, and slept on the floor until a competitive model was shipped". This is extreme firm-level competition that works alongside state direction.[27]

In short, financial repression provides the subsidised gas tank, while fierce domestic rivalry provides the fire. In a market flooded with hundreds of heavily capitalised competitors waging a scorched-earth price war, prices naturally crash to zero, making open-weight release the default equilibrium for survival.

3.5 Reason 5: The Infrastructure Play

The most interesting explanation concerns what happens downstream of model release. Free models accelerate global AI adoption. Global AI adoption drives demand for energy, physical infrastructure, and hardware. China dominates the global supply of all these downstream domains.

If we adopt Jensen Huang's five-layer AI stack classification—energy, chips, infrastructure, models, and applications—Chinese open-weight models are at layer four. China's export strengths are concentrated at layers one and three. The International Energy Agency reports that China controls over 80 per cent of every stage of solar panel manufacturing.[28] Chinese solar exports to Africa and Southeast Asia have ballooned. Moreover, a large chunk of Chinese renewable energy engagement takes the form of engineering, procurement, and construction contracts.

Joe Tsai articulated the business logic explicitly: "The way we benefit from open source is that... it will drive demand for AI, it will drive training needs, and we see in the future a lot of needs for inference."[29] At the World Government Summit 2026, he distinguished between "pure-play model developers" who struggle to monetise open source, and "full-stack technology providers like Alibaba and Google, which operate across the entire value chain from cloud infrastructure, foundational models, to applications."[30] Alibaba's cloud revenue grew 34 per cent year-on-year in the September 2025 quarter, with AI-related revenue posting triple-digit growth — driven in significant part by developers building on free Qwen models using Alibaba's paid cloud infrastructure.[31]

The USCC's "Two Loops" report (March 2026) provides the most rigorous articulation of this mechanism. It argues that China's open-source strategy operates through two mutually reinforcing feedback loops: a digital loop (open-model innovation, community-driven iteration) and a physical loop (large-scale deployment across manufacturing, logistics, and robotics that generates real-world data feeding back into model improvement). The report concludes that this is "not a second-best adaptation to semiconductor export controls, but a coherent doctrine."[32] Crucially, the physical loop generates proprietary industrial data that compounds China's advantage over time, a mechanism that US export controls, designed to target the digital loop by restricting training chips, are structurally incapable of addressing.

3.6 A Counter-Argument: The Inference Scarcity Myth

Before synthesising how these five motivations interact, it is necessary to discuss another oft-cited mechanism that Chinese labs open-weight their models primarily because US export controls deny them the compute needed to serve these models via APIs. According to this view, open-weighting is not a deliberate strategy, but a desperate workaround to outsource inference hosting to the global developer community.

To evaluate this claim, we must adopt the conceptual distinction articulated by Amit Kumar, which separates AI compute into a stock (training) and a flow (inference). In this schema, training a model is a stock problem. A lab accumulates a fixed pool of hardware, runs an intensive workload over several weeks or months, and produces a static set of weights. A lab can build a training stockpile through domestic production, stockpiling, and grey-market procurement.[33]

Whereas serving a model (inference) is a flow problem. Compute requirements scale linearly with active users and query volume. Western giants like OpenAI and Meta have assembled serving fleets approaching or exceeding one million GPUs. The serving fleet for a mass-market consumer AI routinely dwarfs its training fleet by ten to a hundred times. Because semiconductor export controls restrict China's access to advanced chips like Nvidia's Blackwell or H100 series, and domestic foundry alternatives like SMIC's 5nm process suffer from lower yields (with rumours suggesting yields of around 30% without EUV tools), Chinese firms face challenges matching the massive physical serving fleets of their American counterparts.

However, while the stock-versus-flow distinction is accurate, inference scarcity does not explain why Chinese labs open-weight their models. The hypothesis falls apart under empirical scrutiny for two reasons:

3.6.1 1. Empirical Overcapacity, Not Structural Shortage

The premise that China lacks inference compute is contradicted by ground realities. Industry analysis has documented significant inference *overcapacity* across Chinese cloud providers, with some describing a mismatch that exposed deep flaws in China's attempt to merge AI ambitions with energy planning. Chinese cloud providers have even faced pressure to increase rates for AI computing services to manage costs amidst this rapid expansion. Furthermore, inference does not require cutting-edge 3nm or 5nm silicon; legacy and export-compliant chips (such as Nvidia H800s, A100s, or Huawei Ascend 910Bs) handle standard inference workloads effectively[34].

When DeepSeek temporarily froze new user registrations in January 2025, it was widely misread as a hardware bottleneck. But it was a capacity-planning failure triggered by an unprecedented viral traffic spike as DeepSeek-R1 rapidly climbed the App Store rankings, surpassing ChatGPT to become the most downloaded free app in the US.

3.6.2 2. Open Weights Are Preferred Even When There Is No Compute Scarcity

If compute scarcity were the predominant driver of open-weighting, we would observe a clear pattern: compute-poor labs would open-weight their models to offload serving costs, while compute-rich infrastructure providers would keep their models closed behind paid APIs.

The observed behaviour in China is the exact opposite. Alibaba, which possesses the largest cloud infrastructure and GPU footprint in China, is the nation's most aggressive and prolific open-weight releaser with its Qwen series. Conversely, smaller, compute-constrained startups have frequently attempted to gatekeep their models to extract direct API rents before being forced to adapt. The firms with the most compute are leading the open-weight charge, invalidating the compute-scarcity causality.

Similarly, Meta's behaviour of open-weighting even while operating under zero export controls and possessing one of the world's largest GPU clusters indicates that it is a strategic choice.

Inference constraints are, at most, a secondary factor that removes one potential operational friction against open release. They are not the strategic driver.

4 The Five Reasons: Overlaps, Tensions, and Predictions

The five explanations are not of equal analytical status. They can be divided into two groups: reasons that rest on assumptions about the same empirical facts but reach different conclusions, and reasons that are complementary.

4.0.1 Contradictory premises but same direction

Reasons 1 and 4 rest on opposing assumptions about how much open-weighting costs. Reason 1 says costs are low enough that open release is rational without state support. Reason 4 says costs may be substantial but absorbable thanks to China's capital mobilisation infrastructure. Both support open-weighting, but for different reasons. Empirically, both may be partially true: distillation and architectural efficiency reduce costs (supporting Reason 1), but the remaining costs are large enough that subsidised capital environments still matter (supporting Reason 4). Similarly, Reasons 2 and 4 rest on opposing assumptions about intent. Reason 2 posits conscious geopolitical strategy. Reason 4 posits structural overinvestment without any explicit geopolitical aim. They can coexist if state-level actors exploit a structural dynamic they did not create, i.e., Beijing recognises the open-weight wave's geopolitical utility and amplifies it without having initiated it.

4.0.2 Complementary Reasons

Reasons 3 and 5 are mutually reinforcing. Reason 3 attacks the opponent's revenue model while Reason 5 builds China's own. Together, they describe a pincer strategy of compressing American AI labs' ability to monetise proprietary models while simultaneously capturing value at the infrastructure layer.

4.1 The system as a whole

Taken together, the five reasons suggest that China's open-weight strategy is overdetermined, i.e., it is supported by more independent logics than would be necessary to justify it. Strategies with multiple reinforcing rationales tend to be durable. Even if one reason weakens, others remain in force. This overdetermination is itself significant for the predictive question we tackle next: under what conditions would Beijing reverse course?

5 What has to be true for the CPC to restrict frontier model exports?

Rather than viewing the five reasons as isolated theories, we can combine them into a causal model that identifies the structural conditions for reversal. The accompanying causal loop diagram (Figure 1) maps the system's reinforcing and balancing dynamics. This section uses that structure to reason about the future.

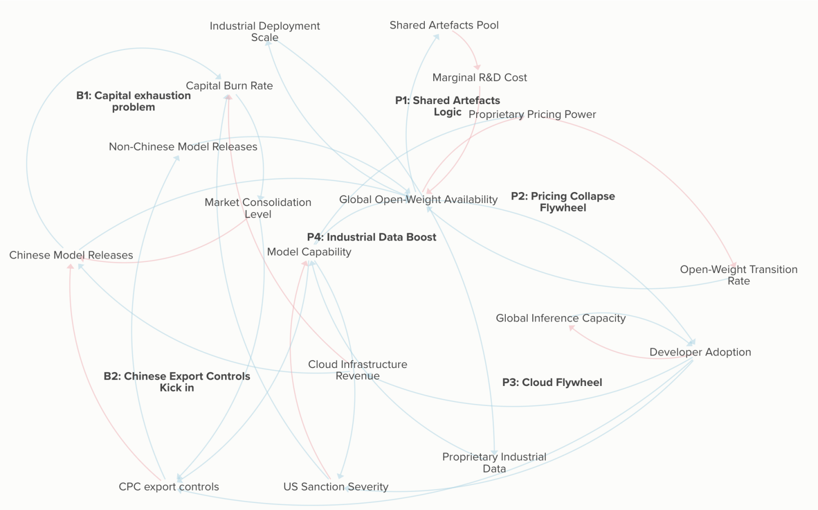

Figure 1: Causal loop diagram mapping the system's reinforcing and balancing dynamics

5.0.1 The current equilibrium is that reinforcing loops dominate

The open-weight ecosystem is currently governed by four reinforcing loops that drive its rapid acceleration. We have extracted them from the figure above to illustrate the logic.

P1 (Shared Artefacts Logic): Open-weight releases allow labs to build on each other's publicly available weights, reducing marginal R&D costs and enabling further releases. This is Kevin Xu's "study group" dynamic. It operates both domestically (Chinese labs building on each other) and globally (any lab building on any open-weight model).

P2 (Pricing Collapse): As open-weight models approach capability parity with closed models, they destroy the proprietary pricing power of American frontier labs. This forces more labs to go open, which increases global open-weight availability, further eroding pricing power. Two second-order effects strengthen this loop beyond what a first-order analysis captures. First, it is not just the availability of open models but their capability that determines the speed of pricing collapse. Second, as more global labs are forced to go open, the loop becomes self-sustaining regardless of any single country's contributions.

P3 (Cloud Flywheel): Open-weight releases attract developer adoption. Developers building on free models use paid cloud infrastructure to deploy them, generating cloud revenue for full-stack providers. This revenue funds further model development, enabling more releases. This loop operates at the firm level; it benefits Alibaba and Tencent specifically, not pure-play model labs.

P4 (Industrial Data): Open-weight models deployed in Chinese manufacturing and logistics generate proprietary industrial data, which feeds back into model improvement. This is the USCC's "physical loop." It is the slowest reinforcing loop—industrial data compounds over years, not months—but also the most durable, as it generates advantages that cannot be replicated through distillation or shared artefacts.

Two balancing loops constrain the system, but neither has yet reached its activation threshold.

B1 (Capital Exhaustion and Consolidation): The *bai mo dazhan* burns capital at an unsustainable rate. Smaller labs exit, the market consolidates, and the competitive pressure that forces open release weakens. A second-order effect accelerates this loop: US sanctions increase the cost of training by forcing reliance on expensive domestic alternatives, paradoxically hastening the consolidation that could eventually reduce open-weighting. Conversely, cloud revenue decelerates this loop for full-stack providers as firms like Alibaba burn capital more slowly because they generate revenue from the very models they give away.

B2 (CPC Export Control Pressure): As Chinese open-weight models achieve global market dominance, pressure builds on Beijing to restrict them for dual-use national security reasons, for diplomatic leverage, or to enable rent extraction by surviving firms. This loop is currently latent but not inactive yet.

The current equilibrium favours openness because all four reinforcing loops are active and the balancing loops have not reached their thresholds. The question is when the balance tips.

5.0.2 Three thresholds for restriction

A second-order analysis of the causal structure (Figure 2) reveals three thresholds that must be crossed, roughly in sequence, before restriction becomes rational.

Figure 2: Analysis of the causal structure. The dotted line indicates the threshold is at an early stage.

Threshold 1: Domestic Consolidation. The *bai mo dazhan* must consolidate from hundreds of labs to an oligopoly of five to eight full-stack giants. Beijing can coordinate five firms; it cannot coordinate seven hundred. Without consolidation, restriction is impractical regardless of strategic desire. This threshold is underway. Xinhua has already reported the shift from the "Hundred Model War" to "Top Five Basic Models." The survivors are precisely the full-stack providers—Alibaba, ByteDance, Tencent, Baidu, and a handful of well-capitalised startups like DeepSeek and Moonshot AI—that have either cloud flywheels or deep-pocketed backers.

US sanctions accelerate this threshold through a counterintuitive second-order effect. Stricter export controls force Chinese labs onto expensive domestic alternatives (Huawei Ascend at approximately 60 per cent of Nvidia H100 inference performance, SMIC fabrication at lower yields), increasing the cost per training run and hastening the capital exhaustion of underfunded labs. The policy implication is that US semiconductor controls, intended to slow Chinese AI, may inadvertently accelerate the market consolidation that makes CPC restriction of open-weight models feasible.

Threshold 2: Ecosystem Lock-in The cloud flywheel (P3) must mature to the point where global developers and enterprise architectures are deeply integrated with Chinese cloud infrastructure. If models are restricted before this threshold, developers migrate to other open weight models, and the cloud flywheel breaks—the most costly outcome for Chinese full-stack providers. If restricted after lock-in, Alibaba and others can transition locked-in customers from free model downloads to paid proprietary API access.

This threshold is at an early stage. Qwen anchors over 200,000 derivative models on Hugging Face, and roughly 40 per cent of all new LLM derivatives on the platform are Qwen-based. Alibaba Cloud's AI-related revenue posted triple-digit growth in the September 2025 quarter. But derivative model counts and cloud revenue growth are not the same as enterprise-grade lock-in. True lock-in requires that switching costs—rewriting application code, retraining fine-tuned models, migrating inference infrastructure—are high enough to retain customers even after the free model advantage disappears. This condition is not yet met.

Threshold 3: Commoditisation Saturation. The marginal strategic value of further Chinese open releases must approach zero. This happens when two conditions converge: American frontier labs' proprietary pricing power has been sufficiently commoditised, and non-Chinese open-weight labs (Meta, Mistral, emerging European and Indian players) are independently sustaining the commoditisation regardless of Chinese contributions.

This threshold is *not yet reached*. American labs still retain meaningful pricing power. Enterprises pay premium prices for the latest proprietary models from OpenAI and Anthropic. But the gap is narrowing. Chinese model providers have compressed the time-to-parity with US labs to roughly six months. The key second-order effect at this threshold is that non-Chinese labs continue commoditising proprietary pricing independently of Chinese decisions. This means China's incremental contribution to the commoditisation shrinks over time. At some point, further Chinese open releases yield diminishing strategic returns because the job is being done by others anyway.

The compensation delay: the critical uncertainty Even after all three thresholds are crossed, the effectiveness of CPC restriction depends on a variable that is difficult to estimate in advance: the compensation delay. This is the period between Chinese restriction and non-Chinese labs filling the gap.

If the compensation delay is short, i.e. non Chinese open weight models scale up within months to fill the vacuum, restriction cedes China's ecosystem position without creating leverage. Global open-weight availability is barely dented. The cloud flywheel shifts from Alibaba to AWS and Azure.

If the compensation delay is long, it takes 12 to 18 months for non-Chinese alternatives to match the capability of restricted Chinese frontier models then China has a meaningful window of leverage. During this window, restricted access to frontier models constitutes genuine bargaining power, particularly vis-à-vis developing countries whose AI ecosystems were built on Chinese model infrastructure.

The solar panel analogy, used earlier in this paper, illuminates the likely outcome. China never restricted solar panel exports despite massive overcapacity and market dominance. The structural reason was that the entire value-capture mechanism depended on global deployment of Chinese panels. Restricting exports would have handed market share to others while collapsing Chinese manufacturing volume. The same logic applies to AI models: restricting them hands ecosystem share to non Chinese models while collapsing Alibaba's cloud flywheel.

There is, however, one critical difference. Solar panels are commodity hardware. Once installed, they generate value regardless of the supplier's future behaviour. AI models are software ecosystems requiring ongoing updates, fine-tuning support, and infrastructure integration. This means lock-in dynamics are potentially stronger in AI than in solar (once developers are in a Chinese stack, switching is genuinely costly), but so is the fragility (if China restricts too early, the ecosystem fragments away permanently, unlike solar panels which remain functional after a supply cutoff).

The most likely pathway The FT reported in July that Chinese regulators led by the Ministry of Commerce have been consulting leading AI companies including Alibaba, Bytedance and Zhipu on two issues: limiting the transfer of training data abroad and whether foreign users should continue to be able to freely download the model weights of Chinese AI systems.[35] The most plausible scenario is not a binary switch from open to closed, but a graduated restriction that preserves ecosystem share while creating a controlled capability gap:

- Delayed release of open weight models. Frontier models are released via APIs and cloud so Chinese companies can continue to monetise demand and open weight models released following a six months embargo. This creates a rolling capability advantage for firms within China's ecosystem.
- Capability threshold. Models above a size or capability threshold require commercial licensing for deployment outside China. Smaller, distilled models remain freely available, preserving the ecosystem on-ramp while gating the frontier.
- Channel-conditioned release. Labs can restrict certain capabilities from the weight-distribution channel while continuing to serve the via API. Modality heads, tool and

execution scaffolding, and specialised post-training recipes are all separable from the base model. A lab can publish weights that reason well while withholding what lets them act, perceive, or persist. This is distinct from US chip export controls as chips are rival and traceable, so they can be governed through licensing and end-user verification. Weights are neither and this is a decision taken at release time.

- Tiered access. Countries participating in WAICO or other Chinese-led consortia can get access to fine-tuning support, technical assistance, evaluation access, and compute partnerships. This converts model access into diplomatic currency. The timeline depends on how quickly the three thresholds mature. Threshold 1 (consolidation) is underway. Threshold 2 (ecosystem lock-in) requires 12 to 18 more months of cloud flywheel maturation at current growth rates. Threshold 3 (commoditisation saturation) is the most uncertain, as it depends on the pace of non-Chinese open-weight development. Our assessment is that graduated restriction is most likely to begin appearing in late 2028.

6 Implications for India

As a consumer and integrator of AI technology rather than a frontier developer, China's open-weight strategy expands access to capable models at zero cost. Nikkei Asia reported in July 2026 that Indian companies are increasingly switching to Chinese LLMs to contain AI costs, with one venture investor noting that startups are cutting costs by "an order of magnitude."[36]

The strategic question for India is to avoid dependence on *any* single source — Chinese or American. Using the framework of *atmashakti* (self-strength through increasing capabilities in a few selective segments rather than full self-sufficiency), Indian strategy should be to maintain access to multiple model ecosystems while building domestic capacity where India can develop genuine comparative advantage. These segments are applications, data, edge inference compute design, and domain-specific fine-tuning.

Simultaneously, even if China were to ban the export of its frontier open-weight models, other labs across the world are likely to pick up the baton. Chinese firms have shown what can be done, and this strategy can't be undone even if China changes its stance.

Endnotes

1. Moneycontrol. 2026. 'How Qwen Overtook Meta's Llama in the Open-Source AI Race- Moneycontrol.Com'. July 9.Link.

2. Jean-Stanislas Denain and Cheryl Wu. 2026. "Final training runs account for a minority of R&D compute spending". March 3. Link.

3. Guo, Daya, Dejian Yang, Haowei Zhang, et al. 2025. 'DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning'. Nature 645 (8081): 633–38.Link.

4. Anthropic Announcements. 2026. 'Detecting and Preventing Distillation Attacks'. Anthropic Blog. Detecting and Preventing Distillation Attacks, February 23.Link.

5. Director Michael Kratsios [mkratsios47]. n.d. 'We have information that Moonshot AI distilled Anthropic's Fable for the development of its K3 model. To do this they developed a sophisticated internal platform to conduct large scale distillation against U.S. models, allowing them to quickly switch between multiple methods of access to avoid detection. Moonshot AI has also acquired GB300-equipped servers and has accessed GB300s in Thailand, likely to train its AI models. The United States strongly supports the free and fair development of AI, including a thriving competitive ecosystem that spans frontier models, specialized systems, open-source frameworks, and open-weight models. Legitimate AI distillation used to create smaller, more efficient models plays a vital role in this open innovation ecosystem. However, large-scale, covert industrial distillation aimed at stealing proprietary U.S. technology and undermining American research is unacceptable.' Tweet. Twitter. Accessed 24 July 2026.Link.

6. Ryan Greenblatt. 2026. "Which model is Kimi K3's identity copied from? Evidence for Claude-labeled training data". July 16. Link.

7. Chenghao Yang et. al. 2026. When Agents Look the Same: Quantifying Distillation-Induced Similarity in Tool-Use Behaviors. April 23. Link.

8. 'Open vs. Closed AI Models: A Conversation with Kai-Fu Lee'. n.d. Capgemini USA. Accessed 24 July 2026.Link.

9. Xu, Kevin. 2025. 'China's Structural Advantage in Open Source AI'. Substack newsletter. Interconnected, June 26. Link.

10. Spotify. n.d. 'Ep 69: Co-Founder of Databricks & LMArena on Current Eval Limitations, Why China Is Winning Open Source and Future of AI Infrastructure • Unsupervised Learning with Jacob Effron'. Accessed 24 July 2026. Link.

11. Mozilla. 2026. The State of Open Source AI – v1.0.1. July. Link.

12. Schneider, Jordan. 2024. 'Deepseek: The Quiet Giant Leading China's AI Race'. Substack newsletter. ChinaTalk, November 27. Link.

13. Ibid.

14. Yanjun Wu, "Thoughts on AI Innovation and Open Source Development: Lessons from DeepSeek," Bulletin of Chinese Academy of Sciences 40, no. 3 (2025): Article 7.

15. Yuntao Long, Haibo Liu, et al., "Suggestions on Building China's Artificial Intelligence Open Source Innovation Ecosystem," Bulletin of Chinese Academy of Sciences 40, no. 3 (2025): 453–458.

16. Mozilla. 2026. The State of Open Source AI – v1.0.1. July. Link.

17. 'Open vs. Closed AI Models: A Conversation with Kai-Fu Lee'. n.d. Capgemini USA. Accessed 24 July 2026.Link.

18. Xu, Kevin. 2023. 'Meta Wants to Make AI "Moatless" '. Substack newsletter. Interconnected, June 23.Link.

19. Zhang, ▯▯▯ Xiaojun. 2026. 'A Conversation with Yang Zhilin of Kimi: Advancing Toward the Endless, Unknown Snow

Mountains'. X (Formerly Twitter), July 19.Link.

20. ▯▯▯ Xiaojun Zhang [zhang_benita]. 2026. 'The Second Interview with Kimi's Yang Zhilin: "Standing at the Beginning of Infinity" '. Tweet. Twitter, July 22. Link.

21. 'Made in China 2025: Evaluating China's Performance'. n.d. Accessed 27 July 2026. Link.

22. Reuters. 2024. 'China Sets up Third Fund with $47.5 Bln to Boost Semiconductor Sector'. Technology. May 27.Link.

23. Chang, Wendy, Rebecca Arcesati, and Antonia Hmaidi. 2025. China's Drive toward Self-Reliance in Artificial Intelligence: From Chips to Large Language Models | Merics. Mercator Institute for China Studies.Link.

24. Lu, Junwei, He Gao, Laifeng Yang, and Zhangxin (Frank) Liu. 2026. 'Government-Guided Funds and the Rise of Corporate AI: Evidence from China'. Pacific-Basin Finance Journal 95 (January): 103004. Link.

25. Xinhua. n.d. 'Major Changes in Domestic Large Model Manufacturing: From "Battle of a Hundred Models" to "Top Five Basic Models" - Xinhua Net Client'. Accessed 31 July 2026.Link.

26. 'AI Regulatory Landscape and Development Trends in China'. 2026. ICLG, July 20.Link.

27. Xu, Kevin. 2025. 'China's Structural Advantage in Open Source AI'. Substack newsletter. Interconnected, June 26.Link.

28. IEA. 2022. 'The World Needs More Diverse Solar Panel Supply Chains to Ensure a Secure Transition to Net Zero Emissions - News'. July 7.Link.

29. South China Morning Post. 2025. 'Joe Tsai Says Open-Source AI Will Boost Alibaba's Cloud Business'. June 11.Link.

30. Joe Tsai, remarks at World Government Summit 2026, Dubai. Reported in Alibaba Group, "Joe Tsai on the Future of Open-Source AI: Why Full-Stack Companies Will Excel," Alibaba Cloud Community / Alizila, February 9, 2026. Link.

31. Alibaba Group, "Alibaba's Investments in AI and Comprehensive Consumption Underpin Solid Q2 Results," Alibaba Cloud Community / Alizila, November 27, 2025. Link. See also Alibaba Group, "September Quarter 2025 Results," November 24, 2025, Link; and "Alibaba shares rise as AI drives 34% cloud sales jump," CNBC, November 25, 2025, Link.

32. Ngor Luong, "Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance," US-China Economic and Security Review Commission Staff Paper, March 23, 2026. Link.

33. Kumar, Amit. 2026. 'Behind China's AI Gift, a Constraint'. World. The Hindu, July 30.Link.

34. Kyle Chan et al. 2025. Full Stack – China's Evolving Industrial Policy for AI. Jun 26. Link.

35. Zijing Wu and Ryan McMorrow. 2026. "China weighs tighter export controls on AI models and chips". Financial Times. July 21. Link.

36. Chakraborty, Sayan. n.d. 'Indian Companies Look to Chinese LLMs as AI Costs Bite'. Nikkei Asia. Accessed 31 July 2026.Link.

TAKSHASHILA
INSTITUTION

The Takshashila Institution is an independent centre for research and education in public policy. It is a non-partisan, non-profit organisation that advocates the values of freedom, openness, tolerance, pluralism, and responsible citizenship. It seeks to transform India through better public policies, bridging the governance gap by developing better public servants, civil society leaders, professionals, and informed citizens.

Takshashila creates change by connecting good people, to good ideas and good networks. It produces independent policy research in a number of areas of governance, it grooms civic leaders through its online education programmes and engages in public discourse through its publications and digital media.