

Internet

AI Model: Scaling Intelligence, Winning On Cost

Highlights

- China's AI model market is entering a hyper-growth phase, driven by accelerating domestic enterprise adoption and growing consumer/SME demand overseas. We are bullish on leading Chinese AI labs that are supported by improving reasoning, easing compute constraints, and frontier R&D, while cost-effective open-weight models support overseas adoption.
- Initiate coverage on Z.ai (2513 HK) and MiniMax (100 HK) with a BUY.

Analysis

- China's AI model market is entering a hyper-growth phase**, with accelerating domestic enterprise and international adoption powering the upside momentum. We forecast AI model application programming interface /subscription revenue growing from Rmb35b in 2026 to Rmb900b in 2030, with daily token consumption rising from 350T tokens to 4,500T (13x). Chinese models should increasingly gain traction in non-US markets as improving intelligence and lower token costs appeal to cost-sensitive SMEs and 24/7 autonomous-agent deployments.
- Chinese models are rapidly narrowing the performance gap with US peers while maintaining a structural cost advantage.** Chinese developers are quickly catching up with US players across reasoning, coding and multimodal AI, with Kimi K3 slightly exceeding Anthropic Opus 4.8 on the Artificial Analysis Intelligence Index just seven weeks post-launch. China's agent and coding token usage has rapidly reached 63% of global usage vs 32% a year ago – evidence of fast global adoption.
- Inference efficiency and frontier R&D are increasingly becoming the key differentiators.** Chinese foundation models have improved coding and agentic performance, with context windows expanding to 1m tokens. Chinese AI labs have significantly improved inference efficiency through efficient mixture of experts architectures, key-value cache optimisation, context compression, retrieval-augmented generation and advanced attention to cost control. We believe the talent, compute and R&D barriers should accelerate industry consolidation, while top models such as GLM-5, M3 and Kimi K3 highlight the fast shifts in model leadership.
- Initiate coverage on Z.AI Co (2513 HK/BUY/Target: HK\$1,830.00) and MiniMax Group Inc (100 HK/BUY/Target: HK\$470.00).** We believe Z.ai remains well positioned with its research talent and we prefer Z.ai for its strong coding performance and better pricing power, while MiniMax stands out with its differentiated multimodal capabilities and superior cost efficiency.

Analyst(s)

Julia Pan

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

MARKET WEIGHT (Maintained)

Segmental Rating

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

Source: Bloomberg, UOB Kay Hian

Key Picks

Company	Ticker	Rec	Share Price (HK\$)	Target Price (HK\$)
Z.ai	2513 HK	BUY	1,030.00	1,830.00
MiniMax	100 HK	BUY	303.00	470.00

Source: Bloomberg, UOB Kay Hian

Figure 1: Peer Comparison

Company	Ticker	Price @ 26 Aug 26 (lcy)	Market Cap (lcy m)	----Calendarised PE----			--Calendarised EV/EBITDA--			--Calendarised EV/Sales--			----- ROE -----		
				2026F (x)	2027F (x)	2028F (x)	2026F (x)	2027F (x)	2028F (x)	2026F (x)	2027F (x)	2028F (x)	2026F (%)	2027F (%)	2028F (%)
China															
Z AI Co Ltd	2513 HK	1,030.00	479,592	n.a	n.a	n.a	n.a	n.a	n.a	94.5	31.9	14.1	7.4	25.3	-8.9
MiniMax Group Inc	100 HK	303.00	105,818	n.a	n.a	n.a	n.a	n.a	n.a	34.6	12.6	5.8	-31.6	-5.1	888.2
Meitu Inc	1357 HK	4.30	19,653	13.8	10.7	8.5	9.2	6.7	4.7	2.8	2.3	1.9	18.0	19.6	21.3
Shenzhen Xunce Technology	3317 HK	122.90	40,554	253.9	64.9	39.5	246.4	67.3	32.1	14.6	9.2	6.3	3.5	14.8	23.8
Alibaba Group Holding Ltd	9988 HK	116.60	2,235,800	18.1	12.8	10.1	11.9	9.1	7.3	1.3	1.2	1.0	8.2	11.1	13.2
Tencent Holdings Ltd	700 HK	445.40	4,054,540	13.0	12.3	11.3	11.0	10.2	9.2	4.3	4.0	3.6	18.7	17.2	16.5
Baidu Inc	9888 HK	90.90	246,656	15.4	12.3	10.1	11.3	9.9	8.1	0.8	0.8	0.7	4.4	5.1	5.8
Kuaishou Technology	1024 HK	33.60	145,472	11.2	11.7	9.8	5.7	5.8	5.1	0.7	0.7	0.6	9.3	9.0	9.5
Avg				54.2	20.8	14.9	49.3	18.2	11.1	19.2	7.8	4.3	4.8	12.1	121.2
US															
Alphabet Inc	GOOG US	343.30	4,221,800	18.1	20.9	17.6	18.4	14.1	10.9	9.5	7.5	6.1	57.7	40.0	29.3
Meta Platforms Inc	META US	570.10	1,452,210	16.2	14.9	12.3	12.0	9.7	7.8	5.8	4.8	4.1	30.1	24.1	21.1
Microsoft Corp	MSFT US	491.70	3,651,210	24.9	20.9	17.4	15.2	12.4	10.0	9.5	8.0	6.6	28.1	26.4	25.3
Apple Inc	AAPL US	309.90	4,522,740	35.2	32.4	29.1	26.1	24.4	22.2	9.4	8.6	8.0	135.8	114.0	88.7
Avg				23.6	22.3	19.1	17.9	15.2	12.7	8.5	7.2	6.2	62.9	51.1	41.1

Source: Bloomberg, UOB Kay Hian

Contents

Investment Highlights	Page 3 →
Sector Outlook	Page 4 →
Valuation	Page 11 →
Appendix	Page 12 →
Stock Picks – Z.AI Co (2513 HK)	Page 19 →
Investment Highlights	Page 20 →
Valuation	Page 21 →
Business Outlook	Page 22 →
Financial Statements	Page 29 →
Appendix I: Risk Factors	Page 30 →
Appendix II: Company Overview	Page 31 →
Appendix III: Management	Page 39 →
Appendix IV: Organisational Structure	Page 40 →
Stock Picks – MiniMax Group Inc (0100 HK)	Page 41 →
Investment Highlights	Page 42 →
Valuation	Page 43 →
Business Outlook	Page 44 →
Financial Statements	Page 51 →
Appendix I: Risk Factors	Page 52 →
Appendix II: Company Overview	Page 53 →
Appendix III: Management	Page 62 →
Appendix IV: Organisational Structure	Page 63 →

This report uses the closing prices of 26 Aug 26

Investment Highlights

China's AI model market is entering a hyper-growth phase, with AI labs reaching a critical intelligence level, driven by accelerating domestic enterprise adoption and growing consumer/SME demand overseas. We forecast AI model application programming interface (API)/subscription revenue growing from Rmb35b in 2026 to Rmb900b in 2030, with daily token consumption rising from 350 tokens to 4,500 (~13x). Chinese models should increasingly gain traction in non-US markets as improving intelligence and lower token costs appeal to cost-sensitive SMEs and 24/7 autonomous-agent deployments. We forecast a three-year revenue CAGR of 123% for MiniMax and 141% for Z.ai in 2026-29, with both breaking even in 2029.

Chinese models are narrowing the performance gap with US peers while retaining a pricing advantage. Chinese AI labs are progressing rapidly across reasoning, coding and multimodal capabilities and the growth is likely to remain non-linear, with capability breakthroughs driving sharp increases in adoption, distribution and revenue before competition shifts toward cost efficiency, reliability and distribution. The latest US models still maintain an advantage at the frontier, while the generally lower pricing of Chinese models supports their competitiveness in high-volume and cost-sensitive deployments.

Inference efficiency and frontier R&D are key differentiators. Chinese foundation models have materially improved coding and agentic performance over the past year, with context windows expanding to 1m tokens. As 1m-token contexts add coding and long-horizon agent use cases, key-value (KV) cache optimisation, context compression, retrieval-augmented generation (RAG) and advanced attention are critical to cost control. In addition, efficient mixture of experts (MoE) architectures, post-training and real-world data flywheels are further improving cost efficiency. Recent model launches highlight continued shifts in model leadership, while Z.ai remains well positioned given its R&D pedigree, Tsinghua affiliation and major internet-platform customers. In our view, Z.ai is well positioned due to its research talent, while Kimi's commercial licensing could help Chinese labs capture more value from global deployment.

Frontier AI remains difficult to replicate, supporting eventual industry consolidation. Sustained leadership requires deep research expertise, specialised talent, improving compute and rapid R&D execution. We expect the frontier to consolidate around fewer players, with large platforms leveraging scale and distribution and focused labs competing on R&D speed. In the near term, we expect frontier coding workloads to remain concentrated among leading models, while lower-end agentic AI stays more fragmented.

Domestic compute adoption expanding into frontier-scale model development. Z.ai has also reportedly completed a 1GW China-chip data centre. Meanwhile, DeepSeek and Z.ai are exploring proprietary inference chips. These developments suggest China's domestic hardware and software stack is reducing its reliance on restricted foreign GPUs, although compute constraints continue to ease gradually rather than be fully resolved, as proprietary-chip efforts are still early-stage and focusing mainly on inference.

Future trend – Bifurcating market. We expect frontier models to retain pricing power, while lower-end models face price competition but benefit from rapidly-growing demand for cost-efficient agentic AI. For 2H26, we foresee: a) multiple Chinese AI model launches with 2T-5T parameters; b) intensifying competition in coding/programming as peers challenge the Z.ai general language model (GLM) through higher-quality real-world coding data and greater model scale; c) leading models shifting from open-source to open-weight licensing, introducing revenue sharing/take rates for commercial use; and d) multimodal and visual reasoning becoming another key upgrade area for models like GLM and DeepSeek, while MiniMax M3 already has a strong multimodal focus. Against this backdrop, R&D capability, pricing power, cost efficiency, financial strength, token scale and market share gains will increasingly differentiate the winners from among large language models (LLM).

Rapid adoption and growing token usage underpin a multi-year monetisation runway

Research intensity and execution barriers favour consolidation at the frontier

Local hardware adoption is strengthening China's AI infrastructure independence

Pricing power should concentrate at the frontier as lower-end competition intensifies

Sector Outlook

China’s AI open-source/open-weight models are approaching a critical intelligence threshold, driven by accelerating domestic enterprise adoption and growing consumer/SME demand overseas. We are bullish on leading Chinese AI labs’ ability to stay near the global frontier, supported by improving reasoning capabilities, easing domestic compute constraints and the “lighthouse effect” of frontier R&D. We expect stronger cost effectiveness and open-source adoption to support domestic and overseas growth. China’s AI market has consolidated from over 200 credible model developers to fewer than 10, with competition centred on tech giants and agile independents such as Z.ai and MiniMax. We prefer Z.ai for its stronger coding performance and pricing power, while MiniMax offers differentiated multimodal capabilities and superior cost efficiency.

Chinese AI models are narrowing the gap with US peers through efficient architectures, post-training and real-world data flywheels. Chinese AI lab growth is likely to remain non-linear, with capability breakthroughs driving sharp increases in adoption, distribution and revenue before competition shifts toward cost efficiency, reliability and distribution.

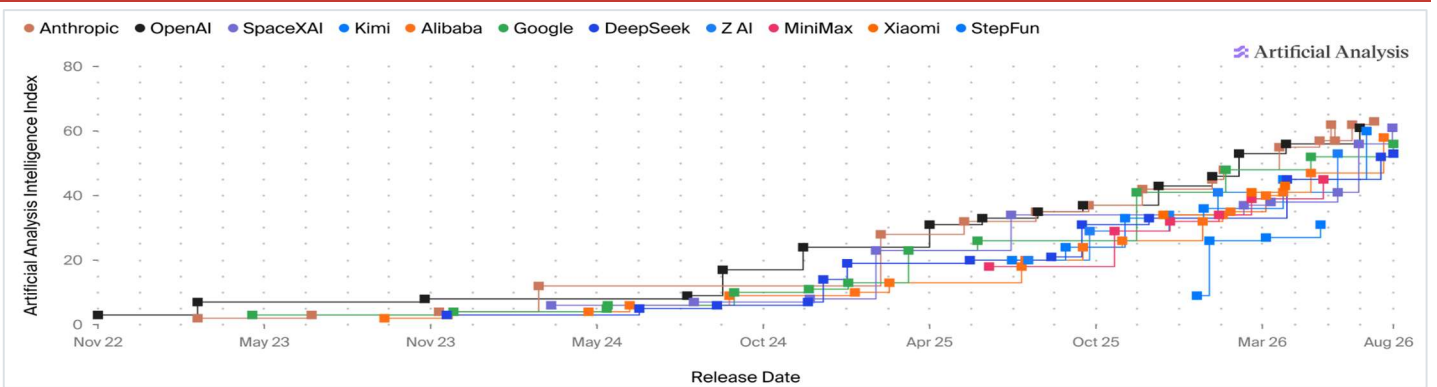
US-developed models continue to define the global intelligence frontier, with leading US laboratories occupying the upper end of benchmark rankings. Chinese models, however, are progressing rapidly across reasoning, coding and multimodal capabilities. Historically, Chinese models have lagged US state-of-the-art (SOTA) by 6-9 months, but GLM-5.3 and Kimi K3 suggest the gap is narrowing despite an order-of-magnitude difference in compute access and pre-training scale in the two countries. On the Artificial Analysis Intelligence Index, Kimi K3 marginally exceeded the performance level of Anthropic’s Opus 4.8 around seven weeks after its launch, indicating a shorter catch-up cycle among leading Chinese developers.

The latest US models still maintain an advantage at the frontier, while the generally lower pricing of Chinese models supports their competitiveness in high-volume and cost-sensitive deployments.

Chinese models are entering a stronger adoption and monetisation phase

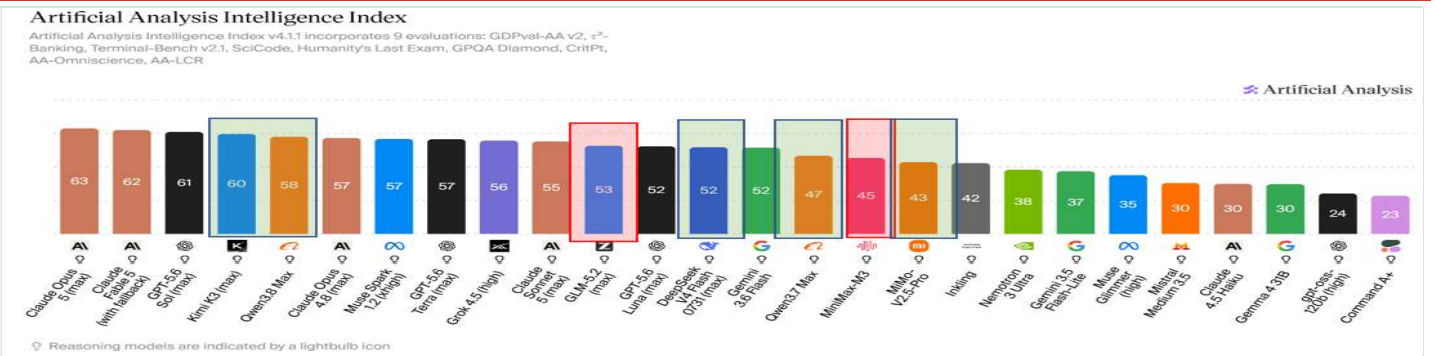
Architecture efficiency and post-training are narrowing the frontier performance gap

Figure 2: Narrowing Gap Between Chinese And US Frontier AI Labs



Source: Artificial Analysis, UOB Kay Hian

Figure 3: Chinese Players Reaching Global Top-tier Positions, Notably In Coding/agentic Tasks



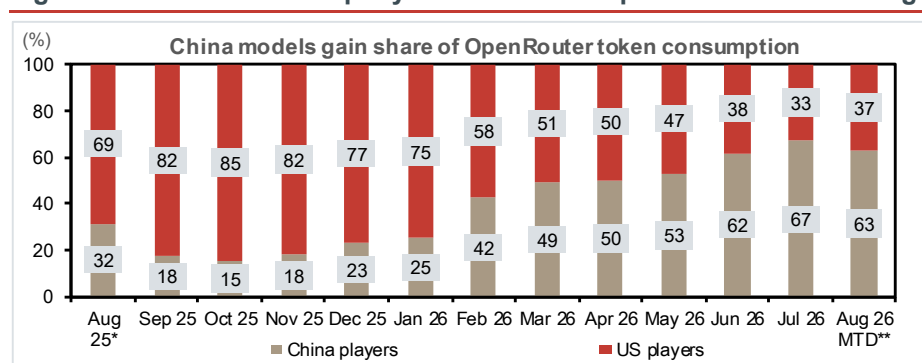
Source: Artificial Analysis, UOB Kay Hian

Chinese models are capturing a growing share of token usage while monetisation remains comparatively low. The global token share of Chinese AI models has risen significantly despite a lower revenue share, reflecting a significant pricing gap vs US models. China accounts for 63% of agent and coding token usage, yet captures only less than 20% of revenue, as substantially lower API pricing limits monetisation per token. We believe this usage-revenue gap provides significant upside as Chinese models continue to improve frontier intelligence, broaden enterprise and overseas adoption, and gradually monetise higher-value agentic and coding workload.

Figure 4: SDLLMTK Index



Figure 5: China Models Rapidly Gain Share Of Open Router Token Usage



Divergence in global foundation model pricing highlights China's cost advantage. Chinese foundation models generally price tokens materially below those of leading US peers due to intense domestic competition, more efficient model architectures and inference optimisation, as well as potentially lower infrastructure and energy costs. Better cloud-resource utilisation and workload optimisation could further improve cost efficiency. While leading US models continue to command a premium for their stronger capabilities, enterprise ecosystems and global positioning, Chinese models are increasingly competitive with their high-volume inference and cost-sensitive enterprise workloads, which support broader international adoption.

Structural cost advantages are supporting broader adoption

Figure 6: Price Comparison Of Selected Top Models From China And US

Provider	Model	Context / input-length tier	Input (US\$)	Cache-hit input (US\$)	Output (US\$)
Kimi	Kimi K3	1m	3.0	0.3	15.0
Qwen	Qwen3.8-Max	1m	2.0	0.25	6.0
Z.ai	GLM-5.2	1m	1.4	0.26	4.4
MiniMax	MiniMax-M3	≤512k	0.3	0.06	1.2
		>512k	0.6	0.12	2.4
DeepSeek	DeepSeek-V4-Pro-0813	1m (OFF-PEAK)	0.66	0.022	1.98
		1m (PEAK)	1.32	0.044	3.96
	DeepSeek-V4-Flash 0731	1m (OFF-PEAK)	0.22	0.007	0.66
		1m (PEAK)	0.44	0.014	1.3
Anthropic	Claude Fable 5	1m	10.0	1.0	50.0
OpenAI	GPT-5.6 Sol	1m	5.0	0.5	30.0
	GPT-5.6 Terra	1m	2.0	0.2	12.0
	GPT-5.6 Luna	1m	0.2	0.02	1.2
Google	Gemini 3.1 Pro Preview	≤200k	2.0	0.2	12.0
		>200k	4.0	0.4	18.0
	Gemini 3.7 Flash (Promo)	1m	0.75	0.075	3.75

Source: Respective companies, industry data, UOB Kay Hian

CHINESE FOUNDATION MODELS ACHIEVING COMPETITIVE PERFORMANCE DESPITE LIMITED COMPUTE RESOURCES

Compute-efficient architectures optimising China's cost. Adoption is expanding across always-on agents and productivity use cases, while global SMEs are increasingly open to Chinese models as their performance has reached a "good enough" threshold, making their cost efficiency more appealing in model selection. Meanwhile, Chinese foundation models have materially

Lean model design sustains China's cost performance edge

improved coding and agentic performance over the past year, with context windows expanding to 1m tokens. Despite constrained access to high-end GPUs, leading models remain relatively compact at 200B-1.6T parameters, or just 2-10% of that of the largest frontier models, while MoE, sparse attention and other architectural innovations limit active parameters to 3-5%. Recent capability gains appear increasingly driven by data curation, distillation and reinforcement learning (RL) post-training rather than parameter scale alone, as demonstrated by DeepSeek's V4 Pro (1.6T), Z.ai's GLM-5.3 (744B) and MiniMax's M3 (428B).

System-level optimisation is further improving economics: DeepSeek's DSpark increased inference speed by 60-85% on V4-Flash and 57-78% on V4 Pro without changing model weights or output quality. As fast followers, they can leverage frontier advances while reducing exploratory R&D costs, supporting a stronger cost-performance.

Inference efficiency and intelligence, rather than parameter scale alone, as key differentiators. Kimi K3 challenges the "closed = intelligent, open = cheap" paradigm, with its 2.8T-parameter model reportedly ranking #1 on WebDev Arena and pricing output at Rmb100/1m tokens, vs DeepSeek V4-Pro (Rmb13.5), GLM-5.3 (Rmb28) and Qwen3.8-Max (Rmb36), suggesting frontier intelligence can support pricing power even for open-weight models. Meanwhile, 1m-token context windows add use cases but raise memory and inference costs, shifting competition toward efficient long-context execution through KV-cache optimisation, context compression, RAG and advanced attention mechanisms. For Moonshot AI, open-weight distribution of its 3T-parameter K3 can boost developer and enterprise adoption while shifting value capture toward chips, cloud, inference platforms and ecosystem services, although model-level pricing power may decline as competition intensifies.

Intelligence and inference efficiency drive pricing power

Real-world adoption creating positive data flywheel. In addition, efficient MoE architectures, post-training and real-world data flywheels are further improving cost efficiency. As Chinese models gain enterprise and global adoption, access to proprietary real-world usage data should increasingly reduce their reliance on synthetic or distilled data from global SOTA models. AI-generated code as a proportion of code cases reportedly reached up to 90% at some Chinese mega-caps, vs 20-30% in 2H25, creating a growing pool of successful and failed coding cases for RL and post-training. We believe this feedback loop could have contributed to the rapid improvement from GLM-5.1 to GLM-5.3 and would further drive capability gains over the next 6-12 months.

Usage scale is turning proprietary data into a compounding competitive advantage

MARKET DIVERGENCE

From token growth to ROI-driven agentic AI. Enterprises are increasingly focusing on task completion and cost efficiency rather than token consumption. Heavy AI users consumed 10x more tokens but generated only 2x more output, while some US internet companies depleted a full year's AI budget within four months of adopting agentic tools. We therefore expect a barbell market, with cheaper flash models handling routine tasks and SOTA models reserved for high-value workloads. Cost per completed task, ROI, daily active agents (DAA) and agentic work units (AWU) should become increasingly important metrics for model selection and monetisation.

ROI replaces token growth as the key metric

Bifurcating market. We believe frontier models retain pricing power, while lower-end models face price competition but benefit from rapidly-growing demand for cost-efficient agentic AI. Frontier models such as Z.ai's GLM-5.3 and Alibaba's Qwen3.8 Max command around US\$1 per 1m blended tokens, about 5x higher than that of low-end Chinese models but still well below US SOTA pricing of US\$4-8. We estimate leading Chinese models can sustain 20-30% gross margins. At the low end, agentic-focused models priced at US\$0.06-0.20 per 1m tokens are expanding adoption among price-sensitive SMEs and one-person businesses, with MiniMax already generating more than 60% of revenue overseas. DeepSeek's peak-hour pricing at 2x non-peak rates, implying blended pricing of US\$0.35/US\$0.12 per 1m tokens for V4 Pro/Flash respectively, also highlights rising compute demand in productivity workloads.

Market bifurcation supports differentiated pricing

COMPETITION: FROM CAPABILITY CONVERGENCE TO INDUSTRY CONSOLIDATION

Frontier coding competition is set to intensify. Following Kimi K3, we expect a new wave of 2T-5T-parameter models from Z.ai, MiniMax and ByteDance, with coding, agentic AI and multimodal reasoning as the key battlegrounds. Faster gains from post-training, supervised fine-tuning (SFT) and RL are narrowing performance gaps and increasing user switching, making pricing, inference cost and latency increasingly important differentiators. Potentially tighter Chinese restrictions on overseas access to Chinese frontier model weights and training data could pose an additional risk to global adoption.

Frontier AI remains difficult to replicate and could lead to industry consolidation. Research pedigree remains a key barrier to replicating frontier AI as sustaining frontier reasoning requires deep research capabilities, specialised talent, large compute capacity and rapid organisational execution. Although US labs retain a substantial compute edge, leading Chinese labs have partly offset this through architectural innovation, talent recruitment and strong R&D capabilities. As the technical and capital barriers rise, we expect to see fewer leading labs at the frontier, with large platforms leveraging scale, funding and distribution while focused independents compete through R&D speed.

Recent launches, from Z.ai's GLM-5.3 and MiniMax's M3 to Moonshot AI's Kimi K3, also illustrate how quickly model leadership and market expectations can shift, implying continued differentiation and volatility across AI names. We believe Z.ai is well positioned given its wealth of foundational science talent, Tsinghua affiliation and broad internet-platform customer base, while Kimi K3's commercial licensing model could provide a template for Chinese labs to capture more value from global adoption.

Low-end agentic AI remains fragmented. Open-source/open-weight models account for around 30% of China's token consumption, vs approximately 70% for ByteDance's closed-source ecosystem. Similar to the US, premium coding workloads in China is concentrated among frontier models, while lower-end agentic AI remains highly competitive given sustained investment and pricing subsidies from both independent AI labs and hyperscalers.

Capability convergence shifts competition toward cost and execution

Rising barriers drive industry consolidation

Pricing remains the key battleground in lower-end agentic AI

Figure 7: Competitive Positioning Across Leading AI Model Companies

Dimension	Z.ai	MiniMax	DeepSeek	Moonshot (Kimi)	ByteDance
Monetisation model	Dual: On-premise + Cloud/API (MaaS); Shifting rapidly toward API	Dual-engine: AI-native consumer apps (To-C) + open platform APIs (To-B)	Primarily open-weight models + API; strong efficiency focus	Open-weight models with commercial licensing push	Closed ecosystem + proprietary models; heavy distribution via own platforms
Model strategy	Open-weight frontier models + commercial API	Open-weight models + proprietary consumer apps/API	Open-weight first	Open-weight frontier models + commercial API	Predominantly proprietary/closed
Geographic exposure	China-led; scaling overseas rapidly	Most global pure play; majority of revenue from overseas	Strong global developer adoption	Growing global adoption	Predominantly China
Core capability	Coding, agentic AI, long-context, enterprise reasoning	Native multimodal (video, speech, music) + cost-efficient LLMs	Inference efficiency, coding/agentic, cost leadership	Frontier intelligence + coding (Kimi K3 highlighted)	Multimodal + closed ecosystem scale
Pricing	Increasing pricing power alongside improving intelligence	Cost-performance focused	Low-price benchmark/price disruptor	Premium pricing for frontier capability	Competitive pricing supported by scale
Distribution	Enterprise installed base + fast-growing coding/API ecosystem	Global consumer apps + developer/API ecosystem	Open-weight distribution (high token share)	Open-weight developer ecosystem	Unmatched domestic consumer + enterprise distribution
Key moat	Intelligence + enterprise monetisation + coding	Multimodal + global user base	Cost efficiency + open distribution	Frontier capability + open distribution	Very strong via own apps and closed ecosystem (~50% of China token consumption cited)
Key constraint	Compute access; execution against high expectations	LLM differentiation outside multimodal	Monetisation/pricing power	Monetisation scale	High capital intensity
Investment read-through	Best listed exposure to coding/agentic monetisation and pricing power	Best listed exposure to AI video/multimodal and overseas growth	Key open-weight benchmark and price disruptor	Key frontier challenger	Structural ecosystem leader

Source: Industry date, respective companies, UOB Kay Hian

Chinese labs may benefit from increasing model commoditisation. With hardware payback in reportedly as short as 10 months, price competition is expected to be driven more by model and inference efficiency than price wars. Given the large total addressable market (TAM), investment in next-generation models should offer a higher ROI than competing purely on pricing, while leading labs are likely to gain greater control over compute capacity.

Efficiency, not price wars, drives competition

COMPUTING POWER: DOMESTIC INFRASTRUCTURE APPROACHING FRONTIER SCALE

Domestic AI compute infrastructure is rapidly approaching frontier-scale capability. Chinese AI labs are increasingly moving beyond inference optimisation toward end-to-end model training on domestic infrastructure. Meituan's 1.6T-parameter LongCat-2.0 completed full pre-training and inference on a 50,000-card domestic cluster, while Z.ai's GLM-Image completed its development pipeline on Huawei's Ascend Atlas 800T A2. Alibaba's M890 Super Node can serve more than 2T parameter models, while Z.ai has reportedly completed a 1GW domestic chip data centre and Tencent Cloud is expanding domestic AI chip deployment. DeepSeek and Z.ai are also exploring proprietary inference chips, highlighting a broader shift toward vertical integration. While compute constraints remain, improving domestic hardware, software and cluster-management capabilities should support larger models and gradually reduce reliance on restricted overseas GPUs.

Domestic hardware is lifting the ceiling for model development

FUTURE TREND – LARGER MODEL; OPEN WEIGHT; HARNESS

Larger but more efficient models are becoming the next scaling paradigm. Chinese AI labs are moving toward trillion-parameter models, with Kimi K3 at 2.8T parameters, Qwen3.8 Max at 2.4T and MiniMax's reported model at about 3T. However, scale alone does not guarantee proportional gains in intelligence or economics. Sparse MoE architectures allow developers to expand total capacity while activating only a fraction of parameters, shifting competition toward data quality, post-training/RL, inference-time compute and system optimisation. We therefore see active parameters, agentic performance, inference efficiency, costs and real-world task execution becoming more relevant measures of model competitiveness.

Model competitiveness is shifting toward larger yet more efficient architectures

Open-weight models broaden distribution. With DeepSeek's V4-Flash and MiniMax's H3 adopting open-weight strategies, model intelligence is increasingly an entry ticket rather than the sole moat. Long-term winners will need to combine frontier capabilities with applications, proprietary data, user feedback, monetisation and self-improving training loops. At the same time, rising R&D, compute and talent requirements should drive eventual consolidation, favouring well-funded platforms and leading independent labs.

Open weights expand distribution and support monetisation

Open weights expand distribution but face geopolitical constraints. DeepSeek's models, Z.ai's GLM and MiniMax's M3 are increasingly distributed through Google Cloud Gemini Enterprise Agent Platform/Model Garden and Amazon AWS Bedrock, which host DeepSeek, MiniMax, Moonshot, GLM and Qwen. Microsoft is also reportedly considering DeepSeek on Copilot via Azure. This could materially expand overseas adoption beyond first-party annual recurring revenue (ARR). However, Western market access, data/compute localisation, high-end GPU access and restrictions on US technology relationships remain key risks.

Geopolitics and commoditisation remain the key constraints

Open-weight monetisation is becoming more commercial, while low-end APIs remain commoditised. Fully open models such as Z.ai's GLM are under an MIT licence to maximise adoption, while MiniMax's M-series Community License enables commercial agreements and revenue sharing. We expect more labs to monetise third-party deployment through partnerships with AWS Bedrock, Alibaba Cloud Bailian, OpenRouter and Cline, while low-end API pricing remains at around US\$0.10-0.20/1m tokens. Financial strength, pricing power and cost efficiency should therefore become key differentiators, benefitting value-oriented models such as Tencent's Hy3 and Xiaomi's MiMo.

Open-weight monetisation broadens as API pricing converges at the low end

Figure 8: China Models Lead Global Token Consumption Rankings

6 Jul 26			20 Jul 26			3 Aug 26		
Rank	Model	Token usage	Rank	Model	Token usage	Rank	Model	Token usage
1	Hy3	6.13T	1	MiMo-V2.5	10.50T	1	DeepSeek V4 Flash 0731	7.88T
2	MiMo-V2.5	5.95T	2	DeepSeek V4 Flash 0423	6.37T	2	Hy3	7.07T
3	DeepSeek V4 Flash 0423	5.22T	3	Hy3 (free)	3.94T	3	DeepSeek V4 Flash 0423	5.42T
4	MiniMax M3	4.26T	4	GLM-5.2	3.29T	4	MiMo-V2.5	4.96T
5	GLM-5.2	3.19T	5	DeepSeek V4 Pro	3.17T	5	GLM-5.2	3.19T
6	DeepSeek V4 Pro	2.83T	6	Nemotron 3Ultra(free)	2.34T	6	DeepSeek V4 Pro	2.38T
			7	MiniMax M3	2.05T			
			8	Step 3.1 Flash	1.92T			
			9	Hy3	1.76T			

Source: Open Router, UOB Kay Hian

AGENTIC HARNESS – THE NEXT STRATEGIC BATTLEGROUND

From token growth to ROI-driven AI. Enterprises are increasingly focusing on task completion and cost efficiency rather than token consumption, with heavy AI users consuming 10x more tokens but generating only 2x more output, while some US internet companies depleted a full year’s AI budget within four months of adopting agentic tools. We foresee a barbell model, with cheaper flash models for routine tasks and SOTA models for high-value workloads, making cost per completed task, ROI, DAA and AWU increasingly important metrics for model selection and monetisation.

Multi-capability models are becoming the new standard, laying the foundation for agentic AI adoption. AI developers are increasingly integrating reasoning, coding, multimodality, agentic execution, tool use and long-context processing into a single foundation model, shifting competition from task-specific capabilities toward more versatile, general-purpose systems. As enterprises move from chatbots to agents, this convergence of capabilities should enable models to handle increasingly complex, multi-step workflows within a unified system.

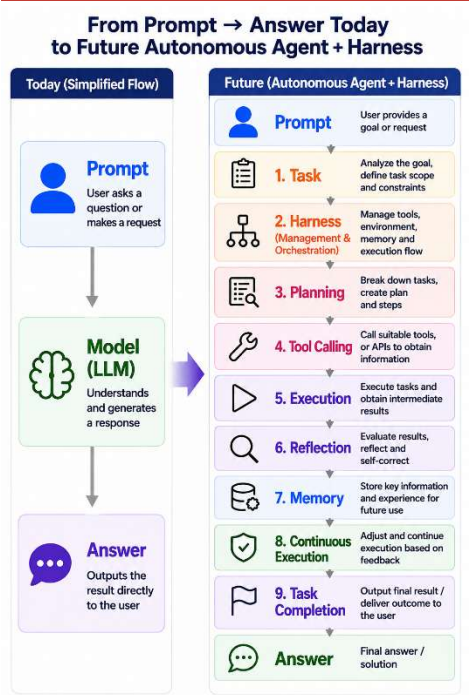
Agentic harness/platforms are becoming the next strategic battleground. OpenAI, Anthropic and Google have launched agent development frameworks covering planning, tool use, workflow orchestration and multi-agent coordination. We expect Chinese AI labs to increasingly use coding and workspace agents as key user entry points, generating proprietary workflow data that can strengthen RL and create a positive data flywheel. MiniMax and Z.ai are also increasing their focus on coding, long-context and long-horizon agent tasks, although the specific architecture and terminology may differ across companies.

Competitive differentiation is shifting from models to agent infrastructure. Mega-cap platforms increasingly support both proprietary and third-party models, monetising through first-party token usage, API resale and revenue sharing. Tencent’s WorkBuddy and Alibaba’s Qoder are evolving into integrated AI workspaces, with co-worker and industry-specific agents likely to be the next focus. Tencent’s WorkBuddy ranked #1 in China’s desktop agent monthly visits in June, highlighting the importance of ecosystem distribution, feature iteration and safety. We believe agent infrastructure and reliable workflow execution will increasingly differentiate providers as enterprises move beyond chatbots toward AI systems capable of completing complex, multi-step tasks.

ADDRESSABLE TAM

China’s AI ecosystem is evolving into a full-stack value chain, with foundation models at the core and value increasingly extending across applications, API and model-as-a-service platforms, and cloud infrastructure. We expect aggressive hyperscaler investment to continue supporting a sizeable AI revenue pool. We also forecast China’s AI capex exceeding US\$1t over 2025-29 with annual capex jumping from US\$130b in 2026 to US\$400b in 2029,

Figure 9: Agentic AI Evolves Beyond Simple Prompt–Response



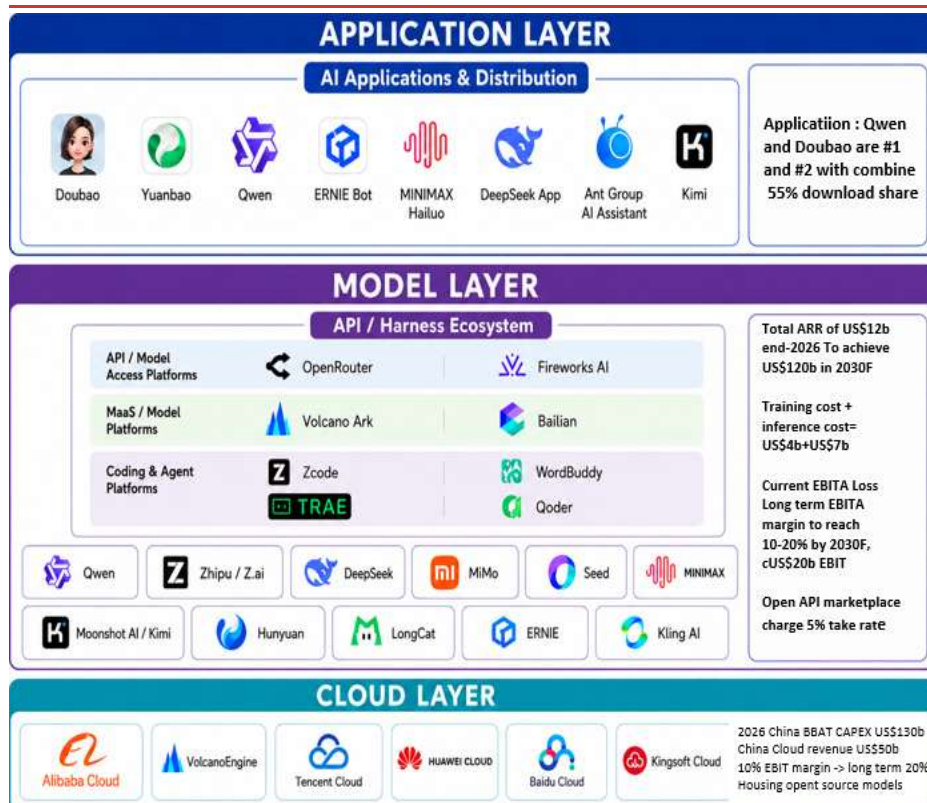
Source: Industry data, UOB Kay Hian

Competitive moats are increasingly forming around agent ecosystems

China’s AI ecosystem is scaling into a multi-layer monetisation opportunity

and estimate China's AI revenue TAM could reach US\$300b by 2030. Leading AI labs could address a revenue TAM of US\$100b-200b by 2030. Alibaba, Tencent, ByteDance and Baidu are projected to account for 50-60% of China's AI capex through 2030. Excluding these platforms, the remaining AI spend implies a substantial enterprise AI revenue opportunity.

Figure 10: Foundation Model Application Market



Source: Industry data, UOB Kay Hian

China's AI model market is in hyper-growth, with international adoption the key upside. We forecast China's AI model monetisation revenue rising from Rmb35b in 2026 to Rmb900b (US\$128b) in 2030, as daily token consumption grows from 350T to 4,500T (~13x). Improving intelligence and lower token costs should drive Chinese-model adoption in non-US markets, particularly among cost-sensitive SMEs and 24/7 autonomous agents, expanding the opportunity beyond China despite geopolitical and compute-access constraints. Frontier labs should capture a disproportionate share of TAM, supported by stronger pricing power and higher-value use cases. In our view, Z.ai is well positioned in coding and agentic workloads, while MiniMax offers differentiated exposure to AI video and multimodal AI. We see meaningful upside to the consensus estimate of US\$15b in 2030 revenue for Z.ai, while MiniMax will offer upside if its video leadership translates into sustained API and subscription monetisation.

Coding is emerging as a key enterprise API demand driver. Beyond general-purpose model APIs, AI coding is shifting from the "Copilot" phase to engineering agent. We estimate the 2030 TAM for coding AI as follows: Rmb150b-200b for China and US\$40b-55b for the global market. We see further upside with unlocking coding agents set to evolve from autocompleting toward a full intent → design → implementation → verification workflow.

Multimodal/AI video remains a high-growth vertical but with a smaller TAM. We estimate a long-term TAM of Rmb80b-120b for China and US\$20b-30b for the global market by 2030 for AI video, driven by content production, advertising, e-commerce, gaming and the creator economy. As video quality converges toward "good enough", competition should shift toward pricing, workflow integration and user experience, making AI video an important commercialisation engine but a smaller value pool than enterprise AI.

Figure 11: China AI Token Generation To Grow 13x By 2030

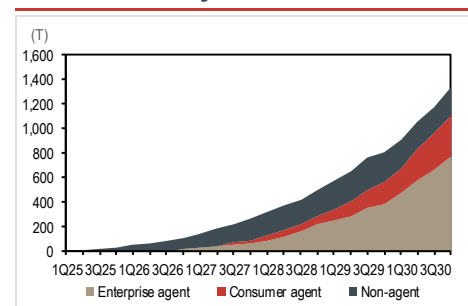


Figure 12: China AI API + Subscription Revenue To Reach About US\$128b By 2030

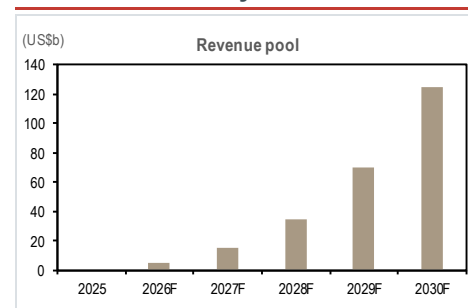


Figure 13: China vs US TAM

	China TAM	Global ex-China TAM
2030F Coding	Rmb150b-200b	US\$40b-55b
AI video	Rmb80b-120b	US\$20b-30b
Other LLM / agents	Rmb580b-670b	US\$120b-165b
Total model monetisation	Rmb900b/US\$128b	US\$180b-250b
Daily token consumption	4.5 quadrillion	4.0 quadrillion

Source: Industry data, UOB Kay Hian

Figure 14: China Coding Agent TAM Expansion

Stage	Product definition	Current China status	Indicative China TAM
Copilot	Code completion, code generation, debugging and Q&A	Maturing/ largely passed	~US\$1b-5b
Agentic coding	Autonomous coding, testing, debugging and tool use	Rapidly scaling/ current focus	US\$10b-30b+
Engineering agent	End-to-end requirements → architecture → coding → testing → deployment	Early stage/ emerging	US\$30b-100b+
Creator	Natural-language software creation by non-developers	Long-term optionality	Potentially much larger

Source: Industry data, UOB Kay Hian

AI video supports attractive monetisation, but with a smaller long-term TAM

Valuation

We prefer Z.ai over MiniMax in the medium term. Z.ai is better positioned in coding and agentic AI, which we see as the largest and most monetisable AI workloads, particularly as usage shifts from simple chatbot queries toward always-on agents with much higher token consumption. Its recent ARR acceleration also suggests improving model competitiveness and pricing power, which should support stronger monetisation. MiniMax remains differentiated in multimodal systems, particularly AI video, and has stronger overseas exposure, which provides meaningful TAM upside. However, we see greater competitive intensity and potentially weaker pricing power in general-purpose LLMs. For MiniMax, a recovery in valuation likely depends on a successful M3 Pro launch.

We believe both stocks' valuations increasingly reflect their potential share of a large medium-term AI revenue pool, while sustainable long-term margins depend on more rational competition and stronger pricing power as the industry consolidates. Both companies' core growth drivers are also converging on API monetisation and overseas expansion: MiniMax benefits from a broad global footprint across consumer apps, generative media and B2B APIs, supporting our forecasts of a 123% three-year revenue CAGR for 2026-29 and breakeven in 2029, while Z.ai combines a sticky on-premise base in regulated Chinese industries with a scalable cloud API business, supporting a 141% three-year revenue CAGR for 2026-29 and profitability in 2029. We expect rising overseas exposure and disciplined pricing to further drive API margin expansion.

Long-term winners will be defined by scale, margins and financial strength. Valuations are increasingly diverging on three dimensions – frontier intelligence, pricing power and capital efficiency, with Z.ai and MiniMax offering attractive exposure to the next phase of AI monetisation. US frontier labs trade at 12-34x P/ARR, while Z.ai and MiniMax span a wider 14-30x range, reflecting a sharp divergence between monetisation quality and strategic/ecosystem optionality. We believe long-term winners in China's AI model industry will be those boasting a mix of scale, gross margin advantage and financial strength.

Figure 15: Growth And Valuation Comparison Across Leading AI Models

Metric	Z.ai	MiniMax	DeepSeek	Anthropic	OpenAI
Market cap/valuation	cUS\$61b	cUS\$13.5b	cUS\$74b	cUS\$965b	cUS\$852b
ARR	cUS\$2.0b	cUS\$1b	cUS\$1b	cUS\$80b	cUS\$25b
P/ARR	c31x	c14x	c74x	c12x	c34x
Gross margin	28%	c18-20%	c44.6%	c63%	c39%
ARR growth	c15x in 6M	>100% recently	Very high	Extremely high	High
Business mix	B2B-heavy	To-C + To-B	API/open-weight	Enterprise-heavy	n.a.
Open-weight	Yes	Yes/restricted	Yes	No	No

Source: Bloomberg, UOB Kay Hian

Figure 16: Z.ai vs MiniMax – Growth And Valuation Comparison

	Z.ai (2513 HK)	MiniMax (100 HK)
Market cap (US\$b)	61	13.5
Current price (HK\$)	1030	303
Price target (HK\$)	1830	470
Upside/downside potential	77.7%	55.1%
Enterprise value (US\$b)	60.96	12.59
Revenue		
2026F (US\$m)	1,141	490
2027F (US\$m)	3,460	1,259
2028F (US\$m)	7,995	2,738
2026-28F CAGR	164.7%	136.4%
EPS		
2026F (US\$)	-1.49	-1.82
2027F (US\$)	-1.93	-1.64
2028F (US\$)	-1.04	-1.13
EV/Sales		
2026F	53.4	25.7
2027F	17.6	10.0
2028F	7.6	4.6

Source: Bloomberg, industry data, UOB Kay Hian

Sustainable AI leadership requires both operating scale and financial discipline

Appendix

KEY RISKS IN AI INDUSTRY

Intensifying competition and pricing pressure. Competition in the AI industry continues to intensify as both specialised AI model labs and large internet platform companies aggressively expand their AI capabilities. Rapid model iteration and performance improvements across reasoning, multimodality, and agentic functionality may shorten product replacement cycles and increase the risk of technological obsolescence. At the same time, large internet platforms with established ecosystems, stronger financial resources, and proprietary cloud infrastructure may possess structural cost advantages that enable them to offer lower-priced AI services amid intensifying industry-wide pricing pressure. This may negatively affect monetisation capability, margins, and market share for smaller or independent AI companies.

Talent scarcity. The AI industry remains highly dependent on top-tier research and engineering talent, particularly in areas such as foundation model architecture, multimodal systems, RL, and agentic AI development. Intensifying competition among global technology companies and AI labs for experienced AI researchers and engineers may continue to drive up compensation costs and employee turnover risks. Failure to attract, retain, and develop key talent could slow model iteration, weaken innovation capability, and negatively impact long-term competitiveness.

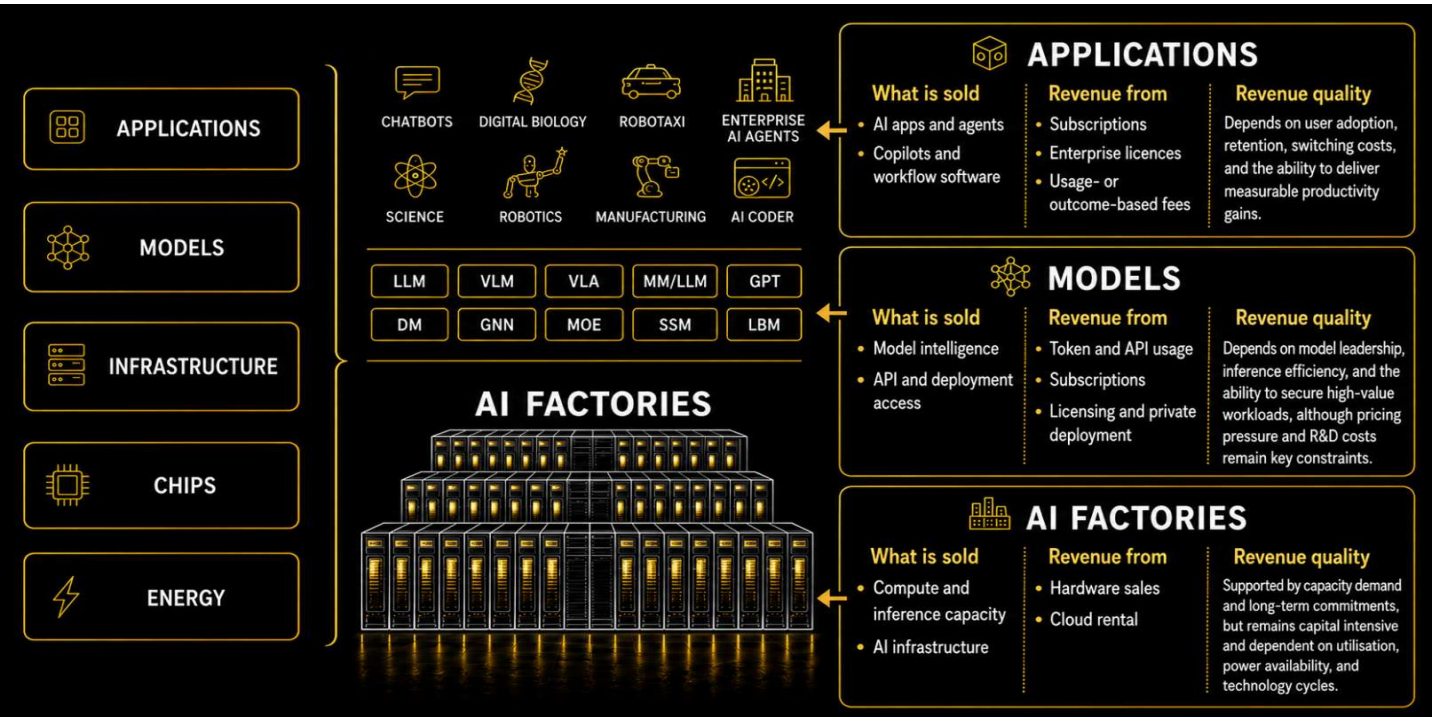
Computation power and chip supply constraints. The development and deployment of advanced AI models require significant computation resources, particularly high-performance GPUs and AI accelerators. Ongoing supply constraints, geopolitical restrictions, export controls, and rising demand for advanced semiconductor chips may lead to computation power shortages and higher infrastructure costs. Limited access to advanced chips could delay model training, reduce inference efficiency, and constrain the scalability of AI services, potentially impacting the pace of product development and commercialisation.

Regulatory and AI safety issues. The AI industry remains subject to evolving regulatory oversight across data privacy, copyright, content generation, model safety, and AI governance. Changes in regulatory requirements or stricter compliance standards may increase operational costs, limit certain AI functionalities, or delay product deployment. Increasing scrutiny of hallucination, misinformation, and autonomous AI behaviour may also create reputational and legal risks for AI companies.

AI ECOSYSTEM: CHINA BUILDS SCALE ACROSS THE STACK, WHILE CHIPS REMAIN THE KEY CONSTRAINT

The five-layer AI stack should be assessed as an integrated value chain, with energy, chips and infrastructure determining the availability and cost of compute; models converting compute into intelligence; and applications translating intelligence into commercial value. In our view, China benefits from its large energy base, rapid infrastructure buildout and improving domestic compute capabilities, while the gap with the US at the model and application layers continues to narrow as Chinese developers improve capability, deployment and price-performance. However, the US retains a clear lead in advanced semiconductors and high-end computing capacity, which remains the principal constraint on the development of China's broader AI ecosystem.

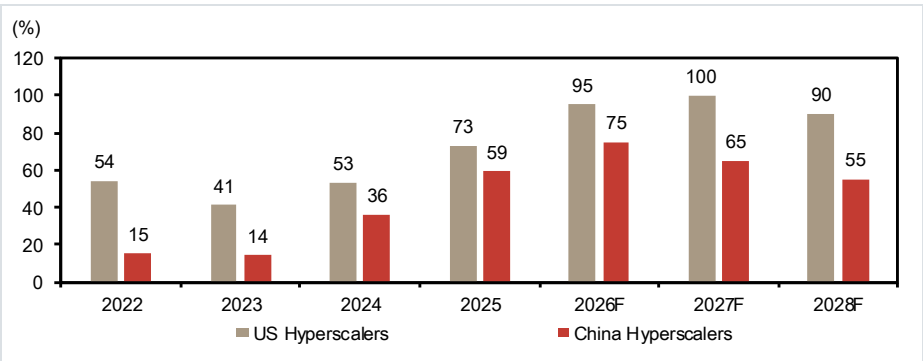
AI Ecosystem



Source: Nvidia, industry data, UOB Kay Hian

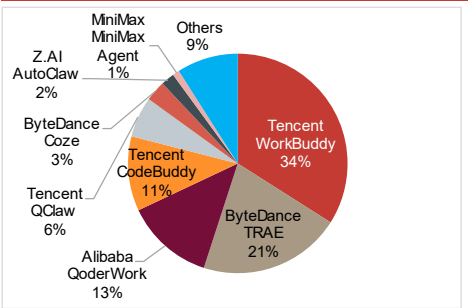
US hyperscalers are entering a substantially more capital-intensive AI investment cycle than their Chinese peers. Aggregate capex among major US technology companies is projected to rise from US\$443b in 2025 to US\$806b in 2026 and US\$1.23t in 2027, significantly outpacing the corresponding increase among Chinese internet platforms from US\$57b to US\$102b and US\$122b respectively. The divergence is also evident in capital intensity, with US capex-to-operating cash flow projected to exceed 1.0x from 2026, vs 0.81x for Chinese peers in 2026 before easing to 0.59x by 2028. We believe this highlights a more measured capex approach among Chinese AI players, with a greater emphasis on compute efficiency and capital discipline despite continued AI infrastructure expansion.

Capex To Operating Cash Flow



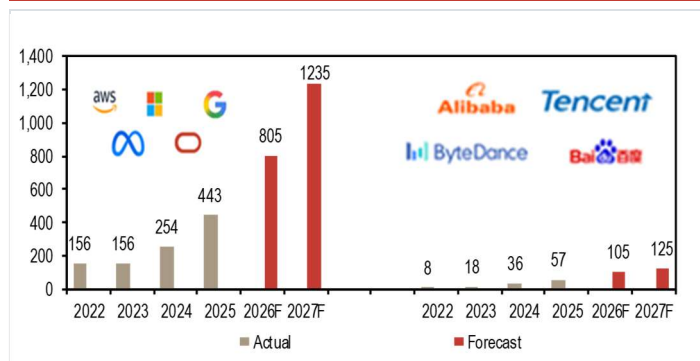
Source: Industry data, UOB Kay Hian

China Capex By Cloud Service Provider



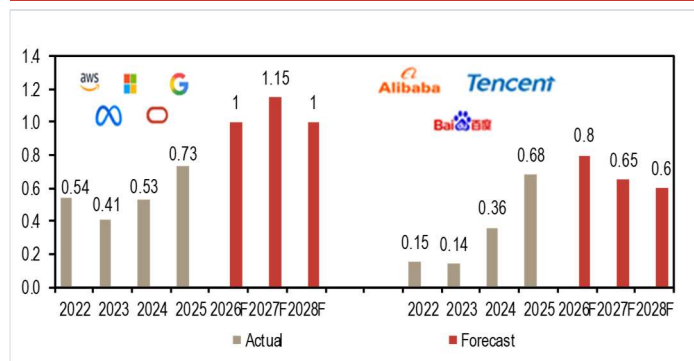
Source: Industry data, UOB Kay Hian

Historical And Projected Capex (US\$b)



Source: Industry data, UOB Kay Hian

Capex To Operating Cash Flow Ratio



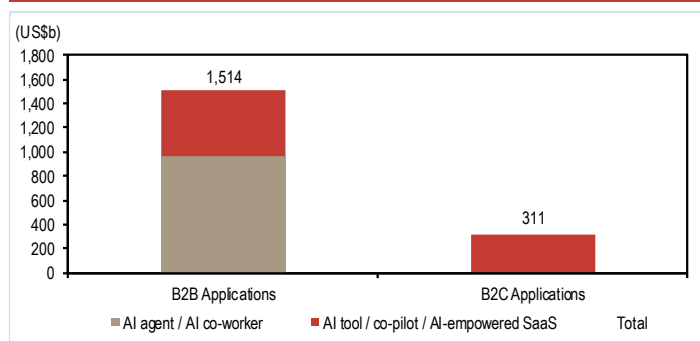
Source: Industry data, UOB Kay Hian

MARKET SIZE

Accelerating B2B and B2C adoption to drive global AI TAM to US\$1.8t by 2030. We forecast the global AI application market reaching approximately US\$1.8t by 2030, supported by the rapid expansion of both enterprise and consumer AI applications. The B2B segment is expected to contribute US\$1.5t, representing around 83% of TAM, as businesses increasingly integrate AI into daily operations, decision-making processes and employee workflows. By region, the US is projected to account for the largest B2B opportunity with US\$771b, followed by the rest of the world with US\$443b and China with US\$299b. Growth will be driven by the adoption of AI agents and AI co-workers that automate or augment labour-intensive tasks, alongside AI-powered SaaS solutions embedded across software engineering, professional services, customer support, sales, manufacturing and other occupational functions.

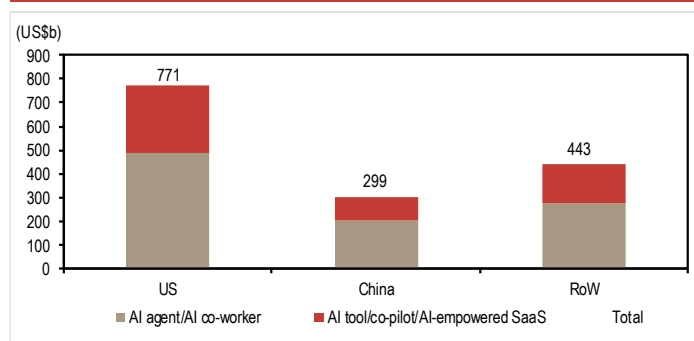
The B2C AI application market is forecast to reach US\$311b by 2030, supported by a growing global user base, improving willingness to pay and broader adoption of paid AI services. Key consumer use cases include general AI assistants and productivity tools, AI companions, creative content generation, education and learning, health and wellness, entertainment and gaming, as well as personal finance and shopping assistants. Overall, we expect B2B applications to remain the main driver of AI monetisation due to their larger addressable spending pool and clearer productivity benefits, while B2C applications should provide an additional growth pillar as AI becomes increasingly embedded in consumers' everyday activities.

Global AI Application TAM By Segment (2030)



Source: Industry data, UOB Kay Hian

Global AI Application TAM By Country (2030)



Source: Industry data, UOB Kay Hian

MODEL EVOLUTION

Acceleration of progress. The AI stack broadly comprises: a) a model layer covering text, visual, audio and multimodal foundation models; and b) an application layer that embeds these models into consumer and enterprise use cases such as productivity, entertainment, coding and content generation. Rapid advances in multimodal and agentic AI are further expanding real-world applications across industries. The intelligence level of foundation models worldwide continues to rise. According to Open AI’s five-level roadmap, current models have reached Level 3. Looking ahead, the trajectory points clearly toward accelerated progress.

Open AI’s Five Levels Of Intelligence

Levels	Name	Description
L1	Chatbots	AI with conversational language
L2	Reasoners	AI with human-level problem solving capabilities
L3	Agents	AI that can take action
L4	Innovators	AI that can aid in invention
L5	Organisations	AI that can do the work of an organisation

Source: Industry data, UOB Kay Hian

MODEL COMPETITION AND MONETISATION

Enterprise LLM API adoption is shifting toward a more diversified, workload-driven model. OpenAI’s market share fell from 50% in 2023 to 27% in 2025, while Anthropic’s increased from 12% to 40% to become the leading provider; Google’s also jumped from 7% to 21%, whereas Meta’s halved from 16% to 8%. We believe this multiple-model adoption suggests enterprise API is unlikely to become a winner-takes-all market, with customers increasingly allocating workloads based on performance, reliability, cost and integration. This should favour models with differentiated strengths in coding, reasoning and long-context tasks, while keeping pressure on pricing and customer retention.

Global LLM Market Share Trends



Source: Menlo Venture, UOB Kay Hian

AI MODELS ARE CONVERGING ON DIFFERENTIATED OPENNESS AND MONETISATION STRATEGIES

The AI model ecosystem is evolving from a simple closed-vs-open debate toward a spectrum of openness and monetisation models. Closed models retain the strongest control over weights, data and pricing, supporting API, subscription and enterprise licensing, while open-weight models broaden distribution and enable self-hosting, fine-tuning and commercial licensing without necessarily releasing training code or data. Fully open-source and open-science models maximise transparency and reproducibility, but typically face greater challenges in capturing direct model-level value. We therefore expect open-weight models to increasingly balance ecosystem adoption with commercial value capture through API, cloud partnerships, enterprise services and licensing.

Comparison Of LLM Model Openness And Monetisation

Model type	Weights released?	Training code released?	Data released?	Can self-host?	Can modify/retrain?	Typical monetisation	Examples
Closed model (API-only)	✗	✗	✗	✗	✗	API usage, subscriptions, enterprise licensing	OpenAI GPT-5, Anthropic Claude, Google Gemini
Open-weight model	✓	Usually ✗	Usually ✗	✓	Usually yes (fine-tuning permitted under license)	API + enterprise support + cloud + commercial licensing	DeepSeek V4, Z.ai GLM-5.2, Moonshot Kimi K3, Meta Llama 3/4, Alibaba Qwen
Open-source model	✓	☑	Often partially or fully	✓	✓	Services, enterprise support, hosting, consulting	OLMo, Pythia, BLOOM, Falcon (depending on release version)
Open-science/fully reproducible model (rare)	✓	✓	✓	✓	✓	Mainly research rather than commercialisation	Allen AI OLMo, RedPajama (partially)

Source: Industry data, UOB Kay Hian

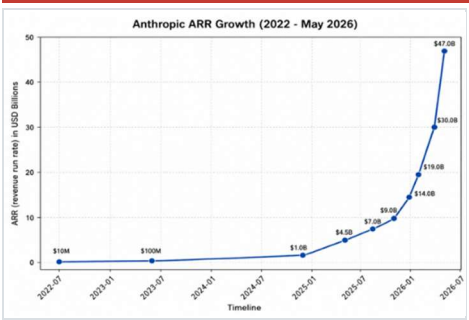
ENTERPRISE USE CASES

Enterprise use cases are becoming the main monetisation driver. AI agents, coding assistants and video generation are among the most commercially attractive categories. Enterprise adoption is increasingly driven by workflow automation, software development and other use cases with measurable productivity benefits. Rising enterprise adoption, with AI-generated code reaching up to 90% in 1H26 at some Chinese mega-caps vs 20-30% in 2H25, should further strengthen the data flywheel.

Coding-led demand is becoming the core driver of enterprise LLM API adoption. Anthropic’s 54% share of the coding market in 2025 underscores its strong positioning in developer productivity and software engineering workflows, including code generation, debugging, testing, documentation and broader enterprise automation use cases. This suggests enterprise adoption is increasingly being driven by practical, workflow-embedded requirements rather than general-purpose chatbot usage alone. As coding workloads are typically high-frequency, token-intensive and closely integrated into developers’ daily operations, they can generate more recurring API consumption and provide stronger monetisation visibility. In this context, model providers with stronger coding performance, enterprise integration capability and developer ecosystem adoption are better positioned to capture sustained usage share growth in the enterprise LLM API market.

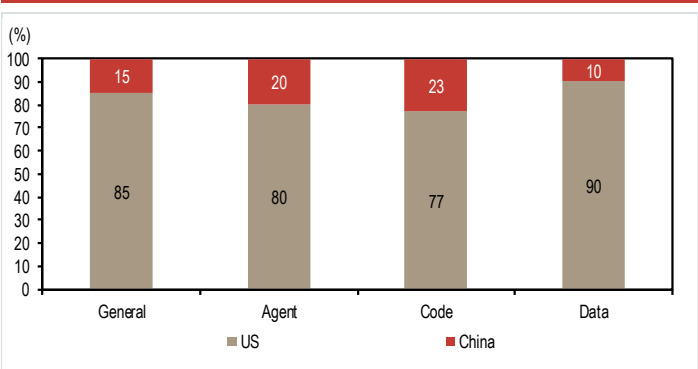
Chinese models are already capturing a disproportionate share of usage despite a smaller proportion of revenue, reflecting a significant pricing gap vs US models. China accounts for 88-92% of agent and coding token usage, yet captures only 20-23% of revenue, as substantially lower API pricing limits monetisation per token. We believe this usage-revenue gap provides significant upside as Chinese models continue to improve frontier intelligence, broaden enterprise and overseas adoption, and gradually monetise higher-value agentic and coding workloads.

Anthropic ARR Growth



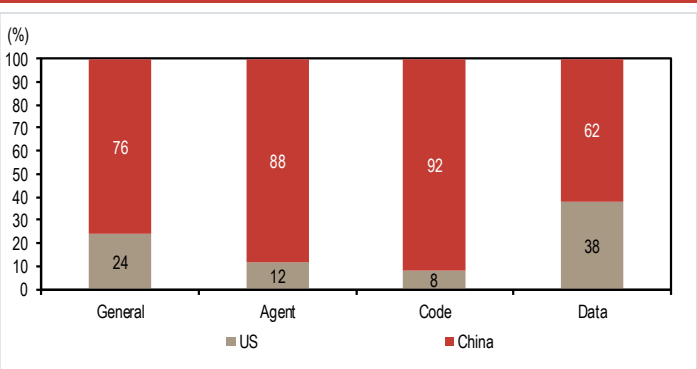
Source: Anthropic, Industry data, UOB Kay Hian

Share Of AI Token Spending



Source: Industry data, UOB Kay Hian

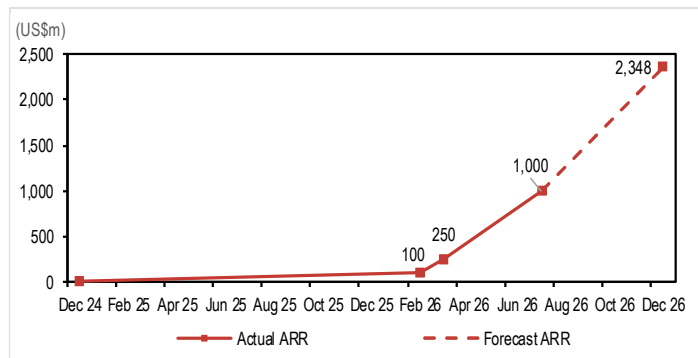
Share Of Token Usage



Source: Industry data, UOB Kay Hian

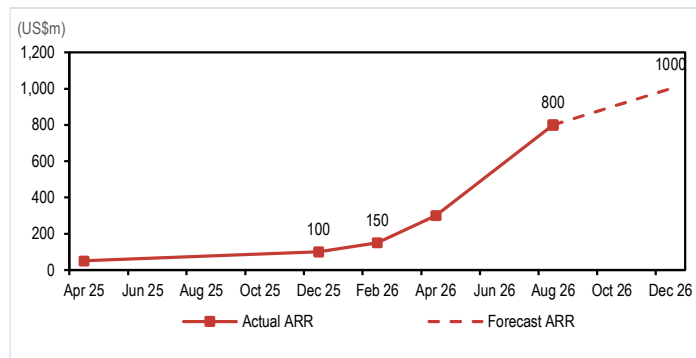
ARR and gross profit growth are important investor metrics. We assess Chinese AI model companies across key dimensions, including: a) ARR scale (token scale x pricing power), driven by token volume and pricing power; b) gross margin advantage, supported by training/inference efficiency and technology; and c) financial strength, including balance sheet capacity and access to compute. Based on this framework, we view Z.ai as the best-placed foundation model player, while Minimax stands out in multimodal AI.

Z.ai ARR Projection



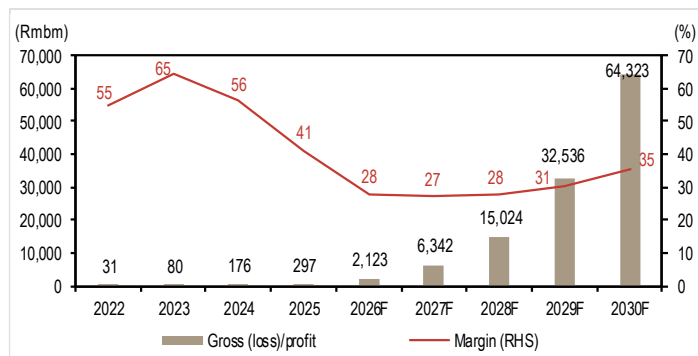
Source: Z.ai, UOB Kay Hian

MiniMax ARR Projection



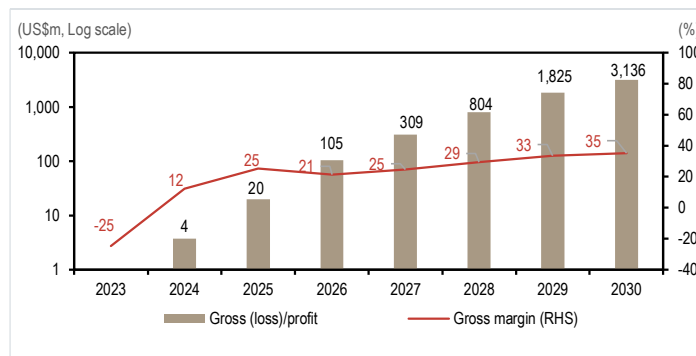
Source: Minimax, UOB Kay Hian

Z.ai Gross Margin Profile



Source: Z.ai, UOB Kay Hian

MiniMax Gross Margin Profile

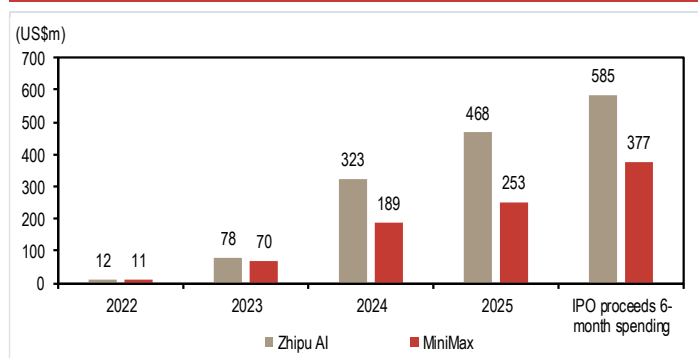


Source: Minimax, UOB Kay Hian

Sustained investment and profitability risk. AI labs need to continue investing substantially in R&D, computing resources, sales and marketing, and talent acquisition to support model iteration and commercial expansion. Given the early-stage nature of the AI industry and the high cost of developing frontier models, these companies may continue to face losses and cash flow pressure before reaching sustainable profitability.

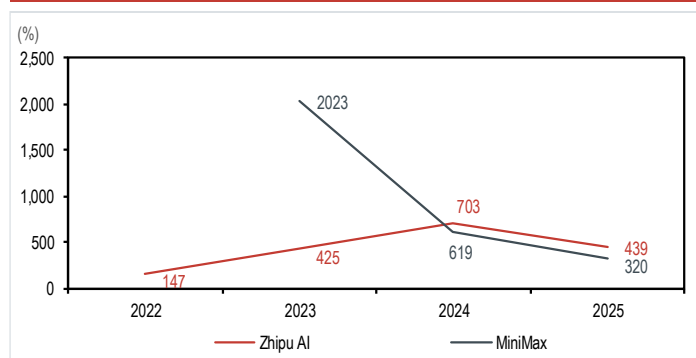
Heavy reinvestment could extend the path of sustainable profitability

R&D Spending/6-month Spending From IPO Proceeds (2026)



Source: Z.ai, industry data, UOB Kay Hian

R&D Spending As % Of Revenue



Source: Minimax, industry data, UOB Kay Hian

Stock Picks

Z.AI Co (2513 HK)

At The Frontier Of China's AI Intelligence Inflection Point

Highlights

- Z.ai is best positioned among domestic peers at the current intelligence inflection point to benefit from explosive demand and global adoption**, in our view, given its pricing power, cost efficiency and financial strength. Rapid iteration of the GLM-5 series has accelerated API adoption, with the latest GLM-5.3 further strengthening Z.ai's positioning in coding, long-context reasoning and agentic workloads, from a relatively compact 744B-parameter architecture. Initiate coverage with BUY. Target price: HK\$1,830.00.

Analysis

- Sustained frontier execution strengthens competitive positioning.** Z.AI Co (Z.ai) has demonstrated strong, sustained frontier model execution, releasing multiple domestic state-of-the-art general language models (GLM) over the past six months. Following the launch of GLM-5.3 on 14 August, we expect a significantly larger next-generation model to be released in the near future. Meanwhile, broad deployment across leading Chinese internet players and emerging technology companies continues to strengthen both commercialisation and proprietary real-world data flywheel.
- Proprietary architecture and frontier coding/agentic capabilities underpin pricing power.** Z.ai's proprietary GLM architecture and GLM-5.3's near-frontier coding and agentic capabilities demonstrate the company's technology differentiation and users' willingness to pay for premium intelligence. GLM-5.3 ranked just behind Anthropic's models on key coding benchmarks at launch, underscoring its frontier capability. Revenue per token has risen 83% ytd while paid token usage has grown 400% yoy, highlighting rare pricing power amid stiff application programming interface (API) price competition. Although GLM-5.2's relative ranking has eased following newer launches, Z.ai's strong iteration track record should support its return to top-tier frontier performance with its next-generation model.
- Clear commercial trajectory and accelerating cloud growth.** We expect the cloud business to remain the primary growth driver, targeting profitability from 2029 onwards. Given the reported 1GW data centre build-out, we believe API annual recurring revenue could approach US\$2.3b by end-26, further reaching US\$6.5b in 2027 as domestic inference capacity expands. We also foresee the next-generation model adopting more restrictive commercial licensing to improve monetisation of overseas demand.
- We initiate coverage with a BUY rating and a target price of HK\$1,830.00**, reflecting strong revenue forecasts driven by the GLM-5 series with its improved capability and higher blended pricing.

Key Financials

Year to 31 Dec (Rmbm)	2024	2025	2026F	2027F	2028F
Net turnover	312	724	7,666	23,252	53,728
EBITDA	-2,261	-3,499	-2,944	-2,775	2,530
Operating profit	-2,541	-3,780	-5,085	-7,312	-4,260
Net profit (rep./act.)	-2,956	-4,698	-5,066	-6,991	-4,194
Net profit (adj.)	-2,466	-3,182	-4,417	-6,217	-3,247
EPS (fen)	(1,542.2)	(1,994.7)	(969.2)	(1,322.0)	(683.6)
PE (x)	n.a.	n.a.	n.a.	n.a.	n.a.
P/B (x)	n.a.	n.a.	21.8	33.9	46.4
EV/EBITDA (x)	n.a.	n.a.	n.a.	n.a.	153.7
Dividend yield (%)	0.0	0.0	0.0	0.0	0.0
Net margin (%)	(946.3)	(648.6)	(66.1)	(30.1)	(7.8)
Net debt/(cash) to equity (%)	42.3	17.1	(121.3)	(90.6)	(43.6)
ROE (%)	n.a.	n.a.	n.a.	(45.6)	(39.4)
Consensus net profit (Rmbm)	-	-	(4,956.3)	(4,737.8)	(2,237.8)
UOBKH/Consensus (x)	-	-	1.11	0.69	0.55

Source: Z.ai, Bloomberg, UOB Kay Hian

Analyst(s)

Julia Pan

juliapan@uobkh.com

+86 21 5404 7225 ext 808

BUY

Share Price	HK\$1,030.00
Target Price	HK\$1,830.00
Upside	77.7%

Stock Data

Sector	Information Technology
Bloomberg ticker	2513 HK
Shares issued (m)	241.1
Market cap (HK\$m)	479,591.8
Market cap (US\$m)	61,177.9
3-mth avg daily t'over (US\$m)	903.6

Price Performance (%)

52-week high/low		HK\$2,980.00/HK\$116.10		
1mth	3mth	6mth	1yr	YTD
(16.7)	(23.4)	83.9	n.a.	n.a.

Major Shareholders (%)

Knowledge Atlas ESOP	14.96
Beijing Lianpai Technology Development	7.63
Jie Tang	6.02

Balance Sheet Metrics

FY26 NAV/Share (HK\$)	40.45
FY26 Net Debt/Share (HK\$)	49.06

Price Chart



Source: Bloomberg

Company Description

Z.AI Co offers technology services. The company provides cognitive intelligence general models, artificial intelligence systems, and other related solutions.

Investment Highlights

Sustained frontier model leadership supports long-term competitive positioning. Z.ai has established one of the strongest records of sustained frontier model execution among Chinese AI labs. In the past six months, it has released multiple domestic state-of-the-art (SOTA) models within its GLM-5.3. This has pushed Z.ai firmly into the frontier tier for coding, outperforming Anthropic's Opus 4.8 on several key benchmarks and approaching the latest closed-source leaders, while establishing open-weight SOTA across multiple coding and agentic evaluations. We believe repeated frontier delivery is a more important competitive indicator than one-off benchmark leadership or short-term usage spikes.

Recurring frontier execution

Proprietary architecture and leadership in coding/agentic AI underpin Z.ai's technology moat and pricing power. Z.ai is one of the few Chinese AI companies developing its proprietary GLM architecture from scratch, with GLM-5.3 approaching leading closed-source frontier models in coding and agentic AI. GLM-5.2 ranked just behind Anthropic's models on key coding benchmarks at launch, highlighting its frontier capabilities and pricing power. While newer launches have since narrowed its relative lead, Z.ai's strong iteration track record should support its return to top-tier performance with its next model. Its strengths in coding, long-horizon reasoning and agentic workflows allow for more higher-value enterprise use cases where task completion and reliability matter more than token cost. This differentiation is reflected in the 83% API price increases implemented ytd, while paid token usage has still grown 400%, demonstrating rare pricing power and supporting higher-quality monetisation.

Coding leadership supports pricing power

API commercialisation is accelerating, while on-premise deployments provide steady enterprise demand. Although on-premise deployments still contributed 73.7% of revenue in 2025, API services are becoming the primary growth engine. Cloud revenue increased 292.6% yoy in 2025, while open platform/API annual recurring revenue (ARR) reportedly surged from US\$250m in late March to US\$1b in July, underscoring a rapid commercialisation momentum. Meanwhile, Z.ai's large installed base across regulated industries provides recurring upgrade opportunities and supports long-term enterprise adoption.

API monetisation is at an inflection point

Strong financial outlook supported by accelerating cloud monetisation. We expect APIs to become the company's primary earnings driver as cloud adoption scales. We forecast a 141% three-year revenue CAGR over 2026-29, with revenue increasing from Rmb9.3b in 2026 to approximately Rmb108.8b by 2029 and profitability being achieved from 2029 onwards. With market reports indicating cloud ARR of US\$1b in Jul 26, we project total ARR surging from US\$2.3b in 2026 to over US\$6.5b by end-27.

Cloud drives the path to profitability

Attractive risk-reward supported by frontier capability and commercial execution. We view Z.ai as a capability-led AI platform combining sustained SOTA execution, differentiated pricing power, expanding enterprise adoption and strong commercialisation momentum. While regulatory risks remain, we believe the company's leadership in coding and agentic AI positions it well to capture long-term enterprise AI demand.

Compelling capability-led AI exposure

Catalysts: a) Resilient demand following the company's effective price increases, despite competitive responses from Kimi K3 and DeepSeek V4 pro; and b) launch of larger parameter models in the new future, which could further strengthen Z.ai's frontier leadership and pricing power.

Valuation

Initiate coverage with BUY and a target price of HK\$1,830.00, based on a DCF valuation methodology, discounted back using WACC of 12.3% and terminal EV/Sales of 5x. Our target price of HK\$1,830.00 implies 11.6x and 5.0x 2027F and 2028F EV/Sales respectively, representing a premium to pure-play peer MiniMax Group Inc and diversified technology companies such as Alibaba, Tencent and other peers. We believe this premium is justified by Z.ai's strong model capabilities, which support superior pricing power. We forecast a three-year revenue CAGR of 141% for 2026-29, and expect the company to turn profitable in 2029.

Figure 1: DCF FCFF Projection

Year to 31 Dec (Rmbm)	2026F	2027F	2028F	2029F	2030F	2031F
Free cash flow	(10,623)	(12,059)	(8,122)	3,256	21,978	45,084
EBIT * (1-t)	(5,051)	(6,985)	(4,176)	5,180	24,535	44,569
+ D&A	2,141	4,537	6,791	9,473	12,421	13,607
+ Capex & investments	(10,014)	(12,015)	(11,777)	(13,543)	(14,762)	(15,943)
+ (Inc) / dec in NWC	2,302	2,404	1,040	2,145	(216)	2,851
Proportionate FCF	(10,623)	(12,059)	(8,122)	3,256	21,978	45,084
Year	0.36	1.36	2.36	3.36	4.36	5.36
WACC	12%	12%	12%	12%	12%	12%
Discount factor	1.04	1.17	1.32	1.48	1.66	1.86
NPV of FCF	(10,183)	(10,292)	(6,171)	2,202	13,237	24,175
Base case WACC calculation:			DCF valuation (Rmbm):			
Risk free rate	1.7%			Total NPV of forecast		23,151
Equity risk premium	11.2%			Terminal Sales		262,070
Equity beta	1.0			Terminal EV/Sales multiple		5 x
Cost of common equity	12.9%			Terminal value		1,311,661
After tax cost of debt	3.7%			NPV of terminal value		703,344
Cost of Debt	4.4%			HKD\$/RMB		1.11
Tax rate	15.0%					
Debt/Total capital (steady state):				Enterprise value		726,495
% common equity	94%			TV: EV		96.8%
% of debt	6%			Net (cash) debt (2026F)		(22,107)
WACC calculated	12.3%			Hypothetical DCF-based fair value:		
Terminal FCFF =	45,084			Equity value (HK\$m)		834,157
Implied perpetual growth rate	8.9%			Per share (HK\$)		1,830

Source: Bloomberg, UOB Kay Hian

Figure 2: Peer Comparison

Company	Ticker	Price @ 26 Aug 2026 (lcy)	Market Cap (lcy m)	-----Calendarised PE-----			--Calendarised EV/EBITDA--			--Calendarised EV/Sales--			----- ROE -----		
				2026F	2027F	2028F	2026F	2027F	2028F	2026F	2027F	2028F	2026F	2027F	2028F
				(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)	(%)	(%)	(%)
China															
Z AI Co Ltd	2513 HK	1030.00	479,592	n.a	n.a	n.a	n.a	n.a	n.a	94.5	31.9	14.1	7.4	25.3	-8.9
MiniMax Group Inc	100 HK	303.00	105,818	n.a	n.a	n.a	n.a	n.a	n.a	34.6	12.6	5.8	-31.6	-5.1	888.2
Meitu Inc	1357 HK	4.30	19,653	13.8	10.7	8.5	9.2	6.7	4.7	2.8	2.3	1.9	18.0	19.6	21.3
Shenzhen Xunce Technology	3317 HK	122.90	40,554	253.9	64.9	39.5	246.4	67.3	32.1	14.6	9.2	6.3	3.5	14.8	23.8
Alibaba Group Holding Ltd	9988 HK	116.60	2,235,800	18.1	12.8	10.1	11.9	9.1	7.3	1.3	1.2	1.0	8.2	11.1	13.2
Tencent Holdings Ltd	700 HK	445.40	4,054,540	13.0	12.3	11.3	11.0	10.2	9.2	4.3	4.0	3.6	18.7	17.2	16.5
Baidu Inc	9888 HK	90.90	246,656	15.4	12.3	10.1	11.3	9.9	8.1	0.8	0.8	0.7	4.4	5.1	5.8
Kuaishou Technology	1024 HK	33.60	145,472	11.2	11.7	9.8	5.7	5.8	5.1	0.7	0.7	0.6	9.3	9.0	9.5
Avg				54.2	20.8	14.9	49.3	18.2	11.1	19.2	7.8	4.3	4.8	12.1	121.2
US															
Alphabet Inc	GOOG US	343.30	4,221,800	18.1	20.9	17.6	18.4	14.1	10.9	9.5	7.5	6.1	57.7	40.0	29.3
Meta Platforms Inc	META US	570.10	1,452,210	16.2	14.9	12.3	12.0	9.7	7.8	5.8	4.8	4.1	30.1	24.1	21.1
Microsoft Corp	MSFT US	491.70	3,651,210	24.9	20.9	17.4	15.2	12.4	10.0	9.5	8.0	6.6	28.1	26.4	25.3
Apple Inc	AAPL US	309.90	4,522,740	35.2	32.4	29.1	26.1	24.4	22.2	9.4	8.6	8.0	135.8	114.0	88.7
Avg				23.6	22.3	19.1	17.9	15.2	12.7	8.5	7.2	6.2	62.9	51.1	41.1

Source: Bloomberg, UOB Kay Hian

Business Outlook

We believe Z.ai is best positioned among domestic peers at the current intelligence inflection point to benefit from explosive demand and global adoption, given its pricing power, cost efficiency and financial strength. Rapid iteration of the GLM-5 series, combined with a strategic focus on agentic AI, tool-augmented reasoning and developer infrastructure, has driven a meaningful acceleration in API adoption. GLM-5.3 has further closed the gap with leading closed-source models, outperforming Anthropic’s Opus 4.8 on several coding and agentic benchmarks while remaining competitive with the latest frontier models. This reinforces Z.ai’s position as a leading open-weight model developer, supporting growing adoption among enterprises, developers and global SMEs.

Best positioned at intelligence inflection point

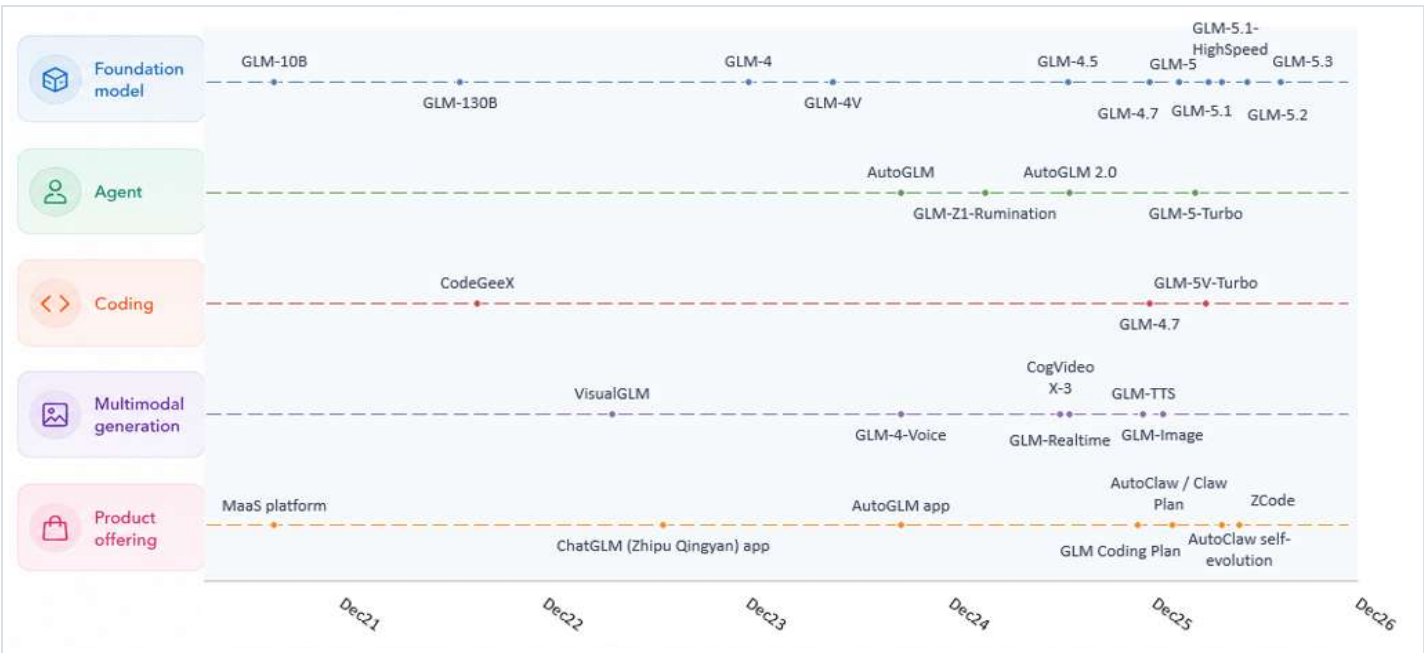
Backed by deep research talent and strong Tsinghua University ties, Z.ai increasingly differentiates itself through post-training, reinforcement learning (RL) and engineering optimisation rather than model scale alone. Following the launch of GLM-5.3 on 14 August, we expect a significantly larger next-generation model to be released in the near future. Meanwhile, broad deployment across leading Chinese internet players and emerging technology companies continues to strengthen both commercialisation and proprietary real-world data flywheel.

Research-led innovation drives differentiation

Z.ai’s long-term AGI vision has shaped GLM into an enterprise-grade foundation model. Since inception, the company has pursued a GLM-centric strategy inspired by GPT-3’s scaling paradigm, while prioritising a controllable and continuously evolving foundation-model architecture. Rather than optimising for individual capabilities, Z.ai has focused on advancing artificial general intelligence (AGI) across reasoning, coding and agentic workflows without compromising the existing performance. In our view, this research philosophy has resulted in a more robust, controllable and production-ready model architecture with lower hallucination rates, well suited to enterprise and government deployment.

AGI-led R&D strengthens enterprise readiness

Figure 3: Model Roadmap



Source: Z.ai, UOB Kay Hian

LEADING POSITION AMONG CHINESE AI MODEL DEVELOPERS

Leading position in coding and agentic AI underpins enterprise differentiation. Z.ai differentiates itself from local peers through stronger agent engineering, tool calling and end-environment interaction, supporting higher-value enterprise workflows and improving customer stickiness. Compared with overseas leaders such as Claude, Z.ai benefits from structural advantages in data sovereignty, on-premise deployment, regulatory compliance and compute cost efficiency, making it well positioned to serve China's government and enterprise markets.

We view coding as the largest and most monetisable large language model (LLM) vertical, with Z.ai following a strategy similar to Anthropic's by using coding as the bridge between frontier model capability and enterprise monetisation. The company serves over 6m developers and users, with nine of China's top 10 internet companies among its customers. This broad deployment creates a proprietary real-world coding data flywheel, where success and failure signals are used to continuously improve model performance through RL, driving stronger product iteration, higher enterprise adoption and a durable competitive moat.

RAPID MODEL ITERATION REINFORCES TECHNOLOGY LEADERSHIP AND ENTERPRISE ADOPTION

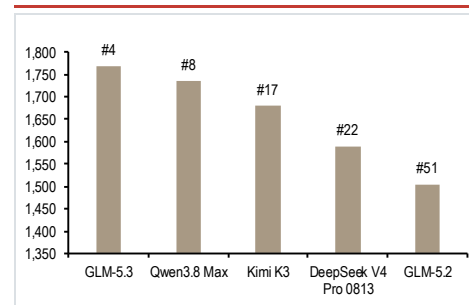
Z.ai maintains a rapid iteration cadence, with minor model upgrades every 1-2 months and major upgrades every 7-8 months (spanning GLM-4.5 to GLM-5.3). The latest GLM-5.3 marks another key milestone, delivering stronger coding and agentic capabilities alongside emerging cybersecurity strengths, having identified 2,436 vulnerabilities across 269 real-world projects, including 1,097 medium-to-high-risk vulnerabilities. We believe the upcoming larger-parameter model release could further help narrow the performance gap and maintain a higher position across key frontier benchmarks. This technical edge has attracted major enterprise adopters such as Alibaba Qoder and Tencent CodeBuddy, validating Z.ai's real-world coding capabilities and further reinforcing its data flywheel.

Beyond model strength, Z.ai's diversified monetisation model – spanning APIs, coding subscriptions, enterprise workflows, and on-premise deployments combined with ZCode 3.0 and proprietary agent frameworks – positions it for expanding enterprise AI adoption across multiple channels. Looking ahead, continued iteration and smaller specialised models are expected to improve inference efficiency while preserving premium capabilities for complex workloads, sustaining pricing power and supporting high-quality revenue growth.

GLM-5.3 reinforces Z.ai's post-training and efficiency edge. GLM-5.3 retains the same 744B-parameter base model as GLM-5.2, but delivers a meaningful step-up in coding and agentic performance through advanced post-training and RL, bringing it closer to leading frontier models while using only 1/10th-1/5th of their parameter scale. Importantly, Z.ai has kept GLM-5.3 at the same API pricing as GLM-5.2, while lower token consumption further reduces the effective cost of completing tasks, highlighting a strong intelligence-cost advantage. We believe this reinforces the company's uniqueness through post-training, RL and engineering optimisation rather than model scale alone, with further upside from next-generation models potentially materialising in 4Q26. Broad deployment across leading Chinese internet and emerging technology companies should also strengthen Z.ai's commercialisation and real-world data flywheel.

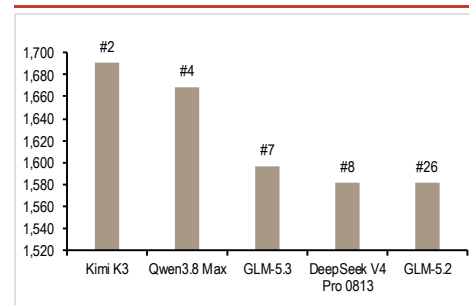
Developer scale is compounding Z.ai's enterprise coding moat

Figure 4: GDPval-AA v2 Model Ranking



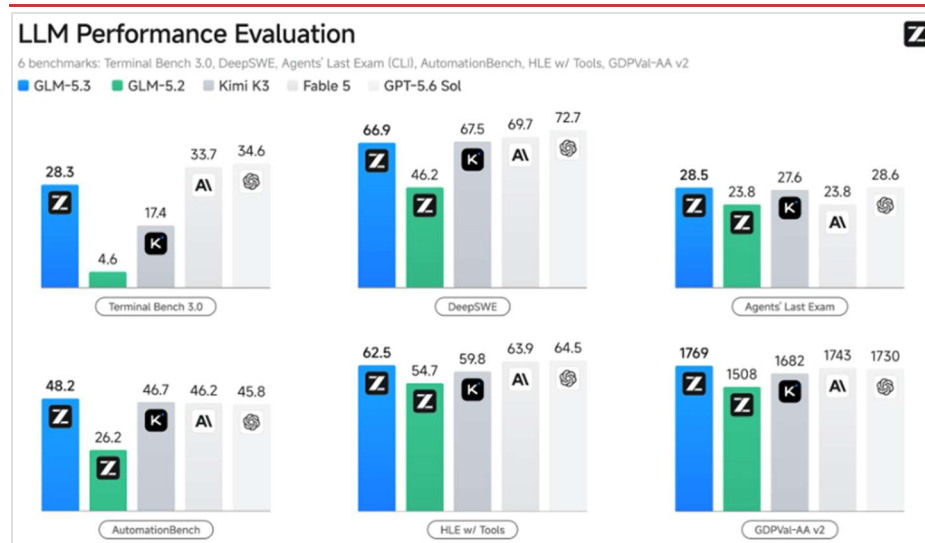
Source: Artificial Analysis, UOB Kay Hian

Figure 5: Code Arena WebDev Model Ranking



Source: Arena ai, UOB Kay Hian

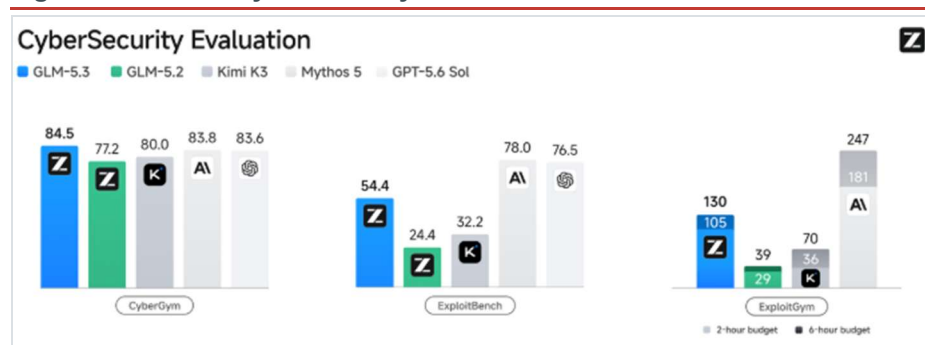
Figure 6: GLM5.3 Performance Evaluation



Source: Z.ai, UOB Kay Hian

Emerging cybersecurity capabilities. GLM-5.3 shows a meaningful step-up in cybersecurity reasoning following the incorporation of vulnerability-discovery data and environments in post-training. As training scaled, the model progressed beyond identifying isolated vulnerabilities to reasoning across multiple stages of exploitation and constructing end-to-end exploit chains. Z.ai evaluated these capabilities across three benchmarks covering different stages of vulnerability analysis and exploitation.

Figure 7: GLM5.3 CyberSecurity Performance Evaluation



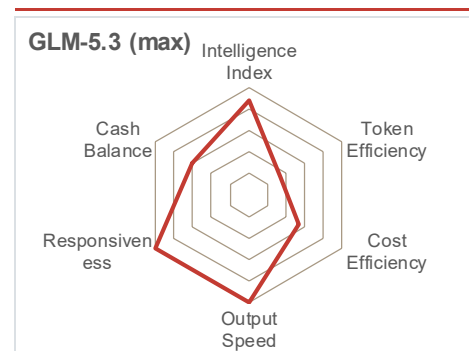
Source: Z.ai, UOB Kay Hian

FRONTIER MODEL CAPABILITIES; IMPROVING INFERENCE EFFICIENCY

Research-driven ecosystem underpins long-term technological leadership. Originating from Tsinghua University's Knowledge Engineering Group (KEG), Z.ai combines strong research capabilities with an engineering-led approach spanning compute, proprietary data and algorithm innovation. Around 70% of its more than 1,000 employees are in R&D, alongside hundreds of PhD interns.

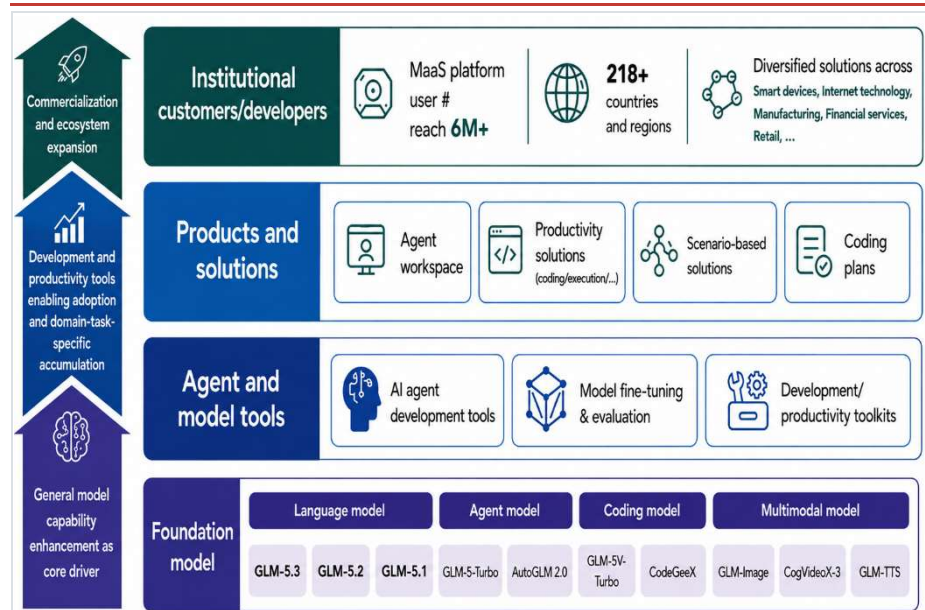
Z.ai sees self-learning agents and multimodal vision-language models as key next-stage scaling drivers. Building on CodeGeeX (2022), the GLM-5 family has reached global frontier coding performance, with GLM-5.3 expected to extend GLM-5.2's inference-efficiency gains. GLM-5.3 delivers a 50.0% uplift on Z.ai Code Bench vs GLM-5.2 and scores 28.3% on Terminal-Bench 3.0, 66.9% on DeepSWE v1.1, 84.5% on CyberGym and 54.4% on ExploitBench, highlighting gains across coding, agentic execution and cybersecurity. Combined with long-horizon agents, these capabilities allow Z.ai to expand from code generation to enterprise-grade agentic engineering, DevOps and complex workflows.

Figure 8: Multidimensional Performance Of GLM



Source: Z.ai, Industry data, UOB Kay Hian

Figure 9: Full-Stack AI Product And Commercialisation Ecosystem



Source: Z.ai, industry data, UOB Kay Hian

IMPROVING INFERENCE EFFICIENCY

Inference efficiency continues to improve. In Apr 26, GLM-5 achieved up to 132% higher system throughput in coding agent workloads through its LayerSplit key-value cache architecture and asynchronous RL infrastructure. The newly-launched GLM-5.3 adopts the same domestic-chip optimisation approach as GLM-5.2, including support for Huawei Ascend and Moore Threads. Further adoption of Huawei Ascend 910C/950 and other domestic AI chips is set to accelerate over 2026-28, with increasingly compute-efficient model architectures helping to offset hardware constraints.

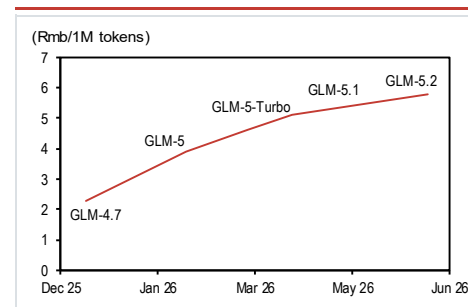
Commercial execution validates monetisation and pricing power. We believe Z.ai's leadership in coding, agents and long-context reasoning supports premium pricing despite increasing industry commoditisation. The GLM Coding Plan reached 240,000 paying developers in 1Q26, while Claw Plan attracted 400,000 subscribers in 20 days, driving 4x yoy paid token growth and 15x Coding Plan token growth in six months. Despite more than 83% cumulative price increases ytd, despite a January supply shortage temporarily reducing sales to 20% of normal levels, new plans continued to sell out, reinforcing our confidence in sustained growth in user numbers, expansion of ARPU and Z.ai's leadership in China's enterprise AI coding market.

ROAD MAP

GLM-5.3 on the price-to-performance frontier. We believe GLM-5.3 remains on the Pareto frontier for coding performance and pricing alongside Kimi K3. Claude Sonnet 5 delivers a comparable benchmark performance at nearly 2x GLM-5.3's token pricing, while GPT-5.6-Luna is cheaper but appears less competitive for long-horizon enterprise workloads. We also view Hugging Face's reported use of GLM-5.2 during a recent security incident as positive validation of its real-world reliability.

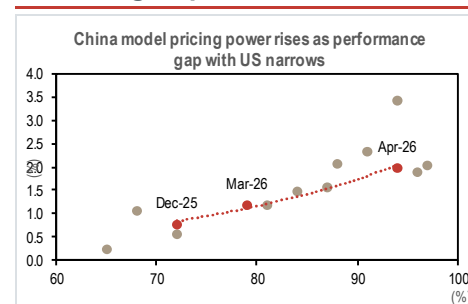
Model roadmap focuses on capability, not parameter size. The newly-released GLM-5.3 retains a similar model scale as GLM-5.2 (~744B parameters), with meaningful capability gains driven by enhanced post-training rather than parameter expansion, particularly across coding, agentic execution and cybersecurity. With a significantly larger next-generation model already under training, frontier competition should ultimately be assessed by real-world capability, efficiency and commercial usability rather than parameter count or short-term benchmark rankings.

Figure 10: Blended Prices Across GLMs



Source: Z.ai, Industry data, UOB Kay Hian

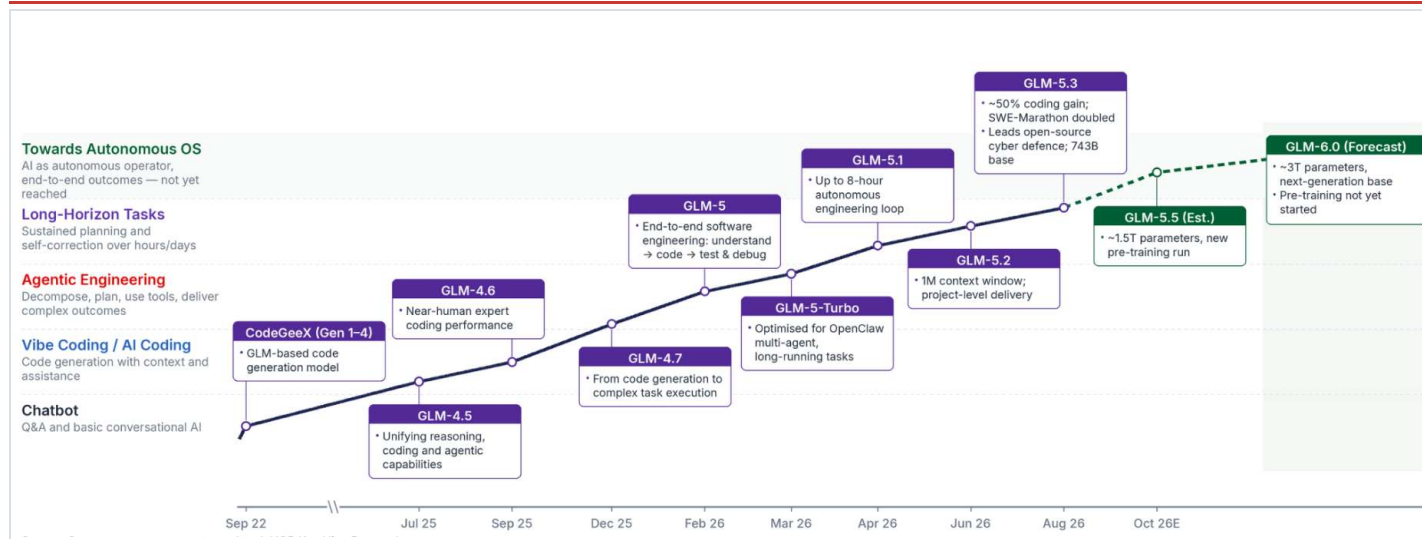
Figure 11: Pricing Power With Narrowing Gap



Source: Z.ai, Industry data, UOB Kay Hian

Execution quality is overtaking scale as the key benchmark

Figure 12: GLM Roadmap – From Coding Assistance To Autonomous Execution



Source: Z.ai, industry data, UOB Kay Hian

Long-horizon agentic tasks as the key growth driver. As AI shifts from conversational assistance toward autonomous execution, complex coding and enterprise workflows could lead to a significant step-up in token usage per user, while also supporting a greater willingness to pay and stronger retention. Z.ai's focus on long-context reasoning, coding and multi-step agent execution positions it well for this transition. We believe the shift from "answering questions" to "completing tasks" could drive a step change in usage intensity, API monetisation and enterprise adoption, making agentic workloads a key driver of Z.ai's medium-term revenue growth.

Agentic workloads drive usage intensity

Open-source ecosystem strengthens long-term competitive moat. Rather than relying solely on a closed ecosystem, Z.ai combines a commercial-friendly open-weight strategy with a strong enterprise focus. The release of GLM-5.3 under a permissive licence, alongside the expansion of its developer ecosystem through ZCode 3.0, should accelerate third-party adoption and deployment, strengthen developer mindshare and generate real-world usage feedback. Over time, this could create a flywheel from open-weight distribution → developer adoption → product improvement → enterprise deployment → commercial monetisation.

Open-weight ecosystem strengthens the flywheel

RISING TOKEN CONSUMPTION UNDERPINS LONG-TERM API DEMAND

Cloud monetisation provides long-term upside. Z.ai is shifting from on-premise toward recurring cloud API and model-as-a-service (MaaS) revenue, targeting coding, financial services, legal and other workflow-intensive sectors, with consumer apps mainly serving as showcases and data channels. Its integrated MaaS platform spans language, multimodal, coding and agent models, although cloud monetisation carries the full serving costs. We expect AI agents, co-workers and coding assistants to drive sustained token consumption as AI increasingly automates 24x7 workflows. Z.ai's token consumption rose 15x within six months of its Coding Plan launch, and we forecast a further 15-20x increase in 2026.

APIs become the long-term growth engine

As Chinese models reduce the frontier gap, rising token demand should support long-term API revenue growth. While on-premise revenue benefits from recurrent upgrades and is expected to grow 50-100% in 2026, we expect its revenue contribution to decline as API monetisation accelerates, making APIs the primary long-term earnings driver. In 2025, on-premise contributed 73.7% of revenue at a 49% gross margin, vs 19% for cloud, with most revenue coming from mainland China. We expect cloud revenue and margins to improve after the release of SOTA models and broader adoption by global developers, while the high-margin on-premise business continues to fund R&D and infrastructure.

Improving cloud margins could support a more scalable revenue model

API margins remain attractive despite near-term monetisation and distribution costs. Open platform/API ARR rocketed from US\$250m in late March to US\$1b in July, with compute capacity, rather than demand, likely to remain the key bottleneck through year-end. Z.ai's open-weight strategy is also driving strong overseas adoption, although a significant portion of self-hosted usage currently generates little direct revenue. We believe about 50% API inference margins are achievable when workloads run on Z.ai's own infrastructure, while testing-as-a-service partnerships dilute blended margins through distribution costs. As domestic compute capacity expands, we expect a greater share of inference to shift to self-operated infrastructure, supporting structurally higher API profitability over time, although near-term margins could remain under pressure due to higher implementation staffing and hardware costs. In parallel, we expect next-generation models to adopt more restrictive commercial licensing, allowing Z.ai to better monetise overseas demand and capture value from the growing open-weight user base.

Self-operated inference supports margin upside

WHAT DRIVES PRICING POWER IN THE LONG TERM?

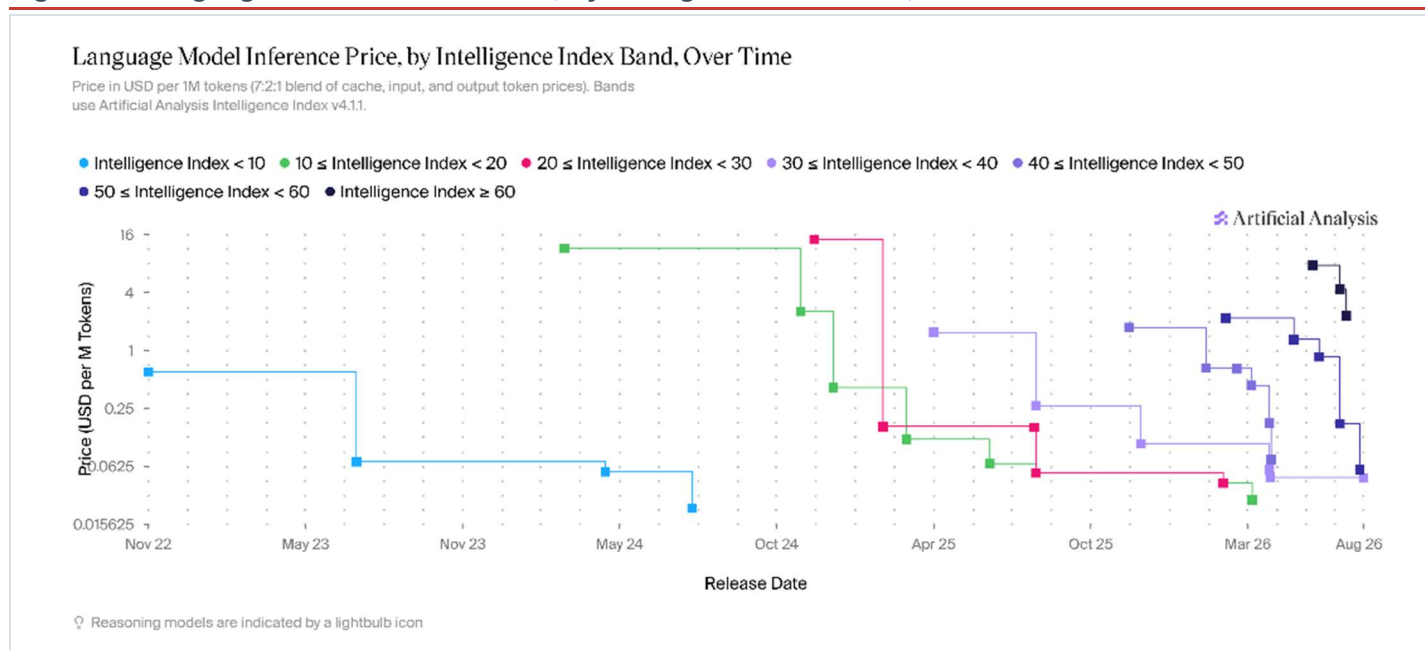
Sustained innovation will drive long-term monetisation. As foundation models commoditise, value creation will increasingly shift toward frontier applications such as coding agents, enterprise workflows, multi-step reasoning and AI for science. Over time, monetisation is also likely to evolve from token-based pricing toward workflow- and value-based pricing tied to business outcomes.

Value capture shifts from model access to business outcomes

Pricing power is increasingly reflected in monetisation and willingness to pay, rather than headline price hikes. Z.ai serves over 6m platform users, with API services contributing about 90% of revenue and Coding Plan subscriptions the remaining 10%. Z.ai has shown its pricing power through price hikes, including over 30% for the domestic Coding Plan (Feb 26), 80-140% for overseas plans (Apr 26), 50% for GLM-5 vs GLM-4.7, 20% for GLM-5 Turbo, and 10% for GLM-5.1. More recently, it has shifted from price hikes to scaling usage after earlier capacity constraints, with the Coding Plan now broadly available, while GLM-5.3 delivers a stronger performance without a price hike like GLM-5.2. We believe the ability to expand capacity and adoption while maintaining pricing, and to deliver higher intelligence at a lower effective cost per task, reinforces Z.ai's differentiated value proposition and provides further headroom for monetisation as enterprise and developer workloads scale.

Monetisation shifts from price to usage, with lower cost per task reinforcing value

Figure 13: Language Model Inference Price, By Intelligence Index Band, Over Time



Higher pricing reflects frontier capability and margin expansion. GLM-5.2 inference cost is materially lower than the prior generation, while cloud gross margin improved from 18.9% in 2025 to ~24.0% currently. At launch, GLM-5.2 implied a ~13% higher realised API price vs GLM-5.1 due mainly to input-length pricing tiers; pricing has since converged, improving unit economics. We see a more segmented AI pricing market, with commoditised models facing deflation while frontier models retain premiums in coding, agents, enterprise automation and long-context reasoning.

Resilient demand after the price increase would further validate Z.ai's pricing power and customers' focus on productivity and ROI rather than token cost. We forecast cloud/on-premise gross margin reaching 35%/60% by 2030, versus 24%/53% in 2026.

BUSINESS MODEL

Reported ARR probably understates underlying demand. We believe cloud ARR (MaaS, API and coding, excluding on-premise deployment) is driven mainly by higher token consumption per customer, supported by better models and user growth. Open platform/API ARR rose from US\$250m in late March to US\$1.0b in July, while compute capacity, rather than demand, should remain the key bottleneck through year-end. Based on media reports of a 1GW data-centre build-out, we estimate ARR could reach Rmb15.5b (US\$2.3b), more than double July's level and above management's original 2026 target, driven by GLM-5.3, future model upgrades and broader coding, agent and co-worker adoption. Open-source distribution through third-party clouds and coding agents should further expand demand, with additional upside as domestic application-specific integrated circuits and inference capacity increase.

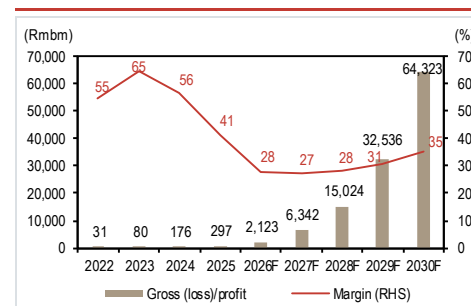
Strong revenue growth and margin expansion support long-term profitability. We project API ARR of US\$6.5b by end-27, driving total revenue to Rmb7.7b in 2026 and Rmb23.3b in 2027. We forecast cloud-based revenue growing 226% yoy in 2027, while cloud-based gross margin could expand to 23.5% in 2026 and 25.0% in 2027. We model R&D expenses rising 90% in 2026 and 98% in 2027 to Rmb6b and Rmb12b respectively, with non-GAAP operating breakeven expected in 2029. By 2030, we forecast a Rmb182.6b revenue, 35% gross margin (Rmb64.3b gross profit), Rmb29.9b GAAP operating profit, about Rmb25.7b in non-GAAP net income and net margin of 14%.

AMPLE FINANCIAL SUPPORT FOR CONTINUED COMPUTE INVESTMENT

Strong funding provides ample capacity for AI infrastructure investment. We believe Z.ai's strong demand and rapid commercial traction should continue to support access to capital, allowing the company to reinvest aggressively in compute and R&D. Before its Jan-26 IPO, Z.ai completed eight private funding rounds, raising Rmb8.4b. The IPO raised Rmb3.6b (HK\$4.3b), followed by a HK\$31.4b (US\$4b) H-share placement on 9 July for model development, compute expansion and commercialisation. Z.ai also plans an SSE STAR Market listing for 2-8% of share capital, targeting up to Rmb15b in proceeds. We forecast Rmb32b net cash by 2026, supported by equity funding, local-government/industrial capital, data-centre and compute partnerships, and potentially supplier financing.

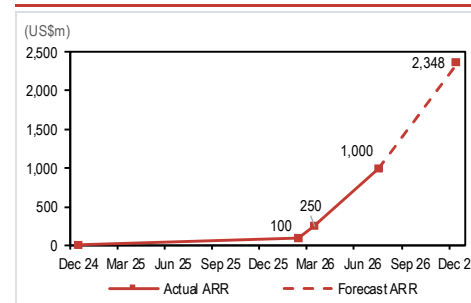
Compute remains a key investment priority. Z.ai has reportedly completed a 1GW-class data centre using Chinese AI accelerators, with Huawei, Cambricon and Alibaba among reported suppliers, while several clusters each contain 10,000+ chips. Full 1GW capacity is expected online by end-26. We see engineering efficiency and token output per unit cost, rather than chip count alone, as the key competitive edge. At full utilisation, the facility could support about US\$3b ARR, although initial run-rate revenue should remain below this level. The facility thus represents not only major capacity growth, but also a crucial real-world test of whether domestic chips can support frontier models at competitive costs in the long run.

Figure 14: Gross Margin Performance Trend



Source: Artificial Analysis, UOB Kay Hian

Figure 15: ARR Projection Roadmap



Source: Z.ai, Industry data, UOB Kay Hian

Robust capital support for compute investment

Infrastructure build-out supports long-term AI competitiveness

Profit & Loss

Year to 31 Dec (Rmbm)	2025	2026F	2027F	2028F
Net turnover	724.3	7,666.5	23,252.3	53,728.3
EBITDA	(3,499.0)	(2,943.7)	(2,774.9)	2,530.1
Deprec. & amort.	281.1	2,140.9	4,537.1	6,790.5
EBIT	(3,780.0)	(5,084.5)	(7,312.0)	(4,260.4)
Total other non-operating income	(878.7)	(0.0)	0.0	(0.0)
Associate contributions	0.0	0.0	0.0	0.0
Net interest income/(expense)	(59.5)	(2.6)	291.1	48.9
Pre-tax profit	(4,718.2)	(5,087.1)	(7,020.9)	(4,211.5)
Tax	0.0	0.0	0.0	0.0
Minorities	20.0	21.5	29.7	17.8
Net profit	(4,698.2)	(5,065.6)	(6,991.2)	(4,193.7)
Net profit (adj.)	(3,182.0)	(4,417.2)	(6,217.0)	(3,246.8)

Cash Flow

Year to 31 Dec (Rmbm)	2025	2026F	2027F	2028F
Operating	(2,246.1)	61.0	759.8	4,619.4
Profit to the year	(4,718.2)	(5,087.1)	(7,020.9)	(4,211.5)
Tax	0.0	0.0	0.0	0.0
Deprec. & amort.	281.1	2,140.9	4,537.1	6,790.5
Associates	0.0	0.0	0.0	0.0
Working capital changes	704.3	2,301.6	2,404.1	1,040.0
Non-cash items	0.0	0.0	0.0	0.0
Other operating cashflows	1,486.7	705.6	839.6	1,000.4
Investing	(375.9)	(10,013.9)	(12,015.3)	(11,776.8)
Capex (growth)	(23.1)	(10,013.9)	(12,015.3)	(11,776.8)
Investment	338.1	338.1	338.1	338.1
Others	(691.0)	(338.1)	(338.1)	(338.1)
Financing	2,614.1	30,927.1	0.0	0.0
Dividend payments	0.0	0.0	0.0	0.0
Proceeds from borrowings	689.1	0.0	0.0	0.0
Loan repayment	(137.2)	0.0	0.0	0.0
Others/interest paid	2,062.2	30,927.1	0.0	0.0
Net cash inflow (outflow)	(7.9)	20,974.1	(11,255.5)	(7,157.4)
Beginning cash & cash equivalent	2,268.1	2,254.9	23,229.0	11,973.5
Changes due to forex impact	(5.3)	0.0	0.0	0.0
Ending cash & cash equivalent	2,254.9	23,229.0	11,973.5	4,816.1

Balance Sheet

Year to 31 Dec (Rmbm)	2025	2026F	2027F	2028F
Fixed assets	655.8	8,520.7	15,991.8	20,972.3
Other LT assets	626.6	634.9	641.9	647.7
Cash/ST investment	2,367.7	23,341.8	12,086.4	4,929.0
Other current assets	1,203.7	3,648.1	6,751.7	10,035.1
Total assets	4,853.9	36,145.4	35,471.8	36,584.1
ST debt	604.8	604.8	604.8	604.8
Other current liabilities	11,801.4	16,547.4	22,055.0	26,378.5
LT debt	378.4	378.4	378.4	378.4
Other LT liabilities	180.1	180.1	180.1	180.1
Shareholders' equity	(8,111.0)	18,434.6	12,253.3	9,042.2
Minority interest	0.0	0.0	0.0	0.0
Total liabilities & equity	4,853.9	36,145.4	35,471.8	36,584.1

Key Metrics

Year to 31 Dec (%)	2025	2026F	2027F	2028F
Profitability				
EBITDA margin	(483.1)	(38.4)	(11.9)	4.7
Pre-tax margin	(651.4)	(66.4)	(30.2)	(7.8)
Net margin	(648.6)	(66.1)	(30.1)	(7.8)
ROA	(101.8)	(24.7)	(19.5)	(11.6)
ROE	n.a.	n.a.	(45.6)	(39.4)
Growth				
Turnover	131.9	958.4	203.3	131.1
EBITDA	n.a.	n.a.	n.a.	n.a.
Pre-tax profit	n.a.	n.a.	n.a.	n.a.
Net profit	n.a.	n.a.	n.a.	n.a.
Net profit (adj.)	n.a.	n.a.	n.a.	n.a.
EPS	n.a.	n.a.	n.a.	n.a.
Leverage				
Debt to total capital	(13.8)	5.1	7.4	9.8
Debt to equity	n.a.	5.3	8.0	10.9
Net debt/(cash) to equity	n.a.	(121.3)	(90.6)	(43.6)
Interest cover (x)	(58.8)	n.m.	n.a.	n.a.

Appendix I: Risk Factors

- a) **Risks of competitive pressure from global foundation model developers and mega-cap technology platforms.** Z.ai operates in a rapidly evolving and highly competitive foundation model market, with frequent model launches and shifting customer needs. Failure to keep pace with technological advances or invest in R&D efficiently could weaken its competitiveness, while early-stage market dynamics make customer demand difficult to predict.
- b) **Limited near-term profitability risks.** High R&D and infrastructure spending will likely constrain near-term earnings visibility. Profitability depends on scaling commercialisation and revenue fast enough to offset continued investment in model development and user experience.
- c) **Sustained investment and profitability risks.** Funding pressure is high due to elevated levels of R&D, training cost, computing resources, sales and marketing, and talent acquisition required to support model iteration and commercial expansion. Models that fail to keep pace with peers may struggle to generate sufficient revenue ($\text{ARR} \times \text{inference gross margin}$) to cover training costs.
- d) **Model performance and technology iteration risks.** Z.ai's premium pricing relies partly on its SOTA model performance. Failure to sustain its technology and iteration lead could weaken its differentiation and pricing power, accelerate commoditisation and drive price declines, customer churn and margin pressure.
- e) **Regulatory and data compliance risks.** Z.ai operates under evolving AI, cybersecurity and data regulations, with growing requirements for content safety, algorithm filing, data processing and training data usage. Non-compliance could increase costs, constrain operations and result in reputational damage or regulatory penalties.
- f) **Geopolitical and supply chain risks.** Z.ai's inclusion on the US Entity List could constrain its expansion in the US and access to US-origin advanced GPUs and computing technologies, potentially raising training and inference costs and slowing model iteration. Availability of advanced AI chips remains a key supply-chain risk.
- g) **Unlock and free-float risks.** Approximately 73.95% of Z.ai's H shares are expected to become eligible for trading in Jan 27, potentially creating a meaningful near-term technical overhang and increasing share price volatility. This risk is particularly relevant given the stock's recent volatility, with the share price retracing about 43% from its high of HK\$2,980 on 22 June to HK\$1,707 one week after the July unlock.

Appendix II: Company Overview

Founded in 2019, Z.ai is focused on advancing cognitive intelligence models with a strong emphasis on innovation in China. The AI company developed the GLM pre-training framework and has achieved breakthroughs in large-scale model development, including the 10B-parameter GLM-10B, a trillion-parameter sparse model based on the MoE architecture, and the open-sourced GLM-130B model. In 2023, Z.ai launched ChatGLM, a 100B-parameter foundational conversational model, followed by multiple upgrades and the release of ChatGLM-6B, which enabled local fine-tuning and deployment for developers. Z.ai continues to demonstrate strong open-source developer traction, with about 18,000 Hugging Face followers and 150 published models, while GLM-5.2 has recorded about 4.8m downloads across its standard and FP8 variants, underscoring strong developer adoption.

Development Milestones

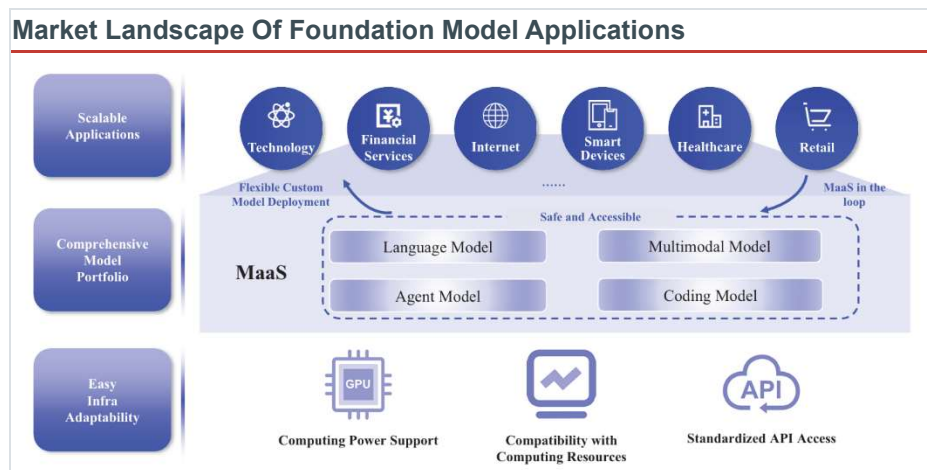


Source: Z.ai, UOB Kay Hian

Model Development Roadmap And Key Breakthroughs

Phase	Model Name	Release Date	Core Positioning	Key Breakthroughs & Features
Foundation & exploration	GLM-130B	2022	China's first open-source 100B+ parameter foundational model	Proved the viability of the in-house GLM architecture; laid the technical groundwork for future development.
	ChatGLM / ChatGLM-6B	Mar 23	China's first 100B+ parameter dialogue model	The open-sourced 6B version lowered adoption barriers and significantly boosted the domestic AI application ecosystem.
Chat & catch-up	GLM-4	Jan 24	Next-gen high-performance foundational model	Performance comprehensively rivalled OpenAI's GPT-4, marking its entry into the top international tier.
	GLM-4.5 / 4.6 / 4.7	2025	Models specialised in coding and agent capabilities	Dense iterations that established code generation and agent task executions as core strengths.
Agent & coding surge	GLM-5	Feb 26	Flagship foundational model (positioned as an "Engineer")	Key breakthrough: AI can autonomously test and iterate code, enabling closed-loop engineering problem-solving.
	GLM-5.1	Apr 26	Long-horizon task-enhanced model	Can work independently and continuously for over 8 hours, autonomously planning and delivering engineering-grade outcomes.
	GLM-5.2	Jun 26	Flagship coding-focused model	Supports 1M lossless context; ranked #1 in coding ability in the global blind test of Code Arena at launch.
	GLM-5.3	Aug 26	Coding, agentic and cybersecurity-enhanced flagship model	Scaling post-training delivered stronger coding and emergent cybersecurity capabilities, demonstrating stronger multi-stage vulnerability analysis and exploitation reasoning.
	GLM-5.3	Aug 26	Coding, agentic and cybersecurity-enhanced flagship model	Scaling post-training delivered stronger coding and emergent cybersecurity capabilities, demonstrating stronger multi-stage vulnerability analysis and exploitation reasoning.
Full-modality portfolio	Vision models (e.g., GLM-4.6V)	Across all periods	Visual understanding & reasoning	Enables the model to "see" images and perform reasoning based on visual input.
	Image generation (e.g., CogView-4)	Across all periods	Text-to-image & text rendering	Excels at generating in-image text (e.g., posters, logos) with high accuracy.
	Video generation (e.g., CogVideoX-3)	Across all periods	Text-to-video	Generates dynamic video content from text descriptions.
	Agent (e.g., AutoGLM)	Post-2025	Cross-application task automation	Can autonomously operate smartphones or computers to complete complex tasks via natural language instructions.

Source: Z.ai, Industry data, UOB Kay Hian



Source: Z.ai, UOB Kay Hian

Comprehensive MaaS ecosystem with flexible deployment. Z.ai's MaaS platform offers a broad portfolio of AI models spanning language, multimodal, agentic and coding capabilities, enabling enterprises and developers to select solutions tailored to different use cases. The platform supports scalable deployment across diverse business workflows and application scenarios, complemented by agent workspaces and customisation tools that facilitate model fine-tuning and agent development. Through collaboration with infrastructure partners, the platform also provides strong infrastructure adaptability, supporting deployment across multiple cloud environments, computing architectures and device types, while delivering standardised API access and integrated computing resources.

AI model portfolio. Z.ai offers a comprehensive portfolio of AI models spanning LLMs, reasoning models, multimodal models, coding models, and AI agents. Its foundation models support capabilities including reasoning, coding, agentic workflows, multimodal understanding and generation, and long-context processing.

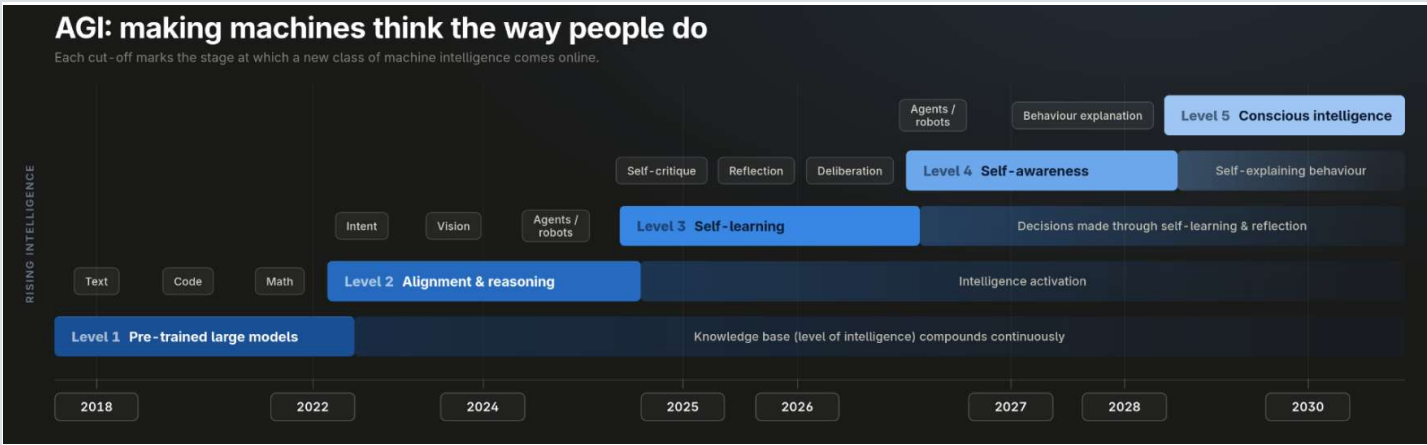
- **Large language models.** Z.ai's foundation models serve as the core platform underpinning its AI ecosystem, supporting a broad range of text generation, reasoning, coding, and agentic AI applications.
- **Reasoning models.** The company develops specialised reasoning models designed to enhance complex problem-solving and deep-thinking capabilities for advanced reasoning tasks.
- **Multimodal models.** Z.ai offers multimodal models covering image generation, visual understanding and reasoning, video generation, real-time visual interaction, and voice applications.
- **Coding models.** It provides AI coding models designed to improve software development efficiency and streamline programming workflows.
- **AI agents.** Z.ai develops AI agents for both consumer and enterprise applications, enabling autonomous task execution, workflow automation, and interaction with digital environments through graphical user interfaces.

Coding Plan Categories

Item	GLM Coding Lite	GLM Coding Pro	GLM Coding Max
Monthly price	Rmb118/month	Rmb538/month	Rmb1,078/month
Annual plan — current promo	~Rmb83/month; Rmb991/year	~Rmb377/month; Rmb4,519/year	~Rmb755/month; Rmb9,055/year
5-hour usage quota	2,000 credits	12,000 credits	28,000 credits
Weekly usage quota	10,000 credits	60,000 credits	140,000 credits
MCP tool quota	Shared credit pool; no separate fixed call quota	Shared credit pool; no separate fixed call quota	Shared credit pool; no separate fixed call quota
Recommended usage	1 active project	1-2 active projects	>2 active projects
Target users	Individual developers and light coding users; 1 active project	High-frequency coding users; 1-2 active projects	Heavy coding users and complex workloads; >2 active projects

Source: Z.ai, industry data, UOB Kay Hian

AI Capability Evolution Roadmap



Source: Z.ai, industry data, UOB Kay Hian

Timeline. Pre-2022, Z.ai explored large models, drawing inspiration from the cognitive dual-system theory (2018), launched the GLM architecture (2020), and persisted through a slow period to release GLM-130B and an MaaS platform (2021-22).

From 2022-26, the company progressed through three strategic phases.

- **Benchmarking (2022-24).** GLM-130B and GLM-4 proved domestic models’ strength in knowledge and bilingual alignment.
- **Deepening (2025).** Z.ai introduced a “thinking” mode and multimodal systems to push reasoning and visual limits.
- **Agentic shift (2026).** With GLM-5, Z.ai prioritised planning, tool stability, and environmental adaptation.

Overall, this trajectory reflects not just model growth, but a fundamental re-envisioning of large models as the backbone of future productivity systems

Dual deployment strategy supports rapid cloud growth while maintaining a resilient on-premise base. Z.ai commercialises its GLM models through both cloud-based APIs and on-premise deployments, allowing enterprises to adopt AI according to scalability, data-security and regulatory requirements. In 2025, on-premise deployment accounted for 73.7% of revenue and generated a 49% gross margin, supported by strong demand from regulated enterprise and government customers.

Looking ahead, we expect cloud revenue to rise more than 30x yoy to Rmb6.6b in 2026, overtaking on-premise deployment as the key revenue contributor as global developer adoption and API usage accelerate. Meanwhile, on-premise deployment should continue to provide a resilient revenue base through recurring model upgrades, capacity expansion and new functionalities,

supported by high switching costs and persistent localisation requirements. While cloud gross margin should improve with scale and utilisation, we expect it to remain below on-premise margin in the near term due to higher inference and computing costs.

Potential path to higher monetisation via licensing shift. A potential shift toward more restrictive commercial licensing for future open-weight models could allow Z.ai to capture greater value from large-scale third-party deployments, particularly through global inference providers and hyperscalers. This could complement its high-margin API business with a relatively asset-light licensing model, supporting broader global adoption and potentially improving long-term monetisation.

AWS partnership strengthens Z.ai's global commercialisation strategy. Z.ai has partnered with AWS to offer its GLM-4 and GLM-5 model families through AWS Marketplace, enabling secure enterprise deployment on AWS' global cloud infrastructure. We view the partnership as strategically important for three reasons: a) it validates Z.ai's frontier model competitiveness internationally; b) it expands access to AWS's global enterprise customers and developer ecosystem, lowering overseas customer acquisition costs; and c) it supports the growth of Z.ai's API and cloud business, accelerating the transition from deployment-led revenue toward scalable recurring cloud monetisation. While the near-term financial contribution is likely modest, we see overseas API revenue and enterprise customer growth as key indicators to watch.

Expanding global reach anchored by China. While revenue is mainly derived from customers in China, the group's AI ecosystem has achieved broad international reach through the deployment of GLM foundation models, AI agents and enterprise-grade large model services across 218 countries and regions worldwide. The company has built a growing developer and enterprise ecosystem, serving over 6m enterprises and developers globally. Supported by its "foundation model + platform + ecosystem" strategy, Z.ai continues to expand the adoption of its AI capabilities across a diverse range of industries and use cases, while maintaining China as its core commercialisation market.

A critical distinction between capability formation and utilisation costs. Z.ai's cost structure separates R&D, which supports capability formation, from COGS, which reflects capability utilisation. R&D covers model pre-training, post-training, RL, alignment, evaluation and related infrastructure, and is therefore less directly linked to near-term usage volumes. In contrast, cloud/API COGS is more usage-linked, driven primarily by inference compute, serving infrastructure, bandwidth and third-party distribution costs. This distinction is important: R&D determines Z.ai's ability to improve model capability, while serving efficiency determines the margin generated from commercialising that capability.

FINANCIALS BY SEGMENT

Z.ai provides AI model capabilities to institutional customers based on its MaaS platform, offering four types of models: language models, multimodal models, agent models and coding models. It also offers integrated tools for model fine-tuning, deployment and agent development. There are primarily two model deployment options and business models for Z.ai's MaaS:

- **On-premise deployment.** Z.ai's on-premise deployment consists primarily of localised deployment of large models and related on-site services to facilitate customer implementation. The company also provides model training and fine-tuning services to support customer-specific requirements.
- **Cloud-based deployment.** Z.ai provides cloud-based deployment through cloud infrastructure. Its cloud-based services include subscription-based contracts and usage-based contracts, enabling customers to access and utilise large model capabilities through cloud deployment according to their contractual arrangements and resource utilisation.

2025 RESULTS REVIEW AND OUTLOOK

Z.ai's revenue grew 132% yoy from Rmb312m to Rmb724m in 2025, driven by continuous iterations of its foundation models, which substantially improved model intelligence and performance. The enhanced capabilities increased model invocation frequency and customer usage, contributing to strong revenue growth. Gross profit rose 69% yoy, although gross margin declined from 56% to 41%, mainly due to a higher contribution from lower-margin cloud-based deployments vs on-premise deployments. Adjusted net loss widened 29% yoy to Rmb3.2b, reflecting elevated R&D investment, higher staff costs from R&D team expansion, and higher computing service fees paid to third-party suppliers as the group continued to iterate its foundation models and invest in more advanced model training infrastructure.

On-premise deployment. Revenue from on-premise deployment increased from Rmb264m in 2024 to Rmb534m in 2025, representing a growth of 102% yoy. The strong performance was driven by continuous model iterations that enhanced model intelligence, expanded model capabilities and supported sustained enterprise demand. While on-premise deployment remained the group's largest revenue contributor, accounting for 73.7% of total revenue in 2025, its contribution declined from 84.5% in 2024 as growth in cloud-based deployment continues to accelerate, resulting in a more balanced revenue mix.

Cloud-based deployment. Revenue from cloud-based deployment increased from Rmb48m in 2024 to Rmb190m in 2025, representing a growth of 293% yoy. The robust growth was primarily attributable to continuous model iterations that significantly enhanced model intelligence and performance, driving higher model invocation volumes and broader customer adoption. Consequently, the contribution of cloud-based deployment to total revenue increased from 15.5% in 2024 to 26.3% in 2025, reflecting growing demand across deployment formats and a gradual diversification of the group's revenue base.

Geographical segments. Z.ai continued to expand its global footprint in 2025, with its programming, agent and enterprise-level large model services covering 218 countries and regions worldwide. It also collaborated with over 6m SMEs and developers to build its AI ecosystem. Despite its large global footprint, revenue continued to be generated mainly by the China market, reflecting the strong domestic adoption of its AI solutions and enterprise services.

OUTLOOK

2026/2027 outlook. We expect revenue to jump 958% yoy to Rmb7.7b in 2026 and 203% to Rmb23.3b in 2027, underpinned by: a) next-generation foundation model upgrades; b) an expanding enterprise customer base across productivity-focused use cases such as business operations, content production and software development; and c) improving monetisation through higher model pricing. Continued advancements in model performance and broader AI adoption are expected to sustain demand growth.

On-premise deployment. We expect revenue from on-premise deployment to grow 104% yoy to Rmb1.09b in 2026 and 64% yoy to Rmb1.8b in 2027, driven by strong demand from data-sensitive customers requiring enhanced security, data sovereignty, operational control, and response stability. Key end markets include the financial services, government and healthcare sectors, where stringent regulatory and compliance requirements continue to support the adoption of on-premise AI solutions.

Cloud-based deployment. We expect cloud-based deployment revenue to grow at an exponential 3,354% yoy to Rmb6.6b in 2026 and 226% to Rmb21.5b in 2027, supported by increasing enterprise adoption of AI cloud services. The pay-as-you-go model allows customers to access LLM capabilities more efficiently, with revenue generated primarily through API consumption and subscription plans. We expect the cloud revenue mix to continue rising, driven by higher AI cloud penetration in China, as enterprises increasingly favour cloud deployment for its lower upfront investment requirements, faster access to model upgrades, greater deployment flexibility, and lower maintenance costs.

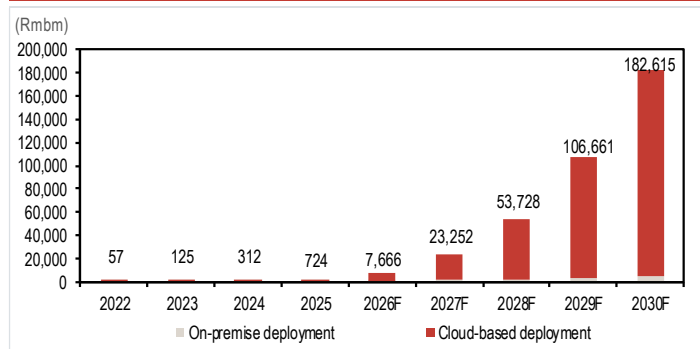
R&D investment is expected to remain elevated in 2026/2027, increasing 90%/98% yoy respectively to about Rmb6b and Rmb12b, as the company continues to invest heavily in next-generation foundation models and agentic AI capabilities. Key R&D priorities are expected to include advancing long-horizon task execution, enhancing memory and reasoning capabilities, strengthening autonomous agent functionality, and deepening software-hardware co-design with domestic AI chips. We believe these initiatives will further improve model performance, broaden enterprise AI applications, and drive higher model invocation volumes, supporting the company's long-term growth trajectory.

Adjusted net profit. We expect Z.ai to remain loss-making in the medium term, with profitability projected to be achieved in 2029, when adjusted net profit is forecast to reach Rmb6.2b. This is likely to be driven by strong monetisation across both the cloud-based and on-premise deployment businesses, with revenue projected to reach Rmb102.8b and Rmb3.8b respectively, implying a four-year CAGR of 382% and 64% over 2025-29. Adjusted net loss is expected to widen by 39% yoy to Rmb4.4b in 2026, reflecting the company's continued commitment to substantial investments in foundation model development, computing infrastructure and AI talent to maintain the competitiveness of the SOTA model.

NET CASH POSITION

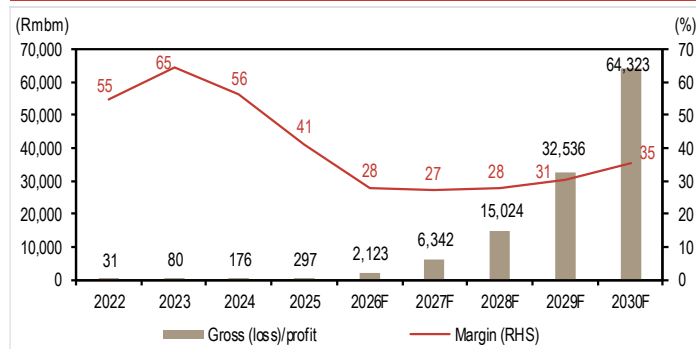
Net cash position of Rmb1.22b is projected to increase to Rmb22.2b in 2026, mainly supported by net proceeds from its IPO and an additional share placement. The strengthened balance sheet is expected to provide an important buffer against continued operating losses and elevated R&D spending, while supporting ongoing investment in foundation model training, computing infrastructure, and AI talent acquisition. We believe Z.ai's solid liquidity position remains critical for sustaining long-term model development and commercialisation expansion amid intensifying competition in the AI industry.

Revenue Trend By Segment



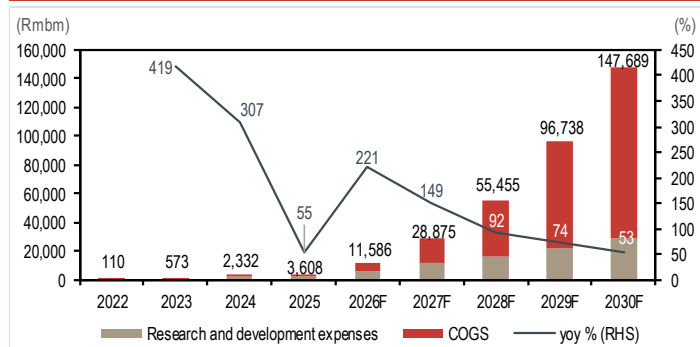
Source: Z.ai, UOB Kay Hian

Gross Margin Performance



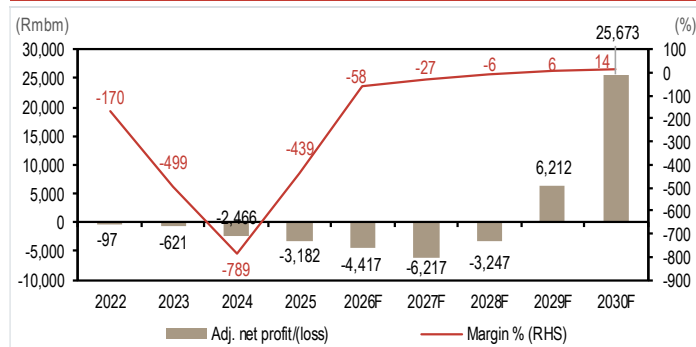
Source: Z.ai, UOB Kay Hian

R&D And COGS Growth Trajectory



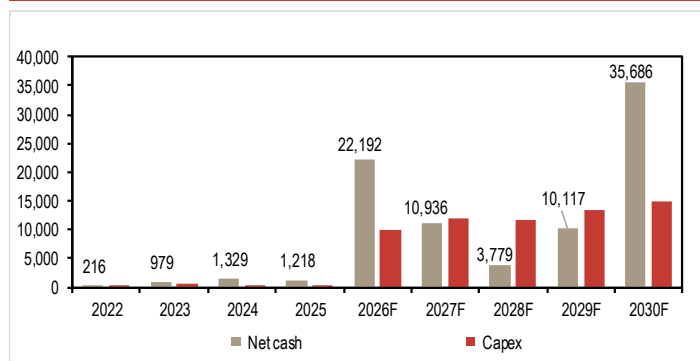
Source: Z.ai, UOB Kay Hian

Adjusted Net Profit/(Loss) Trend



Source: Z.ai, UOB Kay Hian

Net Cash vs Capex Trend



Source: Z.ai, UOB Kay Hian

Revenue Mix And Margin Trend By Deployment Model

	2023	2024	2025	2026F	2027F	2028F	2029F	2030F
Revenue								
On-premise deployment	113	264	534	1,091	1,794	2,666	3,826	5,284
Cloud-based deployment	12	48	190	6,576	21,459	51,062	102,835	177,331
Revenue (yoy %)								
On-premise deployment		134%	102%	104%	64%	49%	44%	38%
Cloud-based deployment		307%	293%	3354%	226%	138%	101%	72%
Revenue - share (%)								
On-premise deployment	90%	84%	74%	14%	8%	5%	4%	3%
Cloud-based deployment	10%	16%	26%	86%	92%	95%	96%	97%
Gross profit								
On-premise deployment	77	174	261	578	978	1,493	2,200	3,144
Cloud-based deployment	4	2	36	1,545	5,365	13,532	30,336	61,179
Gross margin (%)								
On-premise deployment	68%	66%	49%	53%	55%	56%	58%	60%
Cloud-based deployment	31%	3%	19%	24%	25%	27%	30%	35%

Source: Z.ai, UOB Kay Hian

Funding runway remains highly sensitive to revenue growth, margin expansion and capex execution, with capex representing a more significant swing factor for Z.ai. In our base case, we estimate more than three years of cash runway, with cumulative adjusted net losses of about Rmb13.8b through 2028, offset by existing liquidity, working-capital support and depreciation add-backs, leaving around Rmb3.6b of liquidity and no external financing requirement through 2028. In our bear case, slower revenue growth, weaker margin expansion and continued capex shorten the runway to about 2.9 years, resulting in a funding shortfall of around Rmb0.5b and implying the first external financing requirement could kick in in 2028. In our stress case, weaker revenue growth, limited margin improvement and elevated R&D spending, together with continued infrastructure investment, reduce the runway further to about 1.9 years and bring the first financing requirement forward to 2027.

We therefore believe Z.ai's liquidity outlook is increasingly dependent not only on revenue scaling and margin normalisation, but also on the pace and funding structure of its capex programme.

Runway Sensitivity Analysis

Scenario	Cash runway (years)	First external funding	Cum. adj. net losses (Rmbb)	Ending liquidity (Rmbb)		
Base	>3.0	n.a.	13.8	3.6	Base case	In our base case, we assume a 133% revenue CAGR over 2026F-28, with gross margin improving from 26% in 2026 to 26% in 2027 and 27% in 2028, while R&D intensity declines from 66% to 50% and 32% of revenue respectively. This results in cumulative adjusted net losses of about Rmb13.8b over 2026-28. After factoring in about Rmb19.1b of net cash, working-capital support from higher payables and non-cash depreciation, and approximately Rmb23.8b of cumulative capex, we estimate Rmb3.6b of excess liquidity remaining by end-28. We therefore do not expect any external financing requirement through 2028.
					First external financing need	No external financing required
Bear	2.9	2028	18	-0.5	Bear case	In our bear case, we assume a 123% revenue CAGR over 2026-28, with gross margin at 24% in 2026 and 2027 before improving modestly to 25% in 2028, while R&D intensity declines from 70% to 54% and 36% of revenue respectively. This results in cumulative adjusted net losses of about Rmb18.0b over 2026-28. After incorporating existing net cash, working-capital support, depreciation add-backs and cumulative capex, we estimate a Rmb0.5b funding shortfall by end-28, implying that the company could require its first round of external financing in 2028.
					First external financing need	External financing required in 2028
Stress	~1.9	2027	17.6	-0.2	Stress case	In our stress case, we assume revenue growth slows to 113% yoy in both 2026 and 2027, with gross margin remaining at 22% in both years, while R&D intensity stays elevated at 90% and 74% of revenue, respectively. This results in cumulative adjusted net losses of approximately Rmb17.6b over 2026-27. After incorporating Rmb19.1b of net cash as of end-26, together with approximately Rmb4.8b of working-capital support from higher payables and Rmb5.5b of depreciation, against Rmb12.0b of capex, we estimate a modest liquidity shortfall of approximately Rmb0.2b by end-27. We therefore expect Z.ai to require limited external financing in 2027 under our stress-case assumptions.
					First external financing need	External financing required in 2027

Source: UOB Kay Hian

Appendix III: Management

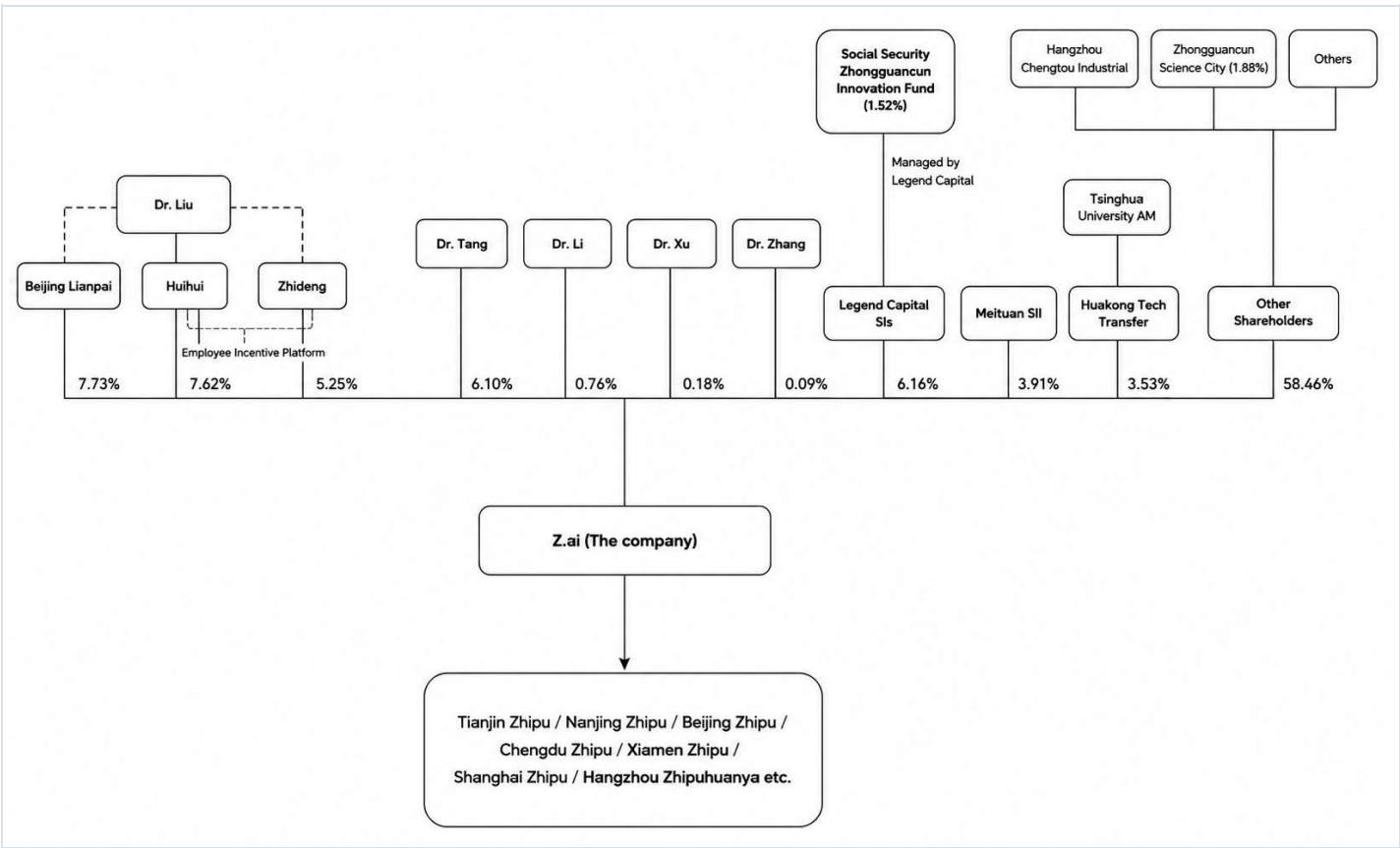
Key Management Team

Name	Age	Position	Experience
Dr Tang Jie	50	Founder, Chief Scientist	Founder and Chief Scientist of Z.ai, responsible for overall technology strategy, AI research and model innovation. He is a professor at Tsinghua University and Director of the Fundamental Model Research Center at the Institute for Artificial Intelligence. His research focuses on AI, knowledge graphs, data mining, social networks and machine learning.
Dr Liu Debing	51	Co-Founder, Chairman & Executive Director	Co-founder, Chairman and Executive Director of Z.ai, overseeing strategic planning, business development and overall operations. He has nearly two decades of experience in computing technology and AI research, with previous roles at Beijing Institute of Technology and Tsinghua University. He holds a PhD in Computer Science and Technology from the Chinese Academy of Sciences.
Dr Zhang Peng	47	Co-Founder, Executive Director, CEO, General Manager	Co-founder, Executive Director, CEO and General Manager of Z.ai, responsible for business strategy, product development and daily operations. Previously engaged in computer science and AI research at Tsinghua University, he has played a key role in the commercialisation and development of the company's AI technologies.
Zhang Xiaohan	29	Executive Director, Employee Representative Director	Executive Director and Employee Representative Director, overseeing data-labelling operations and supporting the development and operation of Z.ai QingYan. She previously studied at Tsinghua University, specialising in Engineering Mechanics and Data Science & Information Technology.
Wang Shaolan	53	Co-Founder, Deputy General Manager	Co-founder and Deputy General Manager of Z.ai, responsible for business development, financing and daily operations. He has over 20 years of experience in the technology industry, with previous roles at Beijing O2Micro Microelectronics, Nebula Sunac, Wankang and Tsinghua University.
Dr Li Juanzi	62	Co-Founder and Non-Executive Director	Co-founder and Non-Executive Director of Z.ai, providing strategic guidance on R&D, business development and corporate strategy. A professor at Tsinghua University and Director of the Knowledge Intelligence Center, she focuses on knowledge engineering, language computing and artificial intelligence.
Li Jiaqing	53	Non-Executive Director	Non-Executive Director providing strategic guidance on business development and corporate strategy. He has over 20 years of experience in venture capital, technology investment and corporate governance, and has held directorships at multiple listed companies and technology enterprises.
Wang Meng	44	Non-Executive Director	Non-Executive Director providing strategic guidance on business development and corporate strategy. He has extensive experience in telecommunications, cloud infrastructure and corporate investment, including management roles at China Mobile Communications Group, China Merchants Innovation Investment Management and Ant Group.
Yang Qiang	65	Independent Non-Executive Director	Independent Non-Executive Director providing strategic and governance advice. A leading AI researcher, he previously held academic roles at the University of Waterloo, Simon Fraser University and Hong Kong University of Science and Technology, and served as Chief AI Officer of WeBank. He holds a PhD in Computer Science from the University of Maryland.
Dr Xie Deren	55	Independent Non-Executive Director	Independent Non-Executive Director providing advice on strategy, governance and financial oversight. He is a professor at Tsinghua University's School of Economics and Management and has extensive expertise in accounting, corporate governance and financial reporting.
Tang Ying	49	Independent Non-Executive Director	Independent Non-Executive Director providing advice on corporate strategy and governance. He has rich management consulting, digital marketing and operating experience, including leadership roles at Roland Berger and Tiens Group.

Source: Z.ai, UOB Kay Hian

Appendix IV: Organisational Structure

Organisation Chart



Source: Z.ai, UOB Kay Hian

MiniMax Group Inc (0100 HK)

Multimodal Breadth Remains A Rare Differentiator

Highlights

- MiniMax has strengthened its position as a leading multimodal AI platform, supported by an integrated model portfolio spanning language, video, speech, music and agentic AI. We believe its born-global strategy, improving cost-performance and increasingly balanced To-C/To-B commercialisation provide a differentiated growth profile, while rapid open platform scaling and higher-margin enterprise monetisation support improving unit economics and a clearer path toward profitability.
- Initiate coverage with BUY; target price: HK\$470.00.

Analysis

- Strengthening its multimodal ecosystem.** MiniMax Group Inc (MiniMax) has built a leading AI portfolio. Its native multimodal architecture integrates video generation with foundation large language models, improving instruction following, semantic understanding and editing. We expect further gains in audio-video synchronisation, editing and multi-shot generation to lower production costs and broaden adoption.
- Diversified global commercialisation across To-B and To-C.** MiniMax has built a dual-engine monetisation model combining AI-native consumer applications with its enterprise open platform. In 1H26, Open Platform revenue surged 703.1% yoy to US\$73.9m, outpacing the 100.9% yoy growth in AI-native products to US\$42.6m, reflecting accelerating enterprise adoption. Overseas markets contributed 60.8% of revenue, with products and services reaching over 230 countries and regions. By August, enterprise customers and developers exceeded 2m, around 10x end-25 levels, while ARR surpassed US\$800m. Token consumption also surged 20x from January to July, underscoring rapidly scaling platform usage.
- Bolstering leadership in frontier models through cost-efficient innovation.** MiniMax sees competition shifting toward price-performance, inference cost and agentic capabilities, given low user switching costs. M3 Pro is scheduled for a later launch as the company developed the dual-model M3 family and expanded self-operated compute clusters only in 1Q26. Following M2 (Oct 25), M2.1 (Dec 25), M2.5 (Feb 26) and M3, we expect MiniMax's continued rapid iteration and reinforcement learning to narrow the gap with US state-of-the-art models, particularly in coding and agentic AI.
- Initiate coverage with BUY and target price of HK\$470.00.** We use DCF with 16.4% WACC and 3.0x terminal EV/Sales, implying 14x 2027F EV/Sales, a discount to pure-play AI peer Z.AI Co, but a premium to Kuaishou and Alibaba given its pure-play exposure to high-growth foundation models.

Key Financials

Year to 31 Dec (US\$m)	2024	2025	2026F	2027F	2028F
Net turnover	31	79	490	1,259	2,738
EBITDA	-285	-319	-681	-792	-615
Operating profit	-287	-321	-684	-796	-620
Net profit (rep./act.)	-465	-1,872	-646	-713	-502
Net profit (adj.)	-244	-251	-571	-571	-401
EPS (US\$ cent)	(224.8)	(230.9)	(182.1)	(163.9)	(112.8)
PE (x)	n.a.	n.a.	n.a.	n.a.	n.a.
P/B (x)	n.a.	n.a.	5.4	5.6	6.8
EV/EBITDA (x)	n.a.	n.a.	n.a.	n.a.	n.a.
Dividend yield (%)	0.0	0.0	0.0	0.0	0.0
Net margin (%)	(1,524.2)	(2,368.0)	(131.8)	(56.6)	(18.3)
Net debt/(cash) to equity (%)	(161.3)	(117.6)	(113.2)	(90.0)	(85.3)
ROE (%)	n.a.	n.a.	n.a.	(30.7)	(22.6)
Consensus net profit	-	-	(482.1)	(523.0)	(324.3)
UOBKH/Consensus (x)	-	-	0.82	0.91	0.77

Source: MiniMax, Bloomberg, UOB Kay Hian

Analyst(s)

Julia Pan

julianpan@uobkh.com

+86 21 5404 7225 ext 808

BUY

Share Price	HK\$303.00
Target Price	HK\$470.00
Upside	55.1%

Stock Data

Sector	Information Technology
Bloomberg ticker	100 HK
Shares issued (m)	268.1
Market cap (m)	105,818.3
Market cap (US\$m)	13,498.4
3-mth avg daily t'over (US\$m)	329.7

Price Performance (%)

52-week high/low			HK\$1,33.0/186.20	
1mth	3mth	6mth	1yr	YTD
54.6	(60.6)	(64.0)	n.a.	n.a.

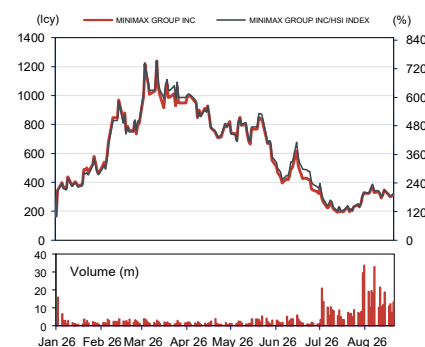
Major Shareholders (%)

Alisoft China Holding Ltd	14.79
Alibaba Group Holding Ltd	14.26
MINIMAX GENE LIMITED	7.79

Balance Sheet Metrics

FY26 NAV/Share (US\$)	7.37
FY26 Net Debt/Share (US\$)	8.04

Price Chart



Source: Bloomberg

Company Description

MiniMax Group Inc focuses on general-purpose large models and AI-native applications, including MiniMax, Hailuo AI, MiniMax Voice and Talkie/Xingye. The AI company serves users in China and overseas.

Investment Highlights

- **Strengthening competitive positioning through a comprehensive multimodal ecosystem.** MiniMax has built a leading multimodal AI portfolio spanning text (M3), video (H3/Hailuo), audio (speech), music and agentic AI, enabling it to address a broad range of high-growth use cases across content creation, voice interaction, coding and enterprise workflows. Its key differentiator is its native multimodal architecture, where video generation is tightly integrated with foundation large language model (LLM) capabilities, delivering stronger instruction following, semantic understanding and editable content than conventional pipelines. We expect future model upgrades to further improve audio-video synchronisation, editing and multi-shot generation, lowering content production costs and expanding adoption by professional creators and mass-market users.
- **Born-global strategy reinforcing overseas commercialisation...** Unlike many China-based AI peers that primarily focus on domestic deployment, MiniMax has pursued a global-first strategy since inception, with more than 61% of 1H26 revenue generated overseas and the US and Singapore emerging as key markets. Its AI-native products and open platform are now deployed across over 230 countries and regions, significantly broadening its addressable market, diversifying its monetisation channels and reducing its reliance on any single regional ecosystem.
- **...with dual-engine commercialisation across To-C and To-B.** MiniMax has built a diversified monetisation model spanning AI-native consumer applications and its enterprise open platform. Consumer products drive subscriptions, in-app purchases and advertising, while open platform benefits from growing API demand for multimodal models across AI agents, video, speech and coding. In 1H26, Open Platform and other AI-based enterprise services revenue surged 703.1% yoy to US\$73.9m, significantly outpacing the 100.9% yoy growth in AI-native product revenue to US\$42.6m. Its revenue contribution consequently increased to 63.4% from 30.3% in 1H25. By August, ARR had surpassed US\$800m, while token consumption surged 220x since January, underscoring rapidly scaling platform usage and monetisation.
- **Bolstering leadership in frontier models through cost-efficient innovation, ...** MiniMax sees competition shifting toward price-performance, inference cost and agentic capabilities, given low user switching costs. M3 Pro is scheduled for a later launch as the company developed the dual-model M3 family and expanded self-operated compute clusters only in 1Q26. Following M3, we expect M3.1 and the 3T-parameter M3 Pro to drive the next phase of rapid iteration, with continued reinforcement learning and long-horizon post-training aimed at narrowing the gap with US frontier models, particularly in coding and agentic AI.
- **...with cost-performance advantage supporting long-term ecosystem expansion.** MiniMax's differentiation is increasingly driven by its ability to deliver frontier coding and agentic capabilities at low inference cost. Its latest M3 model combines a 1m-token context window, native multimodality and a sparse-attention architecture that reduces per-token compute by about 95% for a 1m context vs the previous generation, while delivering more than 9x faster prefilling and over 15x faster decoding. We believe this combination of long-context capability, multimodal intelligence and inference efficiency strengthens MiniMax's positioning in high-value developer, coding and agentic workloads, supporting open platform adoption and long-term ecosystem formation.
- **Well positioned to capture global AI demand across LLMs, multimodal AI and digital labour.** We continue to view MiniMax as one of the best-positioned Chinese AI model companies, supported by its comprehensive multimodal portfolio, strong commercialisation capabilities, low per-token inference costs and high organisational efficiency. As China enters the next wave of 3T-parameter foundation models, we believe pricing, speed and cost

Comprehensive multimodal ecosystem

Global-first strategy broadens overseas monetisation

To-C and To-B dual-engine

Rapid frontier-model iteration

Cost-performance as a key differentiator

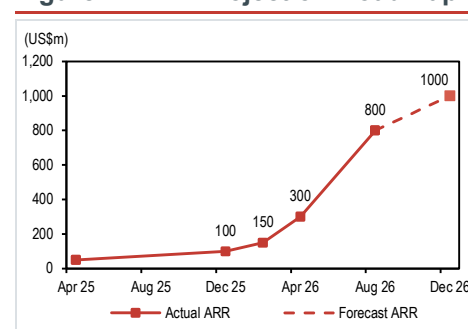
efficiency will become increasingly important differentiators alongside frontier performance, further supporting MiniMax's expansion across text/coding, multimodal generation and agentic co-worker applications.

- **Catalysts** include: a) Rapid model iteration, with M3.1 and M3 Pro further closing the gap with US frontier models at a fraction of the cost; b) broader adoption of AI agents and digital workers, enabled by an estimated operating cost of just US\$1/hour at 100 tokens/second; and c) continued multimodal innovation, with H3's unified architecture advancing AI video generation and further model upgrades expected.

Valuation

- **We initiate coverage with a BUY rating and a target price of HK\$470.00.** We adopt a DCF-based valuation methodology, assuming WACC of 16.4% and terminal EV/Sales of 3.0x. Our target price implies 36x 2026F EV/Sales and 14x 2027F EV/Sales, representing a discount to pure-play AI peer Z.AI Co (Z.ai), but a premium to diversified technology companies such as Alibaba and Tencent. We believe the discount to Z.ai is justified by MiniMax's lower pricing power and greater exposure to commoditised models amid intensifying price competition. Nevertheless, we believe MiniMax warrants a premium to diversified technology peers given its pure-play exposure to high-growth foundation models, multimodal AI and agentic applications. We forecast a three-year revenue CAGR of 123% over 2026-29, and we expect the company to turn profitable in 2029.
- **Key investment ideas** are as follows: a) MiniMax's integrated multimodal capabilities remain relatively scarce and broaden its addressable market beyond text-based LLMs; b) H3's strong performance, competitive pricing and open-weight release validate MiniMax's position in frontier multimodal models; and c) its born-global strategy supports faster overseas adoption, diversified monetisation and access to a larger international market.

Figure 1: ARR Projection Roadmap



Source: Industry data, UOB Kay Hian

Figure 2: DCF FCFF Projection

Year to 31 Dec (US\$m)	2026F	2027F	2028F	2029F	2030F	2031F
Free cash flow	(710)	(585)	(435)	150	913	2,321
EBIT * (1-t)	(581)	(677)	(527)	57	820	2,228
+ D&A	3	4	5	5	6	7
+ Capex & investments	(1)	(2)	(2)	(3)	(3)	(4)
+ (Inc) / dec in NWC	(130)	90	90	90	90	90
Proportionate FCF	(710)	(585)	(435)	150	913	2,321
Year	0.35	1.35	2.35	3.35	4.35	5.35
WACC	16%	16%	16%	16%	16%	16%
Discount factor	1.05	1.23	1.43	1.66	1.93	2.25
NPV of FCF	(674)	(477)	(305)	90	472	1,032
Base case WACC calculation:			DCF valuation (US\$m):			
Risk free rate	1.8%					Total NPV of forecast 813
Equity risk premium	10.2%					Terminal Sales 13,604
Equity beta	1.5					Terminal EV/Sales multiple 3 x
Cost of common equity	17.2%					Terminal value 35,030
After tax cost of debt	2.6%					NPV of terminal value 15,577
Cost of Debt	3.0%					HKD\$/RMB 7.8
Tax rate	15.0%					
Debt/Total capital (steady state):						Enterprise value 16,390
% common equity	94%					TV: EV 95.0%
% of debt	6%					Net (cash) debt (2026F) (2,522)
WACC calculated	16.4%					Hypothetical DCF-based fair value:
Terminal FCFF =	2,321					Equity value (HK\$m) 147,518
Implied perpetual growth rate	9.7%					Per share (HK\$) 470

Source: UOB Kay Hian

Business Outlook

Comprehensive multimodal ecosystem driving commercialisation. MiniMax’s latest multimodal portfolio spans H3 (video), Speech 2.8, Music 3.0 and Image-01, monetised through APIs and subscriptions. H3 supports multimodal inputs and native stereo-audio video generation at up to 2K resolution and 15 seconds, with API pricing of US\$0.13/sec at 2K and US\$0.08/sec at 768p. This compares with US\$0.40/sec for Google Veo 3.1 Standard at 720p/1080p and US\$0.30/sec for Sora 2 Pro at 720p, highlighting H3’s competitive cost-performance. Meanwhile, Speech 2.8 is priced at US\$60/US\$100 per 1m characters for Turbo/HD, while Image-01 costs US\$0.0035/image.

Management believes native multimodal generation has pushed video AI to an inflection point, with usable output rates reaching 80-90% through integrated text, image, video and audio generation, positioning MiniMax well for creator-economy and e-commerce applications. This innovation is underpinned by an agile R&D team, with 73.8% of its 385 employees in R&D as of Sep 25, the broader team averaging below 30 years old, and core R&D members coming from leading AI and technology companies. Its flat organisational structure, with no more than three reporting layers below the CEO, and project-based teams facilitate rapid execution and continuous model iteration across AI modalities.

Global-first strategy and multimodal leadership support long-term growth. MiniMax has pursued an overseas-first strategy since inception, with over 60% of 1H26 revenue generated outside China and AI-native products and open platform services deployed across more than 230 countries and regions. We view MiniMax as one of the best-positioned Chinese AI model companies to capture global growth across text/coding, multimodal AI and agentic digital labour, supported by its native multimodal platform, strong commercialisation, low per-token costs and high organisational efficiency.

Commercialisation is accelerating, increasingly led by enterprise demand. ARR exceeded US\$800m in Aug 26, with To-B contributing over 80%, while enterprise customers and developers surpassed 2m, around 10x end-25 levels. In 1H26, Open Platform and other AI-based enterprise services already accounted for 63.4% of revenue, overtaking AI-native products at 36.6%. Management remains highly confident in achieving its US\$1b year-end ARR target, supported by rapid adoption of M3 across coding, agents and productivity use cases, although compute availability and model-release timing remain key constraints.

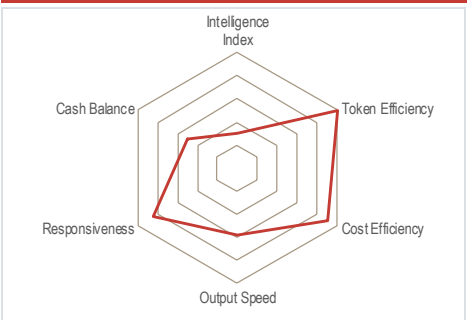
Rapid model iteration supported by strong cost-performance advantages. MiniMax has accelerated its language-model roadmap from M2.7 to M3, with M3.1 and the 3T-parameter M3 Pro next in the pipeline. M3 has emerged as a key commercial driver, with July token consumption reaching 20x January levels, supported by improving adoption across coding, agents and productivity workloads. Following the launch, inference optimisation lifted output speed from 20 tokens/sec to nearly 100 tokens/sec within a month, while steady-state M3 API gross margin has already surpassed M2 despite roughly doubling model parameters, with per-token inference cost falling below M2. M3 also remains highly competitively priced at roughly one-tenth of higher-priced leading models, reinforcing MiniMax’s cost-performance advantage. We expect M3.1 and M3 Pro to further improve model intelligence and unit economics, supporting broader enterprise adoption.

Product breadth supports scalable monetisation

Global-first strategy is expanding TAM

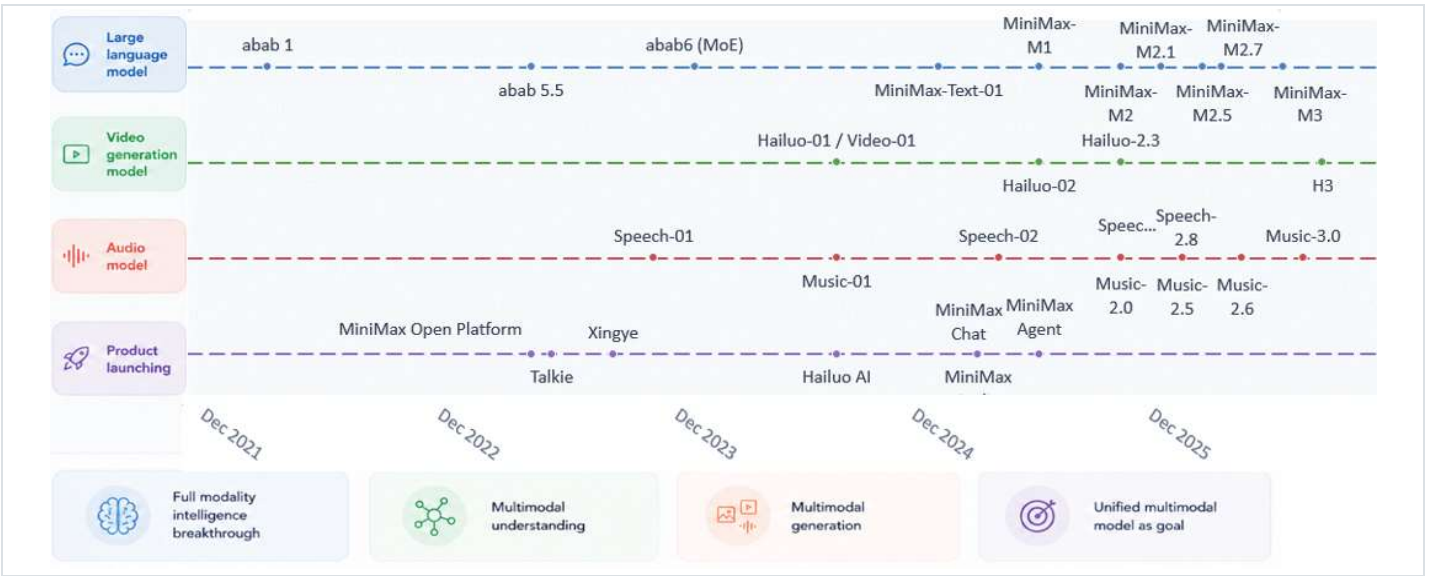
Enterprise adoption is reshaping the revenue mix

Figure 3: Multidimensional Performance Of M3



Source: Industry data, UOB Kay Hian

Figure 4: Production Roadmap



Source: MiniMax, industry data, UOB Kay Hian

AI video remains a growth driver, although long-term differentiation may narrow. Video generation offers significant total addressable market (TAM) expansion by lowering the cost, time and technical barriers to content creation, potentially broadening the creator base well beyond today's more than 200m global creators. We expect AI video to remain an important ARR contributor for MiniMax, supported by H3's strong benchmark performance and pricing at around one-third of Seedance 2.0's, particularly in overseas markets where monetisation opportunities are greater.

AI video is still structurally attractive, though differentiation could diminish over time

Over the longer term, AI video may prove more susceptible to commoditisation than agentic coding. As visual quality converges toward “good enough”, competitive differentiation is likely to rely increasingly on price, workflow integration and user experience rather than model quality alone. Compared with agentic AI, AI video also sees a smaller long-term TAM, suggesting MiniMax will need to continue differentiating its products from well-funded rivals including Google Veo, ByteDance Seedance and Kuaishou Kling.

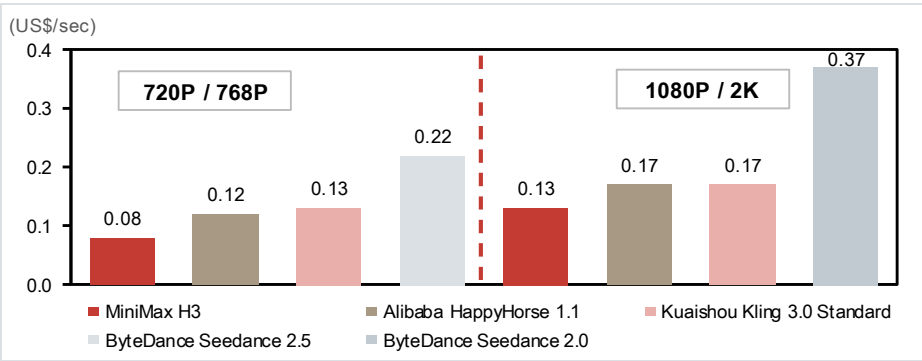
Figure 5: Selected Video Generation Models Comparison

Selected model	MiniMax H3	Seedance 2.0 720p	HappyHorse 1.1	Kling VIDEO 3.0	Gemini Omni Flash
Company	MiniMax	ByteDance	Alibaba	Kuaishou / KlingAI	Google
Release date	Jul 26	Feb 26	Jun 26	Feb 26	May 26
AA T2V Elo (with audio)	1238	1224	1148	1110	1243
AA I2V Elo (with audio)	1187	1196	1107	1074	1194
Price (US\$/min)	7.8	9.07	9.9	20.16	6
Maximum output	Up to 2K; up to 15s	720p; up to 15s	Up to 1080p; 3-15s	Up to 4K; 3-15 s	720p; up to 10s
Input/audio	Text, image, video and audio inputs; native stereo audio	Text, image, audio and video inputs; native dual-channel audio	Text, first-frame image and reference-image inputs; native generated audio	Text, image, start/end frames and image/video element references; native audio with multilingual dialogue	Text, image, audio and video inputs; native audio output; conversational video editing
Main highlights	Frontier open-weight omni-modal model with unified multimodal context, instruction-based editing and video-to-video motion transfer.	Strong multimodal referencing, complex-motion stability, multi-shot generation, editing control and audio-visual synchronisation. Seedance 2.5 extends output to 30s.	Strong reference-image consistency, motion expressiveness, instruction following, visual detail and audio-visual synchronisation; suited to advertising, brand content and gaming cinematics.	Multi-shot storyboarding, strong subject consistency, native-level text rendering, and precise multi-character dialogue with multilingual, dialect and accent support.	Strong instruction following, real-world knowledge and physics, multimodal reference-to-video, and iterative conversational editing.

Source: Industry data, respective companies, UOB Kay Hian

H3 strengthens MiniMax’s multimodal and video leadership with a strong cost-performance proposition. Launched on 31 July alongside ByteDance’s Seedance 2.5, H3 is a unified multimodal model integrating text, image, audio and video, capable of generating up to 15-second 2K videos and ranked #1 on Artificial Analysis’s video-editing-with-audio benchmark list. Its open-weight release should support enterprise adoption, developer ecosystem expansion and domestic-chip optimisation, while its Contextual Omni Representation, H3-VAE (4x sequence-length efficiency), H3-Omni Transformer (30% higher training throughput) and In-context Regeneration improve efficiency and editing capabilities. At Rmb0.80/second, H3 costs around one-third of flagship peers, less than one-third of mainstream models for 2K generation, and less than half for 768P, offering a compelling cost-performance advantage. We expect H3 to support broader commercialisation, with Hailuo’s revenue reaching US\$167m by 2027, although intensifying competition from ByteDance’s Seedance 2.5 and Kuaishou’s Kling raises the bar for sustained ARR growth.

Figure 6: Video Model Pricing By Service Provider



Source: Industry data, respective companies, UOB Kay Hian

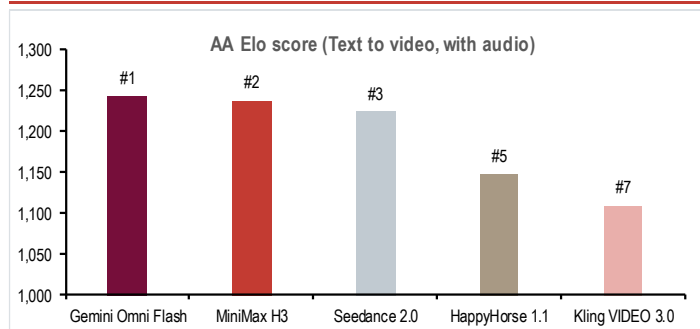
Architecture-driven cost and workflow advantages support commercial adoptions. H3 is priced at around one-third of peers at 2K and one-half at 768P, enabled by innovations including H3-VAE (4x sequence-length efficiency), the H3-Omni Transformer and in-context regeneration. It supports 15-second video generation at 24 frames per second, native stereo audio, prompts of up to 7,000 characters, and up to 12 multimodal reference files (9 images, 3 videos and 3 audio clips), significantly improving character, motion and style consistency. Enhanced instruction following, video editing and motion transfer also reduce regeneration costs, with estimated production costs of only Rmb10 per advertising video and Rmb20 per short-drama episode on LibTV, supporting broader enterprise and creator adoption.

Cost-efficient generation improves commercial scalability

H3 marks MiniMax's return to the frontier of AI video generation. We view H3 as a meaningful step-up in multimodal capabilities, combining open weights, unified multimodal generation and materially lower pricing. Rather than competing solely on visual quality, H3 targets production-ready creative workflows by jointly understanding and generating text, images, video and audio within a single model. Early ecosystem adoption has been strong, with over 24m downloads within three weeks of its open-weight release and more than 300 derivative models developed, highlighting growing developer traction and supporting broader commercialisation of its video-generation capabilities.

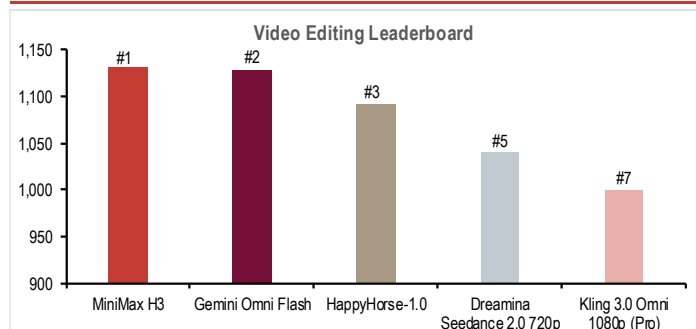
H3 marks a renewed competitive push in AI video

Figure 7: Artificial Analysis Performance Scores (T2V)



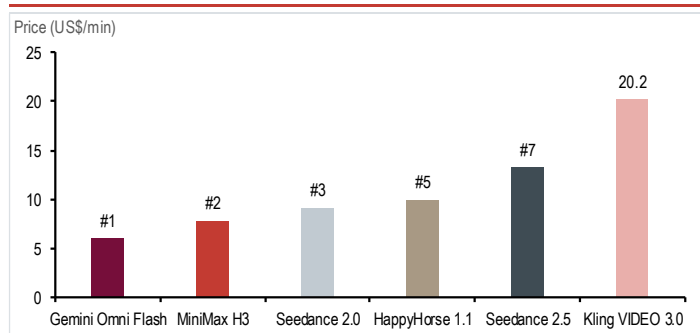
Source: Artificial analysis, UOB Kay Hian

Figure 8: Artificial Analysis Video Editing Leaderboard



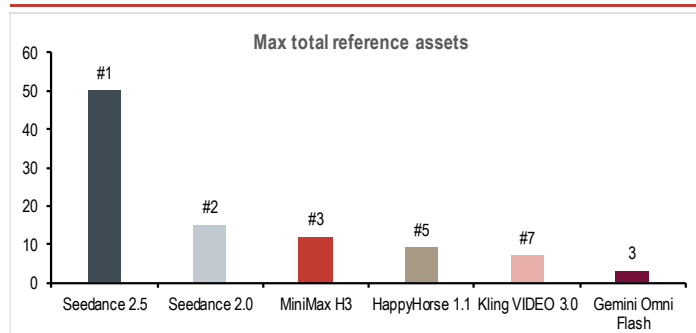
Source: Artificial analysis, UOB Kay Hian

Figure 9: Cost Of Selected Models



Source: Respective companies, UOB Kay Hian

Figure 10: Max Number Of Reference Input



Source: Respective companies, UOB Kay Hian

Multimodal capabilities remain a key differentiator. MiniMax continues to enhance M3's native multimodal capabilities, which will be further integrated into the upcoming M3 Pro. Beyond video generation, the imminent H3 launch introduces a unified omni-model architecture capable of understanding and generating text, images, video, audio and music in a single framework. H3 could reshape the multimodal landscape by offering a compelling alternative to existing AI video-generation models while expanding into broader multimodal creation and editing use cases.

Native multimodality reinforces MiniMax's model advantage

Multimodal and agentic AI remain key roadmap priorities. MiniMax also plans to advance multimodal integration, RL, attention architectures and algorithms, focusing on coding, agents and generative media. Management expects more efficient architectures to improve token economics and better learning algorithms to raise model intelligence, supporting stronger pricing as AI ROI improves. Its multimodal ecosystem has helped drive average monthly active users (MAU) nearly 9x from 3.1m in 2023 to 27.6m in 9M25.

Multimodal and agentic AI remain central to capability expansion

M3 underperformed expectations despite incremental progress. Launched on 31 May, M3 combines native multimodality, coding and the MiniMax Sparse Attention architecture, improving over M2.7 on public benchmarks. Its MoE architecture has 428B total/23B active parameters and a 1m-token context window, alongside improved inference efficiency and deployment flexibility. However, MiniMax subsequently implemented a permanent 50% token price cut to drive adoption. While DeepSeek has recently raised V4-Pro-0813 pricing, narrowing its previous pricing advantage, intense competition among Chinese frontier models continues to constrain pricing power. Despite OpenRouter usage growth, the lower pricing limited the ARR uplift versus GLM-5.3, while Tencent's Hy3 offers stronger reasoning at lower cost. Management acknowledged M3 fell short of expectations, with higher hallucination and error rates than leading peers.

M3 strengthens the model stack, though monetisation remains under pressure

Figure 11: Top Apps Consuming MiniMax M3 API On Open Router Platform



Source: Open Router, UOB Kay Hian

Pricing discipline and inference efficiency are key to margin scalability. As MiniMax shifts from capability validation to scaled monetisation, workload intensity will increasingly matter more than user growth for unit economics. M3’s broader range of use cases across coding, agents, multimodality and 1m-token context could expand TAM but also raise compute consumption under fixed-price subscriptions. MiniMax’s hybrid pay-as-you-go API + Token Plans with five-hour rolling and weekly limits provide some cost control, but overly generous quotas could pressure margins while tighter limits may hurt user experience. We therefore see pricing discipline, inference efficiency and usage control as key margin drivers.

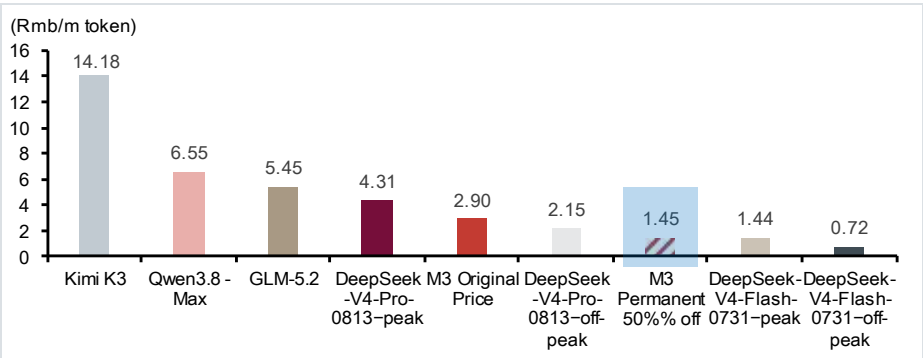
Disciplined pricing strengthens margin scalability

Figure 12: M3 Cutting Prices

Model	Input	Output	Prompt caching Read
MiniMax-M3 ≤ 512k input tokens Permanent 50% off	\$0.60 \$0.30 / M tokens	\$2.40 \$1.20 / M tokens	\$0.12 \$0.06 / M tokens
MiniMax-M3 > 512k input tokens* Permanent 50% off	\$1.20 \$0.60 / M tokens	\$4.80 \$2.40 / M tokens	\$0.24 \$0.12 / M tokens

Source: MiniMax, UOB Kay Hian

Figure 13: Blended Prices For Selected Models



Source: Respective companies, UOB Kay Hian

M3 Pro to be the key catalyst for restoring frontier positioning. Ahead of M3 Pro, M3.1 will serve as an interim upgrade, building on the M3 430B base with more extensive pre-training and scaled post-training to improve task completion, stability, inference efficiency and agent compatibility. Management then plans to launch M3 Pro at 3T parameters, representing a significant step-up in model scale and targeting a higher intelligence ceiling through larger-scale reinforcement learning and long-horizon task training. Despite the substantial increase in parameter count, management expects architectural and inference optimisation to limit the increase in per-token costs. We view M3 Pro as the key catalyst for narrowing the frontier-intelligence gap while preserving MiniMax’s cost-performance advantage.

Next large-scale model launch could be a key catalyst

HIGH-MARGIN OPEN PLATFORM SUPPORTS PROFITABILITY

Open platform scales enterprise adoption and high-margin multimodal APIs. MiniMax's Open Platform provides APIs across text, speech, video, image and agentic workloads, with Open Platform and other AI-based enterprise services revenue surging 703.1% yoy to US\$73.9m in 1H26, accounting for 63.4% of total revenue, vs 30.3% in 1H25. Growth was driven by higher paid-user and enterprise adoption, rising API call volumes and rapid Token Plan penetration. By August, enterprise customers and developers exceeded 2m, around 10x end-25 levels. Current offerings include Speech 2.8-HD at US\$100/1m characters and H3 video generation, alongside M3 for text, coding and agentic workloads.

Open platform supports high-margin growth

High-margin multimodal APIs support profitability. Overseas pricing is modestly higher than in China but is broadly offset by higher inference costs, resulting in similar gross profit levels. MiniMax's open platform achieved a 69% gross margin in 9M25, well above Z.ai's 0% cloud deployment margin in 1H25, despite lower pricing for its M2.1 text model. We attribute this to multimodal models contributing about 75% of API revenue, with higher-value offerings such as Speech-2.6-HD (US\$100/1m characters) and Hailuo-2.3 (US\$0.56 per 768P, 10-second video). In comparison, Anthropic projected a 40% gross margin in 2025 with a projected 86% of revenue from APIs, vs OpenAI's estimated 46%.

Higher-value API mix strengthens unit economics

TAM is expanding beyond coding into vertical AI. MiniMax is targeting knowledge-intensive applications across the legal, healthcare, security and financial services sectors, supported by vertical-specific training data and workflow optimisation. The speech model has already served 2,000+ enterprise customers since Nov 23, while adoption is expanding across coding, agents, video generation, advertising, e-commerce and voice. We estimate MiniMax's open platform gross margin at 66% in 2025, materially above Z.ai's 19% gross margin for cloud-based deployment, highlighting MiniMax's stronger unit economics in enterprise API monetisation. We believe this is supported by its higher-value multimodal revenue mix, particularly video and speech APIs, alongside improving model and inference efficiency. Continued scaling of open platform revenue should support further group-level margin expansion and strengthen MiniMax's path toward profitability.

Vertical AI supports enterprise TAM growth

Cost-efficient innovation and multimodal leadership underpin MiniMax's competitive edge. We estimate MiniMax invested only about US\$420m in training during 2022-25, supported by early adoption of MoE and linear attention to improve inference efficiency and long-context processing, while later reverting to full attention on M2 to enhance reasoning and coding. This supported healthy To-B API gross margins despite being priced at about 1/10th of leading peers. Its native multimodal strategy has also driven strong adoption, with average MAU rising nearly 9x from 3.1m in 2023 to 27.6m in 9M25, reinforcing its platform and commercialisation potential.

Cost-efficient R&D supports differentiation

Enterprise monetisation is inflecting. MiniMax's commercialisation momentum accelerated further in 1H26, with Open Platform and other AI-based enterprise services revenue surging 703.1% yoy to US\$73.9m, accounting for 63.4% of group revenue and overtaking AI-native products. By August, ARR exceeded US\$800m, with To-B contributing over 80%, while enterprise customers and developers surpassed 2m, around 10x end-25 levels. Management remains highly confident in achieving its US\$1b year-end ARR target, supported by rapid M3 adoption across coding, agents and broader productivity use cases, with H3 providing incremental upside.

Pricing pressure and litigation risks remain key overhangs. While MiniMax's cost-performance advantage supports developer and enterprise adoption, its pricing power may come under increasing pressure as competition shifts increasingly toward price and cost efficiency. Aggressive pricing by Chinese peers such as Xiaomi's MiMo has intensified LLM API competition, while OpenAI's recent 80% price cut for GPT-5.6 Luna and 20% cut for GPT-5.6 Terra suggest that pricing pressure is also broadening across global frontier-model providers. We believe continued industry-wide price compression could limit MiniMax's ability to monetise solely through inference-cost advantages and increase the importance of differentiated model capabilities and higher-value multimodal applications. In addition, ongoing copyright litigation related to alleged use of copyrighted materials in AI model training could create financial liabilities, regulatory uncertainty and reputational risks, weighing on MiniMax's longer-term commercialisation outlook.

Pricing pressure remains a key risk

High R&D intensity requires funding. Adjusted operating losses widened from US\$98m in 2023 to US\$280m in 2024 and US\$297m in 2025, while R&D spending reached US\$253m in 2025, about 3.2x revenue. Since its Hong Kong listing, MiniMax has rolled out around 10 major model releases and upgrades across language, speech, music and video, while its enterprise and developer customer base expanded to over 2m by August, around 10x end-25 levels. Rising training and inference demand has accelerated infrastructure spending, with around 77% of IPO proceeds allocated to AI infrastructure and model R&D already deployed within six months, underscoring continued reliance on external funding.

Rapid model iteration sustains elevated capital needs

Strong funding position supports continued long-term expansion. MiniMax raised about US\$1.56b across seven private rounds before its Jan 26 IPO, followed by HK\$5.29b in net IPO proceeds. In Jul 26, it further raised about HK\$16.0b through a 35.6m share placement at HK\$268/share and HK\$6.5b of zero-coupon convertible bonds due 2027, with the transaction about 7x covered, highlighting strong institutional demand.

Strong funding extends runway

MiniMax has also commenced preparations for a potential renminbi-share listing on the SSE STAR Market, although the timing remains uncertain and any offering is subject to market and regulatory approvals. The latest round of fundraising provides near-term liquidity to support foundation model R&D, compute-capacity expansion, commercialisation and overseas growth ahead of any potential A-share financing. While the stronger capital base reduces near-term funding constraints, external financing is likely to remain important until operational breakeven, given rising cloud-service, training and inference costs and the capital intensity of scaling frontier models.

Profit & Loss

Year to 31 Dec (US\$m)	2025	2026F	2027F	2028F
Net turnover	79.0	489.8	1,258.5	2,737.9
EBITDA	(318.7)	(681.1)	(792.0)	(615.4)
Deprec. & amort.	2.7	2.9	3.9	4.8
EBIT	(321.4)	(684.1)	(795.9)	(620.2)
Total other non-operating income	(1,549.5)	38.4	83.7	118.5
Associate contributions	0.0	0.0	0.0	0.0
Net interest income/(expense)	(0.7)	(0.2)	(0.3)	(0.3)
Pre-tax profit	(1,871.6)	(645.8)	(712.5)	(501.9)
Tax	0.0	0.0	0.0	0.0
Minorities	0.0	0.0	0.0	0.0
Net profit	(1,871.6)	(645.8)	(712.5)	(501.9)
Net profit (adj.)	(250.9)	(571.2)	(570.5)	(400.5)

Cash Flow

Year to 31 Dec (US\$m)	2025	2026F	2027F	2028F
Operating	(279.6)	(703.6)	(341.4)	(449.0)
Profit to the year	(1,871.6)	(645.8)	(712.5)	(501.9)
Tax	0.0	0.0	0.0	0.0
Deprec. & amort.	2.7	2.9	3.9	4.8
Associates	0.0	0.0	0.0	0.0
Working capital changes	6.3	(129.8)	224.9	(53.6)
Non-cash items	0.0	0.0	0.0	0.0
Other operating cashflows	1,583.0	69.0	142.3	101.7
Investing	72.3	128.8	(1.8)	(2.1)
Capex (growth)	(1.2)	(1.5)	(1.8)	(2.1)
Investment	76.2	76.8	76.8	76.8
Others	(2.7)	53.5	(76.8)	(76.8)
Financing	423.3	2,712.9	(1.4)	(1.5)
Dividend payments	0.0	0.0	0.0	0.0
Proceeds from borrowings	65.6	0.0	0.0	0.0
Loan repayment	(2.2)	(1.1)	(1.2)	(1.3)
Others/interest paid	360.0	2,714.0	(0.3)	(0.3)
Net cash inflow (outflow)	216.0	2,138.1	(344.7)	(452.7)
Beginning cash & cash equivalent	288.9	507.6	2,645.7	2,301.1
Changes due to forex impact	2.7	0.0	0.0	0.0
Ending cash & cash equivalent	507.6	2,645.7	2,301.1	1,848.4

Balance Sheet

Year to 31 Dec (US\$m)	2025	2026F	2027F	2028F
Fixed assets	1.6	74.9	2.8	3.4
Other LT assets	79.5	181.5	93.3	108.4
Cash/ST investment	521.4	2,659.8	2,315.1	1,862.4
Other current assets	486.0	513.3	617.6	1,013.1
Total assets	1,088.4	3,429.5	3,028.7	2,987.3
ST debt	3,634.3	135.6	135.9	136.9
Other current liabilities	99.3	1,060.7	468.3	825.0
LT debt	0.6	1.8	3.6	5.0
Other LT liabilities	2.3	3.2	3.2	3.2
Shareholders' equity	(2,648.2)	2,228.2	2,417.7	2,017.1
Minority interest	0.0	0.0	0.0	0.0
Total liabilities & equity	1,088.4	3,429.5	3,028.7	2,987.3

Key Metric

Year to 31 Dec (%)	2025	2026F	2027F	2028F
Profitability				
EBITDA margin	(403.3)	(139.1)	(62.9)	(22.5)
Pre-tax margin	(2,368.0)	(131.8)	(56.6)	(18.3)
Net margin	(2,368.0)	(131.8)	(56.6)	(18.3)
ROA	(187.3)	(28.6)	(22.1)	(16.7)
ROE	n.a.	n.a.	(30.7)	(22.6)
Growth				
Turnover	158.9	519.8	156.9	117.5
EBITDA	n.a.	n.a.	n.a.	n.a.
Pre-tax profit	n.a.	n.a.	n.a.	n.a.
Net profit	n.a.	n.a.	n.a.	n.a.
Net profit (adj.)	n.a.	n.a.	n.a.	n.a.
EPS	n.a.	n.a.	n.a.	n.a.
Leverage				
Debt to total capital	368.4	5.8	5.5	6.6
Debt to equity	n.a.	6.2	5.8	7.0
Net debt/(cash) to equity	n.a.	(113.2)	(90.0)	(85.3)
Interest cover (x)	n.m.	n.m.	n.m.	n.m.

Appendix I: Risk Factors

- a) **Overseas regulatory and geopolitical risks.** Overseas customers contributed 61% of 1H26 revenue, exposing MiniMax to regulatory, data privacy, export control and geopolitical risks. Actions against DeepSeek in Italy and South Korea highlight the potential for restrictions to constrain user acquisition, enterprise adoption and monetisation.
- b) **Third-party compute dependence risks.** MiniMax relies heavily on external cloud providers; its two largest suppliers accounted for about 24% and 23% of purchases in 9M25, while its Alibaba Cloud agreement caps annual transactions at US\$115m/US\$125m/US\$135m for 2026-28 respectively. Rising usage or compute shortages could pressure capacity and inference margins.
- c) **Copyright and IP risks.** Disney, Universal and Warner Bros Discovery sued MiniMax in Sep 25 over Hailuo AI, with a US federal judge rejecting its dismissal motion in May 26. An adverse outcome could not only increase damages, licensing and compliance costs, but also constrain model development and content generation.
- d) **Intensifying price competition risks.** Industry API pricing remains under pressure, with OpenAI cutting GPT-5.6 Luna/Terra pricing by 80%/20%, respectively, and subsequently reducing GPT-5.6 Sol pricing by over 20%, alongside aggressive pricing by Chinese peers such as Xiaomi. Further industry-wide price compression could weaken MiniMax's pricing power and limit margin expansion despite its inference-cost advantage.
- e) **Unlock and free float risks.** Approximately 26.2% of MiniMax's shares are expected to become eligible for trading in Jan 27, potentially increasing near-term share supply and share-price volatility. The technical risk was evident around the previous unlock, with MiniMax's share price retracing about 60% from an intraday high of HK\$640.50 on 23 Jun 26 to HK\$256.00 on 16 Jul 26, one week after the 9 July lock-up expiry. Its planned A-share listing could further increase shares outstanding and dilution risk if new shares are issued. We view these as primarily technical risks, while the key medium-term focus remains on sustained revenue growth, gross margin expansion and progress toward profitability, which should help mitigate potential selling pressure and dilution concerns.

Appendix II: Company Overview

Founded in 2022, MiniMax is a leading AI foundation model company focused on advancing artificial general intelligence (AGI) through its mission to “co-create intelligence with everyone”. The company develops proprietary multimodal models spanning text, audio, image, video, and music generation, with capabilities in coding, agentic AI, and ultra-long context processing. MiniMax was also an early adopter of the MoE architecture and developed an innovative linear attention mechanism, helping to significantly reduce computational resource requirements and improve model efficiency. Its core products include MiniMax Agent, Hailuo AI, MiniMax Audio, Talkie, and an open API platform serving enterprise and developer customers globally.

As of 1H26, MiniMax’s AI-native products had cumulatively served over 300m users globally, while its products and services had been deployed across more than 230 countries and regions. Its enterprise and developer customer base had also expanded to over 2m, approximately 10x the level at end-25.

Development Milestones



Source: MiniMax, UOB Kay Hian

BUSINESS MODEL AND MONETISATION

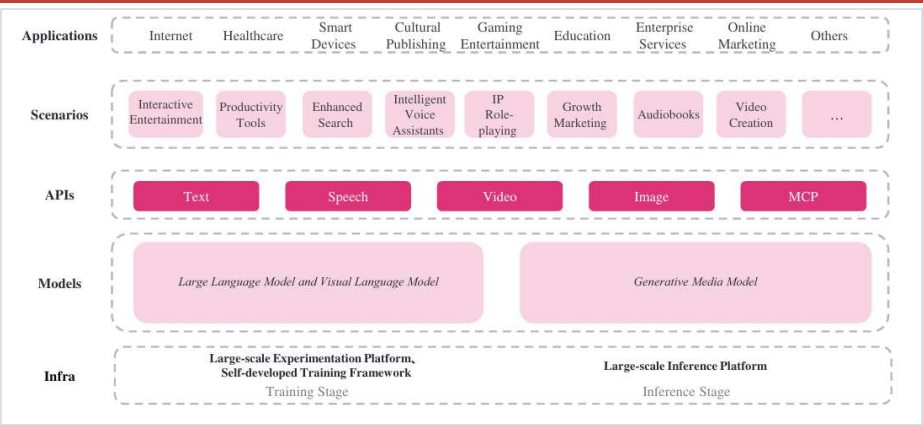
MiniMax monetises its foundation model capabilities through two main channels: AI-native consumer products and open platform services for enterprises and developers.

For individual users, MiniMax generates revenue from products that primarily adopt a freemium model, allowing users to access basic functions for free while charging for higher usage limits, additional credits and premium features. Monetisation is mainly generated through recurring monthly, quarterly or annual subscriptions, as well as standalone purchases of credits or virtual items. Subscription prices vary according to usage allowances, technical capabilities and feature availability.

For enterprise and developer customers, MiniMax monetises its models through the open platform, which enables customers to integrate its language, video and speech models into their own applications and workflows through APIs. Pricing is primarily usage-based but varies by modality: language-model APIs are charged according to input and output token consumption, while video-generation customers purchase prepaid packages containing video points, with each generated clip deducting a specified number of points based on the model, resolution and duration. Speech-generation APIs are offered through subscription packages with bundled monthly credits, with higher tiers providing larger credit allowances, additional voice slots and higher requests-per-minute capacity; actual credit consumption depends on the volume of text processed and the selected model. Accordingly, enterprise revenue is primarily driven by customer usage, model mix and package-tier selection.

Overall, MiniMax operates a foundation-model-as-a-service and AI-application business model. Consumer applications provide direct subscription and credit revenue, while the open platform monetises model usage by third-party enterprises and developers.

Full-Stack AI Capabilities From Infrastructure To Applications



Source: MiniMax, UOB Kay Hian

MiniMax Agent is the company’s AI agent and productivity assistant platform, designed to support complex task execution, reasoning and workflow automation through LLM capabilities. The product focuses on agentic AI applications, enabling users to perform multi-step tasks, content generation, and AI-assisted productivity workflows.

Hailuo AI. This is MiniMax’s AI video creation platform, enabling users to generate and edit videos using text, image, video and audio inputs. The platform integrates MiniMax’s latest video models, including H3, and targets content creation use cases such as short-form video, advertising creatives and AI-assisted visual storytelling.

MiniMax Audio is the company’s AI voice and speech generation platform, providing capabilities such as text-to-speech, voice synthesis, and AI audio generation. The platform supports both consumer and enterprise applications across entertainment, digital assistants, and interactive AI experiences.

Talkie is MiniMax’s AI-native social and character interaction application, enabling users to engage with AI-powered virtual characters through conversational and roleplay-based interactions. The platform focuses on AI companionship, entertainment, and highly personalised user engagement experience.

Monetisation Model

Product Name	Target users	Key underlying models	Primary use cases	Monetisation model
MiniMax	To C	Text, video, speech, and music	Intelligence agent application	Freemium, subscriptions, token-based in-app purchases.
Hailuo Ai	To C	Video	Flagship visual generation platform	Freemium, subscriptions, token-based in-app purchases.
MiniMax Audio	To C	Speech and music	Audio generation tool	Freemium, subscriptions, token-based in-app purchases.
Talkie/Xingye	To C	Text, video, speech, and music	Multi-model entertainment platform	Freemium, subscriptions, online marketing service, in-app purchases.
Open Platform	To B	Text, video, speech, and music	API platform	Freemium, token-based API billing, subscription (coding plan), enterprise licensing.

Source: MiniMax, industry data, UOB Kay Hian

BUSINESS SEGMENT

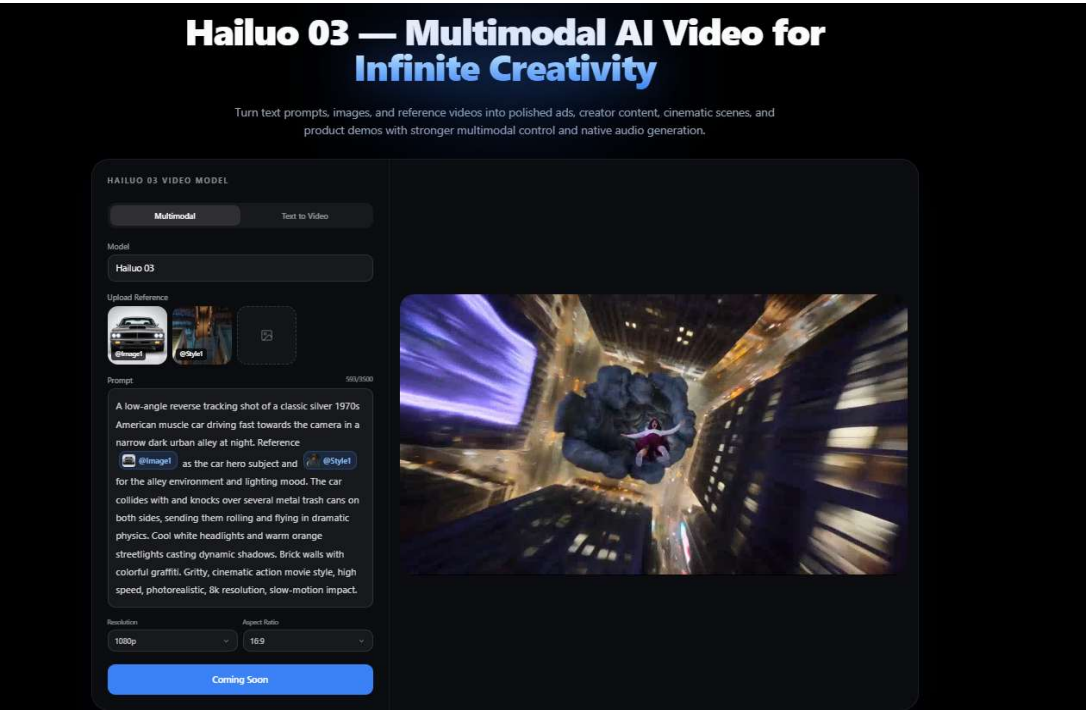
MiniMax operates a dual-engine business model comprising AI-native products and enterprise-focused AI services. Leveraging its proprietary multimodal foundation models, the company serves both consumer and enterprise customers through a combination of AI-powered products and model-as-a-service (MaaS) offerings, enabling the commercialisation of its AI capabilities across a broad range of use cases.

AI-native products. MiniMax develops a portfolio of AI-native applications built on its proprietary multimodal foundation models, spanning text, audio, image, video, and music generation. Its core products include Talkie, Hailuo AI, MiniMax Agent, and MiniMax Audio, serving both consumer and professional use cases across content creation, entertainment, productivity, and AI interaction. The company’s products are designed around AI-native capabilities rather than traditional software workflows, enabling highly interactive and personalised user experience. As of today, MiniMax’s AI-native products have cumulatively served over 236m users across more than 200 countries and regions. Revenue from AI-native products increased 143.4% yoy to US\$53.1m in 2025, contributing 67.2% of total revenue. Growth was mainly supported by higher user engagement, improving user willingness to pay, and continued adoption of flagship consumer AI applications, including Hailuo AI. The segment remains the company’s key revenue contributor.

Open platform and other AI-based enterprise services. Besides consumer applications, MiniMax operates an open platform providing enterprises and developers with API access to its proprietary multimodal models. The platform supports a broad range of AI capabilities, including text generation, speech synthesis, video creation, coding, and agentic AI applications, allowing customers to integrate AI functionalities into their own products and workflows. Through its MaaS-based business model, MiniMax monetises through API usage, enterprise deployment, and AI-based services tailored to commercial customers.

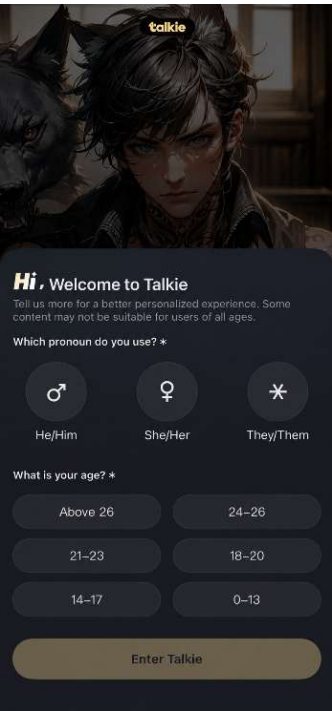
As of 1H26, MiniMax had served over 2m enterprise customers and developers, approximately 10x the level at end-25, reflecting rapid enterprise adoption of its models. Revenue from Open Platform and other AI-based enterprise services increased 703.1% yoy to US\$73.9m in 1H26, accounting for 63.4% of total revenue, up from 30.3% in 1H25. Growth was primarily driven by the expansion of paying users and enterprise customers, higher API call volumes, and rapid adoption of the Token Plan. This made Open Platform the company’s largest revenue contributor in 1H26, overtaking AI-native products, highlighting the rapid scaling of To-B monetisation and enterprise AI deployment.

MiniMax H3 Interface On Topview Ai



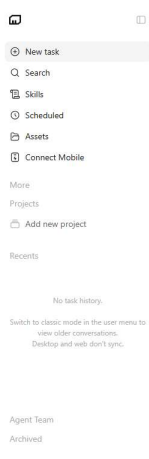
Source: Topview Ai, UOB Kay Hian

MiniMax Talkie



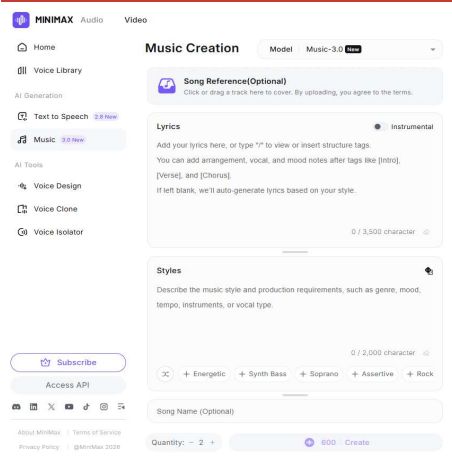
Source: MiniMax, UOB Kay Hian

MiniMax M3



Source: MiniMax, UOB Kay Hian

Minimax Audio/Music

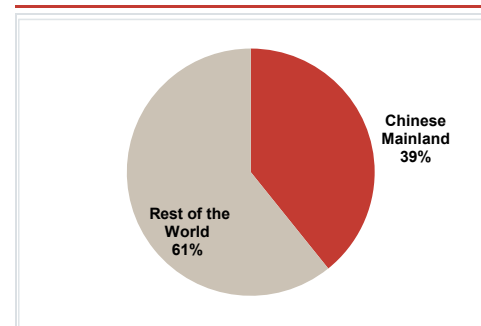


Source: MiniMax, UOB Kay Hian

FINANCIALS AND EARNINGS OUTLOOK

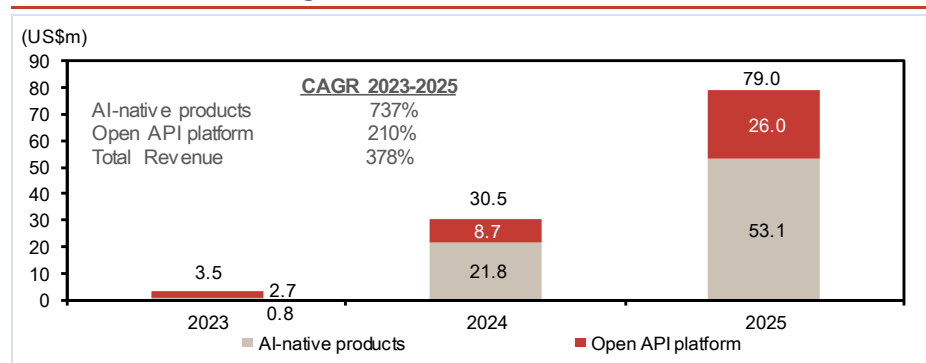
a) 2025/1H26 Results Review And Outlook

- **Overall performance.** MiniMax delivered robust revenue growth in 2025 and 1H26, with total revenue increasing 159% yoy from US\$30.5m in 2024 to US\$79.0m in 2025, before surging 283% yoy to US\$116.6m in 1H26 from US\$30.4m in 1H25. Growth was driven by stronger model capabilities, broader adoption and rapid Open Platform expansion. Gross profit increased 437% yoy in 2025 and 465% yoy in 1H26, supported by a higher contribution from the higher-margin Open Platform business and improving infrastructure efficiency. However, MiniMax remained loss-making, with adjusted net loss widening 6% yoy to US\$251m in 2025 and a further 111% yoy to US\$293m in 1H26, mainly due to persistently high R&D spending, which continued to exceed group revenue.
- **AI-native products.** Revenue from AI-native products increased 143.4% yoy to US\$53.1m in 2025 and 100.9% yoy to US\$42.6m in 1H26, contributing 67.2% and 36.6% of total revenue, respectively. Growth was primarily driven by higher user engagement, stronger willingness to pay, and continued adoption and monetisation of products such as Hailuo AI and other AI-native offerings.
- **Open platform and other AI-based enterprise services.** Revenue from Open Platform and other AI-based enterprise services surged 197.8% yoy to US\$26.0m in 2025 and 703.1% yoy to US\$73.9m in 1H26, accounting for 32.8% and 63.4% of total revenue, respectively. Growth was primarily driven by an increase in paying individual and enterprise users, higher API call volumes, and rapid adoption of the Token Plan. Open Platform has now become MiniMax's largest revenue contributor, overtaking AI-native products in 1H26, highlighting the rapid scaling of To-B monetisation and enterprise AI deployment.
- **Geographical segment.** MiniMax continued to expand internationally, with its products and services deployed across more than 230 countries and regions as of end-1H26. Revenue generated outside Mainland China reached US\$70.8m in 1H26, contributing 60.8% of total revenue, compared with 71.8% in 1H25 and 73% in 2025. Meanwhile, revenue from Mainland China increased to US\$45.7m, accounting for 39.2% of total revenue, vs 28.2% in 1H25. While overseas revenue remained the majority contributor, the higher Mainland China mix reflects faster domestic growth during the period, alongside continued expansion of MiniMax's global commercialisation footprint.
- **1H26 Recap:** 1H26 results were overall in line our expectations. Revenue grew 283.1% yoy, in line our forecast of 340% yoy growth, while gross margin came in at 17.9%, materially below our 30.0% estimate, mainly due to M3 launch compensation, inference ramp-up and a higher mix of lower-margin Token Plans. Adjusted net loss widened 72.7% yoy, vs our forecast of a 22% yoy deterioration, mainly due to weaker margins and elevated R&D and compute spending. Geographically, overseas revenue contribution moderated to 60.8% in 1H26 from 73% in 2025, while the revenue mix shifted materially towards To-B monetisation. Open Platform and other AI-based enterprise services revenue surged 703.1% yoy from US\$9.2m to US\$73.9m, with its revenue contribution rising from 30.3% in 1H25 to 63.4% in 1H26, overtaking AI-native products to become the largest revenue contributor and core growth driver, while July token consumption reached around 20x January levels, underscoring accelerating usage.

Revenue By Geographical Segment
(1H26)

Source: MiniMax, UOB Kay Hian

Revenue Trend And Segment Mix



Source: MiniMax, UOB Kay Hian

b) Outlook

- 2026/27 outlook.** We expect total revenue to reach US\$490m in 2026 and US\$1.3b in 2027, representing yoy growth of 520% and 157% respectively. Growth will be driven by robust expansion across both AI-native products and the open API platform, which are projected to grow 228%/1116% yoy in 2026 and 135%/169% yoy in 2027 respectively. The company's rapid model iteration cycle and focus on high-value use cases, including coding agents, agentic AI and video-generation models, should further accelerate its revenue flywheel.
- Open platform and other AI-based enterprise services to become the key growth engine.** We expect the open API platform to be the key growth driver of group revenue, with revenue surging 1116% yoy to US\$316m in 2026 before moderating to 169% yoy to US\$849m in 2027. This represents absolute revenue additions of US\$290m and US\$533m respectively. Growth will be supported by a strong momentum across text and multimodal offerings, with text revenue expected to increase 4,100% yoy and multimodal revenue rising 219% yoy in 2026, followed by 148% and 251% yoy growth respectively in 2027.
- AI-native products to maintain solid growth momentum.** This segment is expected to maintain solid growth, with revenue increasing 228% yoy to US\$174m in 2026 and 135% yoy to US\$410m in 2027, representing absolute revenue additions of US\$121m and US\$235m respectively. Growth will be primarily driven by a larger MAU base, improving paying-user conversion and higher ARPU, collectively supporting stronger in-app top-ups, subscription revenue and marketing income. Within the segment, MiniMax Agent is expected to be the fastest-growing product, with revenue surging 3,007% yoy in 2026, driven by strong traction in coding and agentic AI use cases. MiniMax Audio, Hailuo AI and Talkie/Xingye are projected to see revenue grow 253%, 174% and 160% yoy respectively in 2026. Growth is expected to moderate to about 139%, 219%, 141% and 120% yoy for MiniMax Agent, MiniMax Audio, Hailuo AI and Talkie/Xingye respectively in 2027.
- Gross margin expected to trough in 2026 before recovering.** We expect gross margin to decline to approximately 21% in 2026 from 25% in 2025, mainly due to one-off token compensation following the M3 launch, the initial inference-optimisation ramp-up and a higher mix of lower-margin Token Plans. Gross margin could recover to around 25% in 2027, supported by improving inference efficiency, normalisation of promotional incentives and a higher contribution from the higher-margin Open Platform and enterprise services segment as To-B adoption scales. Management also noted that M3's steady-state API gross margin has already surpassed M2 despite roughly doubling model parameters, with per-token inference cost falling below M2, highlighting continued improvement in inference efficiency and supporting further margin expansion.

- **Operating expenses to remain elevated amid continued AI investment.** We expect MiniMax's total operating expenses to increase 131% yoy to about US\$789m in 2026, from US\$341m in 2025, before rising a further 40% yoy to US\$1.1b in 2027. The increase is primarily driven by sustained R&D investment in foundation model training, computing infrastructure and AI talent acquisition, alongside higher selling and distribution expenses to support enterprise customer acquisition, international expansion and product commercialisation amid intensifying competition across the AI industry. R&D and selling and distribution expenses are expected to account for approximately 91%/89% of total operating expenses in 2026 and 2027 respectively.

- **R&D investment to remain elevated amid intensifying AI competition.** R&D investment is expected to remain high at approximately US\$632m in 2026 and increase further to US\$822m in 2027, mainly driven by continued investment in foundation model training, computing infrastructure and AI talent acquisition amid intensifying competition across China's AI industry. Competition for top-tier AI researchers and engineers has increased significantly as major technology companies and AI start-ups accelerate large-model development and product commercialisation.

Meanwhile, MiniMax has maintained a rapid iteration cycle across text, video, speech, music and agentic AI models, having released more than 10 major model releases since the beginning of 2026. We expect the company to sustain this pace of model development. Continued R&D investment remains necessary to preserve MiniMax's competitiveness in multimodal foundation models and support long-term commercialisation across both AI-native products and open platform services.

- **Adjusted net profit expected to turn positive by 2029 as scale benefits improve.** MiniMax is likely to remain loss-making through 2028, although its adjusted net margin is projected to improve substantially as revenue scales. We forecast adjusted net losses widening from US\$251m in 2025 to US\$615m in 2026 and peaking at US\$724m in 2027, primarily reflecting elevated investment in model training, computing infrastructure, AI talent acquisition and commercialisation. Despite the higher absolute losses, adjusted net margin is expected to improve from -317% in 2025 to -126% in 2026 and -57% in 2027 on rapid revenue growth and stronger operating leverage. Adjusted net losses are projected to narrow to US\$544m in 2028, with the margin improving further to -20%, before adjusted net profit turns positive at about US\$147m in 2029, representing a 3% margin. We estimate adjusted net profit at US\$1b by 2030, with the margin expanding to 12%, driven by stronger monetisation, a more favourable revenue mix, and greater scale benefits across AI-native products and open platform services.

c) Balance Sheet

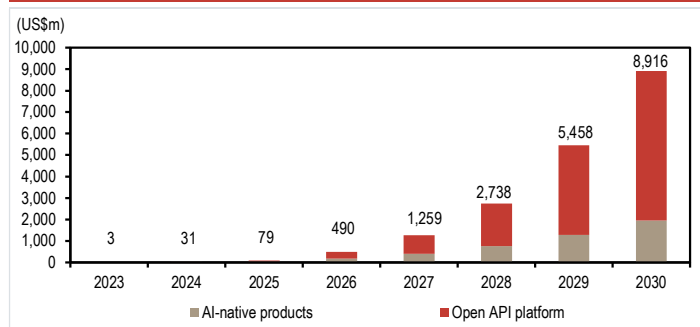
- **Strengthened net cash position to support continued R&D investment and commercialisation.** Net cash position stood at about US\$484m at the end of 2025 and is projected to increase to US\$2.5b in 2026, mainly supported by net proceeds from MiniMax's IPO, placement of new shares and issue of zero coupon guaranteed convertible bonds. The strengthened balance sheet is expected to provide an important buffer against continued operating losses and high R&D spending, while supporting ongoing investment in foundation model training, computing infrastructure, and AI talent acquisition. We believe the company's solid liquidity position remains critical to sustaining long-term model development and commercialisation expansion amid intensifying competition in the AI industry.
- **Current ratio expected to remain healthy.** MiniMax's current ratio is projected to remain healthy at 2.9x in 2026 and 5.0x in 2027, supported by relatively low short-term obligations, as most of its funding is expected to come from additional share issuances to support ongoing operations.

d) 2H26 Outlook

- Outlook:** We expect revenue growth to remain strong in 2H26, supported by accelerating To-B monetisation and continued model upgrades. August ARR exceeded US\$800m, with To-B contributing more than 80%, and management remains highly confident in achieving its US\$1b year-end ARR target, although upside could be constrained by model-release timing and industry-wide compute supply. Recent ARR acceleration was mainly driven by M3 text-model adoption, with demand broadening from coding and agents into office and productivity use cases, while H3 also began contributing in August. We expect gross margin to improve sequentially in 2H26 as one-off M3 compensation rolls off, Token Plan economics improve and inference costs decline. M3.1, M3 Pro and H3.1 should be the key catalysts for the next phase of revenue and ARR growth, while R&D investment remains elevated. Liquidity should remain sufficient, supported by over US\$3b cash, with no current financing plan.

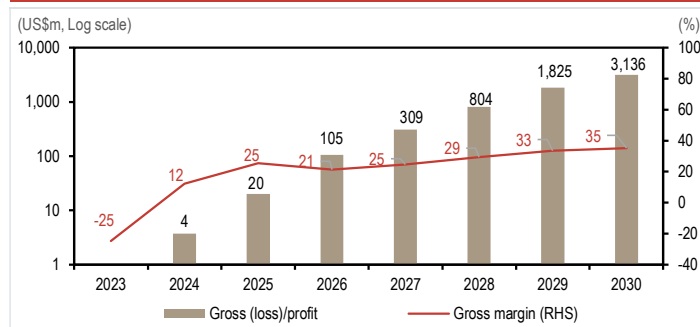
We thus believe MiniMax's current liquidity provides a meaningful near-term buffer, although the timing of external funding remains highly sensitive to the pace of revenue scaling, gross-margin improvement and R&D normalisation.

Revenue Trend By Segment



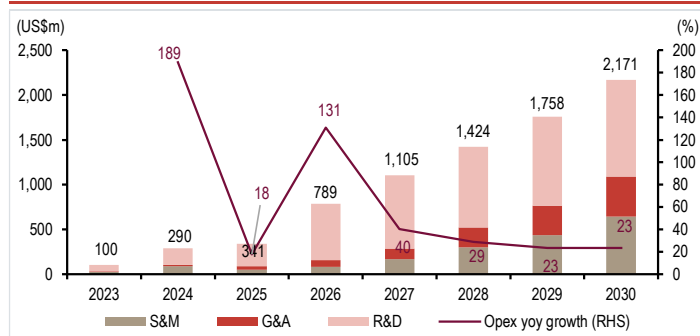
Source: MiniMax UOB Kay Hian

Gross Margin Performance



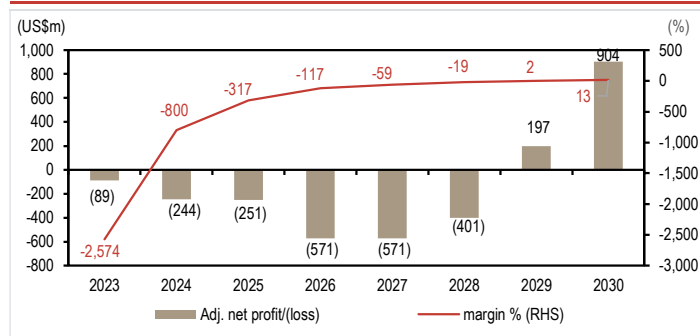
Source: MiniMax UOB Kay Hian

Opex Breakdown



Source: MiniMax UOB Kay Hian

Adjusted Net Profit/(Loss) Trend



Source: MiniMax UOB Kay Hian

Key Operating Metrics

	2023	2024	2025	2026F	2027F	2028F	2029F	2030F
Revenue								
AI-native products	1	22	53	174	410	764	1,289	1,964
Open API platform	3	9	26	316	849	1,974	4,169	6,953
Revenue (yoy %)								
AI-native products		2777%	143%	228%	135%	86%	69%	52%
Open API platform		223%	198%	1116%	169%	133%	111%	67%
Revenue - share (%)								
AI-native products	22%	71%	67%	36%	33%	28%	24%	22%
Open API platform	78%	29%	33%	64%	67%	72%	76%	78%
Gross profit								
AI-native products	(3)	(2)	3	3	12	53	116	216
Open API platform	2	6	17	101	297	750	1,709	2,920
Gross profit - margin (%)								
AI-native products	-380%	-8%	5%	2%	3%	7%	9%	11%
Open API platform	75%	63%	66%	32%	35%	38%	41%	42%
Paying customers (000')								
AI-native products	120	650	1,880	4,037	7,313	11,069	16,213	20,307
Open API platform	0.1	0.7	3.3	9.9	21.8	38.1	61.0	88.4

Source: MiniMax, UOB Kay Hian

Funding runway remains sensitive to revenue growth, margin expansion and R&D discipline. In our base case, we estimate MiniMax has more than three years of cash runway, with cumulative adjusted net losses of about US\$1.8b over 2026-28 sufficiently covered by approximately US\$2.9b of net cash, leaving around US\$1.1b of liquidity and no external financing requirement through 2028. In our bear case, slower revenue growth, more modest gross margin expansion and elevated R&D intensity shorten the runway to about 2.8 years, with cumulative losses of around US\$3.1b exceeding available net cash by approximately US\$0.25b and implying a first external funding requirement in 2028. In our stress case, weaker growth, limited margin improvement and persistently high R&D spending further reduce the runway to about 1.9 years, bringing the first external financing requirement forward to 2027.

We thus believe MiniMax's current liquidity provides a meaningful near-term buffer, although the timing of external funding remains highly sensitive to the pace of revenue scaling, gross-margin improvement and R&D normalisation.

Figure 14: Runway Sensitivity Analysis

Scenario	Cash runway (years)	First external funding year	Cum. adj. net losses (US\$b)	Ending liquidity (US\$b)		
Base	>3.0	n.a	1.8	1.1	Base case	In our base case, we assume a 137% revenue CAGR over 2026-28, with gross margin expanding from 30% in 2026 to 34% in 2027 and 39% in 2028, while R&D intensity declines from 150% to 72% and 38% of revenue respectively. This results in cumulative adjusted net losses of about US\$1.8b over 2026-28, which are sufficiently covered by approximately US\$2.9b of net cash, leaving around US\$1.1b of remaining liquidity. We therefore do not expect any external financing requirement through 2028.
					First external financing need	No external financing required
Bear	2.8	2028	3.1	-0.25	Bear case	In our bear case, we assume a 130% revenue CAGR over 2026-28, with gross margin expanding only modestly from 30% in 2026 to 31% in 2027 and 32% in 2028, while R&D intensity declines from 150% to 100% and 80% of revenue respectively. This results in cumulative adjusted net losses of about US\$3.1b over 2026-28, exceeding approximately US\$2.9b of net cash by around US\$0.25b. We therefore expect MiniMax to require its first round of external financing in 2028 under the bear-case assumptions.
					First external financing need	External financing required in 2028F
Stress	1.9	2027	3.1	-0.26	Stress case	In our stress case, we assume revenue growth slows to 120% yoy in 2027, with gross margin remaining at 30% in both 2026 and 2027, while R&D intensity remains elevated at 250% and 230% of revenue respectively. This results in cumulative adjusted net losses of about US\$3.1b over 2026-27, exceeding approximately US\$2.9b of available net cash by around US\$0.26b. We therefore expect MiniMax to require its first round of external financing in 2027 under the stress-case assumptions.
					First external financing need	External financing required in 2027F

Source: UOB Kay Hian

Appendix III: Management

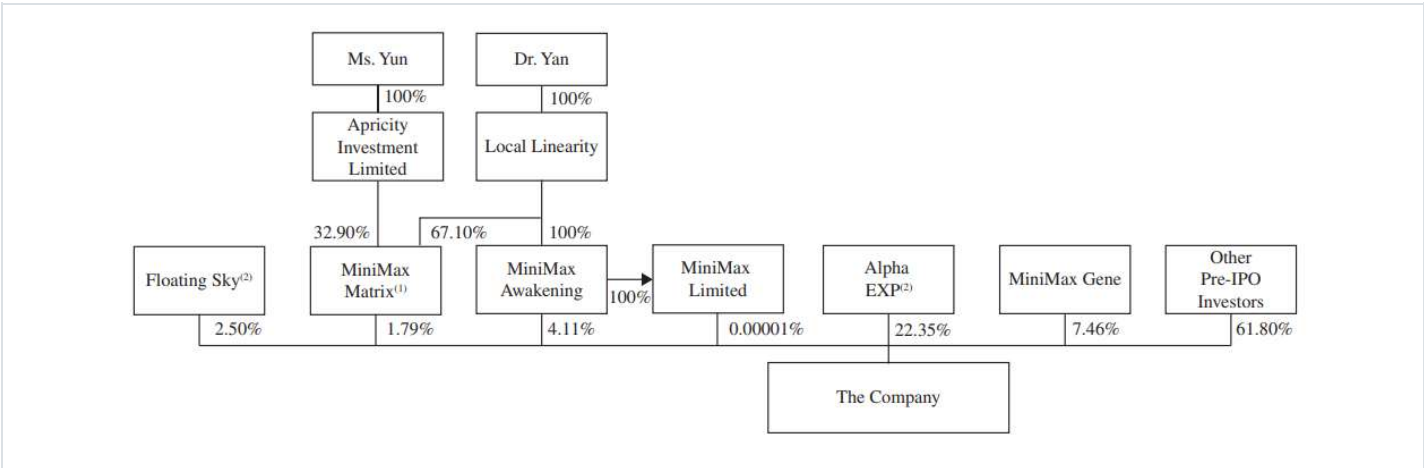
Key Management Team

Name	Age	Position	Experience
Dr Yan Junjie	37	Founder, Chairman of the Board, Chief Executive Officer, Chief Technology Officer	Founder, Chairman, CEO and CTO, overseeing strategy, operations, finance and AI R&D. Previously Vice President and vice-head of SenseTime's Research Institute, he holds a doctorate in AI from the Chinese Academy of Sciences and has published more than 200 academic papers with over 30,000 citations.
Yun Yeyi	33	Executive Director, President	Executive Director and President, overseeing management, product strategy and commercialisation. Ms Yun previously held fundraising, strategic investment, strategy and innovative business roles at SenseTime. She holds a bachelor's degree in Electrical Engineering from Johns Hopkins University.
Zhao Pengyu	31	Executive Director, Large Language Model Research & Engineering Leader	Executive Director and LLM Research & Engineering Leader, responsible for LLM R&D. Mr Zhao previously worked on recommendation algorithms at Beijing Hulu Technology. He holds bachelor's and master's degrees in Computer Science and Technology from Peking University.
Zhou Yucong	34	Executive Director, Visual Model Research & Engineering Leader	Executive Director and Visual Model Research & Engineering Leader, responsible for visual-model R&D. He previously worked on algorithms at SenseTime and Huawei, and holds degrees in Mathematics and Computer Science from Beihang University.
Chen Yingjie	50	Non-Executive Director	Non-Executive Director providing strategic advice. He currently serves as Managing Director of Alibaba Group's Strategic Investment Department and previously worked at PwC and in investment-related roles. He is a Canadian CPA.
Liu Wei	40	Non-Executive Director	Non-Executive Director, and also co-founder and President of miHoYo, where he has contributed to the company's management, growth and global expansion. He holds bachelor's and master's degrees from Shanghai Jiao Tong University.
Huang Guobin	58	Independent Non-Executive Director	Independent Non-Executive Director with extensive investment banking experience, including senior roles at J.P. Morgan, Goldman Sachs and CICC. He is also Chairman of NEXX Global and PEC International Group.
Wang Pengcheng	56	Independent Non-Executive Director	Independent Non-Executive Director with extensive accounting and audit experience. Formerly a Greater China assurance managing partner at EY and financial services managing partner at Deloitte, he is currently a professor at Beijing Technology and Business University and a PRC CPA.
Zhu Huaxing	43	Independent Non-Executive Director	Independent Non-Executive Director and theoretical physicist. Currently a Boya Distinguished Professor at Peking University, he previously conducted research at Zhejiang University, MIT and SLAC. He holds bachelor's and doctorate degrees in Physics from Peking University.

Source: MiniMax, UOB Kay Hian

Appendix IV: Organisational Structure

Organisation Chart



Note:
(1) MiniMax Matrix was established at the early stage of the company's development in order to hold relevant beneficial interests in the company. There is no side arrangement in terms of MiniMax Matrix's control between Dr Yan and Ms Yun.
(2) For estate planning purposes and by Nov 25, all the shares controlled by Dr Yan in Alpha EXP and by Ms Yun in Floating Sky were held under the trusts of Dr Yan and Ms Yun respectively.
Source: MiniMax, UOB Kay Hian

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."