Not for redistribution without written consent of Morgan Stanley

Morgan Stanley | RESEARCH

IDEA

August 12, 2026 04:01 AM GMT

Global Thematics | North America

The Model Race Gets More Complicated: Chinese Model Pricing & Open-Weights Adoption

We address three key open-weights model developments from the past week. The landscape is evolving quickly, but evidence increasingly points to a multi-model world, with open-weights driving tangible enterprise ROI. We see the US/China model price differential as a key swing factor to watch.

Key Takeaways

- Investors remain hyper-focused on what model type prevails – closed, open, or hybrid – and what it means for AI adoption, capex, and winners/losers.
- This is an evolving debate, so we address three key developments from this past week.
- Chinese model price increases show we're entering an era of better monetization. We think the US/China pricing gap will narrow, as US players cut prices.
- Recent US enterprise data points show the world is multi-model. Open models are driving ROI/saving costs, helping to dispel fears of a tokenmaxxing reduction.
- Meta re-entering the open-weight game through Muse Spark 1.2 reads positively for the US ecosystem in the case of a Chinese model ban/sovereign AI world.

Tech Diffusion

A Morgan Stanley Research
Key Theme of 2026

The open vs. closed-weight model plot thickens. As we wrote here last week, a key debate is whether closed or open-weight models win, or we live in a hybrid world. We believe this is one of the most important questions right now, as it has implications for 1) token costs, 2) Jevons Paradox / the speed of enterprise adoption, 3) hyperscaler capex, and 4) what segments of the tech stack benefit most (Exhibit 2). Investors are still hyper-focused on the end-state of the world and remain concerned that compute and AI infrastructure demand will fall with the proliferation of open-weight Chinese models. Since publishing our note last week, **we've seen three important developments: 1)** rising Chinese open-weights model prices and commercial licensing agreements, **2)** greater open-weights adoption data points from US enterprises, and **3)** Meta re-entering the open-weights game, each of which we address below. Our view continues to be that open-weights models are good for

MORGAN STANLEY & CO. LLC

Stephen C Byrd
Equity Analyst
Stephen.Byrd@morganstanley.com

Sabrina E Hao
Equity Analyst
Sabrina.Hao@morganstanley.com

Brian Nowak, CFA
Equity Analyst
Brian.Nowak@morganstanley.com

Morgan Stanley does and seeks to do business with companies covered in Morgan Stanley Research. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of Morgan Stanley Research. Investors should consider Morgan Stanley Research as only a single factor in making their investment decision.

For analyst certification and other important disclosures, refer to the Disclosure Section, located at the end of this report.

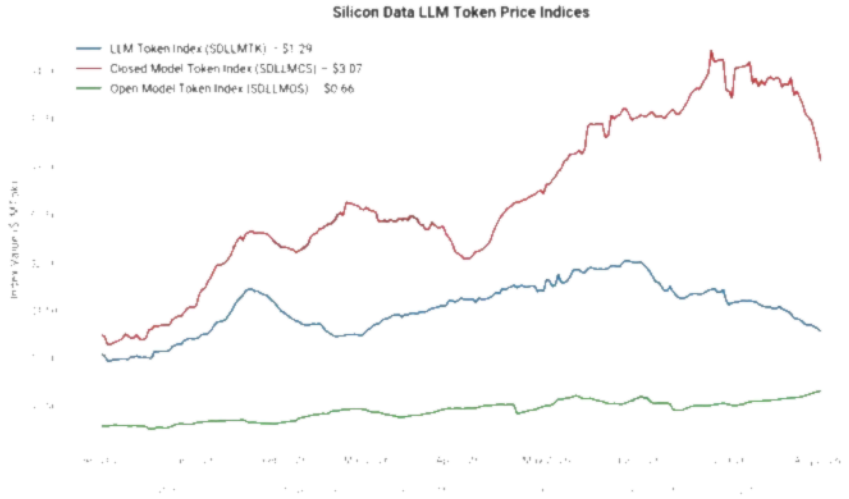
competition and can enable faster AI diffusion, supported by Jevons Paradox. We are seeing that the world is increasingly multi-model, with open-weight models starting to drive tangible ROI and save costs, but the US/China model price differential will be a key swing factor to watch.

First, we are in a unique moment where prices for Chinese models – which were historically viewed as cheap and small – are rising, while frontier model prices are declining. As our China Internet team led by Gary Yu outlined, we are entering an era where Chinese models are becoming larger, with better monetization, higher barriers to entry, and vendors offering both high-end and low-end model options (read more here: China AI Foundation Models: Intelligence War Over Price War (9 Aug 2026)). These Chinese model price increases are occurring through two avenues:

- **Direct price increases from players like DeepSeek** to become more price rational (price increase is undisclosed, but reported to be "significant"). To contextualize, our analysis indicates that a 2x increase in DeepSeek V4-Flash prices (from $0.14 per 1M input tokens / $0.28 per 1M output tokens) would lift margins for a 200MW data center running DeepSeek V4-Flash in-line with Kimi K3 and other frontier models, which we see as an effort to get closer to industry margin profiles. We'd note that the DeepSeek price increases don't apply to customers who download the model and fine-tune it, only to customers consuming via API.
- **Commercial licensing agreements from companies like Moonshot and Alibaba.** These require enterprise customers making $20M+ off these models to negotiate separate commercial terms with model providers. Reuters reports that Moonshot can take "up to 30% revenue share" from customers, which would largely apply to hyperscalers serving these models rather than individual enterprise customers. This is a departure from the industry's historically more permissive open-weight licenses.

We believe the US/China model pricing gap will continue to narrow, and we're already seeing overall token prices decline as a result. Chinese LLM APIs remain priced at roughly 15–20% of U.S. peers, but this recent news and the launch of Kimi K3 show Chinese models are not as cheap as they once were. At the same time, we expect U.S. providers to continue lowering prices, particularly for low/mid models (as we saw a frontier lab do recently). We view greater model segmentation positively, as lower-cost models should expand usage, while premium models can serve enterprises' most complex workflows. The mix shift towards open-weights models combined with closed model price declines are already leading to declines in overall token pricing, per Silicon Data's LLM Token Price Index (Exhibit 1). Overall, we view this as a positive for enterprise adoption and Jevons Paradox. A risk of these dynamics is that higher profitability at the Chinese labs could fund a reinvestment flywheel, allowing them to more quickly narrow the capability gap with US frontier models. We believe it's too early to make a call on how this plays out, but we will be watching these developments carefully.

Exhibit 1: The combination of 1) mix shifting to open-weights models, and 2) closed-weights model price declines is causing a decline in overall token prices.



Source: Silicon Data

And if prices do converge, we believe enterprise adoption of open vs. closed-weight models hinges on customization, enterprise trust, and other factors. Should pricing converge more aggressively, we believe the open-versus-closed debate for enterprise adopters becomes more about enterprise ROI and flexibility — the ability to use the *right model for the right task at the right cost*. We believe there will likely always be a use-case for open-weights models for domain-specific use cases requiring training on company/industry data. However, for standard workflows, factors like convenience, enterprise trust, and geopolitical dynamics could keep enterprises on the frontier models if prices come down far enough (i.e., if the differential is small enough, switching costs might be high enough to keep them on frontier models). We will be tracking these developments carefully.

Second, recent data shows US enterprise adoption of open-weights models is real and generating ROI, which 1) shows the world is increasingly multi-model and 2) can help dispel fears of a reduction in tokenmaxxing. Earnings commentary has given tangible evidence that US companies are utilizing and generating ROI from open-source models – both fine-tuned and off-the-shelf. Companies are citing key cost savings benefits from open-weights models and use of centralized routing layers (*see examples below*). A particularly interesting case study was **Rippling**, which was able to use open-weights models (GLM 5.2) and a custom model routing layer to reduce their token spend from 40% of R&D headcount to 15% without cutting token usage (held stable at ~600B tokens/month). We think examples like this should further dispel fears of a reduction in tokenmaxxing, as it shows enterprises can use open-weights models to "do more with less."

Other notable company references to open-weights models during earnings:

- **Pinterest** is utilizing open-source models post-trained on their data with applications like Pinterest Assistant and a model router layer to optimize model use. This is helping them achieve *"cost per transaction at less than 8% of the cost of comparable closed proprietary models."* They cited that this opportunity becomes even more compelling as open-source models,

including those from the US labs, continue to improve.

- **Revolve** is utilizing a model routing layer and commented: *"open source and some of these lower-cost models is going to be a huge part of the mix. It's already a huge part of our mix."*
- **Spotify** is allowing users to *"switch models mid-task and route every job to the best available performance,"* including to open source models.
- **Duolingo** noted *"Our view inside the company is that it is in our best interest as a company to use open weight models as much as possible, so if I had a magic wand, I would try to move everything to an open weight model"* due to cost considerations.

Third, in a sovereign AI world, Meta (covered by Brian Nowak) re-entering the open-source game is a positive development for the US ecosystem. As we wrote about here, we are seeing multipolarity in motion, with both the US and China considering AI policy intervention. Sovereign AI and a potential ban of Chinese models have ripple effects for global AI adoption, and our conversations indicate that business continuity risk is a large consideration for companies considering adopting Chinese open-weights models. Should we enter a US/China decoupling scenario, development of the US open-weights ecosystem is key. We view Meta releasing the weights for Muse Spark 1.2 positively for the ecosystem, as it provides a strong US alternative to Chinese offerings.

From a stock perspective, we'd direct investors to our "3 States of the World" framework (Exhibit 2). The open vs. closed-weight model debate is an evolving situation, and we will be updating our thoughts frequently. Our stock framework highlights ways to play each state of the world – where closed models win, open models win, and hybrid wins – across the whole stack: semis/hardware/networking, data center operators, cloud providers, model providers, and SaaS. *See Ways to Play* below.

Ways to Play

Exhibit 2: Stocks we like across our "3 States of the World" — where closed models win, hybrid wins, and open models win.

	Scenario 1 (Closed Models Win)	Scenario 2 (Hybrid)	Scenario 3 (Open Models Win)
	Small number of frontier labs win. Prefer names levered towards hyperscale data centers or custom chips for frontier labs.	*Many model architectures win and workloads run both on-prem/cloud. Prefer names exposed to multi-model workflows, orchestration, observability, and/or distributed workflows.*	*Open models proliferate, driving down token costs and driving faster enterprise adoption. More workloads are run on-prem or at the edge, benefitting hardware names.*
Biggest Beneficiaries	**Cloud Providers:** GOOGL, AMZN **Model Providers:** GOOGL, META **Security Software:** PANW, CRWD, ZS, NTSK, OKTA **Optical/Networking:** ANET, LITE, COHR **On-Site Power Providers:** BE, INIO, SEI, WMB, LBRT **Semis:** NVDA, AVGO	**Cloud Providers:** AMZN, GOOGL, MSFT **Infrastructure Software:** DDOG, PLTR, APPN **Security Software:** PANW, CRWD, FTNT, ZS, NTSK, OKTA, SAIL, VRNS **SaaS:** SAP, NOW **Optical/Networking:** CSCO, FFIV **On-Site Power Providers:** BE, INIO, SEI, WMB, LBRT **Semis**: NVDA	**Cloud Providers:** MSFT **Model Providers:** MiniMax, Knowledge Atlas, BABA, Tencent **Infrastructure Software:** PLTR **Security Software:** PANW, CRWD, FTNT, OKTA, SAIL **SaaS:** SAP, NOW, SHOP **On-Prem Infrastructure:** DELL, HPE, NTAP, P **Edge Devices:** DELL, HPQ, AAPL **VARs/Distributors:** SNX, INGM, CDW **On-Site Power Providers:** BE, INIO, SEI, WMB, LBRT **Semis:** NVDA
Other Beneficiaries	**Cloud Providers:** MSFT **Colocation Data Centers:** DLR **Neoclouds:** CRWV, NBIS **Security Software:** SAIL	**Model Providers:** MiniMax, Knowledge Atlas, BABA, Tencent **Colocation Data Centers:** EQIX, DLR **Neoclouds:** CRWV, NBIS **SaaS:** SHOP, WDAY, CRM, INTU **On-Prem Infrastructure:** DELL, HPE, NTAP, P **Edge Devices**: DELL, HPQ, AAPL **VARs/Distributors:** SNX, INGM, CDW	**Cloud Providers:** AMZN, GOOGL **Colocation Data Centers:** EQIX **Neoclouds:** CRWV, NBIS **SaaS:** WDAY, CRM, INTU

Source: Morgan Stanley Research

Given it's still yet to be seen how the mix of open and closed weights model trends for enterprises, we believe it's useful to frame three potential scenarios we see playing out over the next few years:

1. Closed-Weights Models Win. Frontier-model performance remains difficult to replicate, which leads to a small number of well-capitalized labs winning. Enterprises prioritize accuracy, reliability, security, indemnification and ease of deployment over model control or transparency, and they choose to partner with the labs rather than fine-tune open-weights models in-house. AI workloads remain concentrated in hyperscaler clouds, and training demand remains highly compute-intensive. Cost per token falls as models become more efficient, but the cost of compute remains higher than it is in the open-weight scenario given the market oligopoly. Security software is a key winn
security burden is handled by the frontier labs vs. on-prem. Further,
demand of training superclusters further exacerbates the demand fo
power: bullish for onsite power providers.

2. Hybrid State of the World. In this scenario, no single model archi
frontier models dominate the most complex reasoning and agentic v
weight and smaller models handle high-volume, cost-sensitive or do



Some workloads will use both, routing tasks to each based on capability and cost. Closed models retain pricing power, but competition from open alternatives limits excess returns. AI deployment spans public cloud, private cloud, on-premise infrastructure and edge devices. Model orchestration, routing, evaluation, and governance benefit software players in these areas. Security software becomes increasingly more important as more workloads are handled on-prem or in a more distributed manner.

3. Open-Weights Models Win. Open-weight models converge with frontier-level performance and base-model intelligence becomes widely available. Model API prices fall, leading to greater adoption across enterprises, in-line with Jevons Paradox. Companies customize models using proprietary data and run them wherever economics, latency or regulation are most attractive. Data sovereignty, privacy and avoidance of vendor lock-in accelerate adoption by governments and large enterprises. Innovation shifts from pre-training toward fine-tuning, inference optimization, agents, tools and application-specific deployment. AI becomes more decentralized, with substantial growth in private data centers, sovereign clouds, on-prem systems and edge devices. This also drives greater demand for security software to secure on-prem workloads, as well as observability and routing offerings.