The cost of cheap AI for US tech

Source:AI generated

Anyone who has hit the usage limit on an AI subscription knows the problem: intelligence may be digital, but it is not free. At enterprise scale, that limit becomes a question of profitability. If agents consume more tokens than expected, the cost of automation can quickly start to look less attractive than the human labour it was supposed to replace.

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com

Hugo Morel
Syz Research Lab Team
hugo.morel@syzgroup.com

Sophia Houghton
Syz Research Lab Team
sophia.houghton@syzgroup.com

Syz PRIVATE BANKING

On 27 January 27 2025, Nvidia lost close to $600bn in a single trading session, the largest one-day loss in corporate history, on fears that a cheap Chinese model had just rendered the entire US AI buildout obsolete. Eighteen months later, the same fear is circulating again, sharper this time. US inference costs have grown expensive enough that firms are visibly migrating to Chinese open-weight models. The risk is that this migration starts to challenge the economics of the $730bn hyperscaler buildout.

Two philosophies, one price gap

The price gap between US and Chinese AI models reflects more than aggressive pricing. It is the direct output of two deliberately opposite strategies, and understanding those strategies shows whether the gap is durable or temporary.

The US approach: closed, capital-intensive, built to defend a premium

OpenAI, Anthropic, and Google DeepMind pursue a closed-weight strategy. The model itself is never released, only sold as a paid, per-use service. Each customer payment helps cover today's infrastructure and fund the next model cycle by training an expensive system, charging a premium, then reinvesting that revenue into the next, larger training run. This cycle is why the strategy is capital-intensive by design, and why the hyperscaler buildout has run into hundreds of billions. The same firms increasingly own the entire stack, training the models, building the datacentres, and designing custom chips like Google's TPU, Amazon's Trainium, and Meta's MTIA.

The China approach: open, efficiency-forced, cloud-funnelled

China takes the opposite route. Around ten Chinese labs now release near-frontier models every few months: DeepSeek, Alibaba's Qwen, Moonshot's Kimi, Zhipu's GLM. Export controls pushed these labs toward efficiency rather than scale, favouring Mixture-of-Experts designs that fire only a slice of a model's total parameters for each query. The business model differs too. For instance, Alibaba gives Qwen away for free and profits by pulling developers onto Alibaba Cloud. The model is bait, and tthe cloud is the business.

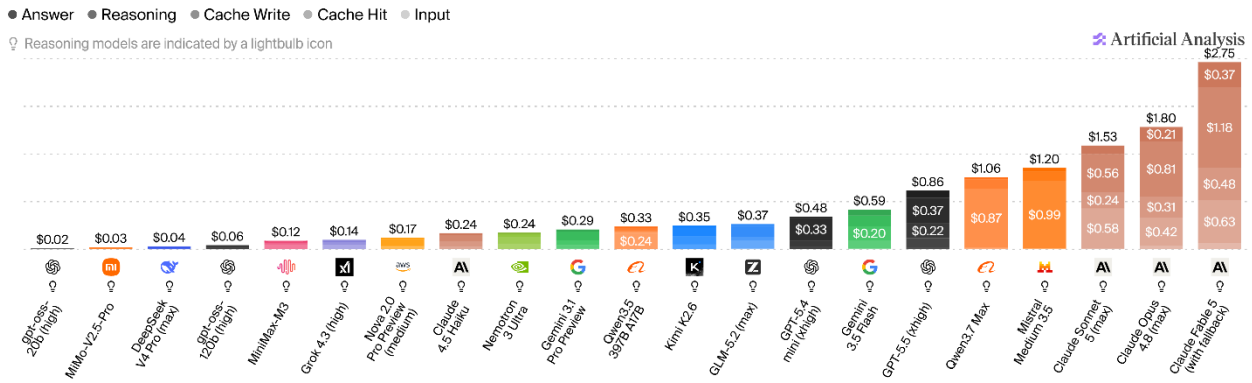
Dimension	United States	China
Weights	Predominantly closed / proprietary	Predominantly open-weight (often permissive/MIT)
Strategic goal	Own the capability frontier; premium pricing	Seed ubiquity; monetise cloud & fine-tuning
Cost posture	High: scale is the moat	Low: efficiency forced by chip constraints
Ecosystem depth	~3-4 serious open efforts	~10 labs at/near frontier, quarterly cadence
State role	Restrictive (export controls, security)	Promotional ("AI+", open-source as policy)
Where value accrues	Full vertical stack	Cloud + adoption tail

Independent benchmarks such as the Artificial Analysis Intelligence Index show Chinese models trailing the US frontier by roughly 3 to 6 months, or about 6 points on a composite scale. That gap is small enough that Chinese models already win commodity workloads, like customer service or routine coding, on price. It stays wide enough that frontier reasoning tasks, where a single error invalidates the output, remain a premium US advantage. If that gap keeps narrowing, the effect climbs into premium-tier work, and the bear case strengthens. If it holds, US pricing power survives.

The real gap at the frontier runs 5 to 15 times per million tokens, well below the widely cited "50x cheaper" headline, which compares Chinese entry-tier models against premium US ones rather than like-for-like. On Artificial Analysis's own cost-per-task measure, Claude Fable 5 runs $2.75 per intelligence-index task against $0.37 for GLM-5.2, roughly a 7x gap, matching the gap described.

Cost Intelligence Index Task

Weighted average cost (USD) per Artificial Analysis Intelligence Index task, segmented by token type. Lower is better



Is the switch real?

The gap is now visible in corporate behaviour. Uber burned through its entire 2026 AI coding budget in four months. Airbnb has moved customer service workloads to Qwen, Alibaba's open-weight AI model family. Pinterest has gone all in on open source and cut costs by roughly 90%, while Coinbase adopted GLM and Kimi and nearly halved its AI bill. Lindy fully switched from Claude to DeepSeek, and Microsoft, Anthropic's own partner, is reportedly evaluating DeepSeek inside Copilot.

OpenRouter's April rankings put three Chinese models atop global token usage: Xiaomi's Mimo, Alibaba's Qwen, and DeepSeek, with output pricing often $0.50 to $3 per million tokens against $15 to $25 for Claude. MiniMax, a smaller Chinese startup, ranked fourth in overall market share, trailing only Alphabet, Anthropic, and OpenAI. Andreessen Horowitz, a top Silicon Valley VC firm, estimates roughly 80% of US startups now build on Chinese base models, and Chinese open weight share of global token usage jumped from about 1.2% in late 2024 to roughly 30% by late 2025. Qwen has surpassed Meta's Llama in cumulative downloads.

Almost all this switching is hitting frontier labs' API revenue, the per-token fees OpenAI and Anthropic charge, plus startup spend. It is not yet touching hyperscaler cloud revenue, the money Microsoft, Amazon and Google make renting infrastructure regardless of which model runs on top. Google Cloud's backlog sits near $460bn. Azure remains capacity constrained. That is the revenue base that funds capex, and so far, it is not cracking.

The Silicon Data LLM Token Expenditure Index nearly doubled between December and May before pulling back about 20%. A softer index doesn't mean AI is getting cheaper outright. It's more likely a sign that buyers' willingness to pay is starting to peak, as demand quietly tilts toward cheaper models. It points to pricing power under strain, even while hyperscaler capex holds firm. Allianz Research puts the growth gap between AI investment and AI sales at nearly 46%, wider than the 32% divergence seen during the 2001 telecom bust.

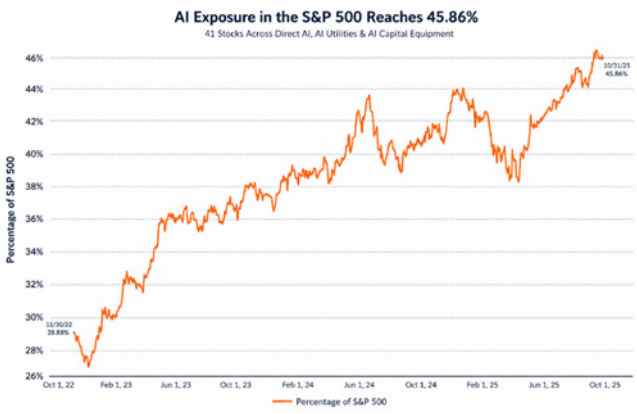
Friction is slowing migration where risk matters most. NIST (National Institute of Standards and Technology) has flagged security issues with DeepSeek. Furthermore, GLM and Kimi were both named in a US congressional probe over data and national security concerns. The EU's AI Act adds compliance burdens on frontier models that cheaper alternatives don't carry. Washington's influence cuts both ways. Export restrictions on Anthropic's Fable 5 were lifted this past week, just as regulators asked OpenAI to phase in its upcoming release.

Meta is taking a different path. Facing the same cost pressure, it is insourcing instead, building MTIA chips and the Prometheus and Hyperion datacentres, a bet made against a tight GPU market with no real relief until 2028. Even so, Meta's guidance raises still triggered a single day stock move of roughly 6 to 9%.

The deflation debate

The rapid decline in AI token prices, which have fallen by as much as 1,000 times since early 2023, has raised concerns that cheaper AI could spread financial stress across the broader economy. The main argument behind this "AI-led deflation" thesis is that lower token costs will force software companies to reduce prices as customers demand that productivity gains be passed on to them. As AI becomes cheaper to deploy, traditional seat-based licensing models are likely to face increasing pressure, reducing profitability across the software industry. This trend is already visible in AI SaaS firms, where gross margins average around 55%, compared with roughly 80% for traditional SaaS businesses.

Investors are increasingly questioning whether hyperscalers will earn sufficient returns on the projected $730bn in capital expenditure planned for 2026, contributing to greater volatility in large technology stocks. Since the "AI Big 10"—Nvidia, Microsoft, Apple, Alphabet, Amazon, Meta, Broadcom, Tesla, Oracle and AMD—accounts for almost 40% of US equity market capitalisation, any significant decline in technology valuations could extend beyond the sector and affect industries closely tied to AI spending, such as semiconductors, industrial equipment, and mining.



Source: Bianco Research

However, this bearish argument overlooks an important economic mechanism captured by the Jevons Paradox. The theory suggests that when the cost of using a resource falls, total demand often increases because activities that were previously too expensive become economically worthwhile. Current evidence points in this direction. As token prices have declined, weekly token consumption has increased thirteenfold over the past year to more than 26 trillion tokens.

Hyperscalers have also continued to report strong revenue growth, with Microsoft's AI business reaching an annualised run rate above $37bn in early 2026. Demand is further supported by the growing adoption of agentic AI systems, which require roughly 23 times more tokens than standard chatbot interactions. New reasoning models are even more computationally intensive, generating between 50 and 100 times more tokens per query.

Unlike firms during the dot-com bubble, today's technology giants also have exceptionally strong balance sheets, supported by large cash reserves and relatively low debt.

That said, the Jevons argument is not without weaknesses. During the middle of 2026, many firms shifted from maximising token usage to minimising it as concerns over rising AI bills encouraged tighter cost controls and more efficient workflows. This change in behaviour contributed to a further 20% decline in token prices during June 2026. At the same time, physical constraints such as power grid bottlenecks and delays in connecting new data centres may limit future growth in AI capacity. Falling token prices therefore create genuine short-term risks for cash flows and investment returns. Even so, the continued expansion in AI demand suggests that lower compute costs alone are unlikely to trigger a systemic market collapse.

Conclusion

So, which is it? Has the migration reached the layer funding the capex cycle, or are markets overreacting again? For now, neither. The switch to Chinese models is unmistakably real, but it is landing on frontier-lab API revenue, not on the hyperscaler cloud contracts that actually underwrite the $730bn buildout. Google Cloud's backlog and Azure's capacity constraints say that base is still intact. What is cracking is pricing power: the softening Silicon Data index, the compressing SaaS margins, and the shift from maximising tokens to minimising them all point to buyers no longer willing to pay a premium for intelligence they can rent for a fraction of the price.

That leaves the debate resting on the 3-to-6-month intelligence gap. If it holds, premium reasoning remains a US franchise while commodity inference shifts gradually to China. If it narrows, the pressure moves from API fees to capex returns. Jevons may keep demand rising, but the next benchmark cycles will decide whether that demand pays frontier prices or Chinese ones.

Welcome to Syzerland®

For further information

Banque Syz SA
Quai des Bergues 1
CH-1201 Geneva
T. +41 58 799 10 00
syzgroup.com

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com

Sophia Houghton
Syz Research Lab Team
sophia.houghton@syzgroup.com

Hugo Morel
Syz Research Lab Team
hugo.morel@syzgroup.com