China Artificial Intelligence

DeepSeek V4: Sector tailwind misread as zero-sum shock – Reiterate OW on Zhipu and MiniMax

DeepSeek's V4 Preview launch on Apr 24 sent Zhipu / MiniMax share prices down 9% (vs. HSI +0.2%), with the market reading V4 as a competitive shock to the listed pure-play LLM names. We see this an over-reaction; on the contrary, V4 reads to us as a sector tailwind instead of a zero-sum shock: the launch reinforces three of the four pillars of our constructive framework: **compute supply unlocked, pricing discipline,** and **structural cost-curve compression**, while leaving the fourth **(relative competitive positioning) tightly contested rather than rebalanced**. The launch also clears the largest negative competitive catalyst on our April–May event calendar. Our framework (report) had flagged V4 as outcome-dependent and as the single largest known competitive overhang into June. With it released and absorbed, the catalyst path from here turns asymmetrically positive: index inclusion (May review, Jun effective), Zhipu Southbound (early Jun), GLM-5.5 and MiniMax M3 launches (~Jun). With fundamentals remaining unchanged, while the largest short-term competition uncertainty is cleared, the next two months sit in favorable corridor between V4 being out of the way and the July lock-up moving into pre-positioning range.

- **Domestic compute validation alleviates the binding constraint on sector ARR.** The DeepSeek V4 launch confirms the practical feasibility of domestic chips for frontier AI model inference. While V4 Pro's throughput is currently limited by Huawei's compute availability, DeepSeek's announcement that Pro pricing will decrease with Ascend 950 supernode scaling in 2H26 demonstrates that domestic compute is not only technically feasible, but could also become cost-competitive with international chips. We see this as a key validation for the Chinese LLM ecosystem that directly benefits Zhipu and MiniMax, as compute constraints have historically limited their ARR conversion. As domestic chips become widely available, these companies will have access to a more cost-effective solution, improving their ability to convert token demand into billable revenue.

- **DeepSeek V4's Pro/Flash pricing reinforces task-based monetization.** A c.12x output-price gap inside DeepSeek's own lineup undermines the bear argument that DeepSeek pricing forces commodity economics across sector, but validates the premium-tier monetization logic Zhipu and MiniMax have been building toward. Combined with the open-source release of V4's token-axis compression and DeepSeek Sparse Attention (DSA) architecture, the sector receives a cost-curve downshift that Zhipu and MiniMax next model generations sit inside the integration window to capture.

- **OW on Zhipu and MiniMax, use the pullback to add ahead of GLM-5.5 and MiniMax M3 in ~June.** DeepSeek V4 leaves GLM5.1 intact on the B-end metrics that drive enterprise ARR, and the two-month corridor between DeepSeek V4 clearing and July lock-up window is asymmetrically positive: index inclusion, Southbound, and next model cycle all run in our favor.

Internet

Olivia Xu AC
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com
SAC Registration Number: S1730525060001

Alex Yao
(86 21) 6106 6505
alex.yao@jpmorgan.com
SAC Registration Number: S1730523020001

Daniel Chen
(86-21) 6106 6205
daniel.q.chen@jpmorgan.com
SAC Registration Number: S1730521040001

J.P. Morgan Securities (China) Company Limited

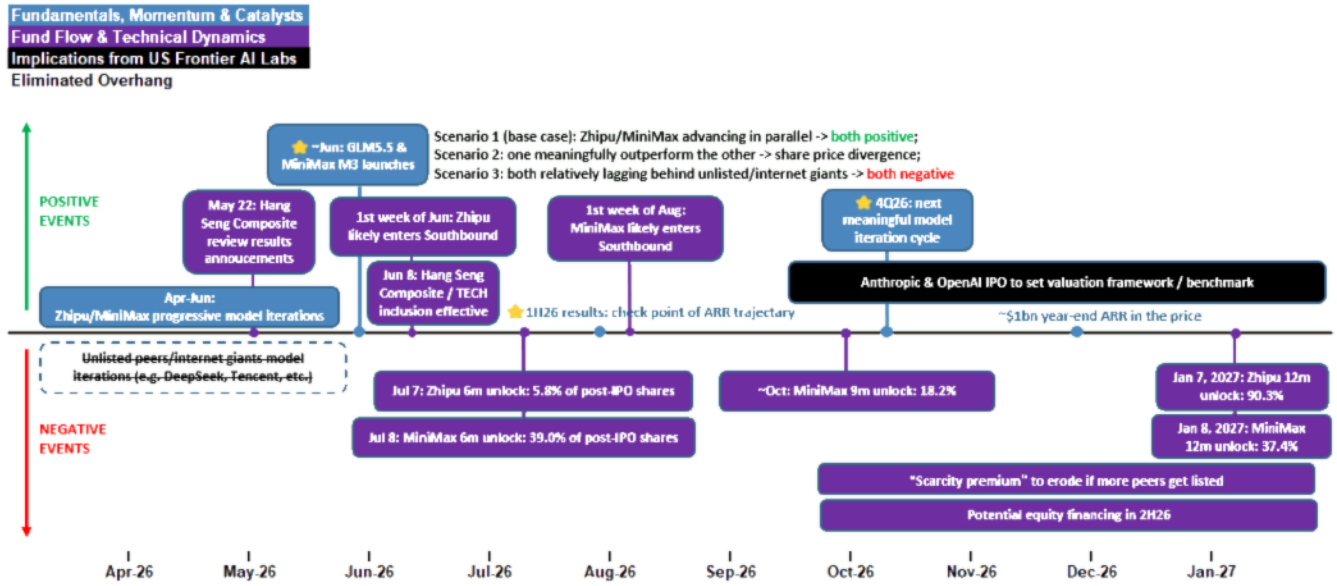
2513.HK, 2513 HK
Overweight
Price: HK$935.00
24 Apr 2026
Price Target: HK$950.00
PT End Date: 31 Dec 2026

0100.HK, 100 HK
Overweight
Price: HK$777.50
24 Apr 2026
Price Target: HK$1,100.00
PT End Date: 31 Dec 2026

See page 8 for analyst certification and important disclosures, including non-US analyst disclosures.

Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Figure 1: Major events timeline for Zhipu AI and MiniMax



Source: Company prospectus (data as of Dec 31, 2025), Hang Seng Index website, Reuters, Sina Finance, Sina News, AAStocks, and J.P. Morgan estimates. (data as of Apr 24, 2026).

Key takeaways

The compute unlocked story is where we see the largest gap between price action and read-through

DeepSeek V4's adaptation of Huawei Ascend chips is a crucial step in reducing China's compute bottleneck for frontier models. As we highlighted, the Chinese LLM industry remains compute-constrained, limiting growth potential and restricting ARR conversion. DeepSeek V4 is an important validation of domestic chips like Ascend, proving they can handle large-scale AI models (1.6T parameters).

This is a key positive for the broader Chinese LLM sector, including Zhipu and MiniMax, in our view, as it alleviates the compute supply constraint — one of the primary obstacles for these companies to fully scale their operations. Per DeepSeek, as Ascend chips become more widely available and optimized, the inference cost curve will improve, creating new avenues for both Zhipu and MiniMax to convert token demand into recognized revenue

DeepSeek V4 Pro/Flash tiering-pricing reinforces task-based monetization

DeepSeek charges a roughly 12x output pricing gap for its top tier (pro version, at $1.7/$3.5 per mn input/output token) versus its commodity tier (flash version, at S0.1/$0.3); i.e., its own pricing book validates the task-based monetization architecture the sector has been building toward: premium pricing for 2B coding and agent workflows, lower pricing for high-throughput simple tasks. This is the same structure Zhipu and MiniMax already operate.

- **Current pricing comparison**: Comparing DeepSeek V4's Pro pricing to other leading models such as GLM-5.1 and Kimi K2.6, DeepSeek V4 Pro does not offer a clear pricing advantage. The standing bear argument that DeepSeek's pricing posture forces all domestic models toward commodity economics sits against the data.

Table 1: Token input / output price comparison for selective Chinese AI models

Model provider	Model	Input Price $/M	Output Price $/M	Contex
Zhipu AI	glm-5.1	$1.05	$3.50	202.8K
Moonshot	kimi-k2.6	$0.74	$4.66	256K
DeepSeek	deepseek-v4-pro-thinking	$1.74	$3.48	1M
Alibaba	qwen3.6-plus	$0.33	$1.95	1M
Xiaomi	mimo-v2-pro	$1	$3	1M
MiniMax	minimax-m2.7	$0.3	$1.2	196.6K
DeepSeek	deepseek-v4-flash	$0.14	$0.28	1M

Source: LMArena, DeepSeek official website (in order of the Code Arena ranking, data as of Apr 24, 2026).

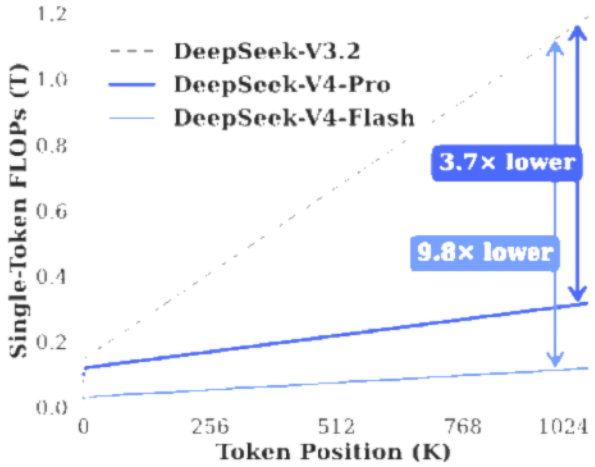
- **V4 Pro's potential price drop:** Even if DeepSeek Pro's price decreases as Ascend 950 supernodes scale in 2H26, Zhipu and MiniMax will likely have newer, more competitive models in the market by then. When new models are launched by Zhipu and MiniMax, the existing models from these companies will move into legacy status, and the associated price adjustments will not materially impact their positioning or pricing strategy in the long run.

Token compression and sparse attention is a sector input rather than a moat

DeepSeek V4's technical innovations, such as token compression and DeepSeek Sparse Attention (DSA), reduce memory and compute costs for long-context processing. These innovations set a new industry standard for long-context inference, which is crucial for coding agents, enterprise knowledge management, and other applications requiring multi-turn reasoning.

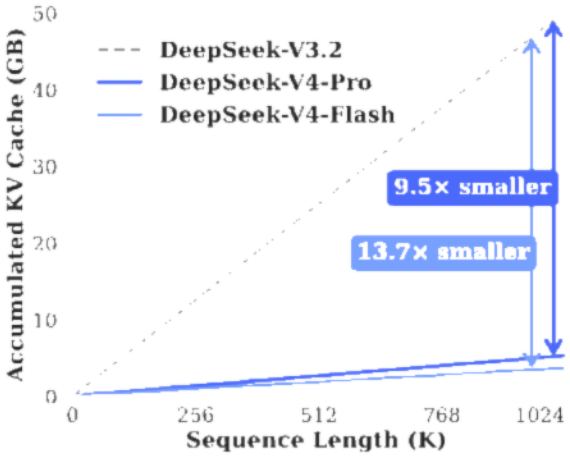
By achieving 1M context as the standard, DeepSeek has lowered the barrier to long-context models, making them more cost-efficient and scalable. This is an important milestone for the LLM industry, and Zhipu and MiniMax will need to incorporate these breakthroughs into their future models to remain competitive. DeepSeek's open-source release accelerates the diffusion of these technical advances, benefiting the industry as a whole: the Chinese LLM ecosystem has historically absorbed DeepSeek architectural innovations within 1-2 model cycles: V3-era MoE routing reached Qwen, GLM and Kimi within roughly 4–6 months.

Figure 2: Inference FLOPS of DeepSeek V4 versus V3.2



Source: DeepSeek official website (data as of Apr 24, 2026).

Figure 3: Accumulated KV cache size of DeepSeek V4 versus V3.2



Source: DeepSeek official website (data as of Apr 24, 2026).

DeepSeek V4 shows meaningful progress, but not yet domestic SOTA in key B2B tasks

While DeepSeek V4 has made meaningful progress, it has not surpassed existing leaders in key B2B tasks like coding and agent workflows. In the benchmark rankings from Artificial Analysis, LMArena Code Arena, and benchmark metrics, DeepSeek V4 performs well in general reasoning but does not exceed Kimi K2.6 / GLM5.1 in critical enterprise metrics such as coding efficiency and agent task completion reliability.

- Artificial Analysis, LMArena Code Arena benchmarks and the benchmark metrics show that DeepSeek V4 is competitive in general-purpose tasks, but in coding and agent tasks, which are the core value drivers for Zhipu and MiniMax, DeepSeek V4 still trails behind GLM-5.1 and Kimi K2.6.

- In this rapidly evolving model market, new model releases drive ranking shifts, and it is common for new models to move to the top of domestic rankings, as new releases benefit from the latest training techniques and benchmark feedback. We believe that DeepSeek V4's position in these rankings reflects a normal industry pattern, where leadership rotates with each new release. Zhipu and MiniMax have their own upcoming releases in/around June, which could again reset the ranking dynamic.

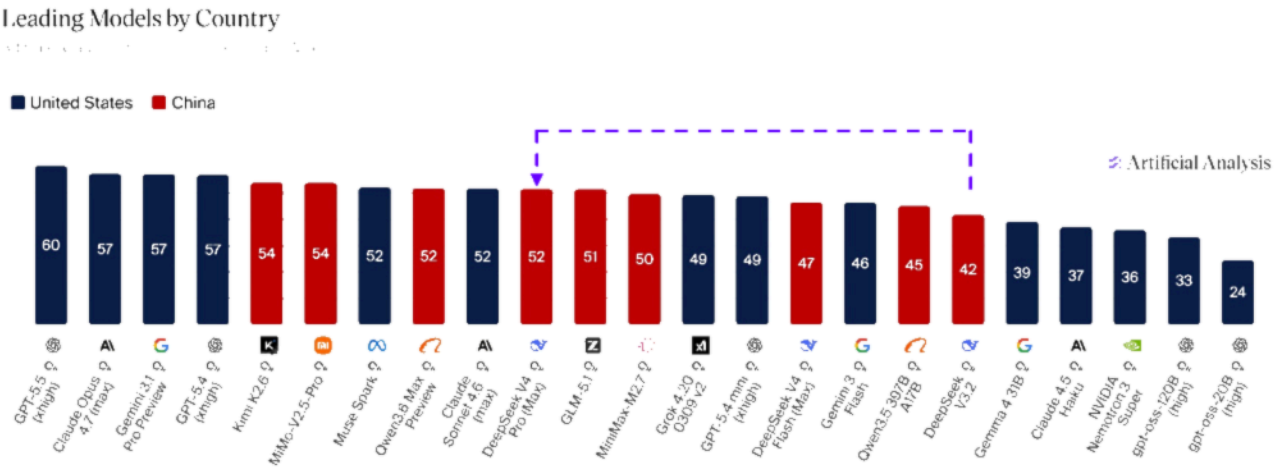
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Table 2: According to Code Arena, DeepSeek V4 is ranked No.14, after GLM5.1, Kimi K2.6 and Qwen3.6 plus from China

Rank	Rank Spread	Model provider	Model	Score	Votes	Input Price $/M	Output Price $/M	Context
1	1<->3	Anthropic	claude-opus-4-7-thinking	1572+18/-18	1,393	$5	$25	1M
2	1<->4	Anthropic	claude-opus-4-7	1566+16/-16	1,782	$5	$25	1M
3	1<->6	Anthropic	claude-opus-4-6-thinking	1552+10/-10	4,922	$5	$25	1M
4	2<->6	Anthropic	claude-opus-4-6	1545+9/-9	5,821	$5	$25	1M
5	3<->8	Zhipu AI	glm-5.1	1534+13/-13	2,478	$1.05	$3.50	202.8K
6	3<->8	Moonshot	kimi-k2.6	1529+17/-17	1,408	$0.74	$4.66	256K
7	5<->8	Anthropic	claude-sonnet-4-6	1527+9/-9	7,751	$3	$15	1M
8	5<->9	Meta	muse-spark	1512+16/-16	1,611	N/A	N/A	N/A
9	8<->9	Anthropic	claude-opus-4-5-20251101-thinking-32k	1491+7/-7	13,065	$5	$25	200K
10	10<->14	Alibaba	qwen3.6-plus	1471+12/-12	2,854	$0.33	$1.95	1M
11	10<->14	Anthropic	claude-opus-4-5-20251101	1468+6/-6	15,283	$5	$25	200K
12	10<->16	Google	gemini-3.1-pro-preview	1457+8/-8	6,813	$2	$12	1M
13	10<->19	OpenAI	gpt-5.4-high (codex-harness)	1457+17/-17	1,482	$2.50	$15	1.1M
14	10<->20	DeepSeek	deepseek-v4-pro-thinking	1456+19/-19	1,066	$1.74	$3.48	1M
15	12<->21	Zhipu AI	glm-4.7	1440+10/-10	4,880	$0.38	$1.74	202.8K
16	13<->21	Google	gemini-3-pro	1438+7/-7	17,169	$2	$12	1M
17	12<->23	OpenAI	gpt-5.4-medium (codex-harness)	1437+16/-16	1,448	$2.50	$15	1.1M
18	13<->21	Google	gemini-3-flash	1437+7/-7	13,276	$0.50	$3	1M
19	13<->22	Zhipu AI	glm-5	1435+9/-9	5,640	$1	$3.20	202.8K
20	14<->22	Moonshot	kimi-k2.5-thinking	1430+8/-8	7,913	$0.60	$3	N/A

Source: LMArena Code Arena (data as of Apr 24, 2026), Chinese models colored in green. Note: Code Arena is an interactive evaluation platform developed to test large language models (LLMs) on real-world coding tasks, it moves beyond static benchmarks by using live, developer-voted evaluations and executable environments to measure how AI agents plan, debug, and build software.

Figure 4: According to Artificial Analysis, DeepSeek V4 improved meaningfully versus V3.2, though its Intelligence Index is not yet the best among Chinese LLMs



Source: Artificial Analysis (data as of Apr 24, 2026).

Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Catalyst path: negatives cleared; positives concentrated into June

The DeepSeek V4 launch has now played out, and the short-term catalyst has already been priced in. The two-month corridor between V4 clearing and lock-up pre-positioning starting is the asymmetric window. Index inclusion provides passive flow; Southbound opens Mainland demand; the model iteration cycle produces the operating catalyst that has historically driven ARR step-ups.

As discussed in our tripwire framework (report), the primary risks for Zhipu and MiniMax remain relative model leadership, pricing discipline, China-US capability gap, ARR conversion, and model release cadence. V4 leaves these intact, in our view, and reinforces several of them: the compute supply story strengthens the ARR-conversion narrative heading into GLM-5.5 and M3, pricing discipline reconciles to V4 strategy, and we did not see a reset in relative model capability positioning.

Table 3: Major events timeline, impacts and expected share price direction for Zhipu AI and MiniMax

Timing	Event	Impact	Share price direction	JPM's event read-through
	Zhipu / MiniMax ongoing model iterations	Focuses on relative model progress, pricing power and enterprise adoption cadence	Positive	
Apr-May 2026	~~DeepSeek V4 launch~~	~~Key read-across for China model competition and for domestic GPU-based training/inference validation~~	~~Outcome-dependent~~	Compute supply unlocked, pricing discipline, structural cost-curve compression, relative competitive positioning tightly contested rather than rebalanced
	~~Tencent Hunyuan 3.0 launch~~	~~A key test of how quickly scaled platforms can close the capability gap~~	~~Outcome-dependent~~	No meaningful impacts to independent LLMs
Jun-26	GLM 5.5 / MiniMax M3	Potential step-function catalyst for ARR; market focus on both absolute capability and relative positioning	Positive, unless relative performance disappoints	
	Potential index inclusion	Passive flow catalyst and broader institutional ownership expansion	Positive	
	Potential Zhipu Southbound inclusion	Expands Mainland investor access and can improve liquidity / valuation support	Positive	
Jul-26	6-month lock-up expiry	Pre-IPO shareholders at 5-7x gains; Zhipu unlocks 5.8% share, MiniMax at 39.0%	Negative	
Aug-26	1H26 results	Check point of ARR trajectary, pricing durability and margin trajectory	Outcome-dependent, positive if execution holds	
	Potential MiniMax Southbound inclusion	Broadens investor base and may partly offset post-lock-up flow pressure	Positive	
Oct-26	9-month lock-up expiry	MiniMax 9m unlock at 18.2%	Negative	
2H26	Potential equity financing	Additional financing is near-certain; R&D = 4-5x revenue	Negative	
	Potential Anthropic / OpenAI IPO	Sets global AI multiple anchor	Positive	
	Potential Kimi / StepFun listing	Validates the sector's investability, but also dilutes scarcity	Mixed to slightly negative	
Jan-27	12-month lock-up expiry	Zhipu 12m unlock at 90.3%, MiniMax at 37.4%	Negative	

Source: Company prospectus (data as of Dec 31, 2025), Hang Seng Index website, Reuters, Sina Finance, AAStocks, and J.P. Morgan estimates (data as of Apr 24, 2026).

Knowledge Atlas Technology Joint Stock Co. Ltd. *(Overweight; Price Target: HK$950.00)*

Investment Thesis

We rate Knowledge Atlas (Zhipu AI) Overweight. Our investment view is anchored in a structural assessment of how value is created and sustained in the foundation-model industry. We believe that long-term economic outcomes are determined primarily by a company's ability to **remain within the global first tier of model capability across multiple technology cycles**, while business model form, deployment modality, and near-term margin structure are largely downstream expressions of that capability.

Zhipu has reached an important inflection point, especially for its global API business. The release of GLM-4.5/4.6/4.7 to GLM5, together with a visible strategic shift toward agentic systems, tool-augmented reasoning, and developer-facing infrastructure, indicates that the company is aligning its technical roadmap with the dimensions of capability that increasingly define the global frontier — particularly production-grade coding, long-context reasoning, and multi-step execution stability. These attributes are commercially important because they enable genuine workflow substitution rather than marginal assistance, unlocking materially larger enterprise and developer budgets.

The company's business architecture reinforces this trajectory. Zhipu has established a substantial on-premise deployment base in regulated Chinese industries, which we view as **a structurally durable demand pool**. As foundation models iterate, this installed base has the potential to evolve into upgrade-driven, recurring economics. In parallel, cloud-based APIs represent a scalable growth engine, and we expect adoption to accelerate as GLM gains recognition among developers, particularly in coding-centric workflows where willingness to pay and usage intensity are higher.

Valuation

Our Dec-26 PT of HK$950 is derived from 30x 2030E P/E, when Zhipu AI will generate normalized earnings, and discounted back using a 15% WACC. The 30x multiple represents a valuation premium over tier-1 Chinese internet players, mostly to reflect the company's 100+% revenue CAGR in 2026-30E.

Price target calculation

2030E Revenue (Rmb mn)	98,832
2030E Adjusted net profit (Rmb mn)	20,360
2030E Adjusted EPS (Rmb/share)	43
Target multiple	30
WACC	15%
Dec-26 PT (HK$)	950

Source: J.P. Morgan estimates.

Risks to Rating and Price Target

Risks include export controls, geopolitical risk, and entity list designation; intensified competition; high and ongoing R&D investment creates execution risk and pressure on profitability; commercialization and customer adoption remain subject to uncertainty; dependence on computing infrastructure and external suppliers exposes Zhipu to cost and availability risks.

MiniMax Group Inc - H *(Overweight; Price Target: HK$1,100.00)*

Investment Thesis

We rate MiniMax OW, for its rare combination of technical strength, multi-modal commercialization potential, and global scalability in the AI foundation model space. The company has established a strong technical track record, with its models delivering strong performance across key benchmarks, while its full-spectrum portfolio lays the groundwork for dual B2B/B2C monetization. Its global footprint further enhances scaling potential and profitability. Our investment theses lie in:

- Model capability will increasingly dominate competitive outcomes in the AI model landscape, where MiniMax has delivered a strong technical track record. MiniMax's models deliver leading performance in global benchmark, its models of text, speech and video generation all ranked among tops globally. The group has achieved high ROI on R&D, cumulatively $450mn R&D investment as of Sep 2025, supporting bi-monthly model iterations. MiniMax's historical behavior — characterized by consistent investment in model training, regular generation upgrades, and early commitment to multimodality — supports the case that it is well-positioned to participate in this AI competition dynamic.
- Full-spectrum models enabling dual B2B/B2C commercialization. The coexistence of text, video, and audio models enables MiniMax to address enterprise and consumer use cases that are increasingly multimodal by design, reducing dependence on any single monetization path.
- Global orientation fueling scaling and profitability. MiniMax's early global orientation, visible in both customer targeting and infrastructure strategy, introduces economic flexibility that is uncommon among peers at a comparable stage.

Valuation

Our Dec-26 PT of HK$1,100 is derived from 30x 2030E P/E, when MiniMax will generate normalized earnings, and discounted back using 15% WACC. The 30x multiple represents a valuation premium over tier-1 Chinese internet players, mostly to reflect the company's 100+% revenue CAGR in 2026-30E.

Price target calculation

2030E Revenue (US$mn)	9,136
2030E Adjusted net profit (US$mn)	2,322
2030E Adjusted EPS ($/share)	7.2
Target multiple	30
WACC	15%
Dec-26 PT (HK$)	1,100

Source: J.P. Morgan estimates.

Risks to Rating and Price Target

Risks include MiniMax's litigation proceedings with US studios; intensified competition; high and ongoing R&D investment creates execution risk and pressure on profitability; commercialization and customer adoption remain subject to uncertainty; dependence on computing infrastructure and external suppliers exposes MiniMax to cost and availability risks.