ROSA & ROUBINI
ASSOCIATES

MACRO PICTURE:

China's AI Strategy:

Infrastructure, Diffusion, and the Limits of Containment

By

Nato Balavadze

17 February 2026

Nato Balavadze

China's AI Strategy: Infrastructure, Diffusion, and the Limits of Containment

17 February 2026

Table of Contents

Nato Balavadze

China's AI Strategy: Infrastructure, Diffusion, and the Limits of Containment

17 February 2026

Executive Summary

- China's 15th Five-Year Plan marks a strategic shift from "dual circulation" toward AI-led development, treating AI as a general-purpose technology embedded across the entire economy rather than a standalone sector.
- Beijing aims to make AI as ubiquitous as electricity by 2030 and to build an "intelligent society" by 2035, supported by rising firm-level investment and the State Council's AI+ Action Plan.
- The core challenge is no longer frontier innovation alone but diffusion—embedding AI into manufacturing, services, and government at scale to deliver productivity gains.
- AI diffusion depends critically on massive, reliable computing infrastructure, prompting China to adopt a state-led approach to cloud and data-centre development.
- The Eastern Data, Western Computing initiative has shifted data centres inland, creating eight national computing hubs and mobilising more than USD 28bn in investment.
- Provinces such as Guizhou have been transformed into national cloud nodes, illustrating how AI infrastructure is being used as a tool of regional development.
- Rapid rollout led to overbuilding, low utilisation rates, and project cancellations, exposing the limits of infrastructure-first planning.
- Beijing has responded by tightening approvals, imposing utilisation thresholds, and exploring a state-run national cloud to pool and allocate computing capacity more efficiently.
- Despite US export controls, Chinese firms are narrowing the AI performance gap through efficiency gains, open-source leadership, and large-scale deployment, as highlighted by the release of DeepSeek-R1.
- Washington's decision to loosen AI chip restrictions reflects growing recognition that containment has been costly and may have accelerated Chinese innovation rather than stopped it.
- The AI contest is evolving into a competition between systems of accumulation: China prioritises low-cost diffusion and observable utility, while the US seeks to export its technology stack and lock in global dependence.

From Dual Circulation to AI-Led Development

China's 15th Five-Year Plan marks a clear evolution in strategy. The 14th Plan (2021–2025) emphasized "dual circulation," a framework designed to reduce vulnerability to external shocks by strengthening domestic demand and technological self-reliance while maintaining access to global markets. The new plan goes further. AI is no longer treated as a sector but as a general-purpose technology to be embedded across the entire economy.

The ambition is explicit. By 2030, AI is expected to be as ubiquitous as electricity or the internet, underpinning industrial processes, consumer products, healthcare, education, and digital government. By 2035, China aims to become what policymakers describe as an "intelligent society." This is not just rhetorical flourish. Survey data show that 87 percent of Chinese firms plan to increase AI investment in 2025, with more than half reporting

faster-than-expected progress. The State Council's AI+ Action Plan is accelerating adoption across manufacturing, energy, finance, retail, and healthcare, even if many firms still struggle to turn pilot projects into measurable productivity gains.

This push for AI diffusion requires something less glamorous than frontier models but far more consequential: massive, reliable, computing infrastructure. That requirement has reshaped China's cloud strategy.

Eastern Data, Western Computing: Cloud by Design

AI diffusion depends less on spectacular breakthroughs than on something far more prosaic: massive, reliable, and affordable computing power. Training frontier models captures headlines, but most economic value comes from running models repeatedly across thousands of applications. That requires data centres, cloud platforms, energy, and connectivity on a scale that few countries can mobilise quickly. This requirement has reshaped China's cloud and data-centre strategy.

In most countries, cloud infrastructure grows where markets point it, toward cheap land, fast connectivity, and reliable power. However, in China, since 2022, the Eastern Data, Western Computing (EDWC) initiative has deliberately shifted data centers away from the densely populated, energy-constrained eastern coast toward inland provinces rich in land and renewable energy.

Eight national computing hubs now sit at the core of this plan, spread across provinces such as Guizhou, Inner Mongolia, Gansu, Ningxia, and Qinghai. Together, they have absorbed more than USD 28 billion in public and private investment and host close to two million server racks. These facilities are not meant to operate in isolation. They are designed to function as a single, nationally coordinated computing network, linking cloud services, big data platforms, and AI workloads.

Guizhou shows how this looks in practice. Long one of China's poorest provinces, it has been transformed into a central node of the country's cloud infrastructure. Domestic hyperscalers have built large campuses there, and foreign platforms such as Apple's iCloud operate locally through Chinese partners to meet data-sovereignty requirements. By 2025, Guizhou's computing capacity had surpassed 90 exaflops, with most of it dedicated to AI. Decisions by Alibaba, Tencent, Huawei, and Baidu over where to expand next now shape which inland provinces attract investment and how quickly new facilities are upgraded for AI workloads.

From Boom to Glut: Correcting the Data-Center Surge

The pace of the EDWC rollout, however, came at a cost. After the initiative was launched in 2022, local governments rushed to build data centers, often on the assumption that demand from state firms and public agencies would naturally follow. In many cases, it didn't. Utilisation rates remain low—typically estimated at just 20–30 percent—and cancellations have mounted as projects struggle to cover costs. What began as a coordinated infrastructure push has, in parts of the country, turned into a familiar story of overbuilding.

Beijing's response has been to tighten control rather than abandon the strategy. The National Development and Reform Commission has launched a nationwide review of the sector, raising the bar for new projects. Minimum utilisation thresholds are being imposed, computing-power purchase agreements are now required in advance, and local governments have been barred from participating in smaller-scale builds. The message is that data centers should no longer be built on hope alone.

At the same time, policymakers are trying to make better use of what already exists. Authorities are exploring the creation of a state-run national cloud that would pool surplus computing power and sell it through a centrally managed platform. Working with the three state telecoms, the industry ministry wants to coordinate and schedule capacity across regions, with nationwide interoperability targeted by 2028.

Whether this will work remains an open question. Many western facilities still fall short of latency requirements for real-time applications, and the hardware mix, combining Nvidia GPUs with domestic alternatives like Huawei's Ascend, makes integration complex. The episode underlines both the power and the limits of China's model: the state can mobilize capital and move infrastructure quickly, but software ecosystems, performance bottlenecks, and effective demand are harder to engineer from the top down.

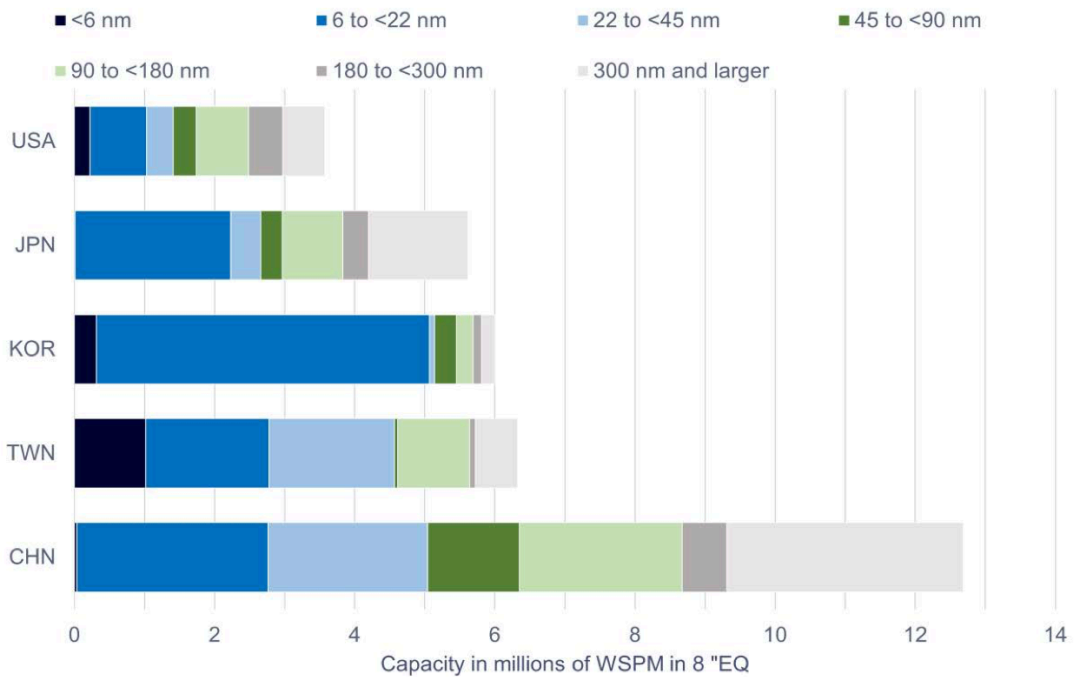
Hardware Reality: Chips, Constraints, and Catch-Up

These constraints matter because AI diffusion ultimately rests on hardware and software stacks. China's semiconductor sector has accelerated domestic R&D and fabrication, particularly for inference and large-scale deployment. Firms like Huawei and Baidu are developing competitive alternatives for many applications, and illegal GPU markets have partially offset export restrictions.

The US retains a clear lead in cutting-edge AI, underpinned by dominance in frontier models and privileged access to Nvidia's most advanced chips. American firms remain at the technological frontier, but their advantage is not unassailable. Chinese companies including DeepSeek and Alibaba are narrowing the performance gap through algorithmic efficiency, system-level optimisation, and leadership in open-source models.

Scale also matters. Asia Pacific has become the world's largest semiconductor market, expanding from USD 39.8 billion in 2001 to USD 333.4 billion in 2024 as electronics production shifted to the region. China dominates as the largest single-country market, accounting for nearly 46 percent of regional demand and 24 percent of global sales, even though shipments have declined since 2022 due to export controls and stronger growth elsewhere.

Figure 1: Feature Size Distribution



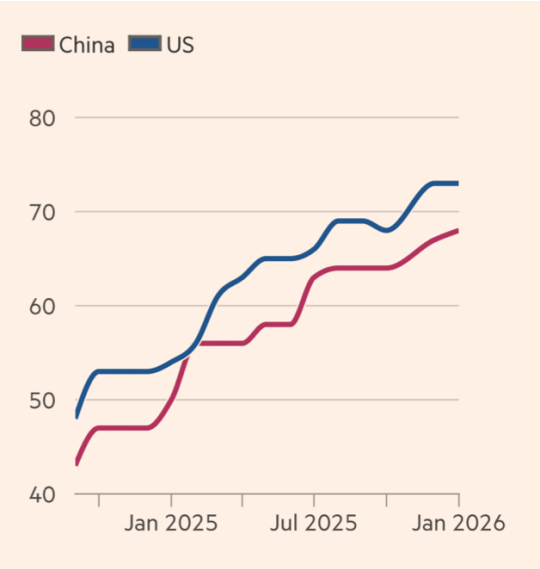
Source: OECD Semiconductor Production Database; Notes: Only in production fabs included here

The January 2025 release of DeepSeek-R1 crystallized this reality. Roughly comparable to GPT-4 and more efficient in some respects, the model demonstrated that Chinese firms can still produce breakthroughs under chip constraints. Whether DeepSeek succeeded without US hardware or by circumventing controls matters less

than the conclusion: export bans have not halted China's AI advance. Investors noticed. In 2025, Chinese tech stocks outperformed US peers, driven by AI breakthroughs and infrastructure spending.

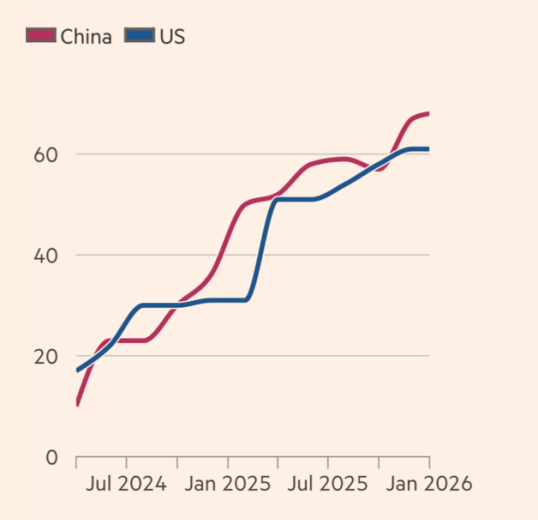
Yet structural bottlenecks remain. Advanced manufacturing equipment is still a constraint, and the absence of a mature, self-sufficient software ecosystem comparable to CUDA limits flexibility. This is where the recent US policy shift becomes critical.

Figure 2: Top LLMs Artificial Analysis' Overall Intelligence Rankings Score



Source: Financial Times

Figure 2: Open Source Models Artificial Analysis' Overall Intelligence Rankings Score



Source: Financial Times

Why the US Loosened AI Chip Restrictions

Until recently, Washington maintained a firm consensus: national security took precedence over tech exports. That consensus is fraying. The Trump administration's decision to allow sales of Nvidia's H20, and later H200, chips to China reflects a move from denial toward managed access.

The drivers are twofold. First, US tech firms are pushing against the profitability limits of export controls. Inference chips like the H20 are not class-leading, but they are commercially valuable. Second, there is growing recognition that restrictions may have accelerated Chinese innovation rather than constrained it.

Big Tech and figures such as David Sacks, Trump's AI and crypto czar, argue that export controls are costly and ineffective. AMD's Lisa Su has acknowledged that some restrictions are necessary, but insists that US firms cannot afford to miss global markets. Even national-security experts like Eric Schmidt have warned that excessive decoupling risks harming US interests, given ongoing American reliance on Chinese resources and markets.

Instead of trying to completely stop others (especially China) from accessing advanced AI technology, the US is changing strategy. It is using its current technological lead to sell and spread American AI systems globally—including chips, cloud infrastructure, and AI models—so that other countries rely on US technology. This mirrors an older historical pattern where the US government and big companies worked together to expand American economic influence abroad. By opening new markets for US firms, they created new sources of profit ("fields for accumulation") while also reinforcing US power and influence.

From Dependence to Decoupling

Blunt US rhetoric about making China "addicted" to American technology backfired, accelerating Beijing's push toward domestic chips and even bans on Nvidia hardware. China has reportedly blocked imports of Nvidia's H200 AI chips, even after Washington cleared them for export, adding to growing confusion around US-China tech trade. According to the Financial Times, Chinese customs halted shipments, prompting parts suppliers to pause production, despite Nvidia expecting more than one million orders from Chinese clients.

Authorities have also warned domestic firms against buying the chips unless strictly necessary, though it remains unclear whether this is a formal ban, a temporary move, or a negotiating tactic. The situation is further complicated by US policy: while exports were approved, the chips must first transit the US for testing, triggering a 25% tarif, also applied to AMD's MI325X. The episode underscores the strategic ambiguity on both sides. Supporters argue that selling the H200 could slow China's domestic chip development and preserve dependence on US technology; critics warn the chips could strengthen China's military capabilities.

Beyond the Bilateral Frame

For much of the global South, the US–China tech contest offers little upside. Competing AI "alliances" export infrastructure and platforms rather than genuine technology transfer. AI systems are embedded in critical infrastructure, costly to replace, and controlled by a handful of hyperscalers, creating powerful lock-in effects. Adoption often brings dependence and rent extraction rather than productivity gains.

Despite massive capital expenditure, real adoption remains uneven. Most users rely on free AI services, and businesses have been cautious about paying for enterprise licenses. Even leading firms acknowledge uncertainty about long-term profitability. Meanwhile, China itself is cooling investment to avoid the kind of overcapacity already visible in data centers and electric vehicles.

What emerges is not a race for technical supremacy but a contest between systems of accumulation. China prioritizes low-cost AI applications with observable utility, while the US relaxes controls on high-end GPUs to facilitate diffusion of its technology stack. Both sides exploit global fears of being left behind.