First Read

China Internet Sector

What are US investors asking about China's AI models

Equities

China

Internet Services

Wei Xiong
Analyst
wei.xiong@ubs.com
+86-21-3866 8883

Kenneth Fong
Analyst
kenneth-kc.fong@ubs.com
+852-3712 3890

Charles Chen
Analyst
S1460524020001
charles-za.chen@ubs.com
+86-21-3866 8907

High interest in China's models amid shift towards token optimization

Following our recent note on cost efficiency of China's AI models, we met 40+ investors in the US with a balanced mix of LOs and HFs. As enterprises become cost conscious and start to optimize token spend by considering Chinese models as an option (Figure 1 and Figure 2), we observe better-than-expected investor interest in Chinese AI labs, in order to understand their relative market positioning and competitive dynamics, as well as to gauge the implications for the AI tech supply chain and global foundation model industry. Notably, Zhipu's recent launch of GLM-5.2 has been well received by the global tech community, leading more investors to believe Chinese models have reached the capability threshold to handle complex, enterprise-grade workloads. DeepSeek, Kimi, Alibaba Qwen and MiniMax also see improving recognition. Despite China's relative strength in multi-modality, this area remains less discussed than AI coding.

Key investor debates and our views

Cost efficiency / token UE: Investors are keen to understand the key drivers behind China models' cost efficiency and whether it can sustain. Many question whether Chinese labs need to step up R&D significantly to continue scaling up parameters and improving model intelligence, and whether US peers can also replicate China's inference optimisation (OpenAI reportedly discovered ways to halve inference cost). We reiterate our view that China's cost advantage is likely structural (see our token UE analysis in Figure 4), mainly supported by continued innovations in model design and AI infra optimisation. We also think the rising adoption of domestic AI chips could help sustain China's cost advantage over time.

Open-source vs closed-source: Investors wonder whether open-source models might face limitations in model capabilities (given restricted access to frontier closed-source models) and if open sourcing hurts monetisation potential. We think distillation can improve training efficiency but does not help much in advancing core capabilities such as long-horizon reasoning and coding, which Chinese models have shown meaningful improvement too, yet the open-source ecosystem in China could expedite tech innovation diffusion to drive R&D synergies. On monetisation, we believe open-source models can still monetise through official API services, on-premise deployment, value-added services (e.g. AI infra optimisation and fine-tuning), and we do not rule out the possibility that Chinese labs may launch closed-source versions in the future.

Competition / market structure: Competition has always been a top concern for Chinese AI labs (vs. internet incumbents domestically, and leaders globally), and the rapid shift towards token optimisation prompts investors to ask about future market structure in the foundation model industry (e.g. how much workload could migrate to Chinese models, especially given data/compliance/geopolitical risks). We note relatively stronger momentum in model development from labs than incumbents globally, while acknowledge the long-term monetization and margin potential remains a key debate, given recurring training workload, increasing hardware costs and pricing pressure in case of model commoditization. While we do not have a TAM estimate with conviction yet, we still expect Chinese models to gain share over time, with more visible ROI advantages in coding and agentic workflows, especially for high-volume, repetitive inference tasks, as evidenced by rising integrations across overseas apps and platforms. (Figure 3). Our preliminary view is that the volume mix of premium vs. mass market workload may follow a 20/80 rule, and the premium segment may have a meaningfully larger revenue share given better pricing power.



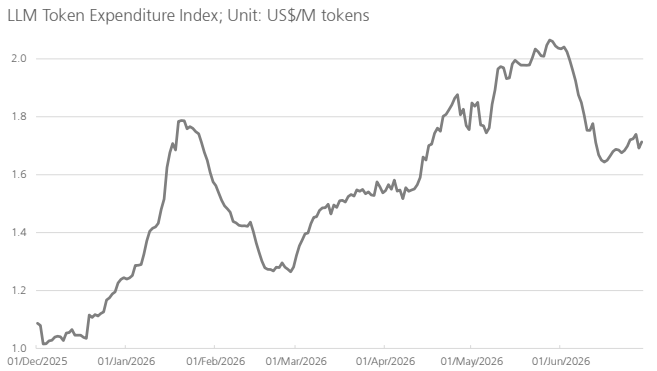
This report has been prepared by UBS Securities Asia Limited. **ANALYST CERTIFICATION AND REQUI**
INCLUDING INFORMATION ON THE QUANTITATIVE RESEARCH REVIEW PUBLISHED BY UBS, BEGIN ON

Compute: Investors question to what extent compute may constrain future model capability improvements and ARR growth for Chinese labs. We agree compute availability remains a key risk, and expect AI labs to diversify their compute sources with limited incentive to engage in price competition. We are also encouraged by the progress of domestic AI chip adoption, with latest example of Meituan's LongCat-2.0 (>1tn parameters, fully trained/served on domestic AI chips), which could gradually ease the compute constraint to fulfil inference demand and then training.

Upcoming catalysts to watch for

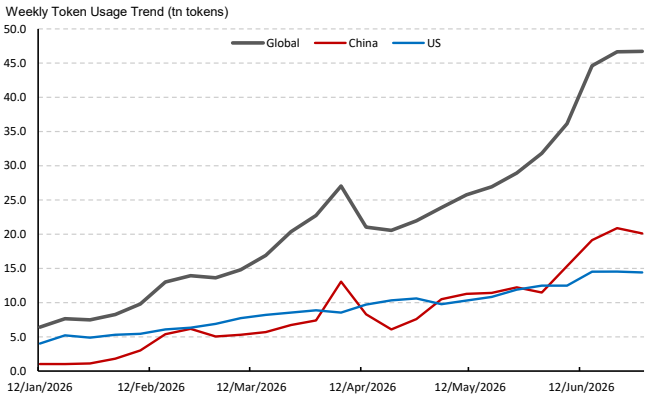
Share prices of listed model companies Zhipu and MiniMax have performed well YTD, and the upcoming catalysts to monitor include 1) model launches (especially from DeepSeek, Kimi and Alibaba); 2) lock-up expiry in early July; 3) mid-year results in August; and 4) A-share listing process (including Zhipu and MiniMax; based on recent STAR Market IPO precedents Figure 5, we expect the earliest potential listings as early as late-2026 or early-2027).

Figure 1: LLM Token Expenditure Index



Source: SiliconData; Note: Data as of 29 June, 2026. Normalized blended rate drawn from observations across frontier API providers, open-weight inference platforms, brokered dedicated-instance markets and self-hosted reference deployments.

Figure 2: China vs US AI model token usage trend on OpenRouter



Source: OpenRouter. Note: data as of 1 July 2026

Figure 3: Overseas apps and platforms are increasingly integrating Chinese AI models

App / Platform	Product category	Key functions	Timing	Integration details
Fireworks AI	AI inference / model hosting platform	**Open-model AI infrastructure** for training / fine-tuning and high-throughput inference, with serverless, on-demand and reserved deployments plus OpenAI / Anthropic-compatible APIs	Kimi / MiniMax / Qwen: 10-12 Jun 2026; GLM-5.2: 16 Jun 2026	Fireworks announced day-zero **GLM-5.2** availability. Its changelog / model library also lists **Kimi K2.7 Code**, **MiniMax M3, Qwen 3.7 Plus** and other Chinese models on the platform.
Vercel AI	Developer tool / AI gateway	**Developer AI gateway** providing one API key for hundreds of models, unified billing / observability, provider routing / fallback, spend controls, BYOK and drop-in SDK compatibility	GLM-5.2: 16 Jun 2026; GLM-5.2 Fast via Wafer: 24 Jun 2026	Vercel added **GLM-5.2** to AI Gateway on 16 Jun under model ID zai/glm-5.2. On 24 Jun, it added **GLM-5.2 Fast** via **Wafer**, a model inference / hosting platform offering serverless and dedicated inference.
Cloudflare Workers AI	Serverless AI / edge inference platform	**Serverless GPU inference on Cloudflare's global networ**k, with open-source model access through Workers / API, pay-as-you-use pricing and integrations with AI Gateway, Vectorize and Workers	16 Jun 2026	Cloudflare announced **GLM-5.2** availability on Workers AI, highlighting agentic coding, function calling and reasoning. The initial context window was 262k, with plans to expand later.
0G Private Computer	Decentralised AI compute / verifiable inference	**Decentralised AI compute network for router / direct inference**, provider marketplace access, wallet-based settlement, TEE-backed verification and chatbot, image and speech service types	18 Jun 2026; Kimi added the following day	0G announced **GLM-5.2** availability on Private Computer, highlighting end-to-end encryption for prompts, code and outputs, as well as verifiable compute. **Qwen-3.7 Plus** was also available, with **Kimi 2.7 Code** added shortly thereafter.
Notion	Notes / AI workspace	**All-in-one workspac**e for docs, wikis / knowledge base, databases, project management, enterprise search, meeting notes and agents	Mid-Jun 2026	Notion announced **GLM-5.2** availability, particularly for long-horizon tasks. The model is served through Baseten, which also confirmed that Notion is offering **GLM-5.2** via its inference platform.
Perplexity	AI answer engine / Enterprise AI / APIs	**AI answer engine and enterprise research platform** spanning web research, internal-file / app search, agentic Computer workflows and developer APIs	Mid- to late Jun 2026	Perplexity's Agent API model list includes **perplexity/glm-5.2** and **perplexity/kimi-k2.7-code**, making both models available through its Agent API.
Devin / Cognition	AI coding / software engineering agent	**Autonomous AI software engineer** for writing, running and testing code, resolving tickets / bugs, migrations / refactors, PR review, codebase Q&A and engineering support	24 Jun 2026	Devin announced that **Kimi K2.7** and **GLM-5.2** are available in Devin Desktop and CLI.
Coinbase	Crypto exchange / digital asset platform	**Retail and institutional crypto / asset trading platform**, with custody, staking, payments, wallets, developer infrastructure and onchain services	Late Jun 2026;	Coinbase co-founder and CEO Brian Armstrong said the company is experimenting with **GLM-5.2** and **Kimi 2.7** as default open-weight models through its internal LLM gateway, aiming to reduce AI spend without restricting engineers' token usage.
Fello AI	Multi-model AI assistant / consumer app	**All-in-one AI chatbot app** for Mac, iPhone and iPad, with multi-model chat, web search, file / PDF analysis, image generation / editing, multilingual chat and cross-device sync	25 Jun 2026	Fello AI 6.7.0 added **GLM-5.2, MiniMax M3, DeepSeek V4 and Qwen 3.7 Max**, alongside new Podcast and Travel Skills and broader file support.
Poe / Quora	Multi-model chatbot / bot platform	**Consumer and developer AI aggregation platform** for accessing multiple models / bots, generating text, images, video and audio, web search, multi-bot chat, bot / app creation, workflows and Poe API	Live on platform; no single announcement date found	Poe pages show access to **GLM-5.2**, **DeepSeek V4 Flash / Pro**, **Kimi K2.7 Code, MiniMax M3** and other Chinese models via providers such as Novita AI.

Source: Company data

Figure 4: Our analysis of inference token UE: China vs. U.S.

US$/M tokens	China	US	China as % of US
Average blended model price	**0.63**	**4.43**	**14%**
Input	0.60	4.21	14%
Output	2.46	19.40	13%
Cached input	0.27	0.51	53%
Gross margin assumption	30%	70%	
Cost of sales	**0.44**	**1.33**	**33%**
Depreciation of compute hardware (US$/M tokens)	0.30	0.78	39%
as % of Cost of sales	*68%*	*59%*	
Power cost (US$/M tokens)	0.03	0.09	30%
as % of Cost of sales	*6%*	*7%*	
Other cost (Facility, network infrastructure etc.) (US$/M tokens)	0.11	0.45	25%
as % of Cost of sales	*25%*	*34%*	

AI data center cost breakdown	China	US	China as % of US
Depreciation of compute hardware (US$/hr)	**76.1**	**97.8**	78%
Capex of compute hardware (US$ M)	3	3	
Depreciation life (years)	5	5	
Utilization	90%	70%	
Power cost (US$/hr)	**7.2**	**11.7**	61%
IT power (kW)	130	130	
Power usage effectiveness (PUE)	1.25	1.14	
Electricity price (US$/kWh)	0.04	0.08	56%
Other cost (Facility, network infrastructure etc.) (US$/hr)	**28.4**	**56.8**	

Source: Epoch AI, UBS estimates

Figure 5: Selected recent STAR Market IPO timeline

Company	Company (CN)	IPO counselling started	Counselling completed/accepted	Listing application accepted	Listing committee approval	CSRC registration approved	Listed	Counselling completed → Listing
Zhipu	智谱	6-Jun-26	17-Jun-26	N/A	N/A	N/A	N/A	N/A
MiniMax	MiniMax	29-May-26	In progress	N/A	N/A	N/A	N/A	N/A
Moore Threads	摩尔线程	Nov-24	8-Jun-25	30-Jun-25	26-Sep-25	30-Oct-25	5-Dec-25	~6.0 months
MetaX	沐曦股份	Jan-25	1-Jun-25	30-Jun-25	24-Oct-25	12-Nov-25	17-Dec-25	~6.5 months
UX IC	优迅股份	Nov-24	Early Jun-2025	26-Jun-25	15-Oct-25	24-Oct-25	19-Dec-25	~6.5 months
SJ Semiconductor	盛合晶微	Jun-23	20-Jun-25	30-Oct-25	24-Feb-26	5-Mar-26	21-Apr-26	~10 months
Unitree Robotics	宇树科技	Jul-25	1-Nov-25	20-Mar-26	1-Jun-26	Submitted for registration	Not yet listed	N/A
CIQTEK	国仪量子	Sep-24	N/A	10-Dec-25	11-May-26	11-Jun-26	Not yet listed	N/A

Source: Shanghai Stock Exchange, CSRC, company filings, public news

Valuation Method and Risk Statement

We believe key sector risks include: 1) an evolving competitive landscape and intensifying competition; 2) fast-moving trends in technology, as well as internet users' needs and preferences; 3) uncertain monetisation; 4) the rising cost of traffic acquisition, content and brand promotions; 5) the upkeep of IT systems; 6) expansion into international markets; 7) adverse changes in market sentiment; and 8) regulatory risks.