UOBKayHian

Sector – Update | Greater China

Wednesday, 16 September 2026

China AI

AI Landscape: Efficiency And Adoption Improve, Cybersecurity Emerges, Funding Risks Diverge

Analyst(s)

Julia Pan
juliapan@uobkh.com
+86 21 5404 7225 ext 808

Highlights

- AI model competition is broadening from capability to efficiency, monetisation and global adoption, with cybersecurity emerging as a high-value AI use case. DeepSeek V4.1 Flash highlights lower compute requirements and more dynamic pricing, while MiniMax expands overseas through deeper model adoption. Funding intensity remains high as Z.AI moves upstream into compute infrastructure, while potential IPOs could weigh on sector liquidity.
- Maintain MARKET WEIGHT. Top BUYs: Z.AI and Alibaba.

Analysis

- **Funding intensity remains high□ although September demand is supported by long-term institutional investors.** Z.AI has raised HK$70.6b since July, comprising HK$31.4b net proceeds from its July placement and HK$39.3b in September, including HK$15.7b from new shares and HK$23.6b from a zero-coupon convertible bond (CB). The September share placement represents 4.5% immediate dilution, while full conversion of the CB would take the combined potential dilution to 9.4%. The September deal was launched immediately after the July placement's 60-day issuer lock up expired, while the latest financing introduces another 60-day restriction through mid-November, highlighting near-term funding needs. That said, the investor mix was supportive, with overseas institutions from Asia, Europe and the US forming the majority of demand. Nearly 30 long-only investors participated, including existing holders, alongside close to 20 technology-focused investors. The top 20 investors accounted for over 85% of the equity allocation and over 88% of the CB allocation, suggesting demand was concentrated among larger institutional investors rather than short-term capital.
- **Cybersecurity is emerging as an increasingly important layer of AI infrastructure.** Rising model autonomy and progress in AI-assisted recursive self-improvement could reduce direct human intervention, increasing the need for stronger safeguards. At the same time, cybersecurity could become an early high-ROI enterprise use case for agentic AI, particularly in vulnerability discovery, secure coding and security operations centre automation. Recent incidents highlight the dual-use nature of AI cyber capabilities: an OpenAI autonomous agent gained unauthorised access to Hugging Face systems, while Hugging Face subsequently used Z.AI's GLM-5.2 to support its investigation. As model intelligence improves, we expect AI's role in cybersecurity to expand, with GLM-5.3's stronger cyber benchmarks positioning Z.AI to benefit. Independent evaluation, red-teaming, runtime monitoring and "trusted checkpoints" should also become increasingly important.

MARKET WEIGHT (Maintained)

Segmental Rating

Segment	Rating
E-commerce	MARKET WEIGHT (Maintained)
Online Game	OVERWEIGHT (Maintained)
OTA	OVERWEIGHT (Maintained)
AI	OVERWEIGHT (Maintained)
Online Property Vertical	MARKET WEIGHT (Maintained)
Local Life Service	UNDERWEIGHT (Maintained)
E-commerce	MARKET WEIGHT (Maintained)

Source: UOB Kay Hian

Sector Picks

Company	Ticker	Rec	Share Price (HK$)	Target Price (HK$)
Z.AI	2513 HK	BUY	672.00	1,400.00
Alibaba	9988 HK	BUY	107.20	188.00

Source: Bloomberg, UOB Kay Hian

Peer Comparison

Company	Ticker	Price @ 15 Sep 26 (lcy)	Target Price (lcy))	Upside To TP (%)	Market Cap (lcy m)	-- Calendarised PE -- 2026F (x)	2027F (x)	2028F (x)	Calendarised EV/EBITDA 2026F (x)	2027F (x)	2028F (x)	Calendarised EV/Sales 2026F (x)	2027F (x)	2028F (x)	---------- ROE ----------- 2026F (%)	2027F (%)	2028F (%)
China																	
Z AI Co Ltd	2513 HK	672.00	1,400.00	108.3	312,899	n.a	n.a	n.a	n.a	n.a	419.1	44.8	15.0	6.9	-0.6	41.7	1.8
Minimax Group Inc	100 HK	233.40	470.00	101.4	81,512	n.a	n.a	n.a	n.a	n.a	n.a	19.7	6.8	3.2	12.0	1.4	-8.7
Alibaba Group Holding Ltd	9988 HK	107.20	188.00	75.4	2,131,671	17.0	12.3	9.3	11.7	8.8	6.8	1.3	1.1	1.0	8.2	11.1	13.2
Tencent Holdings Ltd	700 HK	439.80	608.00	38.2	4,003,571	12.8	12.1	11.0	10.5	9.9	9.1	4.3	3.9	3.6	18.7	17.1	16.5
Baidu Inc	9888 HK	89.20	136.00	52.5	243,697	15.8	12.9	11.0	10.1	10.0	8.3	0.8	0.8	0.7	4.2	4.8	5.4
Avg						15.2	12.4	10.4	10.7	9.6	110.8	14.2	5.5	3.1	8.5	15.2	5.6
US																	
Alphabet Inc	GOOG US	345.70	UR	n.a	4,251,164	18.2	21.6	17.6	18.2	14.2	11.0	9.6	7.6	6.2	59.5	42.6	31.5
Meta Platforms Inc	META US	665.60	UR	n.a	1,695,620	19.0	17.4	14.1	14.0	11.3	8.9	6.8	5.6	4.7	29.9	24.1	21.1
Microsoft Corp	MSFT US	505.40	UR	n.a	3,752,945	25.5	21.5	17.7	15.6	12.8	10.4	9.7	8.2	6.7	28.5	26.7	25.5
Avg						20.9	20.1	16.5	16.0	12.7	10.1	8.7	7.1	5.9	39.3	31.1	26.0

Source: Bloomberg, UOB Kay Hian

- **Z.AI is moving upstream into compute infrastructure as its business shifts towards cloud-based APIs.** We think the rising funding requirement partly reflects Z.AI's increasingly capital-intensive expansion beyond the model layer, with 60% of September proceeds earmarked for next-generation GLM R&D and training as well as inference infrastructure, alongside plans for a 1GW AI data centre and the acquisition of XCore Sigma. The infrastructure buildout should also support the company's planned trillion parameter model, expected to launch by end-26, which will require substantially greater training and inference capacity. As Z.AI shifts from on premise deployment towards cloud-based API services, greater control over its own compute infrastructure should improve cost visibility, utilisation and margin control, while supporting more scalable and reliable inference services. However, greater vertical integration also increases capex and liquidity requirements, suggesting external fundraising could remain an important funding source as Z.AI scales.
- **Potential IPOs of DeepSeek，Kimi and Anthropic could create a near-term liquidity vacuum for listed LLM names.** We see some risk that new AI listings could divert investor capital away from existing listed names as investors reassess relative growth, monetisation and valuation. Anthropic is targeting a potential valuation of up to US$2t, equivalent to around 20x its estimated US$100b 2026 annualised revenue run rate (ARR), vs roughly 18x for Z.AI. A successful IPO at this valuation could establish a higher valuation benchmark and support a broader sector rerating. We therefore see the impact as two-sided: near-term liquidity dilution from a larger investable universe, but potentially positive valuation read-through for existing LLM names. Near-term selling pressure should also be partly contained by shareholder lock-ups, although liquidity could remain a key overhang.
- **US AI safety concerns are slowing frontier development and delaying IPO plans，potentially benefitting Chinese AI labs.** Recent incidents involving OpenAI agents, including the May RubyGems incident, where researchers said agents uploaded hundreds of malicious packages, and the July Hugging Face breach, where OpenAI agents accessed external systems, have intensified concerns over the controllability of increasingly capable AI agents. Anthropic's Dario Amodei has called for a slower pace of frontier capability development, with Sam Altman and Elon Musk broadly supporting stronger pacing and safety safeguards. OpenAI has also put its IPO plans on hold amid rising safety concerns. In contrast, China continues to advance frontier model development alongside tighter safety oversight. We believe any moderation in US frontier model development could give Chinese LLM developers more time to narrow the capability gap, potentially improving their relative positioning in the global AI landscape.
- **At the 2026 Bund Summit，Ant highlighted agent security and trusted AI as key themes.** Ant Group (Ant) expanded its Ling 3.0 model family into multimodal and vertical-finance use cases, launching the 124B Ling-3.0-flash-VL and Ling-3.0-flash-Fin. Rather than competing purely on frontier-scale intelligence, Ant appears to be emphasising efficient sparse architectures and agentic workflows across financial services, payments and enterprise applications. Agent security and trusted AI were also clear themes at the Bund Summit, with Ant highlighting authorisation, identity, capability assessment, fund security and auditability as key requirements as agents increasingly execute transactions on users' behalf. This positions Ant more around secure agent infrastructure and trusted execution, ensuring agents can safely access accounts, payments and external tools, rather than cybersecurity as a model capability, such as vulnerability discovery, penetration testing or cyber reasoning.
- **DeepSeek launched V4.1 Flash with performance surpassing V4 Pro at materially lower inference costs.** DeepSeek released V4.1 Flash on 10 September, a 552b parameter MoE model using a new Causal Encoder Decoder architecture, while activating only 8b parameters for input and 16b for output, equivalent to just 1.4% and 2.9% of total parameters, respectively.

Despite the low activation ratio, internal and third-party testing indicates that V4.1 Flash matches or exceeds V4 Pro across performance, response speed, usage cost and total processing time. Infrastructure requirements were also materially reduced, with high bandwidth memory (HBM) demand cut to 1/4 and solid-state drive (SSD) demand to 1/8 vs the previous generation. Pricing remains highly competitive at Rmb1/Rmb4 per 1m uncached input/output tokens respectively during off peak hours and Rmb2/Rmb8 during peak hours, while cached input costs only Rmb0.02/Rmb0.04. We think V4.1 Flash highlights how architectural efficiency can deliver higher-than-flagship-level performance at significantly lower compute and inference costs.

- **DeepSeek adopts peak/off-peak pricing while cutting Flash prices.** DeepSeek introduced peak/off-peak API pricing in August, with off-peak prices set at 50% of peak rates to better balance compute demand. Following the launch of V4.1 Flash, it further cut Flash API prices from 10 Sep 26, with off-peak cache-hit input, cache-miss input and output prices reduced by 60%/33%/11% to Rmb0.02/Rmb1.0/Rmb4.0 per 1m tokens, respectively; peak-hour prices remain 2x off-peak rates. Our take: the move suggests LLM pricing is becoming more dynamic rather than simply trending lower, with lower off-peak pricing stimulating usage and shifting flexible workloads away from peak periods, while higher peak pricing helps monetise scarce compute capacity. We see this as supportive of the view that differentiated models can retain pricing power even as headline token prices decline.
- **Pricing pressure is diverging between frontier and commoditised models.** We believe OpenAI's sharp price cut for GPT 5.6 Luna partly reflects the cost performance advantage offered by Chinese models, as capability gaps narrow at the lower end of the market. In contrast, newly launched SOTA models continue to command substantial pricing premiums, suggesting that stronger capabilities still support pricing power. GPT 6 Astra is priced at US$10/US$50 per 1m input/output tokens respectively for short context, vs US$1.4/US$4.4 for GLM 5.3, leaving Chinese frontier models such as GLM meaningful pricing headroom if they continue to close the performance gap with leading US models. Meanwhile, pricing pressure remains more intense among price-performance-focused models from DeepSeek, MiniMax and Xiaomi's MiMo, where continued capability gains alongside aggressive API pricing could limit pricing upside across the broader market.
- **MiniMax is expanding overseas through deeper model-level adoption beyond traditional API distribution.** Its open-weight M3 is increasingly being used as an underlying foundation model for overseas applications and localised model development. Genspark adopted M3 as the base model for its Gen-1 Slides product, while Saudi Arabia's HUMAIN further trained M3 to launch an Arabic-focused HUMAIN M3 model. We see this as an important evolution in MiniMax's international expansion, as adoption is extending from API usage into application integration, post-training and local deployment, which could broaden its developer ecosystem and support longer-term enterprise/API monetisation.
- **OpenAI pushes the frontier towards agentic AI with GPT-6 Astra.** OpenAI launched GPT-6 Astra, positioning the model around deeper reasoning and autonomous task execution rather than another broad-based chatbot upgrade. Astra delivered significant gains in scientific reasoning, computer use and workflow automation, including 72.6% on OSWorld2.0 and 41.4% on Automation Bench, with the latter more than doubling GPT-5.6 Sol's 18.1%. OpenAI's president described the launch as the arrival of the artificial general intelligence (AGI) era, although we would view this more as a signalling point than a formal AGI milestone.

Foundation Model Launches And Architecture

Company/model	Release	Update type	Parameters and architecture	Commercial implication
DeepSeek / V4.1 Flash	10 Sep 26	Model refresh / multimodal upgrade	552B MoE; 8B active for input / 16B for output. Causal encoder-decoder architecture.	Different compute allocation for input processing and generation.
OpenAI / GPT-6 Astra	3 Sep 26	New generation / flagship model	Not publicly disclosed.	Premium general-purpose agent model.
Alibaba / Qwen3.8-Max-0902	2 Sep 26	Model refresh	2.4T/95B active; MoE and hybrid attention; 1M context.	Model capability supports MaaS and cloud demand.
Google / Gemini 3.8 Flash	2 Sep 26	Variant expansion	Parameter count not disclosed; text, image, audio and video; 1M context.	Aggressive introductory pricing and broad distribution.
Anthropic / Fable 5.1; Mythos 5.1	1 Sep 26	Deployment / safety update	Parameters not disclosed. Mythos access is restricted to vetted uses.	Access policy differentiates products as well as model capability.
Tencent / Hy4 preview	28 Aug 26	New-generation preview	770B/49B active backbone; gated DSA, IndexCache and iHC.	Serving efficiency complements enterprise and consumer distribution.
Z.ai / GLM-5.3; 5.3 Flash	14 Aug 26	Model refresh / variant expansion	5.3: HF artifact 753B, MoE/DSA; active count not restated. Flash: 320B total / 18B active; hybrid sparse/linear attention, mHC.	Flagship refinement and a distinct low-cost serving model.
Moonshot / Kimi K3	17 Jul 26	New generation	2.8T / 104B active; KDA and gated MLA; native low-precision training.	Large active network targets demanding coding and agent tasks.
ByteDance / Seed2.1	23 Jun 26	Model refresh / variant expansion	Parameter count not disclosed; multimodal architecture supporting long-context reasoning, coding and tool use.	Supports productivity agents and enterprise/API distribution through Volcano Engine and ByteDance products.
MiniMax / M3	1 Jun 26	New generation	About 428B/23B active; MiniMax Sparse Attention; text, image, video; 1M context.	Long-context serving plus a broader media product portfolio.
Baidu / ERNIE 5.1	9 May 26	Model refresh	Official emphasis: asynchronous RL and agentic post-training. Exact current size not established here.	Monitor adoption and Qianfan conversion, not release cadence alone.

Source: Respective companies, UOB Kay Hian

Frontier LLM API Pricing Comparison

Model	Input (US$/1m tokens)	Cache read Input (US$/1m tokens)	Output Input (US$/1m tokens)
GPT-6 Astra	10.0	1.0	50.0
Claude Fable 5.1	10.0	0.25	50.0
Kimi K3	3.0	0.3	15.0
Qwen3.8-Max-0902	2.0	0.25	6.0
GLM-5.3	1.4	0.26	4.4
Gemini 3.8 Flash	0.75	0.075	3.75
Tencent Hy4 preview	0.834	0.042	2.501
MiniMax M3, >512K	0.6	0.12	2.4
MiniMax M3, ⩽512K	0.3	0.06	1.2
DeepSeek V4 Pro (Peak)	1.32	0.044	3.96
DeepSeek V4.1 Flash (Peak)	0.3	0.006	1.2

Source: Respective companies, UOB Kay Hian

Selected Chinese LLMs – Flash Version: Model Configuration□Pricing and Use Cases

	Alibaba	Z.ai	DeepSeek	DeepSeek
Model	Qwen3.8-Flash	GLM-5.3 Flash	V4-Flash (Peak)	V4.1-Flash (Peak)
Release date	Aug-26	Aug-26	Jul-26	Sep-26
License	Qwen Community License 1.0	MIT	MIT	MIT
Total parameter (B)	125	320	284	552
Active parameter (B)	6	18	13	—
Prefill	—	—	—	8
Decode	—	—	—	16
Context window (K)	1,000	1,000	1,000	1,000
API price (Rmb/1m tokens)				
Input (uncached)	0.8	0.8	3.0	2.0
Input (cached read)	0.1	0.23	0.1	0.04
Output	2.7	2.8	9.0	8.0
AA Intelligence Index (v4.3)	40	42	35	40
Target workload	High-volume agents, coding & coworking	Multimodal coding, long-horizon agents & professional work	Low-cost reasoning & agentic coding	High-throughput multimodal agents & coding

Source: Respective companies, UOB Kay Hian

Valuation/Recommendation

- **Z.AI (2513 HK).** Maintain BUY with a lower target price of HK$1,400.00. Self-built data centres should provide greater control over costs and compute capacity while reducing reliance on external cloud providers, while the planned trillion-parameter model could help narrow the capability gap with global frontier models.
- **MiniMax (100 HK).** Maintain BUY with a target price of HK$470.00. Broader overseas adoption should support ecosystem expansion and monetisation, while its native multimodal capabilities and planned trillion-parameter model could further strengthen model capabilities and competitive positioning.
- **Alibaba (9988 HK).** Maintain BUY with a target price of HK$188.00. Its full-stack AI capabilities, spanning data centres, cloud infrastructure and models, provide greater control over compute capacity and costs, while its model portfolio covers both frontier performance and cost-efficient deployment. Ant's focus on trusted agent infrastructure could further broaden secure agent adoption across financial and enterprise use cases.

Sector Catalyst/Risk

- **Catalysts:** a) Stronger frontier-model adoption and monetisation, b) easing domestic compute constraints, and c) improving model/inference efficiency and broader overseas adoption.
- **Risks** include: a) placement/CB dilution, b) sector liquidity diversion from new AI IPOs, c) higher funding intensity, d) intensifying model pricing competition, and e) tighter regulatory/geopolitical constraints.

IMPORTANT NOTICE - DISCLOSURES AND DISCLAIMERS

This report is provided subject to various disclosures and disclaimers (the "Disclosures / Disclaimers") which form an integral part of this report and are available at this link or by scanning the QR code below:

The Disclosures / Disclaimers contain important information, including without limitation, (a) exclusions of liability, (b) confidentiality obligations, (c) restrictions on publication, circulation, reproduction, distribution and use of the report, (d) potential conflicts of interest, and (e) disclosures and requirements specific to recipients in the United States and other applicable jurisdictions.

Specifically, this report is intended for general circulation and informational purposes only and does not take into account the specific investment objectives, financial situation, or particular needs of any individual person. It is not intended to constitute personal investment advice or a recommendation to buy or sell any investment product or security. You should independently evaluate the information and, where necessary, seek advice from a qualified financial adviser regarding the suitability of any investment, taking into account your specific objectives, financial situation and needs, before making any investment decision. Analyst certifications required under applicable regulations, including SEC Regulation AC (where relevant), are included in this report.

Recipients of this report must carefully read, review and understand the full Disclosures / Disclaimers before using or relying on any information in this report. By accessing, receiving or using this report, you acknowledge and confirm that you have read, understood, accepted and agreed to be bound by the Disclosures / Disclaimers (as may be amended or updated from time to time) in full."