China AI Intelligence

Evolving competition dynamics on capability and cost efficiency

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

Mycroft He
Associate
S1460126080003
mycroft.he@ubs.com
+86-21-3866 5000

Broader trends reshaping the global AI model competitive landscape

While a narrowing capability gap and cost advantages have increased recognition of Chinese models' global competitiveness (note), we note recent developments that have complicated the competitive outlook and contributed to weak momentum in listed model companies (MiniMax/Zhipu down c.20/40% MTD): **1) Regulatory landscape continues to evolve** (note), with debate over potential restrictions on open-weight models adding uncertainty to overseas adoption, and tighter anti-distillation measures by global leaders and growing US policy focus on industrial-scale distillation raising questions over the pace of capability catch-up. **2) Cost-performance advantages are becoming more task-dependent**, particularly in complex, long-horizon tasks, where we observe overseas frontier models partly offsetting their per-token price premium through improved token efficiency and task success rates, while more developers launch smaller/flash models to optimise inference efficiency for volume-driven tasks. While we expect such debates to remain fluid, we note that **Chinese AI labs are expanding their overseas reach**, including MiniMax's global collaborations spanning enterprise adoption, vertical fine-tuning and cloud deployment, Zhipu's recent partnerships with global CSPs, and Kimi K3's availability on AWS Bedrock. This could help them to better monetise rising global adoption of open-weight models, in our view.

Model pricing to diverge further across task tiers

We expect pricing trends to diverge further between SOTA and non-SOTA models as they increasingly serve distinct task tiers. **At the volume end, we expect continued per-token price deflation amid competition**. We note Chinese and overseas developers increasingly offer Flash/smaller-parameter models with near-flagship performance at substantially lower prices, while global peers increasingly emphasize cost efficiency through direct price cuts, free token-plan quota resets and lower KV caching costs. **At the capability frontier, pricing power is likely more durable**, as stronger capabilities and higher token efficiency can lower cost per successful task despite higher per-token pricing, allowing frontier models to retain or even widen their price premium without compromising customer ROI (e.g., GPT-6 Astra).

Race for frontier capability remains intense

Amid growing discussion about the pace of frontier AI development, we could see a more measured approach to model releases, while internal R&D intensity is likely to remain high, as seen in early progress in AI-assisted R&D and RSI (recursive self improvement) exploration among AI labs (e.g., OpenAI's and Anthropic's use of AI across coding, experiments and research workflows, and Zhipu's Infra Agent). We expect Chinese labs to continue expanding compute capacity following recent funding rounds to support the next generation of larger-scale model training. Meanwhile, we expect Chinese models' future capability gains to be increasingly driven by diversified data sources (e.g., GitHub issues/Pull Requests and synthetic data) and accumulated domain-specific environment data (e.g., cybersecurity).

Upcoming model launches to watch

While concerns on intensifying global competition recently drove a sector de-rating, we think Chinese labs' progress in overseas distribution, architectural innovation and open-source R&D synergies remains underappreciated. We think continued model launches (e.g., Step 5 Preview, Xiaomi MiMo-V2.6) could help restore investor sentiment by extending the intelligence-cost Pareto frontier, with potential further catalysts ahead including DeepSeek V4.1 Pro, Kimi K3.1, MiniMax M3 Pro, Zhipu GLM-5.5/6.0, etc.

This report has been prepared by UBS Securities Asia Limited. **ANALYST CERTIFICATION AND REQUIRED DISCLOSURES, INCLUDING INFORMATION ON THE QUANTITATIVE RESEARCH REVIEW PUBLISHED BY UBS, BEGIN ON PAGE 7.**

Expanding overseas reach supports further adoption

We are encouraged by Chinese AI developers' expanding overseas reach, including MiniMax's global collaborations spanning enterprise adoption, vertical fine-tuning and cloud deployment; Zhipu's recent partnerships with global CSPs; and the availability of Kimi K3 on AWS Bedrock. We think these developments could strengthen monetisation opportunities as global adoption of open-weight models accelerates.

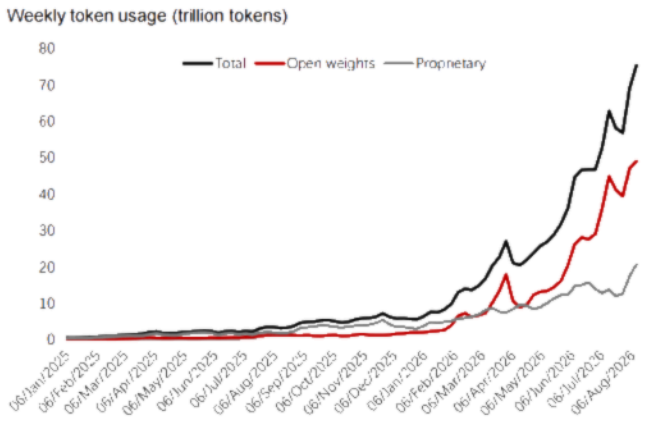
Usage traction for China/open-weight models remains strong. OpenRouter data indicates continued usage-share gains, supported by leading Chinese models and a narrowing capability gap with leading US frontier models, in our view.

Figure 1: Chinese AI developers continue to broaden their overseas reach

Developer	Selected recent overseas expansion
MiniMax	• **Singapore**: Partnered with Singtel to include MiniMax Agent, Hailuo AI and MiniMax Audio in Singapore's public AI-skills programme. • **Saudi Arabia**: HUMAIN continued pre-training M3 on >1tn Arabic tokens to build Humain M3 • **Japan**: Opened a Tokyo office and launched a Japan IP/AI collaboration with KAGAMI AI and content-industry partners • **Post-training**: Genspark on partnership with Fireworks AI, post-trained M3 for Gen-1 S ides • **Cloud**: M3 launched on Nebius Token Factory under an exclusive strategic partnership; MiniMax continues chip-level deployment optimisation with SambaNova
Kimi	• **Cloud distribution**: Kimi K3 is now available to overseas customers through Amazon Bedrock with enterprise-grade access encryption and audit controls, as well as prompt caching to reduce latency and input costs for repeated long-context workloads
Zhipu	• **CSP partnerships**: Zhipu is advancing API hosting and revenue-sharing arrangements with domestic and overseas CSPs, leveraging their inference infrastructure, enterprise customers and sales channels to accelerate global distribution

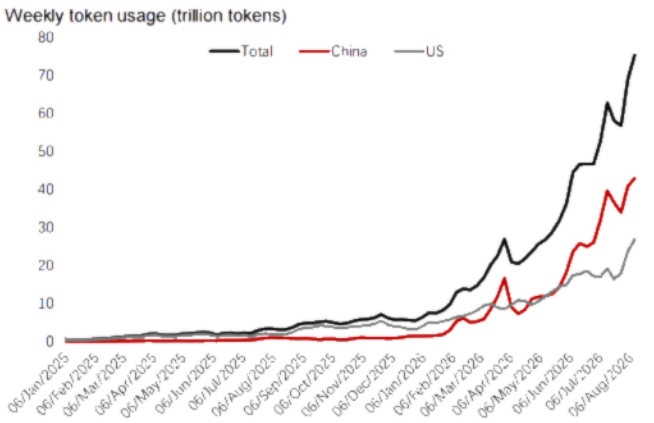
Source: Company data, Amazon Web Services, UBS

Figure 2: AI model token usage trend on OpenRouter – open weights vs proprietary



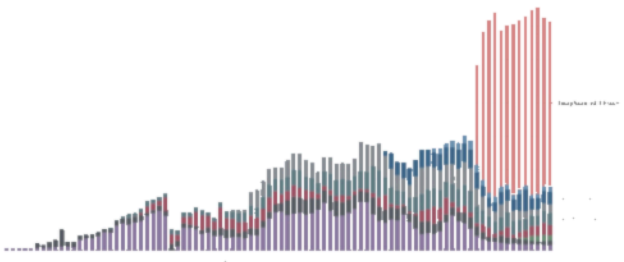
Source: OpenRouter

Figure 3: AI model token usage trend on OpenRouter – China vs US



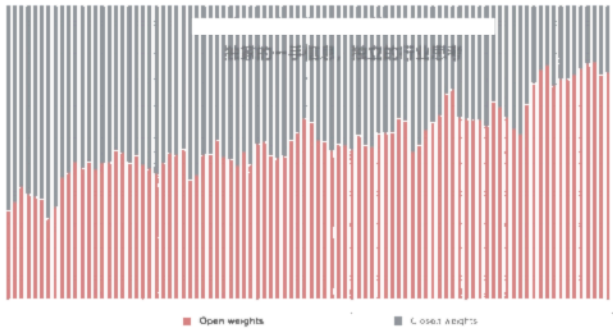
Source: OpenRouter

Figure 4: Share of AI Gateway token volume across the top text models



Source: Vercal

Figure 5: Open vs. Closed token volume share on Vercel's AI Gateway platform



Source: Vercal

Model pricing to diverge further across task tiers

Per-token price deflation at the volume end. At the volume end, where available models already meet task requirements and incremental intelligence offers limited marginal value, we expect continued per-token price deflation to help broaden AI adoption, as evidenced by:

- **Expansion of tiered model portfolios**: Chinese and overseas developers are introducing Flash and smaller-parameter models with near-flagship performance at substantially lower prices, including GLM-5.3-Flash, Kimi K2.8 Preview, Qwen3.8-Flash and Gemini 3.8 Flash.
- **Increasing overseas focus on cost efficiency**: In response to open-weight models' cost advantage, overseas developers are using direct price cuts (e.g., OpenAI's GPT-5.6 series), free token-plan quota resets and lower caching costs (e.g., Claude Fable 5.1's KV cache pricing is 75% lower than 5.0).
- **Recent launches further reinforce this trend.** On 22 September, OpenAI introduced **GPT-6 Sol and Luna** at roughly half the API prices of their GPT-5.6 counterparts, broadening its lower-cost offerings across task tiers. Anthropic launched **Claude Opus 5.5** on the same day, with stronger performance than Opus 5, 20% lower input/output prices and 60% lower cache-read prices. Anthropic estimates c.40% lower cost for typical tasks.

Figure 6: AI developers are launching more Flash models, offering near-flagship performance at lower prices

Developer	Tier	Model	Release date	Parameters (total / active)	AA Index	Input US$/mn tokens	Cached input US$/mn tokens	Output US$/mn tokens	Blended price 7:2:1 blend
Zhipu	Flash	GLM-5.3-Flash	26 Aug 2026	320B / 18B	41.8	$0.15	$0.026	$0.50	$0.10
	Flagship	GLM-5.3	18 Aug 2026	753B / ~40B	45.0	$1.40	$0.26	$4.40	$0.90
Moonshot	Flash	Kimi K2.8 Preview	11 Sep 2026	Not disclosed	-	$1.00*	$0.25*	$4.00*	$0.78*
	Flagship	Kimi K3	16 Jul 2026	2.8T / 104B	50.0	$3.00	$0.30	$15.00	$2.31
Alibaba	Flash	Qwen3.8-Flash-Next	26 Aug 2026	125B / 6B	40.0	$0.15	$0.016	$0.47	$0.09
	Flagship	Qwen3.8 Max	3 Aug 2026	2.4T / 95B	40.0	$2.00	$0.25	$6.00	$1.18
Google	Flash	Gemini 3.8 Flash (high)	2 Sep 2026	Not disclosed	41.0	$0.75	$0.075	$3.75	$0.58
	Flagship	Gemini 3.1 Pro Preview	19 Feb 2026	Not disclosed	46.5	$2.00	$0.20	$12.00	$1.74
SpaceXAI	Flagship	Grok 4.7 (high)	21 Sep 2026	Not disclosed	46.3	$2.00	$0.50	$6.00	$1.35
	Flash	Grok Build 0.1 0616	16 Jun 2026	Not disclosed	27.2	$1.00	$0.20	$2.00	$0.54
DeepSeek	Flash	DeepSeek V4.1 Flash	10 Sep 2026	552B / 8B, 16B	39.0	$0.30	$0.006	$1.20	$0.18
	Flagship	DeepSeek V4 Pro 0424	24 Apr 2026	1.6T / 49B	30.0	$0.435	$0.0036	$0.87	$0.18
OpenAI	Balanced	GPT-6 Sol	22 Sep 2026	Not disclosed	48.0	$2.00	$0.20	$10.00	$1.54
	High-volume	GPT-6 Luna	22 Sep 2026	Not disclosed	37.0	$0.10	$0.01	$0.50	$0.08
	Flagship	GPT-6 Astra	3 Sep 2026	Not disclosed	53.0	$10.00	$1.00	$50.00	**$7.70**
Anthropic	Efficient	Claude Opus 5.5	22 Sep 2026	Not disclosed	58.0	$4.00	$0.20	$20.00	$2.94
	Flagship	Claude Fable 5.1	1 Sep 2026	Not disclosed	53.0	$10.00	$0.25	$50.00	**$7.18**

Source: Artificial Analysis, company data, UBS. Note: Prices are in US$/mn tokens. Blended price represents a 7:2:1 blend of cached-input, input and output token prices. Parameters refer to total/active parameters where disclosed. *Kimi K2.8 Preview pricing is based on a third-party API provider. Data as of 22 September 2026.

Figure 7: OpenAI directly cut GPT-5.6 prices, while Anthropic has improved cost efficiency by lowering KV cache prices

Developer	Model / transition	Input US$/mn tokens: old → new	Cached input US$/mn tokens: old → new	Output US$/mn tokens: old → new	Key change
OpenAI	GPT-5.6 Sol	$5.00 → $4.00 (-20%)	$0.50 → $0.40 (-20%)	$30.00 → $20.00 (-33%)	Broad price reduction
	GPT-5.6 Terra	$2.50 → $2.00 (-20%)	$0.25 → $0.20 (-20%)	$15.00 → $12.00 (-20%)	20% across token types
	GPT-5.6 Luna	$1.00 → $0.20 (-80%)	$0.10 → $0.02 (-80%)	$6.00 → $1.20 (-80%)	80% across token types
Anthropic	Claude Fable 5	$10.00	$1.00	$50.00	Previous generation
	Claude Fable 5.1	$10.00	$0.25 (-75% vs 5)	$50.00	Cache-led cost reduction

Source: Artificial Analysis, company data, UBS

Figure 8: LLM Token Expenditure Index

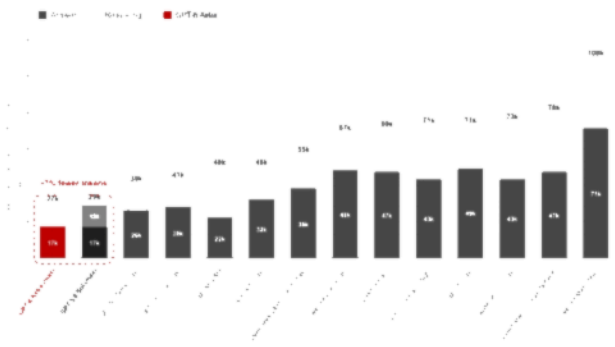


Source: SiliconData. Note: Data as of 21 September 2026. Normalized blended rate drawn from observations across frontier API providers, open-weight inference platforms, brokered dedicated-instance markets and self-hosted reference deployment

More durable pricing power at the capability frontier, as suggested by GPT-6 Astra, with frontier models potentially retaining or even widening their price premium while preserving customer ROI.

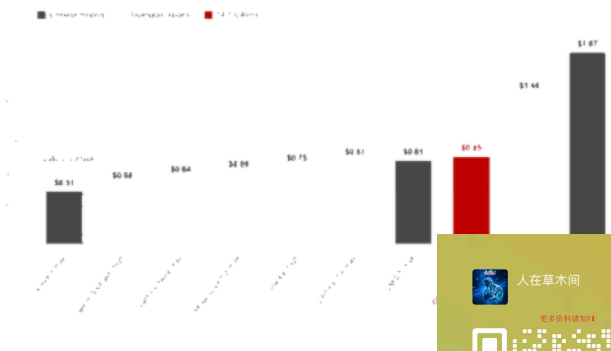
- **Stronger task-solving capabilities**: Frontier models can address more complex, long-horizon tasks across computer use, software engineering, scientific research and professional workflows, supporting higher task success rates and greater value per successful task.
- **Higher token efficiency**: Lower token usage and fewer retries can reduce cost per successful task despite higher per-token pricing. At maximum reasoning effort, Astra used about one-third the number of tokens required by GPT-5.6 Sol per coding task, according to Artificial Analysis. OpenAI also reports lower estimated API cost per task across multiple benchmarks, despite Astra's higher per-token pricing.

Figure 9: GPT-6 Astra uses fewer output tokens than GPT-5.6 Sol on Intelligence Index tasks



Source: Artificial Analysis, UBS

Figure 10: GPT-6 Astra's terminal-bench task cost is comparable with leading Chinese models



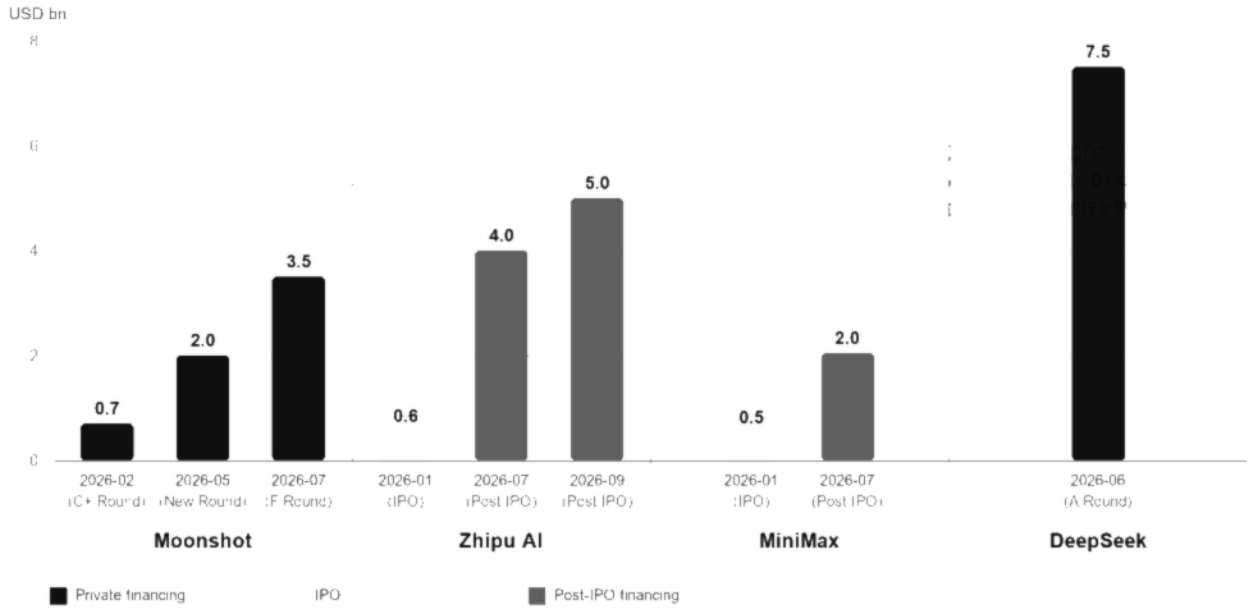
Source: Artificial Analysis, UBS

人在草木间

Recent financing to support compute capacity expansion

We expect recent financing activities to accelerate compute capacity expansion across Chinese AI labs, providing the infrastructure required for next-stage training of larger and more capable foundation models.

Figure 11: Summary of China AI labs' financing in 2026



Source: IT Juzi, company data

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.

Knowledge Atlas Technology (Z. AI/Zhipu): We adopt price-to-sales (P/S) as our valuation methodology for Zhipu, cross checked with SOTP analysis.
Key downside risks include: macro and geopolitical uncertainty that may constrain AI adoption; competitive and operating challenges delaying commercialisation and pressuring margins; potential reduction in key account (KA) customers as their in-house model capability improves; data-related risks affecting model training and performance; regulatory risks related to data security and AI governance; and potential disruptions in access to key computing infrastructure and service providers.
Key upside risks include: easing geopolitical tension and export restrictions that may improve access to advanced hardware and accelerate model upgrades; faster-than-expected improvement in model capability, both in magnitude and iteration speed, supporting higher usage and pricing; technological breakthroughs in model training and inference efficiency that could enhance model capability and margins; stronger-than-expected enterprise adoption, particularly in the to-B/to-G segments, driving faster commercialisation; and the emergence of new AI use cases supporting higher monetisation and revenue growth.

MiniMax: We adopt Price/ARR as our valuation methodology for MiniMax.
Key downside risks include: macro and geopolitical uncertainties that may constrain AI adoption; competitive and operational challenges delaying monetisation and pressuring margins; data-related model training risks; user-generated content governing risks; and disruption of key service providers.
Key upside risks include: easing geopolitical tensions and export restrictions that may improve access to advanced hardware and accelerate model upgrades; technological breakthroughs in AI training and infrastructure efficiency that could enhance model capability and margins; and faster user adoption and commercialisation driving stronger monetisation and profitability.