Morgan Stanley | RESEARCH

ASIA PACIFIC INSIGHT

April 27, 2026 09:00 PM GMT

China's Emerging Frontiers

China's AI Path: More Bang For The Buck

WHAT'S CHANGED		
Knowledge Atlas Technology JSC Ltd (2513.HK)	**From**	**To**
Price Target	HK$560.00	**HK$990.00**
MiniMax (0100.HK)		
Price Target	HK$990.00	**HK$1,100.00**

Foundation models continue to scale, driven by US compute and China's engineering efficiency. We believe a recent surge in token usage and pricing power points to non-linear ARR growth and higher valuations for frontier models with high intelligence density per token cost (MiniMax, Zhipu, Alibaba).

Scaling Law still holds, but so does China's engineering efficiency "scaling": In 10 Thematic Predictions for 2026, we wrote that American frontier LLMs will achieve a step-change increase in capabilities in 1H26 that will "not be matched within the same timeframe by Chinese competitors". Although US AI labs continue to see breakthroughs in foundation model capabilities that exceed our expectations, the US–China AI race remains tight, with China following by just 3-6 months. This is counterintuitive if Scaling Law still holds, as compute remains a key constraint in China, which shows that China's engineering efficiency has not yet hit its own "scaling wall".

China's AI path – High intelligence density per token cost: Beyond smaller model sizes (parameters), we highlight engineering-efficiency gains in: (1) architecture (dense vs. MoE, attention mechanisms), (2) post-training (reinforcement learning, model distillation), and (3) inferencing infrastructure (hardware optimization, KV cache efficiency). The recent launch of DeepSeek-V4 further proved Chinese models' efforts in efficiency. As a result, China AI labs deliver similar model intelligence to the US at only 15-20% of inference costs.

Upsurge in token usage and pricing power: According to OpenRouter, top Chinese LLMs grew their token-usage market share from 5% in April 2025 to 32% in March 2026 (vs. top US models down from 58% to 19%). MiniMax, Zhipu, and BABA reported 4-6x token-usage increases in February and March vs. December on breakthroughs in model intelligence, particularly coding and agentic capabilities. Contrary to concerns that foundation models are being commoditized, they are showing pricing power: Zhipu has roughly doubled prices YTD, and MiniMax doubled its KV cache price. We expect MiniMax to implement a major price hike after the M3 model upgrade.

(Continued)

Gary Yu
Equity Analyst
Gary.Yu@morganstanley.com +852 2848-6918

Lydia Lin
Equity Analyst
Lydia.Lin@morganstanley.com +852 2239-1572

Yang Liu
Equity Analyst
Yang.Liu@morganstanley.com +852 2239-1911

Rebecca Xu
Equity Analyst
Rebecca.Xu@morganstanley.com +852 2848-7359

Joanne Lau
Research Associate
Joanne.CY.Lau@morganstanley.com +852 3963-1592

Tom Tang
Equity Analyst
Tom.Tang@morganstanley.com +852 3963-1860

Eddy Wang, CFA
Equity Analyst
Eddy.Wang@morganstanley.com +852 2239-7339

Kathy Zhu
Research Associate
Kathy.Zhu@morganstanley.com +852 3963-2618

Laura Wang
Equity Strategist
Laura.Wang@morganstanley.com +852 2848-6853

Chloe Liu
Equity Strategist
Chloe.Liu1@morganstanley.com +852 2848-5497

Vicky Wu
Equity Strategist
Vicky.Wu@morganstanley.com +852 3963-3928

GREATER CHINA IT SERVICES AND SOFTWARE

Asia Pacific
Industry View **In-Line**

We believe AI and LLM names will become a much bigger driver of Hong Kong equity markets, reshaping index composition, performance, liquidity, and fund flows. Knowledge Atlas and Minimax are likely to enter the Hang Seng Tech Index on June 8 with a combined 5–7% weight, which would have lifted HSTECH YTD performance by ~5ppt if included at listing. The rebalance could drive US$1.25–1.75bn of passive inflows, while Southbound inclusion (June 8 / August 6) may lift mainland ownership to 9–20% of free float within six months. Strong regulatory support is evident, with tech accounting for 40% of HK IPO fundraising YTD and 43% of the pipeline, reinforcing AI as a durable force in Hong Kong's equity market.

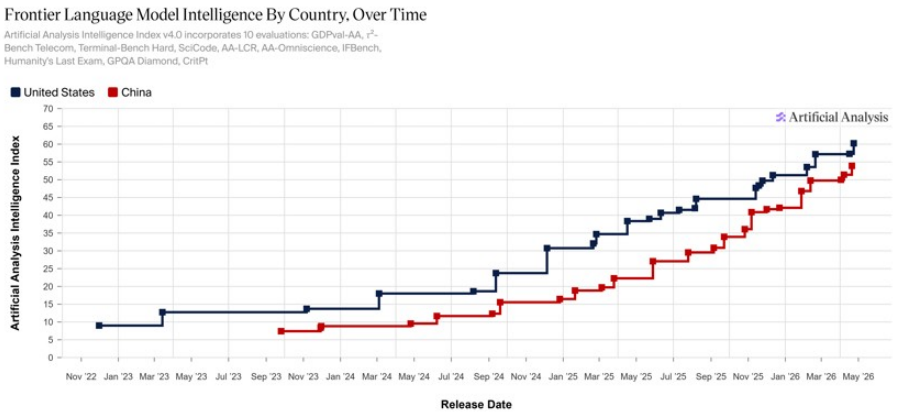
Investment implications: We forecast frontier models' ARR to reach US$1-1.5bn each by end-2026 and US$2.5-5bn by end-2027 (base-bull ranges), implying 3-5x annual growth. Key OWs: **MiniMax** and **Zhipu** for pure-plays, **BABA** for full AI stack capabilities.

The Story in Charts

Exhibit 1: Latest leading Chinese LLMs

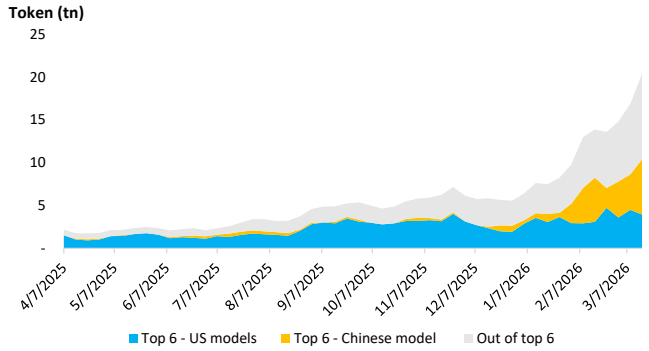
AI lab	Alibaba	MiniMax	Z.ai	Moonshot	DeepSeek
Model	**Qwen3.6-Plus**	**M2.7**	**GLM-5.1**	**K2.6**	**V4**
Release date	Apr-26	Mar-26	Apr-26	Apr-26	Apr-26
License	Proprietary	Open Weight	Open Weight	Open Weight	Open Weight
Total parameter (B)	n.a	230	754	1000	1600
Context window (K)	1000	205	200	256	1000
API price (Rmb/mn tokens)					
Input (uncached)	5.00	2.10	7.00	6.50	12.00
Input (cached read)	n.a	0.42	1.65	1.10	0.10
Output	30.00	8.40	24.28	27.00	24.00
AA Intelligence Index	50	50	51	54	52
ARR (US$mn)	n.a	150 (Feb)	250 (Mar)	n.a	n.a
Valuation (US$bn)	n.a	37	56	18	n.a

Exhibit 2: Frontier LLM intelligence trend by country



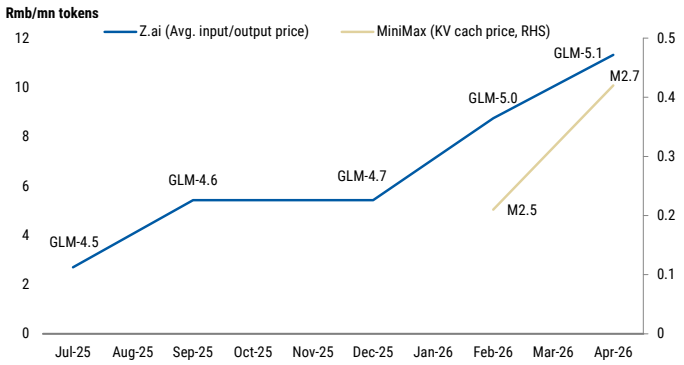
Source: Artificial Analysis, updated by Apr 2026

Exhibit 3: Weekly token consumption of models across OpenRouter



Source: OpenRouter, updated Mar 2026

Exhibit 4: MiniMax and Z.ai flagship model price



Source: Company official pricing, Z.ai API price blended with input and output (3:1)

Exhibit 5: Frontier Chinese LLMs average API price as % of US LLMs

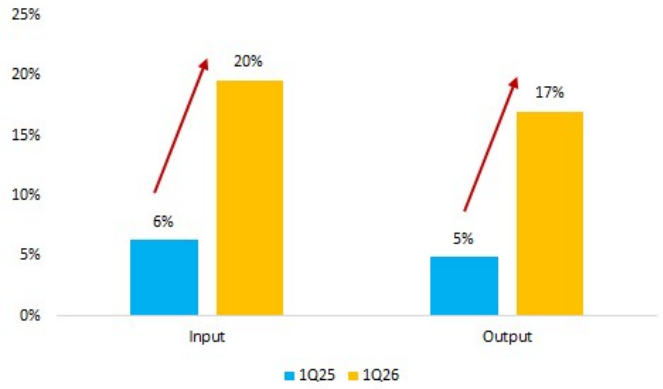
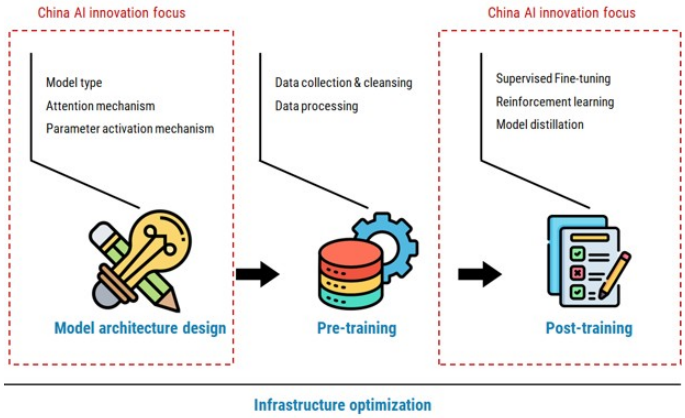


Exhibit 6: China AI innovation focus



Executive Summary

What the market assumes...

- US AI labs and hyperscalers will win long term by outspending on compute.
- China competes mainly on cheaper compute and energy costs.
- Chinese models are commoditized and margin-poor.

What our analysis reveals instead...

- Algorithmic innovation, not cheap energy, is the core Chinese advantage, enabling higher intelligence density per unit of compute. Chinese algorithmic efficiency highlights that compute advantage alone is not a sufficient moat.
- Pricing power has returned for Chinese frontier models, with API prices rising 80% (input) and 36% (output) from 2Q25–1Q26 while margins expanded. Chinese AI labs have also started to launch proprietary models, closing the commercialization loop.
- Pure-play AI labs can compete without hyperscale-level capex by "scaling" up engineering efficiency. Organizational efficiency as an invisible factor is also underrated, and here pure plays outperform hyperscalers.

Algorithm innovation leading to high efficiency

While US AI labs continue to focus on scaling the model size and outspending others, Chinese labs with high AI talent density are exploring algorithm innovation in model architecture design, post-training enhancement and inference efficiency. Techniques such as MoE parameter activation, more efficient attention mechanisms, low-precision training, and novel Reinforcement Learning such as GRPO allow Chinese LLMs to deliver higher intelligence with lower compute intensity. And we believe the cost optimization headroom is far from exhausted with algorithm-level innovation. With only 15% of US hyperscalers' capex in 2025, China continues to close the gap with frontier US AI models.

Leading Chinese AI models starting to show pricing power

We have observed a clear inflection in pricing power among leading Chinese AI labs. Chinese LLM price competition started in 2Q24 and peaked in 2Q25, followed by major AI labs raising their input/output API prices by an average 80%/36% from 2Q25 to 1Q26. Even as US AI labs reduce pricing, Chinese LLMs' API price as a % of US LLMs' API price has gone from 5-6% in 1Q25 to 17-20% in 1Q26. The price hikes of Chinese LLMs are not driven by rising costs but by performance improvements, as shown in an expanding API gross margin, indicating a meaningful shift away from pure commoditization towards performance-based monetization.

Non-linear token consumption growth

Along with gradual price rises, we are seeing exponential growth in token usage by Chinese LLMs, driven by the emergence of agentic and workflow-based applications that are inherently far more token-intensive than traditional chat use cases. Chinese models have been gaining share on OpenRouter, with top LLMs from DeepSeek, Alibaba, MiniMax, Z.ai, Moonshot, and Xiaomi seeing their combined market share rise from 5% in Apr 2025 to 32% in Mar 2026. At the same time, average daily token consumption in China surged more than 1000x to over 140 trillion in just two years.

Investor FAQs

Will foundation models be commoditized? Partially. Yes for average-level models, but no for best-in-class models. Models with superior intelligence/user experience will enjoy a price premium. And the industry is still in a tech-driven phase where major iterations can materially expand capability boundaries, which delays full commoditization.

Are Chinese models simply exporting their energy advantage? No. First of all, computing costs are not simply driven by energy but are related to model algorithms, and Chinese AI labs have invested a lot of effort in lowering cost/token. Second, model pricing is performance-driven rather than cost-driven.

Can Chinese models catch up through model distillation? No. Model distillation may narrow the performance gap at the margin but is not a core path to improving model intelligence. It also carries significant risks including inherent performance ceilings, poor generalization to complex edge cases, model collapse and recursive error amplification, and potential detection or poisoning by frontier AI labs.

Can Chinese model companies break even? Yes, in principle. The gross margin has been positive and continues to rise for two listed names—MiniMax and Z.ai—with revenue outgrowing training costs. But breakeven also requires 1) continuous token consumption scaling and 2) sustained pricing power.

How can pure plays with limited resources compete against hyperscalers? Not by outspending but through innovation and focus. Pure plays can compete on organizational efficiency and concentration on model iteration, while hyperscalers can be distracted by other priorities (i.e. their own ecosystem/infra).

Stock implications

MiniMax (OW): We believe the market is underestimating MiniMax's near-term ARR growth, which we see in line with/stronger than peers at US$1bn, supported by multimodal capabilities, global exposure, and infra advantages. We see multiple catalysts over the next 2-3 months, including product launches, model upgrades, price hikes, and potential index inclusion. We raise our DCF-based PT by 18% to HK$1,100 with upward revenue revisions of 104%, 71% and 99% in 2026-28. Our PT implies 36x 2027 P/S.

Z.ai/Knowledge Atlas (OW): Z.ai demonstrated exceptional near-term momentum with ARR guided to surge 10x within a year from <US$100mn by end-2025 to US$1bn ARR by end-2026, strengthening our OW investment thesis. The company's ability to raise API prices across model iterations signals high recognition of its foundation model intelligence

by customers. Key catalysts includes model upgrades, potential index inclusion and A-share listing. Compute constraints remain the key to watch. We raise our DCF-based PT by 77% to HK$990 with upward revenue revisions of 138%, 138% and 163% in 2026-28. Our PT implies 38x 2027 P/S.

Alibaba (Top Pick in China Internet): We view BABA as the key China AI winner with a comprehensive full stack offerings (chips, infrastructure, model and applications). BABA has been stepping up AI developments recently with the establishment of the Alibaba token hub and key AI model launches/updates. The newly launched Qwen3.6 has topped daily, weekly, and trending leaderboards on OpenRouter. BABA sees Alicloud external revenue of US$100bn in five years (40%+ CAGR) and a long-term cloud margin of 20% (vs high single digit % now), while MaaS will become the key longer-term growth driver (estimated >50% of cloud revenue). In our high-end SOTP analysis of US$350, we estimate Qwen ARR at US$1bn, and value Qwen at US$27bn based on 30x EV/S, implying US$12/share.

Equity market implications

We believe LLMs and other AI/tech related companies will become a much bigger driving force and help reshape Hong Kong market dynamics in terms of index composition, performance, as well as liquidity and fund flows. We expect both Knowledge Atlas and Minimax to be included into Hang Seng Tech index on June 8th, with a joint index weight of 5-7%. Their inclusion would have lifted HSTech Index YTD cumulative performance by 5ppts were they included upon listing. **On the fund flow side, we forecast the Hang Seng Tech rebalancing to bring US$1.25-1.75bn of passive inflows into Knowledge Atlas and Minimax.** We also expect both names to soon be available for Chinese mainland investors through Southbound inclusion, on June 8th for Knowledge Atlas, and on August 6th for Minimax. **We believe Southbound cumulative ownership could become 9-20% of these two companies' free float market cap within the first 6 months of inclusion**, based on historical experience. **Looking forward, we see clear, determined regulatory support for AI-related IPOs in Hong Kong** from the government. As evidence, we note that 40% of IPO fundraising in Hong Kong year to date has been in tech-related sectors, with another 43% of the current IPO pipeline also in tech. We expect AI and tech-related capital market activities and opportunities to stay as a durable force, with meaningful long-term implications for the market's index composition, sector weights, and structural attractiveness to both domestic and international investors.

Exhibit 7: Qwen sensitivity (US$bn)

	US$mn	ARR (US$mn)				
		1,000	2,000	3,000	4,000	5,000
EV/S	30	27,273	54,545	81,818	109,091	136,364
	35	31,818	63,636	95,455	127,273	159,091
	40	36,364	72,727	109,091	145,455	181,818
	45	40,909	81,818	122,727	163,636	204,545
	50	45,455	90,909	136,364	181,818	227,273

Exhibit 8: Qwen sensitivity (US$/share)

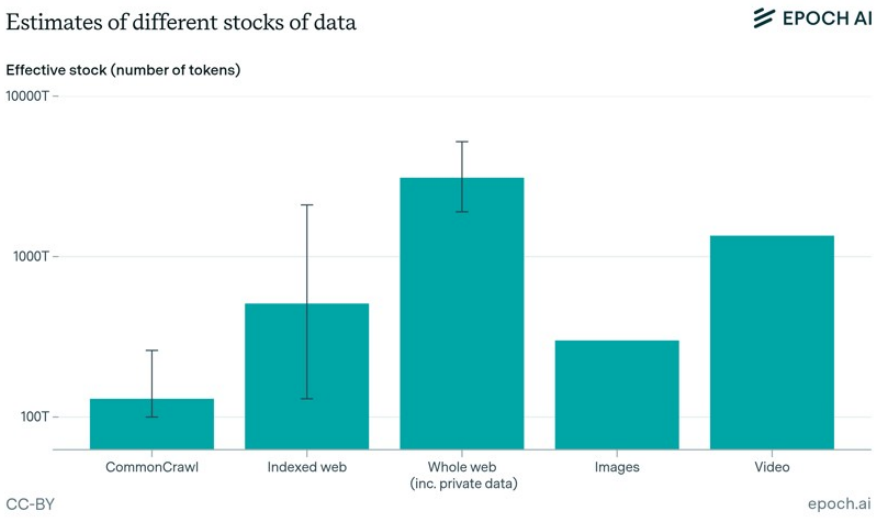
	US/share	ARR (US$mn)				
		1,000	2,000	3,000	4,000	5,000
EV/S	30	11.5	23.0	34.4	45.9	57.4
	35	13.4	26.8	40.2	53.6	67.0
	40	15.3	30.6	45.9	61.2	76.6
	45	17.2	34.4	51.7	68.9	86.1
	50	19.1	38.3	57.4	76.6	95.7

Foundation Model Technology – Where the Intelligence Comes From

How is an AI foundation model trained? The AI foundation model training and development process is directly associated with the intelligence level of AI. Imagining AI as a person, foundation model training can be divided into steps that are similar to the growing and education process of a person:

- **Model design – brain structure:** A person's intelligence development begins with the biological setting of the brain. Developers design the "brain structure" of the model and common settings includes the architecture family (neural system layout - Transformer), attention mechanism design (how neurons route signals - Self Attention/Linear Attention/Bi-directional Attention/Multi-head Latent Attention/ Sparse Attention...), total parameters (brain size - model capacity of 7B/32B/230B/400B/600B/1T...) and attention strategy (how neurons are activated - Dense/MoE).
- **Pre-training – primary/middle school student:** When the model is born with its "biological setting", it needs to go to "school" for education. The pre-training is the process that trains the model with massive general datasets from all kinds of data including website articles, books, papers, codes structured data, multilingual corpora, and sometimes synthetic or filtered mixtures. In this process, the foundation model will pick up the contexts of languages just like a primary/middle school student learning how to read and exploring the common sense and basic knowledge about the world via textbooks. This step involves the most computing power in the entire training process because the general datasets (textbook) for pre-training are massive.

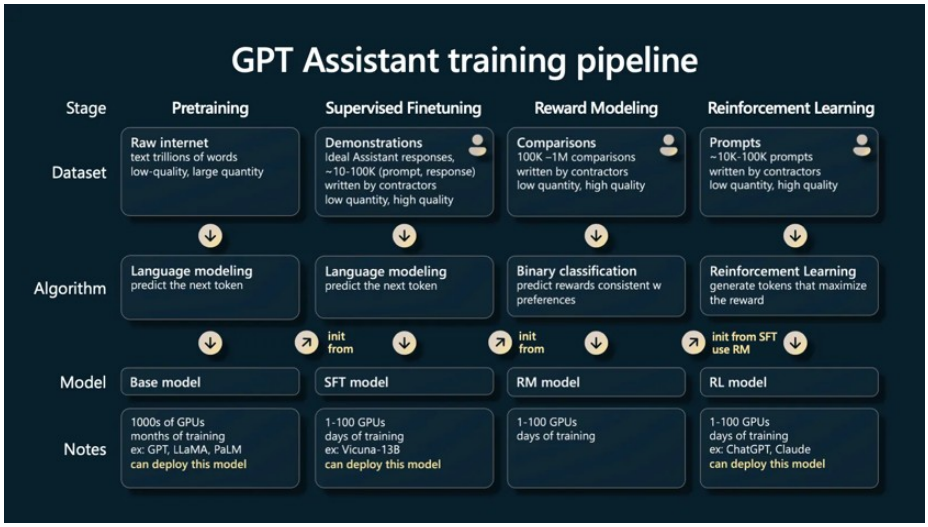
Exhibit 9: Estimates of different stocks of data



Source: Epoch AI

Post-training – college student: As the model graduates from pre-training, it is ready to receive higher-level education in "college". In post-training, developers utilize specific and labeled datasets to the improve depth of model intelligence and make sure the model "thinks" properly, aligned with human values. This process involves Supervised Fine-Tuning/SFT using human-made Q&A data for training the answers, Reward Modeling/RM to evaluate the answers, and Reinforcement Learning (RL) to optimize the model towards preferred behaviors (helpfulness, instruction-following, safety), which can indirectly improve performance on some complex tasks. This step is similar to a college student taking all kinds of exams to improve the answers. After "graduating from the college", the foundation model has got comprehensive abilities to answer complicated questions or execute complex tasks. The model can be further trained in multi-rounds back and forth between SFT and RL with particular data continuing to improve its intelligence level, just like a graduate student pursing a higher degree equipped with professional knowledge in specific verticals. The model is then ready to be put into productivity scenarios – getting a real job. Post-training is typically less compute-intensive than full pre-training, though the cost can be material depending on context length, data volume, and the number of alignment iterations.

Exhibit 10: LLM training process



Source: Microsoft "State of GPT", 2023

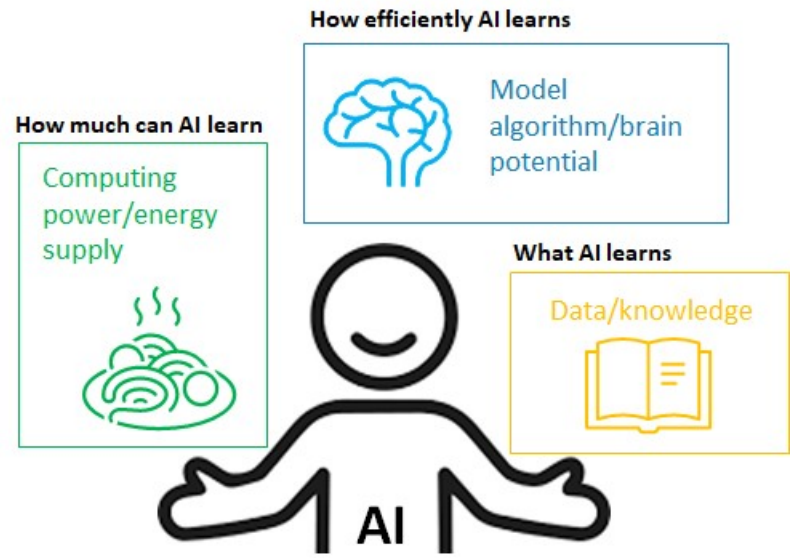
Key factors of foundation model success: computing power, data and algorithms. These three are essential components in the model training process that directly affect the model intelligence level.

- Computing power determines feasible training scales and inference scales of the model, similar to the energy that supports a person to learn, think and complete tasks.
- The data determines the quality and variety of the model, which can be considered as the knowledge base that a person has.
- The algorithm determines the innovation and efficiency of the model, which is similar to a person's brain potential or how efficiently the person can learn, think and complete tasks.

To have a higher intelligence level, a person needs stronger brain potential, a larger

knowledge base (or more specific knowledge base to become professionals) and higher energy supply, making algorithm design, quantity/quality of data, and computing power the key factors in AI model success.

Exhibit 11: AI foundation model key factors



The universal question – are we hitting the scaling wall? Scaling Law is a fundamental principle applied in foundation model training where model performance improves with more parameters (larger brain capacity) and larger dataset sizes (larger knowledge base) which needs to be supported by more computing power (energy supply) during training. However, as model size expansion's marginal return was lower for models growing from 175B to 1T vs. 1.5B to 175B, there was concern of the scaling law hitting a wall, meaning diminishing returns from scaling up the parameters/dataset size/compute in model pre-training. In addition to diminishing return, scaling up models with larger datasets and more computing power also faces constraints from power supply, chipset manufacturing capacity, data scarcity and the latency wall, according to Epoch AI.

Scaling Law still holds but is not the only factor impacting model intelligence anymore. Scaling up the model size still generates a positive return on model intelligence improvement as shown in the Artificial Analysis statistics where larger models with more total parameters generally scored higher in the intelligence index.

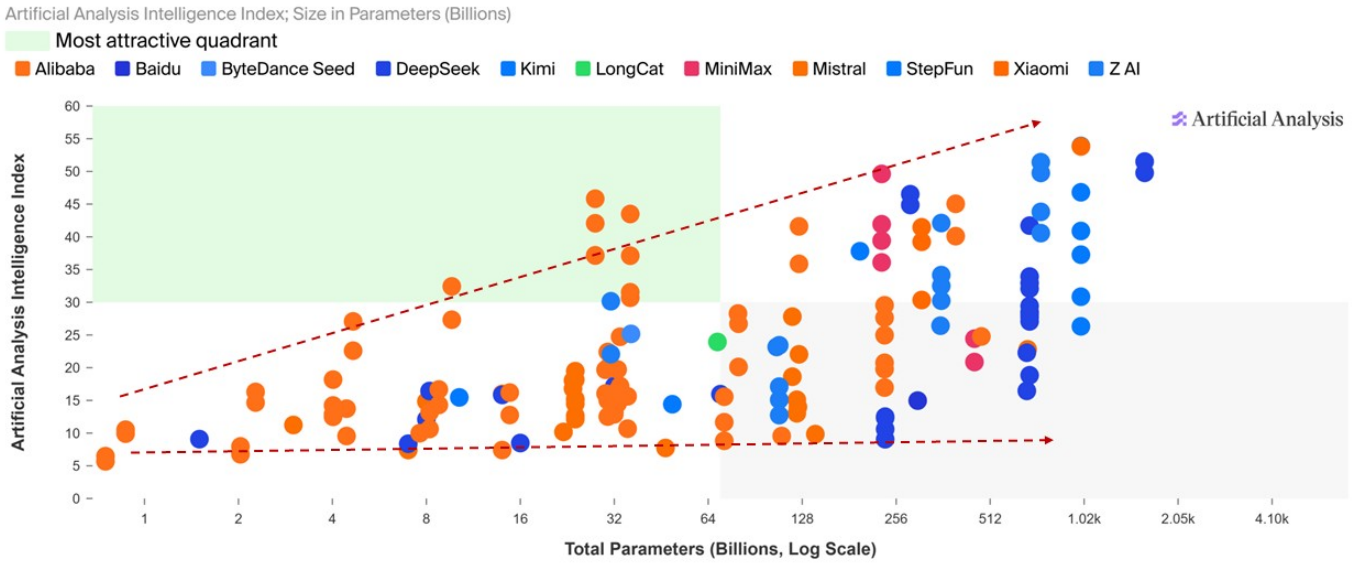
However, the intelligence index range (Exhibit 3) also expands when scale increases, and some smaller models perform even better than large models. Scaling relationships show that, holding other factors constant, performance often improves with larger models and more training data, but frontier gains increasingly depend on data quality, architecture/optimization, post-training strategy, and compute/data allocation.

Funding for model upscale training is also becoming so massive that developers will need to evaluate ROIC to decide on whether to allocate spending to model scaling or somewhere else to improve model intelligence. Hence, developers are exploring more approaches to improve model performance alongside building bigger models and training with larger datasets. As Russ Salakhutdinov, leading AI researcher and professor at

Carnegie Mellon University, said: "Big tech excels at large-scale engineering and the scaling of foundation models, but further progress towards superintelligence will require new breakthroughs in architecture, optimization, and the efficient use of data."

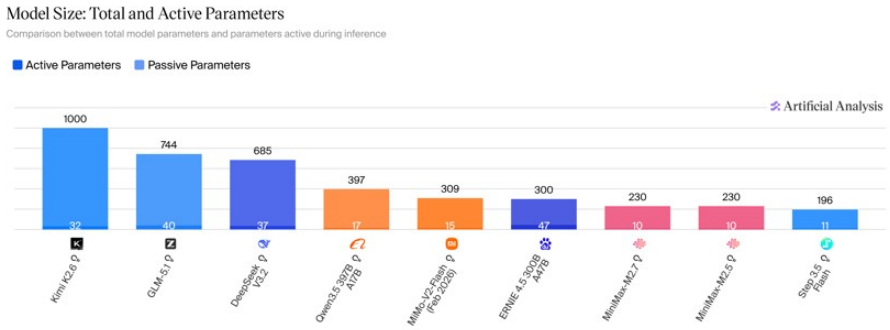
We identified some of the approaches adopted by frontier model labs throughout pre-training, post-training and inference side for model intelligence improvement in China's Way: More Bang For The Buck .

Exhibit 12: Frontier LLMs' intelligence vs. total parameters



Source: Artificial Analysis. Updated Apr 2026.

Exhibit 13: Major Chinese open-source LLM size



Source: Artificial Analysis. Updated Apr 2026

Talent-laden Chinese AI labs. To improve model intelligence without massive scaling via new breakthroughs in architecture, optimization, and the efficient use of data requires sufficient support from AI talent. We see AI talent supply as a key advantage of Chinese AI labs, and leading persons play a crucial role in this fast-changing industry.

Key persons at China AI labs

AI lab	Key person	Age	Role	Introduction
Alibaba	Junyang Lin	33	Former head of Qwen team	Junyang Lin is a Principal Researcher at Alibaba, where he plays a key role in the development of large-scale foundation and multimodal models within the Qwen team at DAMO Academy. He received his education from Peking University, where he built a strong foundation in computer science and artificial intelligence. Lin joined Alibaba in 2019 as a Researcher at DAMO Academy and has since advanced to Principal Researcher. His work centers on large language models (LLMs), multimodal systems, and scalable AI architectures, contributing to Alibaba's core AI research and commercialization initiatives. He left Alibaba in Mar 2026.
Alibaba	Jingren Zhou	N/A	Head of Qwen BG, Chief AI Architect	Jingren Zhou is a Partner at Alibaba Group, having joined the company in 2015. He is a senior executive with extensive experience spanning both technology and business. Early in his tenure, he led Alibaba Cloud's Institute of Data Science and Technologies (iDST, the predecessor to DAMO Academy), before moving to the e-commerce division, where he oversaw search, recommendation, and advertising systems. In late 2020 he transitioned to Ant Group and, after more than a year, returned to Alibaba Cloud as CTO and Deputy Dean of DAMO Academy. In his latest role, Zhou leads the Tongyi Large Model Business Unit (upgraded from the Tongyi Lab) and serves as Chief AI Architect of the Group's Technology Committee, driving Alibaba's development in large models and AI infrastructure.
Bytedance	Yonghui Wu	N/A	Head of Foundation AI Research (Seed)	Yonghui Wu leads foundational AI research at ByteDance, heading the company's large-model organization known as Seed. He earned his Ph.D. from the University of California, Riverside in 2008. Following his doctoral studies, Wu joined Google, where he spent more than a decade in senior technical leadership roles. His work spanned Google Search ranking systems, Google Brain, and large-scale deep learning research. He later held a Vice President–level research leadership role associated with Google DeepMind. At ByteDance, he is responsible for advancing core large-model research and next-generation AI systems.
Tencent	Shunyu Yao	28	Chief AI Scientist, Head of AI Infra Department and LLM	Shunyu Yao serves as Chief AI Scientist at Tencent, where he leads strategic large-model research and AI infrastructure initiatives. He received his undergraduate education at Tsinghua University and completed graduate studies at Princeton University, focusing on artificial intelligence and machine learning. Prior to joining Tencent, Yao worked as a researcher at OpenAI, contributing to cutting-edge research in large language models and reasoning systems. At Tencent, he oversees foundational AI research and works closely with executive leadership to shape the company's long-term AI strategy.

Baidu	Haifeng Wang	55	CTO	Haifeng Wang is the Chief Technology Officer of Baidu and a central architect of the company's AI and search technology strategy. He received his B.S., M.S., and Ph.D. degrees in Computer Science from Harbin Institute of Technology. Before joining Baidu in 2010, Wang served as Chief Research Scientist at Toshiba's R&D Center. At Baidu, he has held multiple senior leadership roles, including Vice President and Senior Vice President, before being appointed CTO in 2019. He has led major advancements in search technology, natural language processing, and large-scale AI platforms, including Baidu's foundational model initiatives.
DeepSeek	Wenfeng Liang	41	Founder, CEO	Liang Wenfeng is the Founder and Chief Executive Officer of DeepSeek. He earned both his Bachelor's and Master's degrees in Engineering from Zhejiang University, specializing in electronic information engineering and information and communication engineering. Prior to founding DeepSeek, Liang co-founded High-Flyer, a quantitative investment firm known for its early adoption of large-scale machine learning infrastructure. Building on his experience in high-performance computing and AI-driven quantitative systems, he established DeepSeek to focus on large language model research and advanced AI infrastructure development.
MiniMax	Junjie Yan	37	Founder, Chair, CEO, CTO	Junjie Yan is the Founder, Chairman, and Chief Executive Officer of MiniMax. He received a Bachelor's degree in Mathematics from Southeast University and earned his Ph.D. in Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences. He subsequently conducted postdoctoral research at Tsinghua University. Before founding MiniMax, Yan spent more than six years at SenseTime, where he served as Vice President and Deputy Head of the Research Institute. His industry experience spans computer vision, deep learning research, and large-scale AI commercialization. At MiniMax, he leads the company's strategic direction in foundation models and generative AI systems.
Z.ai	Jie Tang	49	Co-founder, Chief Scientist	Jie Tang is the Co-Founder of Z.ai AI and a Professor of Computer Science at Tsinghua University. He received his Ph.D. in Computer Science from Tsinghua University in 2006. Tang founded the Knowledge Engineering Group (KEG) at Tsinghua, where his research focuses on knowledge graphs, social network analysis, and artificial intelligence. Z.ai AI originated from his academic research initiatives and has grown into a leading foundation-model company in China. Tang bridges academia and industry, contributing to both fundamental AI research and large-scale commercialization.
Moonshot	Zhilin Yang	34	Founder, CEO	Zhilin Yang is the Founder and Chief Executive Officer of Moonshot AI. He earned his Bachelor's degree from Tsinghua University and completed his Ph.D. at Carnegie Mellon University, where his research focused on natural language processing and machine learning. Prior to founding Moonshot AI, Yang held research positions at Google Brain and Meta AI, contributing to advancements in large language models and representation learning. At Moonshot AI, he leads the development of large-scale foundation models and the Kimi AI assistant, positioning the company as a prominent player in next-generation AI systems.

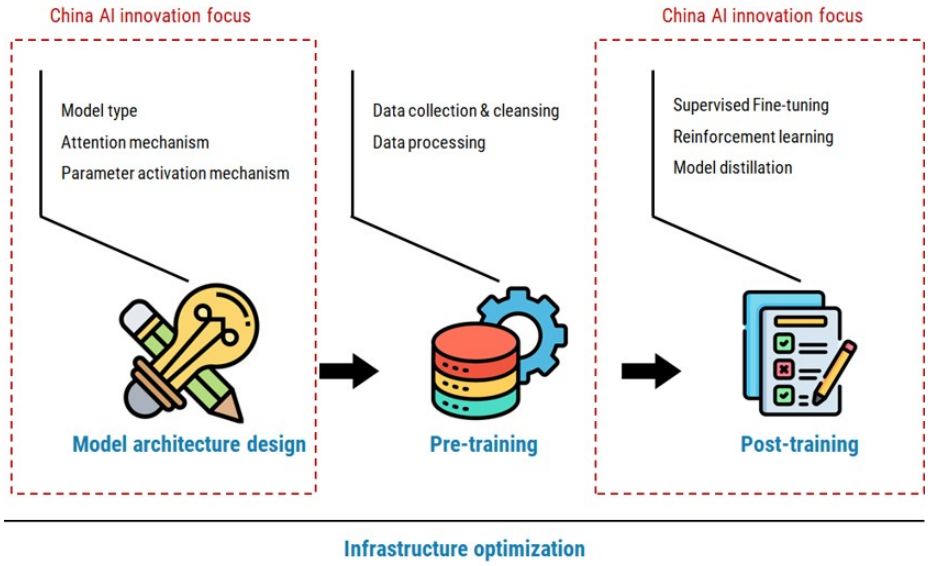
Xiaomi	Fuli Luo	31	Head of AI Model team	Fuli Luo is a leading AI researcher at Xiaomi's AI Lab, where she leads development of the MiMo foundation model. She received her bachelor's degree from Beijing Normal University and a master's degree in Computational Linguistics from Peking University. Luo began her career at Alibaba DAMO Academy through the "Ali Star" program, where she led the development of the VECO multilingual pretraining model and contributed to the open-sourcing of the AliceMind framework. She later joined DeepSeek in 2022, serving as a key contributor to the DeepSeek-V2 Mixture-of-Experts model. In late 2024, she joined Xiaomi to lead its LLM research efforts.

China's Way: More Bang For The Buck

Relative compute constraints and ROI discipline increase the premium on algorithmic and engineering efficiency of Chinese AI labs, though leading players still invest materially in training/serving infrastructure. Meanwhile, organizational efficiency is also becoming more important to AI competition.

AI model development shifting emphasis from pre-training scaling to model architecture innovation + post-training RL + inference efficiency. Developers are increasingly exploring algorithm-based approaches to improve model performance instead of simply building bigger models and training with larger datasets. Below we highlight select algorithm approaches adopted by frontier model labs and especially Chinese AI labs to show how model intelligence can be improved in a cost-effective way.

Exhibit 14: China AI innovation focus



Summary of major Chinese AI lab algorithm innovations		
Chinese AI lab	Model	Algorithm innovation
Model architecture design		
MiniMax	abab 6	**Attention strategy:** First adopter of Mixture of Experts (MoE) with large parameter in China
MiniMax	abab 6.5	**Attention mechanism:** Lightning Attention
DeepSeek	DeepSeek-V2	**Attention mechanism:** Multi-head Latent Attention
DeepSeek	DeepSeek-V3	**Attention strategy:** DeepSeek MoE with Shared Experts and Routed Experts
DeepSeek	DeepSeek-V3	**Attention mechanism (improvement):** Manifold-Constrained Hyper-Connections (mHC)
DeepSeek	DeepSeek-V3.2	**Attention mechanism:** DeepSeek Sparse Attention (DSA)
Moonshot	Kimi K2.5	**Attention mechanism:** Kimi Linear/Kimi Delta Attention (KDA)
Moonshot	Kimi K2.5	**Attention mechanism (improvement):** Attention Residual
Z.ai	GLM-5	**Attention mechanism:** MLA (Multi-latent Attention)-256 + Muon Split
DeepSeek	DeepSeek-V4	**Attention mechanism:** Hybrid attention architecture that combines Compressed Sparse Attention (CSA)and Heavily Compressed Attention (HCA)
Pre-training		
Z.ai	GLM-1	**Pre-training task design:** Autoregressive Blank Infilling
DeepSeek	DeepSeek-V3	**Model quantization:** Dual-scale FP8 Training
Alibaba	Qwen 3	**Data curation & diversity**: Instance-level mixture optimization via automatic dataset evaluation system on 10T-level token
Alibaba	Qwen 3	**Synthetic data:** Multi-modal data expansion with Native Dynamic Resolution of Qwen-VL
Moonshot	Moonlight	**Model scaling techniques:** Muon (MomentUm Orthogonalized by Newton-schulz) optimization on large scale models
Post-training		
DeepSeek	DeepSeel-V3 and R1	**Reinforcement Learning:** Group Relative Policy Optimization (GRPO)
Moonshot	Kimi K1.5	**CoT scaling:** long2short
Z.ai	GLM-5	**RL optimization:** Scalable Lightweight Infrastructure for Multi-agent Environment (SLIME)
Whole process enhancement		
Alibaba	Qwen-Math/Qwen2.5	**SFT+RL:** Rationale Consistency
Inference efficiency management (KV cache compression)		
DeepSeek	DeepSeek-V2	**Memory module:** Engram
Moonshot	Moonshot-v1	**KVCache-centric disaggregated architecture:** Mooncake

Model architecture design: Transformer enhancement

Transformer is the mainstream LLM model architecture. Transformer originated from Google's "Attention is All You Need" which provides a Self-Attention-based mechanism to model the relationships and dependencies between all words (tokens) in a sequence simultaneously, regardless of their distance from one another, enabling AI models to understand and learn from human text. Model developers improve the details in Transformer to deliver better intelligence and efficiency.

Attention strategy. The attention strategy is the part of the model architecture design which determines which subsets of parameters are active per token.

- Dense model – a generalist: Dense means the model activates all parameters for every input involving the full network for every forward pass, resulting in intensive computation but straightforward and thorough results. It's like a person that

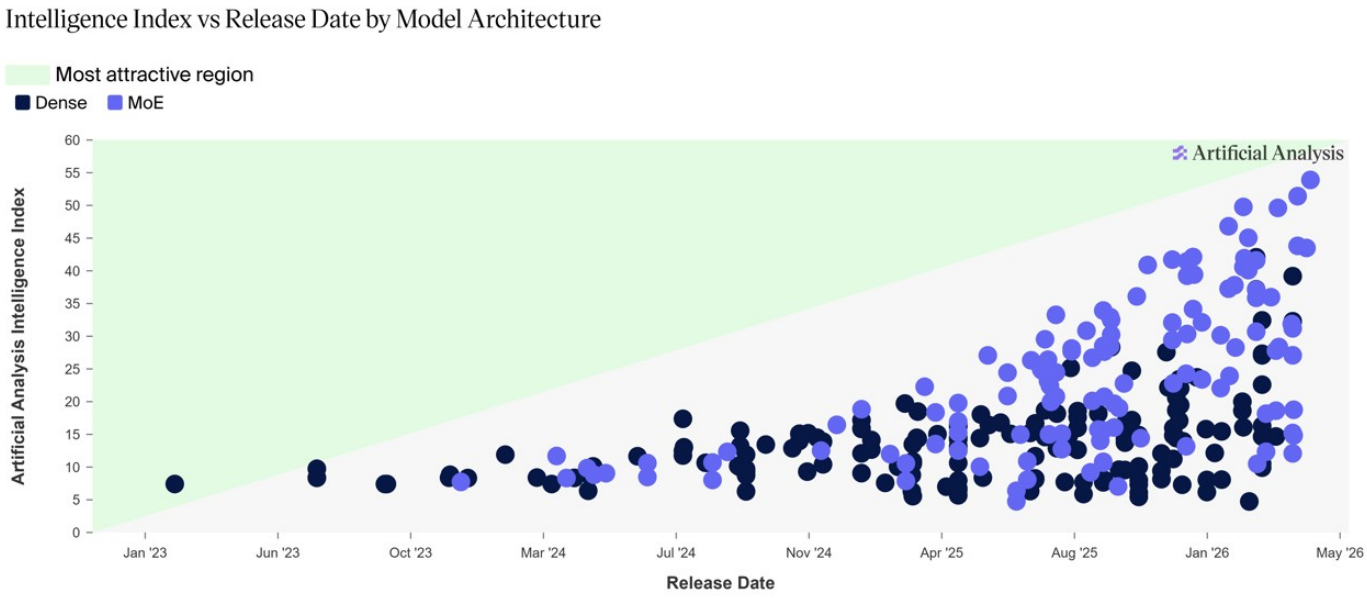
knows about everything and uses their own knowledge base to answer questions. Typical dense models includes Llama 3.

- MoE (Mixture-of-Experts) model – a person supported by many experts: MoE means the model only activates a subset of parameters ("experts") per input by routing mechanism, improving compute efficiency per token. It's like a person with a limited knowledge base but access to experts to help. DeepSeek-V3 and MiniMax-M1 are early adopters of MoE.

MoE is an increasingly common design due to its outperformance in model intelligence density to inference costs. MoE-style conditional computation was introduced in prior research in 2017 and later scaled in transformer systems (e.g., Switch Transformers, 2021), and further optimized by DeepSeek-V3 in 2024.

In the contrary, despite high inference costs, dense models can be simpler to train and serve (no routing).

Exhibit 15: Model intelligence vs. model architecture



Source: Artificial Analysis, Updated as of Apr 2026

Attention mechanisms. Attention mechanisms are the core technology of Transformer-based AI models in determining how a model "reads" the words and "understands" the context of the inputs. Imagining the inputs as a book, the original transformer attention mechanism (Softmax Attention) requires the model to read all of the words and sort out the relationship between each word in the book to understand the content, which is slow and consumes exponential amounts of computing power. AI model developers adjusted the attention mechanism to improve this "reading" efficiency and prevent excess computing power use.

- Sparse Attention: Instead of reading the whole "book", the model focuses on key chapters to summarize the content. A typical Sparse Attention mechanism is Sliding Window Attention (SWA) represented by Mistral 7B and DeepSeek Sparse Attention (DSA) represented by DeepSeek.

- Linear Attention: Applying mathematical transformations of the Softmax attention formula to make sure computation grows linearly with input length. A typical Linear Attention mechanism is Lighting Attention represented by DeepSeek V2, MiniMax M series and Qwen 3.5.
- Multi-head Latent Attention: Compresses the words in the "book" into more concise information and saves them for future reading, which accelerates computing speed and releases more memory space, enabling the model to read more context. This approach was invented by DeepSeek.

To balance the accuracy and efficiency of the attention mechanism, more and more models are adopting Hybrid Attention to apply different attention mechanism to different layers.

Pre-training: data quality improvement and model quantization

Dataset quality improvement. Dataset quality has become more important than dataset size in model pre-training. Dataset quality and variety determine the model's quality and variety, hence datasets are an essential and highly confidential asset of AI labs.

- Data cleaning, curation and diversity: This is the key step to improving data quality. Developers utilize algorithm and engineering methods to de-duplicate, label, evaluate and diversify raw data to fetch a high-quality dataset.
- Synthetic data: Since human-generated data has largely been explored and most is low-quality data that may even hamper model performance, developers use AI models to generate textbook quality synthetic data for training at the late stage of pre-training to improve model intelligence. Typical models are Llama 3 and Kimi K2.

Post-training: novel RL

Reinforcement Learning (RL). RL has become a core process to improve model intelligence in post-training with all kinds of novel algorithms being developed. RL trains models by evaluating and rewarding model answers that are considered "better" to teach the model how to think and output. RL can be divided into RLHF (Reinforcement Learning from Human Feedback) and RFVR (Reinforcement Learning from Verifiable Rewards), depending on whether human preferences are involved in evaluating the model's answers.

1) Reinforcement Learning from Human Feedback (RLHF). As the name suggests, RLHF uses human feedback to refine a model's output through manual data labeling. RLHF is often utilized in general-purpose foundation models which need to align with human values in all kinds of output.

- PPO (Proximal Policy Optimization) – hiring a lot of AI tutors to ensure high quality exam results: PPO is a mature algorithm mostly applied in RLHF (can also be applied in RFVR) where developers developed Reference Model (for the student to take a reference), Reward Model (to evaluate exam results) and Critic Model (to forecast future exam result) to train the Policy Model (the foundation model being trained). This algorithm can largely ensure the stability of model performance with high quality. But the cons is the high training cost and massive

memory space needed. Typical models using PPO are Llama 2/3.

- DPO (Direct Preference Optimization) – learning and answering all based on human evaluation: DPO removes the Reward Model and directly updates models offline with data based on human preferences after each "exam". DPO is a cost-saving algorithm with much less memory space needed compared to PPO. However, models trained with DPO will be highly limited by the human-provided data and perform poorly on logic questions. Typical models are Zephyr-7B, Llama 3.1 and Mistral.

2) Reinforcement Learning from Verifiable Rewards (RLVR). Thinking of the foundation model being trained as a student, RFVR assesses answers based on objective standards such as mathematical rules without human intervention. RFVR works very well in solving complex logical problems such as mathematical questions and coding. It is in the RLVR process that foundation models can evolve into reasoning models with the Chain-of-Thought (CoT) and Self-Correction capabilities. DeepSeek-R1 is an early adopter of RLVR.

- MCTS (Monte Carlo Tree Search) – evaluating every thinking step to secure high quality exam results: The model being trained will explore all possible ways to solve problems, evaluate each step to follow the optimal moves and repeat the exploration. MCTS rewards right step instead of right outcome and encourage models to expand the thinking territory, reaching higher level of intelligence. However, it's also extremely costly.
- GRPO (Group Relative Policy Optimization) – rewards answers that are better than average. The GRPO algorithm removes the Critic Model on the basis of PPO. It instead asks the model to output a range of answers to a query and ranks the answers that are better than the average of total answers. The model derives strong thinking capabilities from this process with lower costs, reaching a balance between model scale and training costs. DeepSeek invented this approach and first adopted it in DeepSeek-V3, followed by further optimization in DeepSeek-R1.

Inference scaling and efficiency improvement

Test-time compute scaling: The more time a model is given to think, the better its answer will be. Inference scaling refers to improving a trained model's performance at inference time by allocating more computing power or using more sophisticated strategies during generation. Through inference scaling, a small-sized model with thinking can outperform a large-sized model without thinking. Inference scaling is the latest focus in foundation model development. There are two major approaches in inference scaling now.

- Search-based scaling: The model will think about multiple answers and choose the best before outputting. Best-of-N and Beam Search/Tree Search are typical methods.
- CoT-based scaling: The model will think in more steps before outputting. Long CoT (extend the steps) and Self-Correction are typical approaches.

Inference efficiency improvement – handling long context: Context length is the maximum amount of information (measured in tokens) that a model can process at once during inference, similar to how much information a person can read/say in each round of

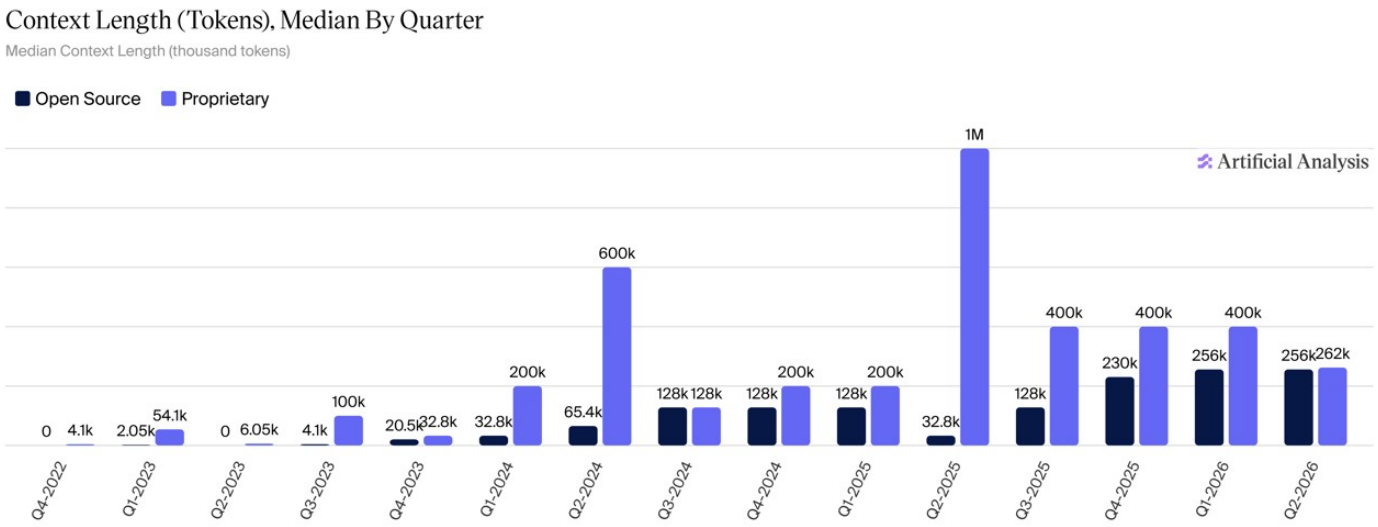
conversation. Longer context length means the model is able to process longer conversations, larger documents (PDFs, codebases) and more retrieval-augmented generation (RAG). Meanwhile short context length often leads to older information being truncated or forgotten.

The context length of foundation models is generally growing longer to reach better model performance, partially driven by inference scaling. Currently proprietary models are at 400k tokens context length and open-weight models are at 200k in median. However, longer context length also consumes more computing power.

The bottleneck of long context – KV Cache. During the inference process, KV (Key, Value, which is intermediate result during token processing) will be cached to save the previous context and avoid repetitive computing. The KV Cache will take up VRAM (Video Random Access Memory) space in GPUs. The longer the context length is, the more space the KV Cache will take in GPUs and the more GPUs will be needed.

Chinese AI labs' efforts in inference efficiency improvement. With limited computing power, Chinese AI labs came out with algorithm innovation in model training including attention mechanisms and model quantization to effectively compress the KV Cache. In addition, Chinese AI labs are also improving the system engineering to enhance inference efficiency, such as Mooncake (a distributed cache management system) created by Moonshot, Engram by DeepSeek, Paging technology enhancement by Alibaba and a combination of distributed caching and searching/recommendation mechanism by Bytedance, thanks to the highly developed internet network and cloud infrastructure in China.

Exhibit 16: Median model context length



Source: Artificial Analysis, updated as of Apr 2026

Model infra optimization

Model quantization. A Floating Point Operation (FLOP) is a unit to measure the amount of computing in model training and inference. Each FLOP represents one basic computation. If we compare a model to a person, the total number of FLOPs is like the number of thinking steps a person needs to learn and use knowledge.

- When calculations are done with higher precision, the model handles information more accurately, but each computation uses more memory, more energy, and takes more time. For example, with FP32, each computation uses 32 bits of memory.
- When calculations are done with lower precision, the model processes information less precisely, but it can learn and respond faster while using less energy and less hardware. For example, with FP16, each computation uses 16 bits of memory—half as much as FP32—which makes it faster and reduces GPU requirements.

To improve computing efficiency, model developers are lowering computing precision via model quantization (a mathematical method) from FP32 to FP8 after the model was being trained. FP8 quantization is the most cutting-edge technology with models being DeepSeek-V3 and DeepSeek-R1. Chipsets that support FP8 are also advanced GPUs including Nvidia's Hopper and Blackwell series.

Other infra optimization methods: Common methods also include: 1) Parallelism (e.g. Data Parallelism, Tensor Parallelism, Pipeline Parallelism, Sequence Parallelism); 2) Memory Optimization (e.g. Activation Checkpoint); 3) Kernel Optimization; 4) Communication Optimization, etc.

Model distillation

What is model distillation? Model distillation is a way of taking a highly skilled, larger but expensive "teacher" model and training a smaller, faster, and more cost-efficient "student" model to behave similarly. Imagine a top equity analyst who has deep expertise but is costly and has limited time for consultation—distillation is like training a junior analyst to replicate most of that senior analyst's judgment at a fraction of the cost and time. It enables companies to deliver AI capabilities at scale with lower infrastructure costs, faster response times, and improved margins, without sacrificing too much performance. Model distillation is a very common practice in the AI industry throughout the model training process, mostly in post-training stage. AI labs would distill their own large-parameter and expensive models to develop small-parameter and cheap models for more general scenarios.

Can Chinese AI models simply distill US frontier models to catch up? No. Model distillation may help to narrow the performance gap between China and the US but is not the core approach needed to improve model intelligence and it also bears risks.

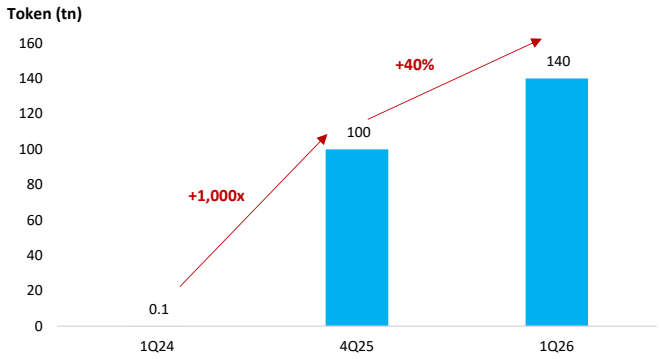
1. The student can never beat the teacher: The essence of distillation is imitation. It doesn't expand the intelligence territory but simply copies existing knowledge.
2. "How to think" can't be learned: The student model can mimic the output of the teacher model but can't distill the reasoning process. When the student model deals with complex edge cases that the model distillation process didn't cover, performance will be poor.

3. Model collapse risks and recursive contamination risks: If model distillation becomes the major way to improve model performance without additional raw data input in pre-training, output would become increasingly monotonous (model collapse) and output losses in teacher models could be amplified (recursive contamination).
4. Model poisoning risks: Frontier AI labs could also detect distillation behaviors via approaches such as Statistical Watermarking, Fingerprinting and Honey-pots. They could even poison the data output of their own models to hinder the distilling models' performance.

Latest Trend of China LLMs: Closing the Commercialization Loop

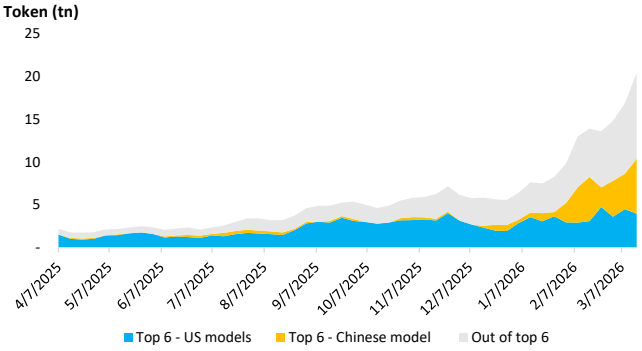
Surging token consumption. China's token usage has scaled rapidly with average daily token consumption standing at approximately 100bn in early 2024, increasing to around 100tn by the end of 2025, and surpassing 140tn by March 2026—representing more than a 1,000x increase over two years, according to National Data Administration. In addition, Chinese models have also been gaining global market share since end-2025, leading the significant token growth on OpenRouter, primarily due to the explosion in demand for OpenClaw. MiniMax disclosed that M2's tokens grew 6x in Feb 2026 vs. Dec 2025.

Exhibit 17: China's daily token consumption



Source: National Data Administration

Exhibit 18: Weekly token consumption of models across OpenRouter



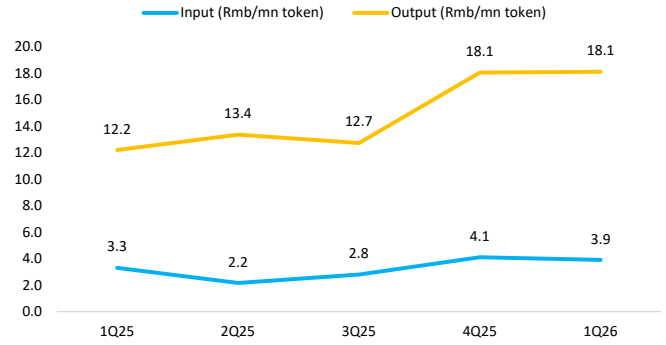
Source: OpenRouter, updated by Mar 2026

Frontier Chinese models have started to raise prices, ending the commoditization trend. While the low API price of Chinese LLMs is key to their growing market share in agentic scenarios, where users are more price-sensitive due to massive token consumption, Chinese LLMs have ended their price war and are in fact raising prices.

A recap of the LLM price war: It began with DeepSeek-V2 priced at Rmb1/mn tokens in 2Q24. Since then, Chinese AI labs have generally cut API prices by 70-90% across their product offerings, and this quick deflation soon made Chinese frontier LLMs 10x cheaper than US peers. The price war also raised concerns of foundation model commoditization.

The price war seems to have peaked in 2Q25, followed by leading Chinese AI labs including Alibaba, Bytedance, Baidu, Tencent, MiniMax, Z.ai, Moonshot, and DeepSeek raising API prices for almost each new flagship model since 3Q25. From 2Q25 to 1Q26, the average input price has been raised by 80% and the average output price went up by 36%. Z.ai is leading the trend with API prices up 200%+ from GLM-4.5 in 2Q25 to GLM-5 in 1Q26. At the same time, US LLMs are generally seeing a downward trend in API prices, narrowing the gap with Chinese LLMs. Chinese LLMs' API pricing was just 5-6% of US LLMs in 1Q25 but has risen to 17-20% in 1Q26.

Exhibit 19: Frontier Chinese LLMs average API price



Qwen3, Qwen3-Max, Qwen3.5), Bytedance (Doubao-1.5-pro, Doubao-seed-1.6, Doubao-seed-1.8, Doubao-seed-2.0-pro), Baidu (Ernie 4.5, Ernie x1, Ernie x1.1, Ernie 5.0-Thinking-Preview), Tencent (Hunyuan-T1, HY-2.0), MiniMax (MiniMax-Text-01, MiniMax-M1, MiniMax-M2, MiniMax-M2.1, MiniMax-M2.5, MiniMax-M2.7), Z.AI (GLM-4-Plus, GLM-4.5, GLM-4.6, GLM-4.7, GLM-5), Moonshot (Kimi-latest-128k, K2, K2.5), DeepSeek (DeepSeek-R1, DeepSeek-V3.2). Input price averaged with cached and uncached and all context length.

Exhibit 20: Frontier Chinese LLMs average API price as % of US LLMs

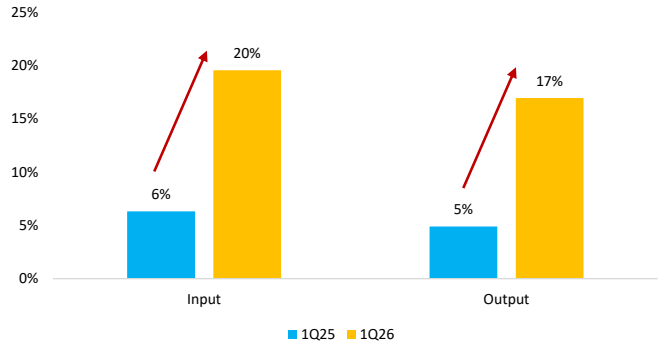


Exhibit 21: Z.ai gross margin forecast

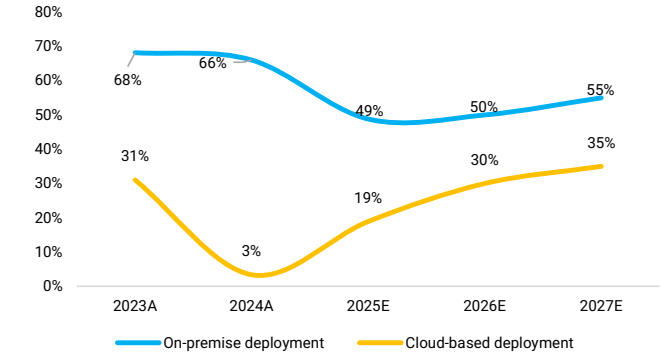
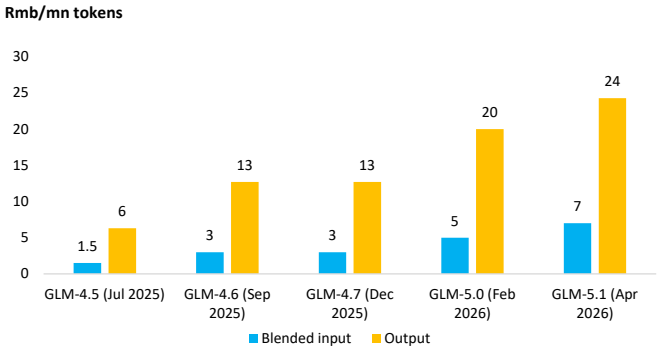


Exhibit 22: Z.ai flagship model API price



Source: Company official pricing, input price is uncached, blended for all context length

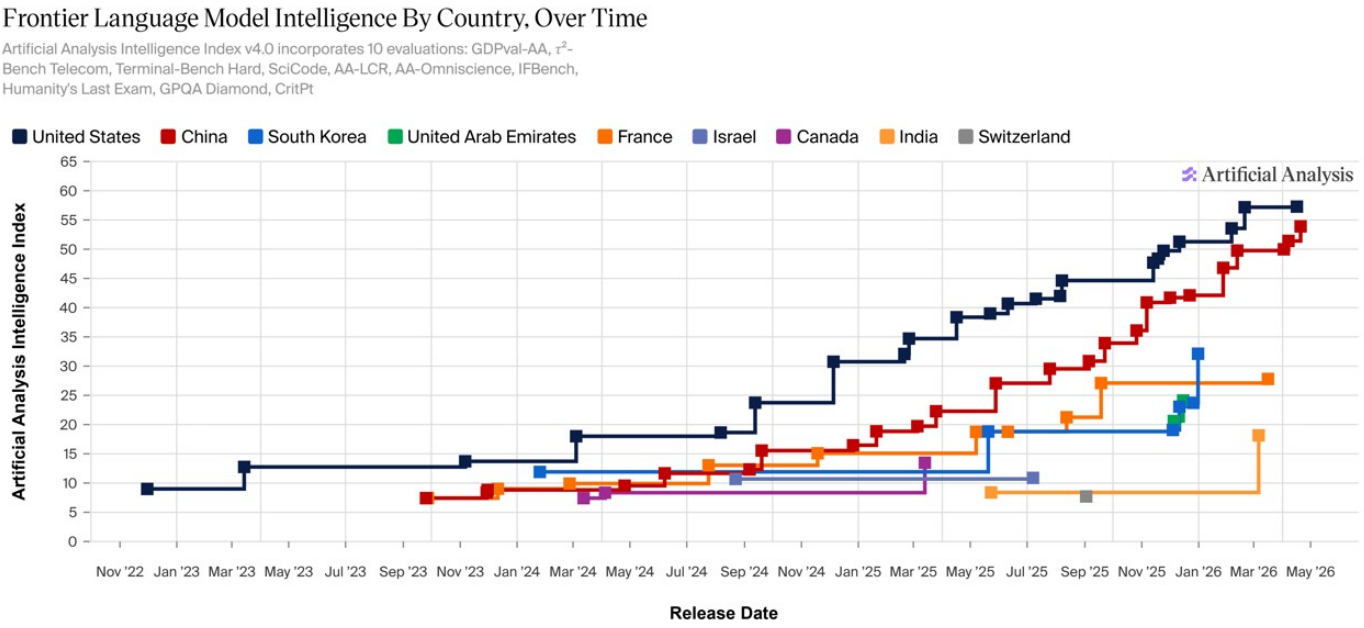
As model intelligence continues to see breakthroughs, we see a reverse trend of model commoditization, as model API prices begin to be driven by performance rather than inference costs – high-performing models can enjoy a price premium because the value-add of best-in-class models is much more than average models for end users. Chinese LLMs being able to raise prices indicates a significant improvement in performance rather than rising inference costs – the rising gross margins of MiniMax and Z.ai demonstrate this.

Launching proprietary model. While Chinese LLMs are still leading the world's open-source community as US LLMs are all proprietary, we have also seen Chinese AI labs start to launch proprietary models in late 2025 and early 2026. These AI labs include both hyperscalers and pure play companies. We believe an open-source strategy is effective for market share gains while proprietary models can optimize commercial value. Chinese LLMs are turning from 100% open-source to a more balanced approach aimed at commercialization, which also indicates the improving competitiveness of Chinese model performance.

Latest proprietary LLMs by Chinese AI labs

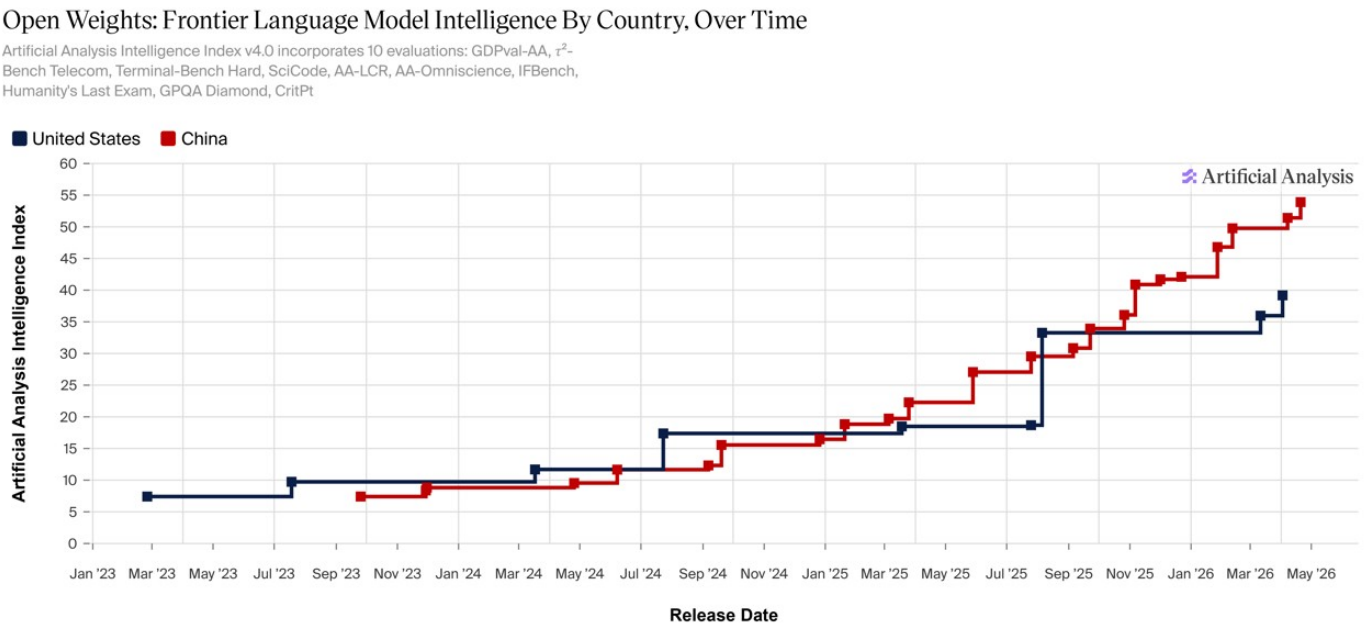
AI lab	Launch date	Model
Bytedance	Feb-26	Doubao-seed-2.0-pro
Alibaba	Jan-26	Qwen3-Max-Thinking
Baidu	Nov-25	Ernie 5.0-Thinking-Preview
Tencent	Nov-25	HY-2.0 Think
MiniMax	Mar-26	MiniMax-M2.7
Z.ai	Mar-26	GLM-5-Turbo
Xiaomi	Mar-26	MiMo-V2-Pro
Alibaba	Apr-26	Qwen3.6-Max
Xiaomi	Apr-26	MiMo-V2.5-Pro

Foundation Model Industry Landscape – China vs. US



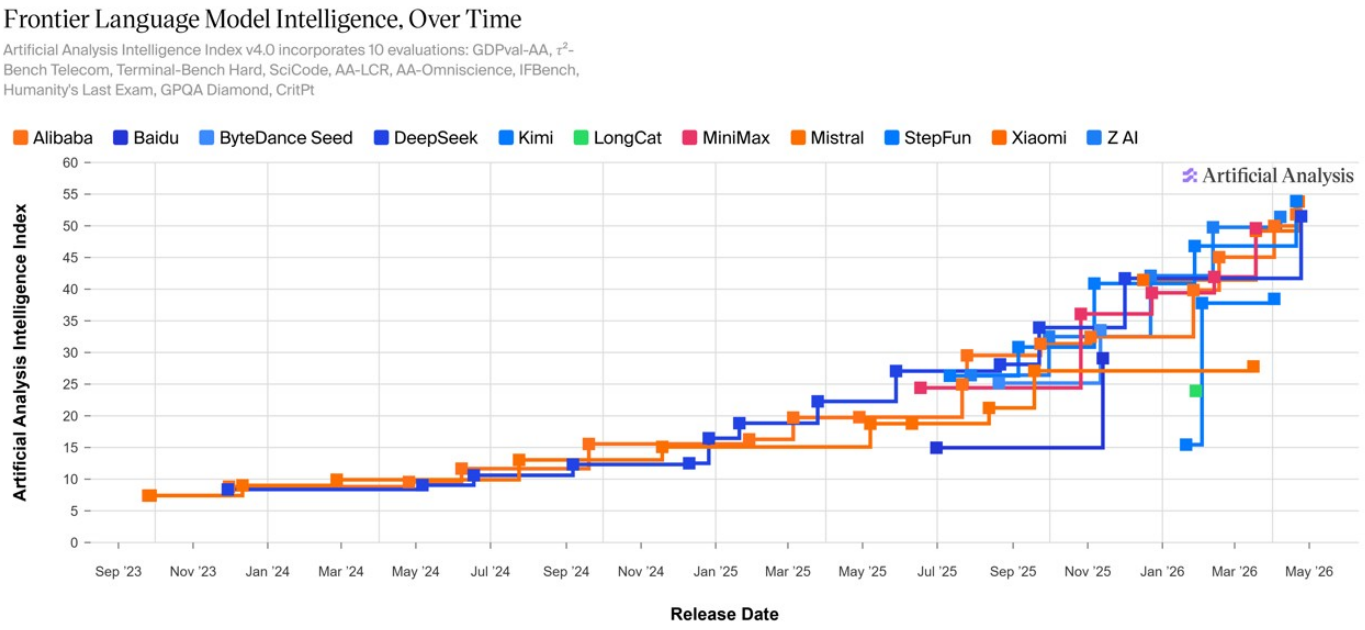
Source: Artificial Analysis, updated by Apr 2026

Exhibit 25: Frontier open weight LLM intelligence by country



Source: Artificial Analysis, updated by Apr 2026

Exhibit 26: Frontier LLM intelligence by AI lab



Source: Artificial Analysis, updated by Apr 2026

Where China differs from the US: Compared to the US, China's AI roadmap is more skewed towards open weight, fast speeds and low pricing via model architecture design, algorithm enhancement and infrastructure optimization, to gain market share and offset computing power constraints. China is also specialized in multimodality strength, which provides higher commercialization upside.

- Low price: At a similar performance level, China's AI models provide much lower API prices compared to US models through algorithm enhancement and infrastructure optimization to reduce inference costs.
- "Closed" US and "Open" China: China leads the world in terms of open-source AI models. Among the global SOTA models, most of the China models are open source, compared to all US models being proprietary.
- Multimodality: China has also developed multi-modal capabilities and is a leader in video models globally. Alibaba, Bytedance and Kuaishou are the current top players in this field.

Drivers of model differences – China vs. US:

- **Computing power:** The US has a vast advantage in computing power with sufficient funding and the largest GPU supplies. Meanwhile China has limited access to advanced computing power for model training owing to US export restrictions. While we estimate China's AI GPU self-sufficiency ratio will rise from 41% in 2025 to 76% in 2030, we forecast Chinese hyperscalers' capex will reach Rmb706bn (US$101bn) in 2027, which would only account for 12% of US hyperscalers' capex.
- **Data:** China and the US are largely equivalent in terms of data as a majority of the training data is from public sources. Meanwhile China's abundance of data – driven by 1.1bn internet users, the widespread use of mobile apps, e-commerce platforms,

and social media – gives the country a significant edge in AI research. Technologies such as face and voice recognition, intelligent robots, virtual reality, and driverless vehicles are widely used in the country and adopted in fields such as transport, education, medical care, science and technology, logistics, agriculture, and entertainment.

- **Algorithms:** A good supply of AI talent is the key to algorithm innovation and enhancement. China is home to some of the world's top AI researchers and has a wealth of AI talent, producing 47% of all AI researchers, according to the World Economic Forum. It has invested heavily in developing its AI workforce, with government-backed programs encouraging students to pursue AI-related degrees and critical skills. Major Chinese universities, such as Tsinghua and Peking University, are driving cutting-edge AI research and producing a growing pool of AI professionals.

Exhibit 27: China vs. US model performance drivers

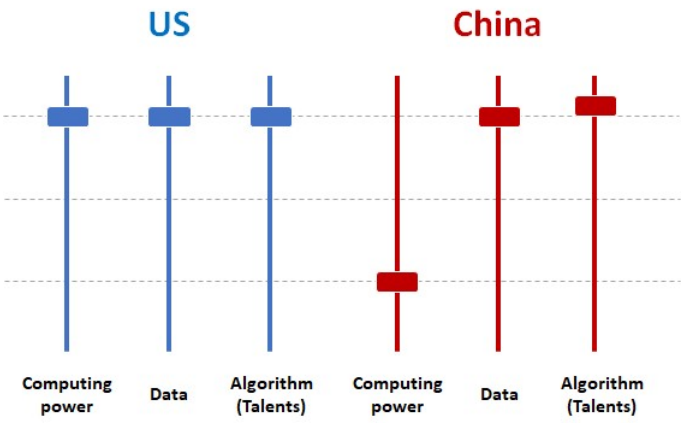


Exhibit 28: Hyperscaler vs. pure play performance drivers

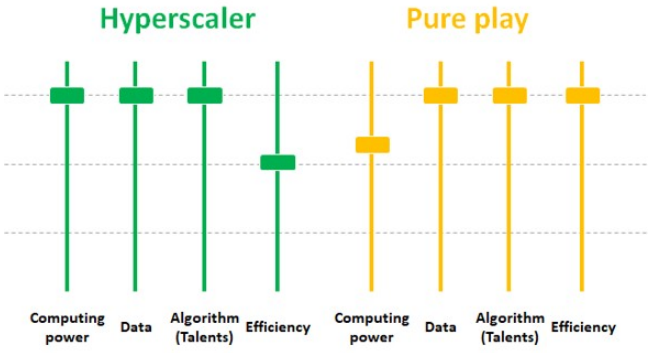


Exhibit 29: China AI chip self-sufficiency rate

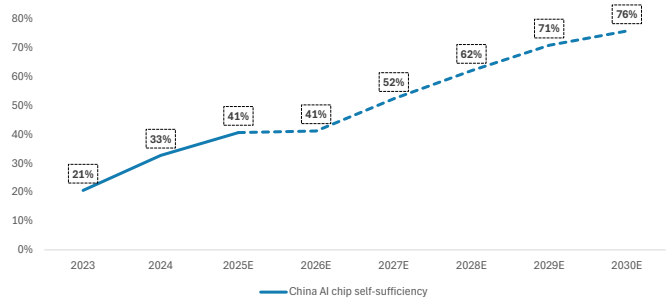
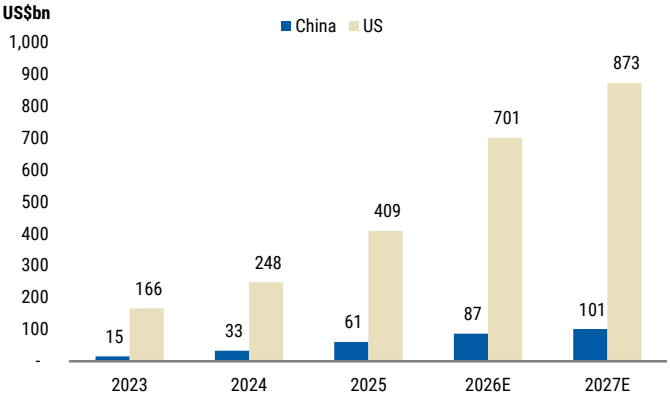


Exhibit 30: China vs. US hyperscaler capex



Tencent, Baidu, Bytedance, Kuaishou. US hyperscalers include Google, Meta, Microsoft, Amazon, Oracle

Equity Market Implications – LLMs Reshaping the Hong Kong Market Landscape

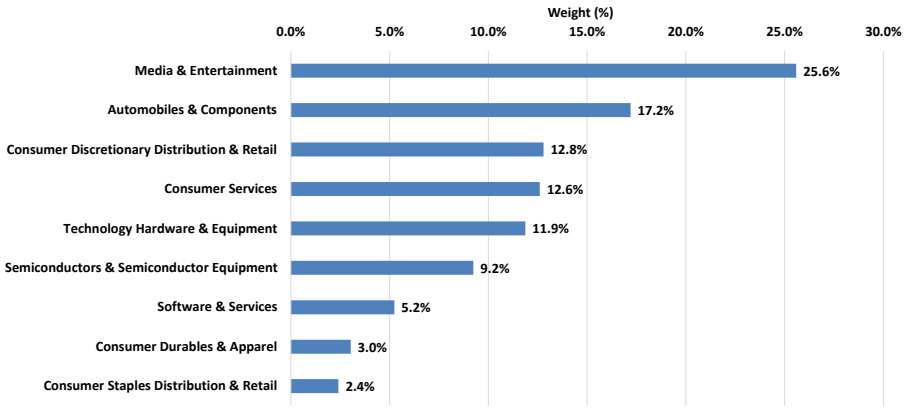
China equity strategy: Laura Wang, Chloe Liu, Vicky Wu

Foundation model listings – A rising force for the Hang Seng Tech Index

Hang Seng Tech Index – Recent performance suggests AI disruption effects, among other challenges faced by conventional internet businesses

Year to date, the Hang Sang Tech Index has has lackluster returns, returning -11% so far. This has been driven by multiple factors including the Middle East geopolitical crisis, which has significantly dampened risk sentiment globally. However the biggest and most fundamental reason, in our view, is that **investors had started to seriously question whether the Hang Seng Tech Index truly represents the most advanced technology innovation happening in China or a China consumption proxy dampened by price competition and macro challenges?** As show in Exhibit 31 , Hang Seng Tech index currently is overly represented by the e-commerce sectors (25.4% of index weight, Baba + Meituan + JD + Trip.com + Tongcheng Travel) and the auto/EV sector (25.1% of index weight, BYD + Xiaomi + NIO + XPeng + Li Auto + Zhejiang Leapmotor), with a combined index weight of 50.5%.

Exhibit 31: Hang Seng Tech Index index weight by GICS Industry Group – Auto, Internet and Consumer related account for >50% of the index weight

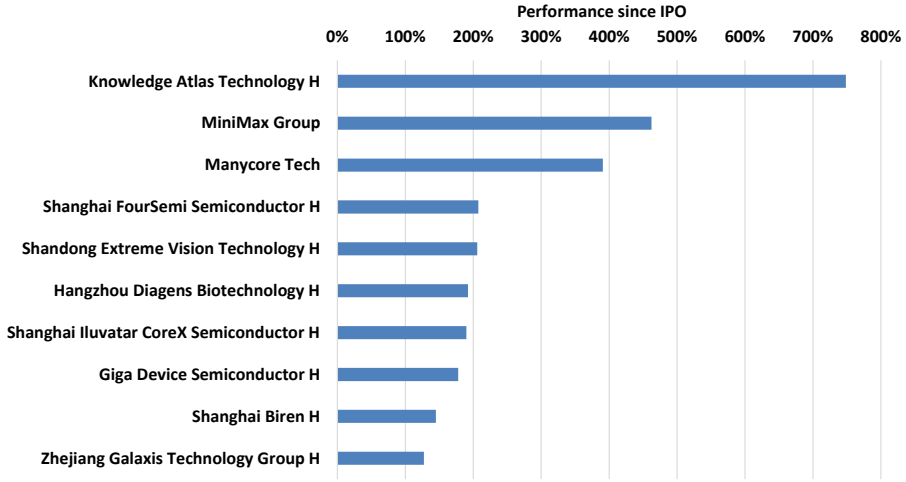


Meanwhile the **"ByteDance Effect" also proves that the AI disruption effects are well reflected at Hang Seng Tech index level while the latest tech breakthroughs and advances are excluded.** ByteDance has been making significant breakthroughs in both e-commerce and AI technology development that are disrupting publicly-traded Chinese internet giants. A lot of the technical details on the progress the company has achieved on the AI/LLM front are discussed in earlier sections of this report. Given that the company is not yet listed, investors have largely chosen to reduce their exposure to the companies'

direct competitors – the broad internet/software space in general.

LLM listings – best performing stocks YTD in HK, but not part of the HSTech Index yet: The AI competitive landscape in China is still very much evolving and we continue to maintain an overall positive view on the sector/theme outlook. We continue to hold Alibaba in our MS China strategy HK/China focus list as a well balanced play to the AI theme. Meanwhile, **Knowledge Atlas and MiniMax are the two best-performing stocks in the HK market year to date (Exhibit 32).** However, they are too new to be part of the Hang Seng Tech index to show the positive opportunities that have come along in the AI innovation and disruption space.

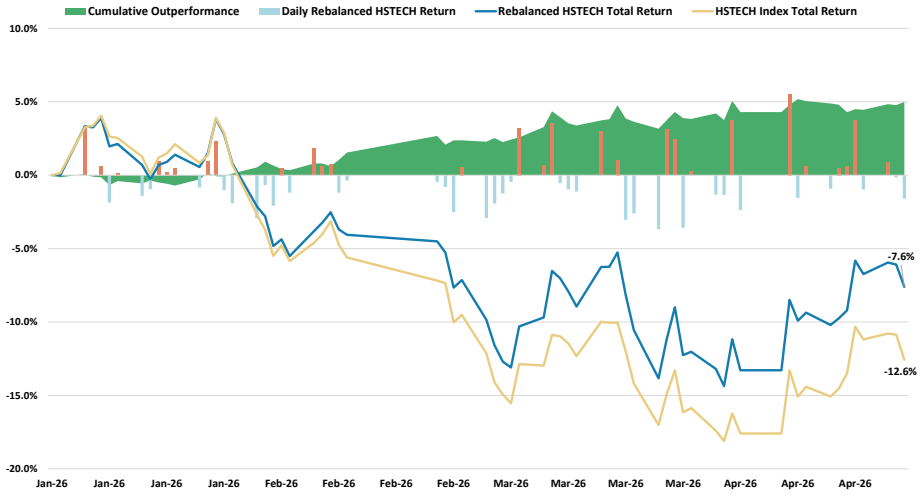
Exhibit 32: Top 10 best performing 2026 HK IPO stocks since IPO (with over US$2bn market cap) – Knowledge Atlas and MiniMax top the chart



HSTech Index YTD cumulative performance would have been lifted by 5ppts were the LLMs in it: We have calculated the proforma Hang Seng Tech Index YTD performance assuming Minimax and Knowledge Atlas were added to the index upon their IPOs, replacing the two smallest index constituent companies Kingdee Soft and Kingsoft (more detailed discussion on our inclusion and deletion rationale in section Knowledge Atlas and Minimax to join HSTech Index and Southbound soon). Our analysis shows that the index's YTD total return would have been lifted from -12.6% to -7.6% by 5ppts. This would have been able to mitigate most of the correction since mid/late Feb around the Chinese New Year period, when ByteDance's Doubao and Seedance had been gaining significant traction.

Addition Forecast in HSTECH		Deletion Forecast in HSTECH	
0100.HK	Minimax - W	0268.HK	Kingdee International Software
2513.HK	Knowledge Atlas Technology	3888.HK	Kingsoft Corp

Exhibit 33: The rebalanced HSTECH, with Minimax and Knowledge Atlas replacing Kingdee and Kingsoft, shows outperformance of 5% vs. original HSTECH from Jan 9, 2026 till now



Note: The portfolio is constructed based on a hypothetical rebalancing of the HSTECH index to include Minimax and Zhipu from January 9, 2026 (the first trading day following their IPOs), replacing Kingdee and Kingsoft. The portfolio follows the same methodology as HSTECH, with market-capitalisation-based weighting, an individual stock weight cap of 8%, and weekly rebalancing. Performance is back-tested under these assumptions.

Knowledge Atlas and Minimax to join HSTech Index and Southbound soon

We expect both stocks to be included in the Hang Seng Tech Index on June 8th, with a joint index weight at 5-7%: We believe both stocks will be able to pass the June review for Hang Seng Tech Index inclusion. The two smallest constituent companies at the moment, Kingdee and Kingsoft, are most likely to be excluded as a result, given that the Hang Seng Tech Index is designed to have a fixed number of 30 constituents only, with maximum index weight per constituent capped at 8%. See Exhibit 34 for the detailed index inclusion criteria for Hang Seng Tech Index.

Hang Seng Tech Index (HSTECH) eligibility criteria

The Hang Seng Technology Index (HSTECH) maintains a fixed universe of 30 constituents, selected through a five-stage screening process.

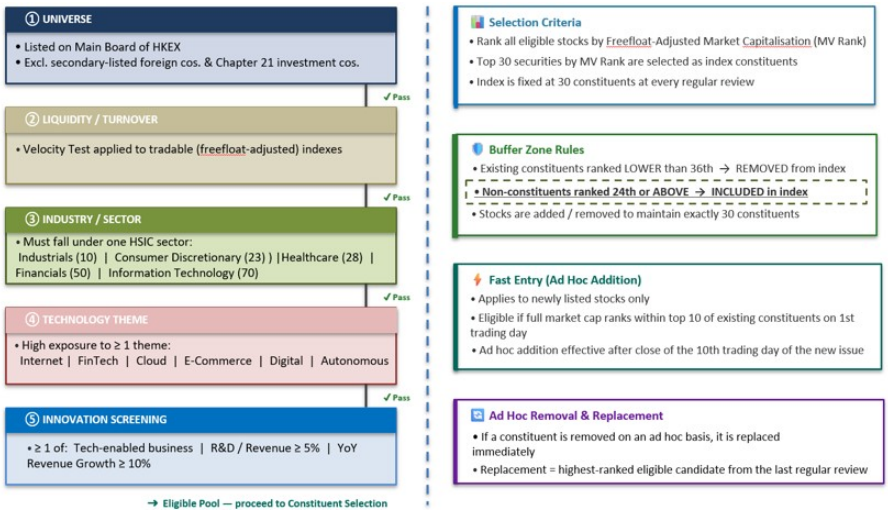
- Universe - The starting universe is confined to stocks listed on the Main Board of HKEX, ensuring a baseline of listing quality and regulatory standards. Secondary-listed foreign companies and Chapter 21 investment companies are explicitly excluded.
- Liquidity / Turnover - Eligible stocks must pass a velocity test applied to freefloat-adjusted indexes, which measures the ratio of traded volume to freefloat-adjusted shares outstanding. This screen filters out stocks with thin or irregular trading activity, ensuring that only sufficiently liquid names are considered for inclusion.
- Industry / Sector - Stocks must fall under one of five designated HSIC sectors: Industrials (10), Consumer Discretionary (23), Healthcare (28),

Financials (50), or Information Technology (70). This sector boundary is intentionally broad, reflecting the cross-industry nature of technology disruption, while still excluding sectors where technology exposure is typically peripheral.

- Technology Theme - Beyond sector classification, stocks must demonstrate high exposure to at least one of six technology themes: Internet, FinTech, Cloud, E-Commerce, Digital, or Autonomous. This thematic filter ensures the index captures companies that are genuinely technology-driven in their business model, rather than simply operating in a tech-adjacent sector.
- Innovation Screening - As a final qualitative and quantitative filter, stocks must satisfy at least one of three innovation indicators: being classified as a tech-enabled business; having R&D expenditure accounting for ≥5% of revenue, reflecting meaningful investment in innovation; or delivering YoY revenue growth of ≥10%, signalling strong business momentum driven by technology adoption.

Stocks that clear all five screening criteria enter the eligible pool, from which the **top 30 securities by free-float-adjusted market capitalisation** are selected as index constituents at each regular review. Under the buffer-zone rule, a **non-constituent must rank 24th or above** to be included and replace an existing constituent.

Exhibit 34: Eligibility assessment criteria for Hang Seng Technology Index



Both companies to be Southbound eligible by August

We expect Knowledge Atlas to be tradable through Southbound on June 8th, while MiniMax will be tradable on August 6th. Both will, in our view, meet the fast entry rules and should get included in Hang Seng Composite Index on June 8th, the nearest review date (serves as a base universe for Southbound eligibility review). Minimax's actual eligible effective date will be delayed to August 6th, because of its WVR structure, which will require a minimum 6 months and 20 days of trading history in the Hong Kong market.

Stock Connect Southbound eligibility fast-entry rule

Beside the two regular semi-annual reviews, the Southbound Stock Connect's **Fast Entry Rule** offers newly listed stocks an accelerated path into the eligible universe via two routes.:

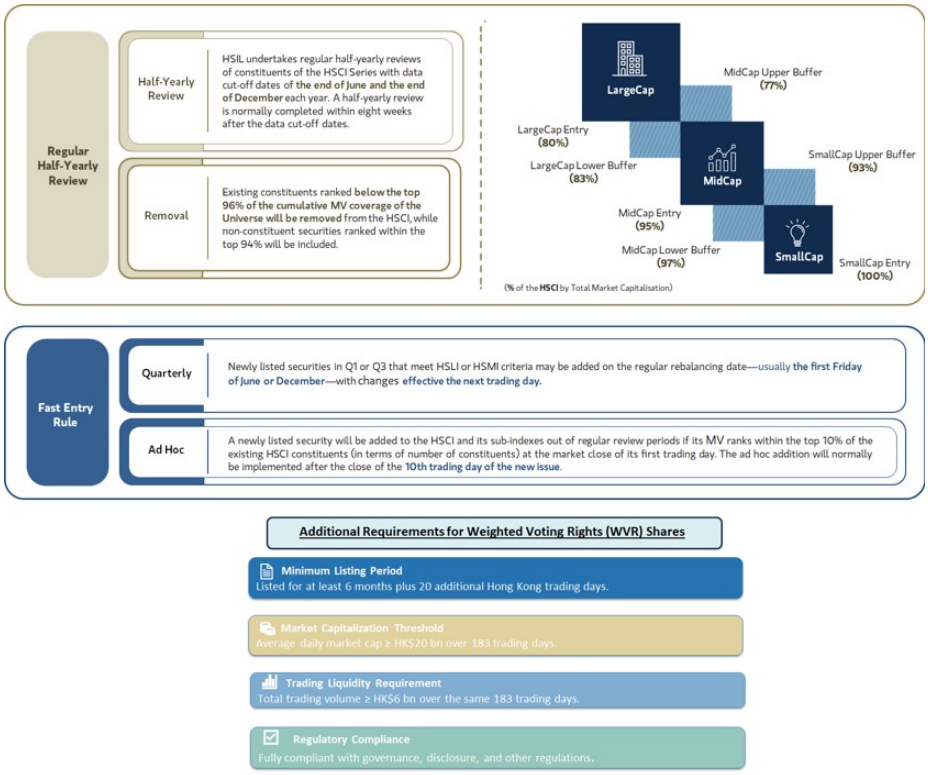
- **Quarterly:** Stocks listed in Q1 or Q3 that meet HSLI (Hang Seng Composite LargeCap Index) or HSMI (Hang Seng Composite MidCap Index) criteria can be added on the next regular rebalancing date (typically the first Friday of June or December), effective the following trading day.
- **Ad Hoc:** If a newly listed stock's market cap ranks in the top 10% of existing HSCI constituents on its first trading day, it can be added outside of regular review periods - normally implemented after the close of the 10th trading day of the new listing.

For stocks with **Weighted Voting Rights (WVR)** structures, four additional requirements must be met on top of standard criteria:

- Minimum listing period - listed for at least 6 months plus 20 Hong Kong trading days
- Market cap threshold - average daily market cap ≥ HK$20bn over 183 trading days
- Liquidity requirement - total trading volume ≥ HK$6bn over the same 183 trading days
- Regulatory compliance - fully compliant with governance, disclosure, and other regulations

These additional hurdles for WVR shares reflect the heightened governance scrutiny applied to dual-class share structures within the Southbound Connect framework.

Exhibit 35: Stock Connect Southbound eligibility assessment including fast entry rules and additional requirements for WVR shares



Strong flow support for Knowledge Atlas and Minimax to come

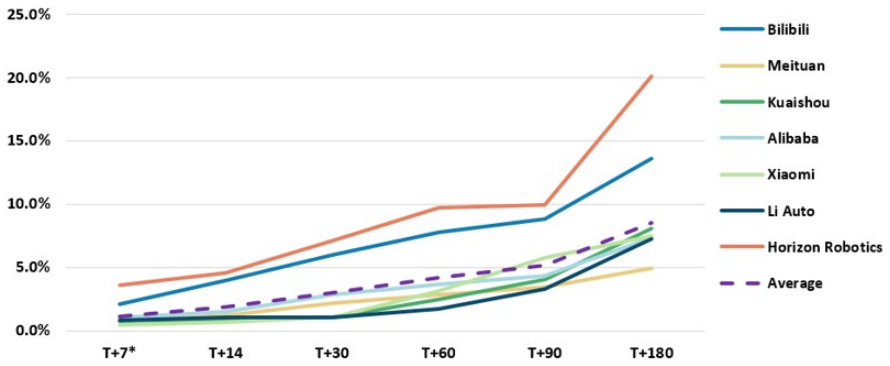
We believe Hang Seng Tech index rebalancing and Southbound eligibility will trigger meaningful liquidity into the two companies, which should help support their stock prices and in turn support the broader Hang Seng Tech index and Hong Kong market dynamics as well.

Hang Seng Tech rebalancing – We expect a total of US$1.25-1.75bn of passive inflows into Knowledge Atlas and Minimax due to ETF/passive rebalancing. Currently the total passive ETF AUM benchmarked against Hang Seng is estimated to be US$25bn, according to Bloomberg.

Southbound eligibility – 9-20% of free float market cap to be owned by onshore investors within first 6 months of inclusion: We have analyzed Southbound (SB) ownership build-up over time since SB inclusion across major internet and technology companies in recent years. Our analysis shows that SB holdings typically rise to an average of ~9% of free-float shares of those companies within six months following their inclusion, with faster accumulation observed among small free-float market cap names (Exhibit 49).

Exhibit 36: Major IT companies' SB holdings as % of free-float shares post-Southbound inclusion — average holdings are expected to reach ~9% after six months

	SB Holdings as % of free-float shares					
	T+7*	T+14	T+30	T+60	T+90	T+180
Bilibili	2.1%	4.0%	6.0%	7.8%	8.8%	13.6%
Meituan	0.8%	1.2%	2.2%	2.9%	3.5%	5.0%
Kuaishou	0.7%	0.9%	1.1%	2.5%	4.1%	8.1%
Alibaba	1.0%	1.5%	2.9%	3.7%	4.4%	7.3%
Xiaomi	0.5%	0.7%	1.0%	3.1%	5.8%	7.5%
Li Auto	0.9%	1.1%	1.1%	1.7%	3.3%	7.2%
Horizon Robotic	3.6%	4.6%	7.1%	9.7%	10.0%	20.1%
Average	1.2%	1.9%	3.0%	4.2%	5.2%	8.5%



We lay out base and bull cases of Southbound inflow support for Minimax and Knowledge Atlas. In our base case, we assume a similar ownership trajectory for the two names with SB holdings reaching 8–9% by T+180, broadly in line with the historical peer average. In a bull-case scenario, we assume a steeper accumulation path—consistent with the upper-end trend experienced by more recent tech listings with smaller free float market cap. size, such as Horizon Robotics—where SB holdings could rise to ~20% of free-float shares within six months.

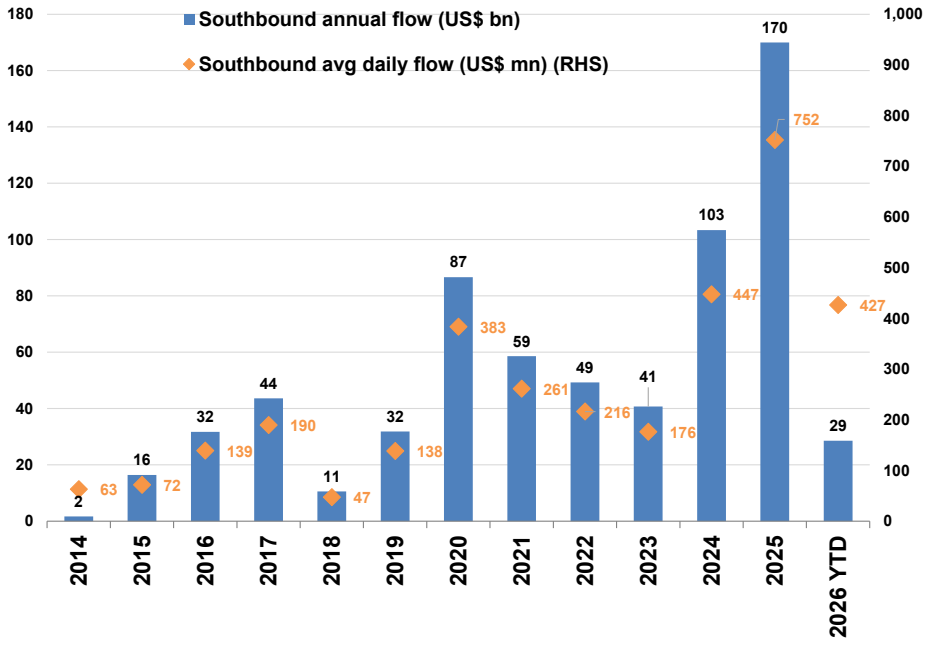
Translating these ownership assumptions into flow terms, **our estimates imply cumulative SB net inflows of HK$18–19bn for each stock in the base case, and HK$43–44bn per stock in the bull case by T+180, respectively.** For Minimax, this HK$18-43bn range is equivalent to 12–30 days of its April 2026 average daily turnover. For Knowledge Atlas, the HK$19-44bn range is equivalent to 10–25 days of its April 2026 average daily turnover.

We strongly believe that the inflows attracted upon inclusion of these two stocks into the HSTech Index as well as into Southbound will in return support the stock price and valuation for them given the scarcity of similar investment targets in the current HK and/or A-share market. **Such heightened interest will also support the broader market sentiment** with potentially more institutional and even retail participation.

Exhibit 37: Potential SB flows into Minimax and Knowledge Atlas – Base case: HK$30-40bn estimated inflows six months after they are included

			T+7	T+14	T+30	T+60	T+90	T+180
Minimax	Base	SB Holdings as % of free-float shares	1.2%	1.9%	3.0%	4.2%	5.2%	8.5%
		Cumulative SB net inflow forecast (HK$bn)	2.5	4.1	6.5	9.1	11.2	18.4
		Implied days of average daily turnover*	1.8	2.9	4.6	6.3	7.8	12.8
	Bull	SB Holdings as % of free-float shares	3.6%	4.6%	7.1%	9.7%	10.0%	20.1%
		Cumulative SB net inflow forecast (HK$bn)	7.7	9.9	15.3	21.0	21.5	43.4
		Implied days of average daily turnover	5.4	6.9	10.7	14.7	15.0	30.3
Knowledge Atlas	Base	SB Holdings as % of free-float shares	1.2%	1.9%	3.0%	4.2%	5.2%	8.5%
		Cumulative SB net inflow forecast (HK$bn)	2.6	4.1	6.6	9.2	11.3	18.6
		Implied days of average daily turnover	1.4	2.2	3.5	4.9	6.1	10.0
	Bull	SB Holdings as % of free-float shares	3.6%	4.6%	7.1%	9.7%	10.0%	20.1%
		Cumulative SB net inflow forecast (HK$bn)	7.8	10.0	15.5	21.3	21.8	43.9
		Implied days of average daily turnover	4.2	5.3	8.3	11.4	11.7	23.5

Exhibit 38: Southbound annual flow and average daily flow, 2014-26 YTD – 2026 daily flow tracking at 57% of 2025, reflecting fading momentum after record inflows last year



What's next – Hong Kong IPO market turning more supportive of AI-related themes

Clear, determined government regulatory support for AI-related IPOs: Over the past 12–18 months, Hong Kong regulators and the HKSAR Government have taken concrete, market-facing steps to position Hong Kong as a supportive listing venue for AI and large-language-model (LLM) companies, combining listing rule liberalisation, dedicated listing infrastructure, and explicit AI-first industrial policy. On the capital-markets side, HKEX and the SFC have materially lowered entry barriers for early-stage and pre-profit technology issuers through targeted reforms to the Specialist Technology Company regime, while simultaneously launching a dedicated Technology Enterprises Channel to streamline and de-risk the listing process for advanced technology companies, including AI. At the policy level, successive budgets have explicitly elevated AI to a "core industry," backed by large,

earmarked funding pools, compute infrastructure, and coordination mechanisms, reinforcing investor confidence that AI listings are aligned with long-term government priorities. These regulatory and policy signals have translated into real outcomes, including Hong Kong hosting the world's first major LLM company IPOs (Knowledge Atlas and Minimax) and an IPO pipeline increasingly dominated by AI and other hard-tech issuers, underscoring Hong Kong's intent to be a primary offshore listing venue for next-generation AI companies.

Key regulatory and policy actions supporting AI/LLM-related listings:

1. Lower listing thresholds for advanced technology companies:

In September 2024, HKEX and the SFC temporarily reduced minimum market-capitalisation requirements under Chapter 18C (Specialist Technology Companies) from HK$6bn to HK$4bn (commercial) and HK$10bn to HK$8bn (pre-commercial) for a three-year period starting September 2024, explicitly to improve the viability of listing high-growth, R&D-intensive tech companies such as AI and LLM developers.

2. Dedicated fast-track listing support for technology issuers:

The SFC and HKEX jointly launched the Technology Enterprises Channel (TECH), a dedicated pathway to facilitate new listing applications from prospective Specialist Technology Companies and Biotech Companies, alongside a new confidential filing option.The Exchange also updated the Guide for New Listing Applicants, under which these companies are presumed to have satisfied the Innovative Company Requirements and external validation requirement for listing with a weighted voting right (WVR) structure under Main Board Chapter 8A.

3. AI formally designated as a core industry with capital support:

The 2025–26 Budget set aside HK$10bn for an Innovation and Technology Industry-Oriented Fund to channel market capital into emerging industries, with AI explicitly prioritised, alongside HK$1bn to establish the Hong Kong AI Research and Development Institute to support upstream R&D and commercialisation. The 2026–27 Budget further escalated AI policy support by establishing a Committee on AI+ and Industry Development Strategy, reinforcing long-term coordination across policy, industry and capital markets, and signalling sustained government backing beyond near-term market cycles.

The confluence of supportive regulatory reforms and targeted listing policies has catalysed a meaningful structural shift in Hong Kong's IPO landscape, with technology and AI-related issuers emerging as the dominant force in 2026.

In 2026 year-to-date, total Hong Kong IPO financing reached HK$139bn, already equivalent to approximately 50% of the full-year 2025 total, underscoring the accelerating momentum in primary market activity. More strikingly, the composition of this capital raising has shifted decisively toward high-technology sectors. Information Technology IPO financing amounted to HK$54 bn in 2026 YTD, representing 39% of total market IPO proceeds. This number has not only surpassed the full-year 2025 IT IPO total of HK$40bn, but also marked a dramatic increase of the sector's share from just 14% of total proceeds in 2025. Within the IT sector, the most pronounced inflection has been in semiconductors, which now account for 21% of total market IPO financing, up sharply from a mere 3% in 2025, reflecting both strong investor appetite and Hong Kong's growing role as a capital-

raising hub for China's semiconductor industry.

Exhibit 39: Total HK market IPO financing by year – 2026 YTD Information Technology IPO financing has surpassed 2025's full-year total

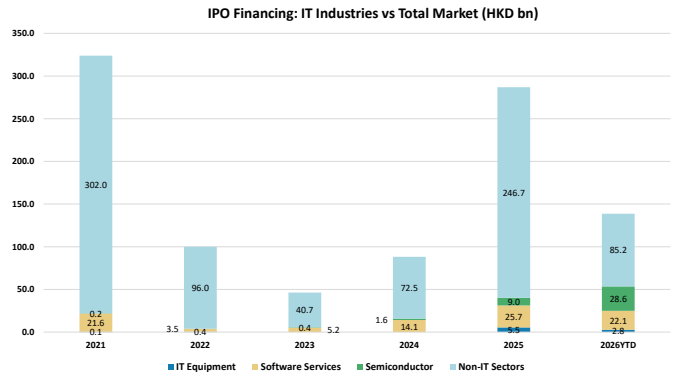
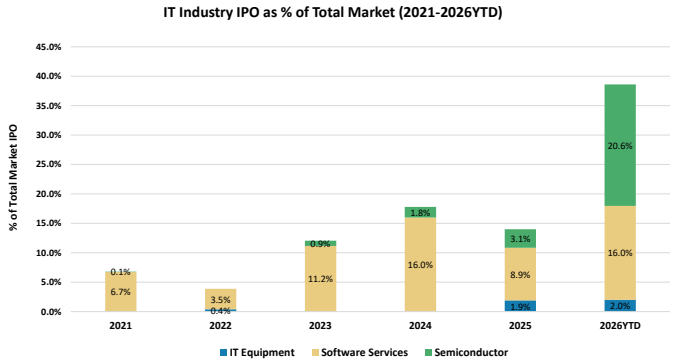


Exhibit 40: IT and semiconductor IPO financing as % of total IPO by year – 2026 YTD marks the highest focus on IT and semi IPOs in years, thanks to the AI theme



The HK market IPO pipeline further reinforces this structural narrative. Of approximately 390 companies currently in the Hong Kong IPO pipeline, around 43% are from the Information Technology sector, followed by Healthcare at 22% and Consumer Discretionary at 13%. The skew towards technology issuers reflects a deliberate and successful repositioning of Hong Kong as a preferred listing venue for high-tech and AI-adjacent companies, underpinned by recent regulatory accommodations including the streamlined Chapter 18C framework for specialist technology companies. If this pipeline converts at a reasonable rate, the technology sector's dominance in Hong Kong's primary market could become a durable feature, with meaningful long-term implications for the market's index composition, sector weights, and structural attractiveness to both domestic and international investors.

Exhibit 41: HK market IPOs by count – IT has accounted for 40% of IPOs by company count...

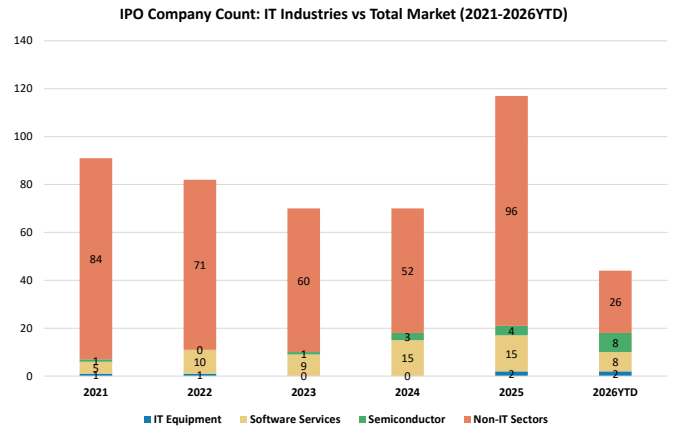
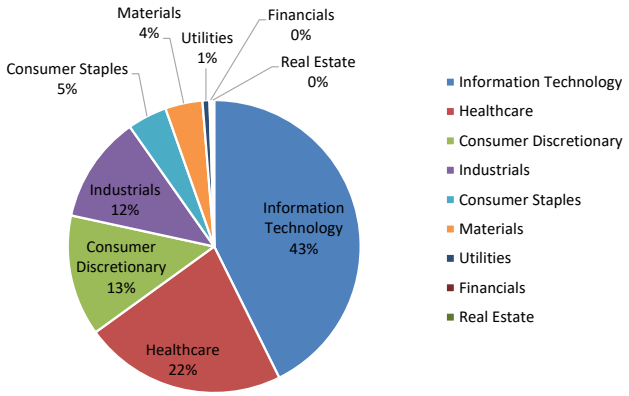


Exhibit 42: 2026 HK IPO pipeline by industry group – IT represents nearly half of the remaining IPO pipeline in the HK market



What's Changed

Earning estimate revisions

MiniMax

We raise our revenue forecasts by 104.2%, 71.1% and 99.2% in 2026-28 respectively to reflect faster-than-expected ARR growth. Our non-IFRS operating loss estimates are changed by -0.7%, +7.9% and -53.7% in 2026-28 respectively as a mixed result of higher revenue and higher R&D expense assumptions. As a result, our non-IFRS loss per share forecasts are changed by +1.4%, +10.7% and -31% in 2026-28 respectively.

Z.ai

We raise our revenue forecasts by 137.5%, 138.2% and 162.7% in 2026-28 respectively to reflect the 2025 result and faster-than-expected ARR growth. Our non-IFRS operating loss estimates are changed by -0.8%, -17.9% and -95% in 2026-28 respectively as a mixed result of higher revenue and higher R&D expenses assumptions. As a result, our non-IFRS loss per share forecasts are changed by -10.7%, -32.8% and -133.8% in 2026-28 respectively.

Exhibit 43: Earnings estimate revisions: MiniMax

Years Ending December 31	New			Old			% change		
US$ '000	2026E	2027E	2028E	2026E	2027E	2028E	2026E	2027E	2028E
Total revenue	398,946	1,197,635	2,991,273	195,418	699,818	1,501,965	104.2%	71.1%	99.2%
Gross profit	113,907	431,631	1,194,701	41,795	222,573	662,156	172.5%	93.9%	80.4%
Operating profit/(loss) - MS denifition	(556,500)	(633,325)	(296,238)	(549,428)	(570,263)	(411,754)	1.3%	11.1%	-28.1%
Operating profit/(loss) - MS denifition (Non-IFRS)	(477,198)	(522,302)	(140,805)	(480,574)	(484,195)	(304,170)	-0.7%	7.9%	-53.7%
Profit/(loss) before tax	(540,890)	(590,275)	(268,473)	(534,646)	(531,876)	(388,272)	1.2%	11.0%	-30.9%
Net profit/(loss)	(540,890)	(590,275)	(268,473)	(534,646)	(531,876)	(388,272)	1.2%	11.0%	-30.9%
Net profit/(loss) (Non-IFRS)	(461,588)	(479,252)	(113,041)	(465,792)	(445,808)	(280,687)	-0.9%	7.5%	-59.7%
EPS (Non-IFRS)	(1.72)	(1.88)	(0.86)	(1.70)	(1.70)	(1.24)	1.4%	10.7%	-31.0%
Price target		**1,100.0**			**930.0**			**18.3%**	

Exhibit 44: Earnings estimate revisions: Z.ai

Years Ending December 31	New			Old			% change		
Rmb mn	2026E	2027E	2028E	2026E	2027E	2028E	2026E	2027E	2028E
Total revenue	3,828	10,044	23,523	1,612	4,217	8,953	137.5%	138.2%	162.7%
Gross profit	1,362	3,868	9,973	738	1,842	4,273	84.5%	109.9%	133.4%
Operating profit/(loss) - MS denifition	(4,300)	(3,716)	(187)	(4,334)	(4,527)	(3,726)	-0.8%	-17.9%	-95.0%
Profit/(loss) before tax	(4,281)	(3,694)	(216)	(4,342)	(4,524)	(3,740)	-1.4%	-18.4%	-94.2%
Net profit/(loss) attributable to equity holders	(4,261)	(3,674)	(206)	(4,327)	(4,509)	(3,724)	-1.5%	-18.5%	-94.5%
EPS (Non-IFRS)	(7.91)	(6.09)	2.37	(8.86)	(9.05)	(7.02)	-10.7%	-32.8%	-133.8%
Price target		**990.0**			**560.0**			**76.8%**	

Valuation

MiniMax: We raise PT by 18% to HK$1100. We value MiniMax using a 10-year DCF model as a more stable approach to capture long-term growth potentials in AI industry. Our key DCF assumptions include a 15% WACC and a 3% terminal growth rate. Our PT implies 36x 2027 P/S. Our bull/bear case values are also revised up to HK$2,480/HK$600.

Z.ai: We raise PT by 77% to HK$990. We value Z.ai using a 10-year DCF model as a more stable approach to capture long-term growth potential in the AI industry. Our key DCF assumptions include a 15% WACC and a 3% terminal growth rate. We lower our WACC assumption to reflect higher visibility into future growth. Our PT implies 38x 2027 P/S. Our bull/bear case values are also revised up to HK$1,730/HK$420.

Exhibit 45: DCF valuation: MiniMax

Years Ending December 31 US$ '000	2023	2024	2025	2026E	2027E	2028E	2029E	2030E	2031E	2032E	2033E	2034E	2035E
DCF													
Operating profit (IFRS)	(101,298)	(286,620)	(321,401)	(556,500)	(633,325)	(296,238)	752,930	2,510,901	4,475,008	6,981,979	9,730,518	12,177,834	13,645,301
D&A	811	1,901	2,679	2,726	3,224	3,767	4,371	5,052	5,804	6,620	7,491	8,405	9,352
Capex paid	(697)	(759)	(1,188)	(1,428)	(1,643)	(1,889)	(2,172)	(2,498)	(2,848)	(3,218)	(3,604)	(4,001)	(4,401)
Lease payments	(757)	(1,506)	(2,222)	(2,236)	(2,417)	(2,704)	(3,073)	(3,516)	(4,015)	(4,557)	(5,133)	(5,734)	(6,350)
Taxes	-	-	-	-	-	-	-	-	(805,501)	(1,256,756)	(1,751,493)	(2,192,010)	(2,456,154)
Working capital	25,900	13,716	6,255	144,370	(29,215)	93,573	(183,406)	(70,586)	(153,574)	(199,027)	(235,571)	(194,950)	(107,244)
Unlevered FCF (FCFF)	**(76,041)**	**(273,268)**	**(315,877)**	**(413,069)**	**(663,375)**	**(203,490)**	**568,650**	**2,439,353**	**3,514,874**	**5,525,041**	**7,742,208**	**9,789,545**	**11,080,504**
% YoY	495.4%	259.4%	15.6%	30.8%	60.6%	-69.3%	-379.4%	329.0%	44.1%	57.2%	40.1%	26.4%	13.2%
% of revenue	-2197.7%	-895.3%	-399.7%	-103.5%	-55.4%	-6.8%	9.5%	24.4%	23.5%	26.4%	28.4%	29.9%	30.8%
WACC				15.0%									
TGR				3.0%									
Year				**-**	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**8.00**	**9.00**
NPV of FCFF				14,434,604	17,263,169	20,056,135	22,495,905	23,430,937	23,430,704	21,420,269	16,891,101	9,635,221	-
Discounted terminal value				27,035,533	31,090,862	35,754,492	41,117,665	47,285,315	54,378,113	62,534,830	71,915,054	82,702,312	95,107,659
Enterprise value				**41,470,136**	**48,354,032**	**55,810,627**	**63,613,570**	**70,716,252**	**77,808,816**	**83,955,098**	**88,806,155**	**92,337,533**	**95,107,659**
Net cash/(debt)				1,392,900	883,701	863,520	1,661,505	4,414,750	8,370,355	14,506,253	23,082,950	33,984,681	46,499,877
Financial assets				-	-	-	-	-	-	-	-	-	-
Equity value				**42,863,036**	**49,237,733**	**56,674,147**	**65,275,075**	**75,131,002**	**86,179,171**	**98,461,352**	**111,889,104**	**126,322,214**	**141,607,536**
No of shares ('000)				313,635	313,635	313,635	313,635	313,635	313,635	313,635	313,635	313,635	313,635
FX rate (HK$/US$)				7.78	7.77	7.77	7.77	7.77	7.77	7.77	7.77	7.77	7.77
Value per share (HK$)				1,063.26	1,219.82	1,404.05	1,617.12	1,861.30	2,135.00	2,439.28	2,771.94	3,129.51	3,508.18
Target NTM EV/S				34.63	16.17	9.33	6.38	4.72	3.71	3.08	2.72	2.57	16.00
Target NTM P/S				35.79	16.46	9.47	6.54	5.02	4.11	3.61	3.42	3.51	15.50
Price target (HK$)				**1,100.00**									

	Rev (2027e)	P/S	US$ '000	
Bull case	2,500,000	40	100,000,000	2,480
Bear case	800,000	30	24,000,000	600

Exhibit 46: DCF valuation: Z.ai

Years Ending December 31 Rmb mn	2023	2024	2025	2026E	2027E	2028E	2029E	2030E	2031E	2032E	2033E	2034E	2035E
DCF													
Operating profit (IFRS)	(616)	(2,541)	(3,780)	(4,300)	(3,716)	(187)	8,150	21,185	37,226	58,149	81,864	104,199	119,543
D&A	69	280	281	226	188	166	155	153	157	167	180	196	215
Capex paid	(509)	(133)	(23)	(27)	(31)	(35)	(40)	(46)	(53)	(60)	(67)	(74)	(82)
Lease payments	(39)	(178)	(211)	(176)	(153)	(138)	(130)	(129)	(131)	(138)	(148)	(160)	(174)
Taxes	-	-	-	-	-	-	-	-	(6,701)	(10,467)	(14,735)	(18,756)	(21,518)
Non-controlling interests	(0)	2	20	20	20	10	-	-	-	-	-	-	-
Working capital	(96)	(0)	704	951	535	2,152	1,316	2,300	2,652	2,254	2,003	605	(499)
Unlevered FCF (FCFF)	**(1,191)**	**(2,570)**	**(3,008)**	**(3,307)**	**(3,156)**	**1,968**	**9,450**	**23,462**	**33,150**	**49,905**	**69,097**	**86,010**	**97,485**
WACC				15.0%									
TGR				3.0%									
Year				**-**	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**8.00**	**9.00**
NPV of FCFF				138,233	162,124	184,474	202,695	209,637	207,933	189,218	148,504	84,769	-
Discounted terminal value				237,855	273,534	314,564	361,748	416,010	478,412	550,174	632,700	727,605	836,745
Enterprise value				**376,088**	**435,658**	**499,038**	**564,443**	**625,648**	**686,344**	**739,391**	**781,203**	**812,374**	**836,745**
Enterprise value (US$ mn)				53,727	64,067	73,388	83,006	92,007	100,933	108,734	114,883	119,467	123,051
Net cash/(debt)				3,985	1,829	5,079	16,172	41,786	77,759	131,257	204,862	296,434	400,620
Financial assets				-	-	-	-	-	-	-	-	-	-
Equity value				**380,074**	**437,487**	**504,117**	**580,616**	**667,434**	**764,103**	**870,648**	**986,066**	**1,108,808**	**1,237,365**
Equity value (US$ mn)				54,296	64,336	74,135	85,385	98,152	112,368	128,036	145,010	163,060	181,965
No of shares ('000)				446	446	446	446	446	446	446	446	446	446
FX rate (HK$/Rmb)				1.11	1.14	1.14	1.14	1.14	1.14	1.14	1.14	1.14	1.14
Value per share (HK$)				**947.47**	**1,121.23**	**1,292.00**	**1,488.05**	**1,710.56**	**1,958.31**	**2,231.38**	**2,527.18**	**2,841.75**	**3,171.23**
Target NTM EV/S				37.4	18.5	10.9	7.5	5.6	4.5	3.7	3.3	3.1	
Target NTM P/S				37.8	18.6	11.0	7.8	6.0	5.0	4.4	4.1	4.2	
Price target (HK$)				**990.00**	**9.6%**								

	Rev (2027e)	P/S	US$ mn	
Bull case	17,250	40	99,281	1,730
Bear case	5,520	30	23,827	420

Financial Summaries

Exhibit 47: Financial summary: MiniMax

Years Ending December 31 US$ '000	2025	2026E	2027E	2028E
INCOME STATEMENT (IFRS)				
MiniMax Agent	1,062	11,942	41,797	200,624
Hailuo AI	24,945	100,312	214,416	523,627
MiniMax Audio	1,592	11,146	23,824	58,181
Talkie/Xingye	25,476	75,632	137,929	220,686
AI-native products	53,075	199,031	417,966	1,003,118
Open Platform and other AI-based enterprise services	25,963	199,915	779,669	1,988,156
Total revenue	**79,038**	**398,946**	**1,197,635**	**2,991,273**
AI-native products	3,463	9,952	41,797	200,624
Open Platform and other AI-based enterprise services	16,616	103,956	389,834	994,078
Gross profit	**20,079**	**113,907**	**431,631**	**1,194,701**
% GP margin	25.4%	28.6%	36.0%	39.9%
S&M expenses	51,896	98,602	138,043	193,261
R&D expenses	252,771	505,542	834,144	1,167,802
G&A expenses	36,813	66,263	92,769	129,876
Total operating expenses	**341,480**	**670,408**	**1,064,956**	**1,490,939**
Operating profit/(loss) - MS denifition	**(321,401)**	**(556,500)**	**(633,325)**	**(296,238)**
% OP margin	-406.6%	-139.5%	-52.9%	-9.9%
Non-operating items	**(1,550,216)**	**15,610**	**43,051**	**27,764**
Other income and gaines, net	40,369	16,595	43,153	27,877
Finance costs	(672)	(984)	(102)	(113)
FV loss on financial liabilities	(1,589,850)	-	-	-
Impairment losses on financial assets, net	(63)	-	-	-
Profit/(loss) before tax	**(1,871,617)**	**(540,890)**	**(590,275)**	**(268,473)**
% PBT margin	-135.6%	-49.3%	-9.0%	13.0%
Income tax expense	-	-	-	-
% of PBT	0.0%	0.0%	0.0%	0.0%
Net profit/(loss)	**(1,871,617)**	**(540,890)**	**(590,275)**	**(268,473)**
% NP margin	-2368.0%	-135.6%	-49.3%	-9.0%
EPS	**(17.23)**	**(1.72)**	**(1.88)**	**(0.86)**
INCOME STATEMENT (NON-IFRS)				
Total revenue	79,038	398,946	1,197,635	2,991,273
Gross profit	20,079	113,907	431,631	1,194,701
Operating profit/(loss) - MS denifition	(290,490)	(477,198)	(522,302)	(140,805)
Adjusted EBITDA - MS denifition	(287,811)	(474,472)	(519,078)	(137,038)
Net profit/(loss)	(250,856)	(461,588)	(479,252)	(113,041)
EPS	(2.31)	(1.47)	(1.53)	(0.36)
% margin				
Gross profit	25.4%	28.6%	36.0%	39.9%
Operating profit/(loss) - MS denifition	-367.5%	-119.6%	-43.6%	-4.7%
Adjusted EBITDA - MS denifition	-364.1%	-118.9%	-43.3%	-4.6%
Net profit/(loss)	-317.4%	-115.7%	-40.0%	-3.8%

Years Ending December 31 US$ '000	2025	2026E	2027E	2028E
BALANCE SHEET				
Cash and cash equivalents	507,621	1,392,900	883,701	863,520
Time deposits	13,787	-	-	-
Restricted cash	20,377	20,377	20,377	20,377
Financial assets (amortized cost)	-	-	-	-
Financial assets (FVPL)	438,525	-	-	-
Trade receivables	10,730	87,640	207,667	529,907
Prepayments, other receivables and other assets	16,319	43,292	90,280	172,556
Current assets	**1,007,359**	**1,544,209**	**1,202,025**	**1,586,361**
PP&E	1,571	2,018	2,463	2,928
Right-of-use assets	2,357	2,763	3,209	3,710
Restricted cash	41	41	41	41
Financial assets (FVPL)	69,965	-	-	-
Financial assets (FVOCI)	6,224	-	-	-
Prepayments, other receivables and other assets	887	2,353	4,907	9,379
Non-current assets	**81,045**	**7,175**	**10,620**	**16,058**
Total assets	**1,088,404**	**1,551,383**	**1,212,645**	**1,602,419**
Bank borrowings and convertible bonds	35,452	-	-	-
Trade and bill payables	57,677	235,658	376,065	793,144
Other payables, accruals and other liabilities	34,068	105,052	104,170	188,740
Contract liabilities	7,541	8,295	9,125	10,037
Lease liabilities	1,318	1,326	1,433	1,604
Convertible redeemable preferred shares	3,597,566	-	-	-
Current liabilities	**3,733,622**	**350,331**	**490,793**	**993,525**
Lease liabilities	638	642	694	776
Other non-current liabilities	2,334	2,334	2,334	2,334
Non-current liabilities	**2,972**	**2,976**	**3,028**	**3,110**
Total liabilities	**3,736,594**	**353,307**	**493,821**	**996,635**
Share capital	-	-	-	-
Share premium	-	4,307,854	4,307,854	4,307,854
Share option reserve	35,269	114,571	225,595	381,027
Other reserves	1,280	1,280	1,280	1,280
Accumulated losses	(2,684,739)	(3,225,629)	(3,815,904)	(4,084,377)
Total shareholder equities/deficits	**(2,648,190)**	**1,198,076**	**718,824**	**605,784**
Total liabilities and equities/deficits	**1,088,404**	**1,551,383**	**1,212,645**	**1,602,419**
CASH FLOW STATEMENT				
Operating cash flow	(279,641)	(313,507)	(505,140)	(15,588)
Investing cash flow	72,261	527,073	(1,643)	(1,889)
Financing cash flow	423,342	671,713	(2,417)	(2,704)
Other adjustments	2,747	-	-	-
Net changes in cash	**218,709**	**885,279**	**(509,199)**	**(20,180)**
Operating cash flow	(279,641)	(313,507)	(505,140)	(15,588)
Capex	(1,188)	(1,428)	(1,643)	(1,889)
Lease payments	(2,222)	(2,236)	(2,417)	(2,704)
Cash burn	**(283,051)**	**(317,172)**	**(509,199)**	**(20,180)**

Exhibit 48: Financial summary: Z.ai

Years Ending December 31 Rmb mn	2025	2026E	2027E	2028E
INCOME STATEMENT (IFRS)				
On-premise deployment	534	1,068	1,762	2,819
Cloud-based deployment	190	2,760	8,281	20,704
Total revenue	**724**	**3,828**	**10,044**	**23,523**
On-premise deployment	261	534	969	1,692
Cloud-based deployment	36	828	2,899	8,281
Gross profit	**297**	**1,362**	**3,868**	**9,973**
% GP margin	41.0%	35.6%	38.5%	42.4%
S&M expenses	391	528	686	892
R&D expenses	3,180	4,453	6,011	8,115
G&A expenses	505	682	887	1,153
Total operating expenses	**4,077**	**5,663**	**7,584**	**10,160**
Operating profit/(loss) - MS denifition	**(3,780)**	**(4,300)**	**(3,716)**	**(187)**
% OP margin	-521.9%	-112.3%	-37.0%	-0.8%
Non-operating items	**(938)**	**19**	**22**	**(29)**
Interest income	15	23	40	18
Finance costs	(74)	(30)	(24)	(21)
Share of profits/(losses) of associates	55	55	55	55
Other income	0	-	-	-
Impairment losses on financial assets	(22)	(29)	(49)	(81)
Changes in FV (FVPL)	25	-	-	-
Changes in the carrying amount of financial instruments issued to investors	(937)	-	-	-
Profit/(loss) before tax	**(4,718)**	**(4,281)**	**(3,694)**	**(216)**
% PBT margin	-651.4%	-111.8%	-36.8%	-0.9%
Income tax expense	-	-	-	-
% of PBT	0.0%	0.0%	0.0%	0.0%
Non-controlling interests	(20)	(20)	(20)	(10)
Net profit/(loss) attributable to equity holders	**(4,698)**	**(4,261)**	**(3,674)**	**(206)**
% NP margin	-648.6%	-111.3%	-36.6%	-0.9%
INCOME STATEMENT (NON-IFRS)				
Total revenue	724	3,828	10,044	23,523
Gross profit	297	1,362	3,868	9,973
Operating profit/(loss)	(3,222)	(3,547)	(2,737)	1,087
Adjusted EBITDA	(2,941)	(3,321)	(2,549)	1,253
Net profit/(loss)	(3,182)	(3,528)	(2,714)	1,058
% margin				
Gross profit	41.0%	35.6%	38.5%	42.4%
Operating profit/(loss)	-444.8%	-92.6%	-27.2%	4.6%
Adjusted EBITDA	-406.0%	-86.8%	-25.4%	5.3%
Net profit/(loss)	-439.3%	-92.1%	-27.0%	4.5%

Years Ending December 31 Rmb mn	2025	2026E	2027E	2028E
BALANCE SHEET				
Cash at bank and on hand	2,255	3,985	1,829	5,079
Time deposits	109	-	-	-
Restricted cash	4	4	4	4
Contract assets	2	3	5	9
Trade and other receivables	699	1,113	1,969	3,177
Inventories and contract costs	126	252	415	664
Short-term investments (FVPL)	377	-	-	-
Current assets	**3,571**	**5,358**	**4,222**	**8,933**
PP&E	656	514	422	365
Intangible assets	51	53	56	61
Goodwill	39	39	39	39
Interests in associates	338	393	448	503
Other non-current assets	198	358	605	1,043
Time deposits	-	-	-	-
Non-current assets	**1,282**	**1,357**	**1,571**	**2,012**
Total assets	**4,854**	**6,715**	**5,793**	**10,945**
Bank loans and convertible bonds	605	-	-	-
Trade and other payables	1,329	2,680	4,106	7,587
Contract liabilities	149	297	491	785
Lease liabilities	251	251	251	251
Financial instruments issued to investors	10,073	-	-	-
Current liabilities	**12,406**	**3,229**	**4,847**	**8,623**
Lease liabilities	294	205	144	106
Deferred revenue	180	360	594	951
Non-current liabilities	**559**	**565**	**738**	**1,057**
Total liabilities	**12,965**	**3,794**	**5,586**	**9,680**
Non-controlling interests	(18)	(38)	(58)	(68)
Share capital	40	45	45	45
Share premium	-	14,556	14,556	14,556
Other reserves	(3,724)	(2,970)	(1,990)	(716)
Accumulated losses	(4,409)	(8,671)	(12,344)	(12,550)
Total shareholder equities/deficits	**(8,093)**	**2,960**	**266**	**1,333**
Total liabilities and equities/deficits	**4,854**	**6,715**	**5,793**	**10,945**
CASH FLOW STATEMENT				
Operating cash flow	(2,246)	(2,348)	(1,973)	3,423
Investing cash flow	(376)	459	(31)	(35)
Financing cash flow	2,614	3,619	(153)	(138)
Other adjustments	(5)	-	-	-
Net changes in cash	**(13)**	**1,730**	**(2,156)**	**3,250**
Operating cash flow	(2,246)	(2,348)	(1,973)	3,423
Purchase/disposal of PP&E	(11)	(13)	(15)	(17)
Purchase/disposal of wealth management products	(312)	377	-	-
Lease payments	(211)	(176)	(153)	(138)
Cash burn	**(2,781)**	**(2,160)**	**(2,141)**	**3,268**

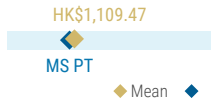
Risk Reward – MiniMax (0100.HK)

A Global AI Foundation Model Leader

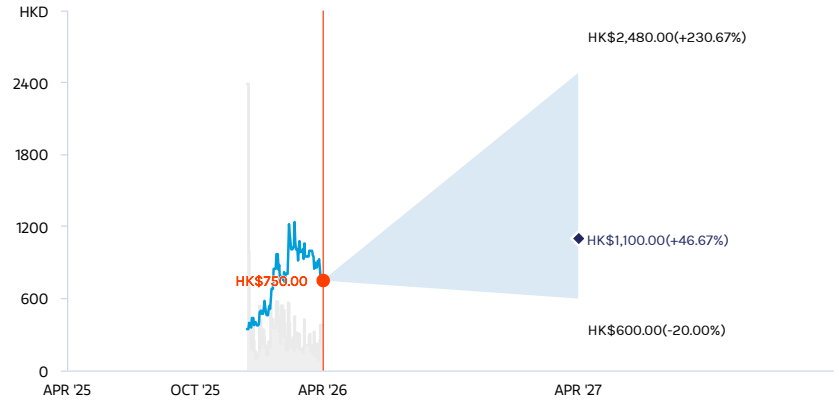
PRICE TARGET HK$1,100.00

We use a discounted cash flow (DCF) methodology to value the company, assuming a 15% WACC and 3% terminal growth. Our PT implies 36x P/S in 2027e.

Consensus Price Target Distribution



RISK REWARD CHART

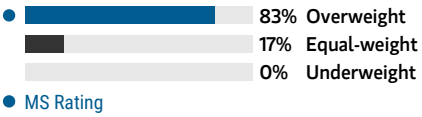


Key: Historical Stock Performance Current Stock Price Price Target

OVERWEIGHT THESIS

- We see significant global opportunities for AI foundation models driven by model performance breakthroughs with non-linear market expansion potential.
- We like MiniMax's strong industry position based on advanced technology, full-modality capabilities, diverse AI applications and highly scalable business model.
- The stock currently trades at 24x 2027e P/S based on our estimates and our DCF-derived PT implies 36x 2027e P/S. We see upside in the stock's valuation.

Consensus Rating Distribution



Risk Reward Themes

Disruption:	*Positive*
New Data Era:	*Positive*
Secular Growth:	*Positive*

View descriptions of Risk Rewards Themes here

BULL CASE HK$2,480.00

40x 2027 P/S

MiniMax's next-generation model, to be launched in mid-2026, becomes the global SOTA model with superior performance vs. peers, leading to rising customer demand, expanding market share and potential price hikes. We project 2027 revenue of US$2.5bn.

BASE CASE HK$1,100.00

36x 2027 P/S

MiniMax's next-generation model, to be launched in mid-2026, matches the performance of global SOTA models, leading to non-linear revenue growth in 2027. We project 2027 revenue of US$1.2bn.

BEAR CASE HK$600.00

30x 2027 P/S

MiniMax's next-generation model, to be launched in mid-2026, fails to outperform peers in terms of performance, resulting in linear top-line revenue growth rather than step-change growth in 2027. We project 2027 revenue of US$800mn.

Risk Reward – MiniMax (0100.HK)

KEY EARNINGS INPUTS

Drivers	2025	2026e	2027e	2028e
Total revenue % (%)	158.9	404.8	200.2	149.8

INVESTMENT DRIVERS

- Model performance iteration
- Enhanced application monetization
- Improving operating leverage

GLOBAL REVENUE EXPOSURE

View explanation of regional hierarchies here

RISKS TO PT/RATING

RISKS TO UPSIDE

- Launching global SOTA model
- Eased market competition

RISKS TO DOWNSIDE

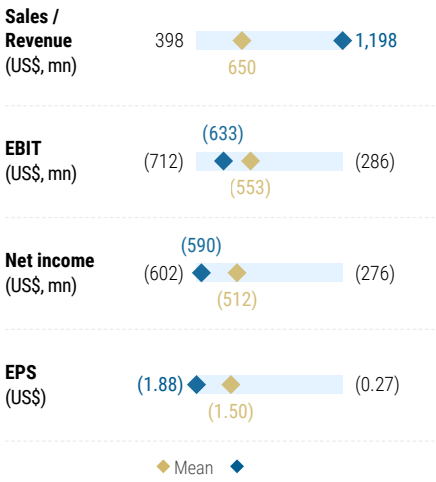
- Geopolitical risk
- Intensified competition and price war
- Model performance lagging peers

OWNERSHIP POSITIONING

Inst. Owners, % Active 100%

MS ESTIMATES VS. CONSENSUS

FY Dec 2027e



Risk Reward – Knowledge Atlas Technology JSC Ltd (2513.HK)

Pure AI Foundation Model Play

PRICE TARGET HK$990.00

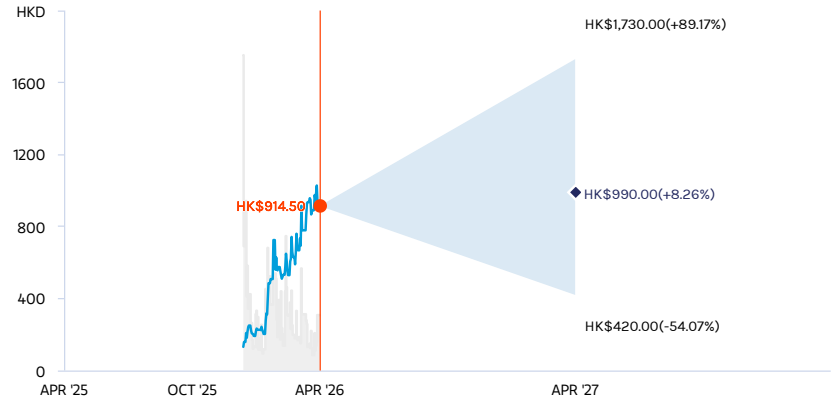
We use a discounted cashflow (DCF) methodology to value the company, assuming a 15% WACC and 3% terminal growth. Our PT implies 38x P/S in 2027.

Consensus Price Target Distribution HK$559.37 HK$927.54 HK$1,220.04

MS PT

Mean

RISK REWARD CHART



Key: Historical Stock Performance Current Stock Price Price Target

OVERWEIGHT THESIS

- Z.ai ranks among global leaders in AI foundation models. Its advanced capabilities are underpinned by consistent world-class flagship model launches, pioneering technology accumulation, a sole focus on LLM development, and a comprehensive model portfolio.
- The company is exposed to the early-stage, fast-growing China AI foundation model market focused on productivity scenarios, with a rising cloud revenue mix and improving model API pricing.
- The company trades at 35x 2027 P/S based on MSe, and our DCF-derived PT implies 38x 2027 P/S. We see upside to the company's valuation.

Consensus Rating Distribution

82% Overweight
18% Equal-weight
0% Underweight

MS Rating

Risk Reward Themes

Disruption:	*Positive*
New Data Era:	*Positive*
Secular Growth:	*Positive*

View descriptions of Risk Rewards Themes here

BULL CASE HK$1,730.00

40x 2027 P/S

- Z.ai launches a global SOTA model with superior performance to peers and expands its customer base through cloud-based API sales with continuous price hikes, leading to non-linear revenue growth in 2027

- Revenue of Rmb17,250mn in 2027

BASE CASE HK$990.00

38x 2027 P/S

- Z.ai consistently launches foundation models that match global SOTA model performance and expand its customer base mainly through cloud-based API sales, leading to non-linear revenue growth in 2027

- Revenue of Rmb9,735mn in 2027

BEAR CASE HK$420.00

30x 2027 P/S

- Z.ai fails to keep up with global SOTA model performance, leading to linear revenue growth rather than step-change growth in 2027

- Revenue of Rmb5,520mn in 2027

Risk Reward – Knowledge Atlas Technology JSC Ltd (2513.HK)

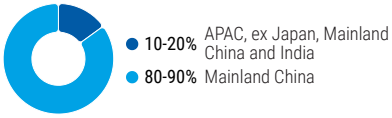
KEY EARNINGS INPUTS

Drivers	2025	2026e	2027e	2028e
Revenue growth (%)	131.9	428.5	162.3	134.2
GP Margin (%)	41.0	35.6	38.5	42.4

INVESTMENT DRIVERS

- Model performance iteration
- Rising cloud revenue
- Improving operating leverage

GLOBAL REVENUE EXPOSURE



View explanation of regional hierarchies here

RISKS TO PT/RATING

RISKS TO UPSIDE

- Launching global SOTA model
- Easing market competition

RISKS TO DOWNSIDE

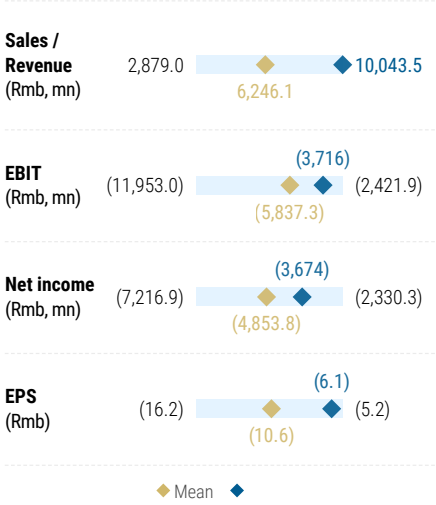
- Geopolitical risk
- Intensified competition and price war
- Model performance lagging peers

OWNERSHIP POSITIONING

Inst. Owners, % Active 100%

MS ESTIMATES VS. CONSENSUS

FY Dec 2027e



Valuation Methodology and Risks

Tencent Holdings Ltd. (0700.HK)

HK$650: Base case, sum of the parts.

Core business - HK$560: DCF (10% discount rate, 3% terminal growth rate)

Associate investments - HK$90/share (based on reported book value) and 30% discount applied to investment value.

Risks to Upside

- Solid execution in new game launches, both domestically and overseas
- Market share gain in social and short video ads
- Resilience in social network and online entertainment competition
- 2C AI adoption ramp-up

Risks to Downside

- Gaming industry regulatory uncertainties
- Intensified competition in social networks and ad budget with emerging entertainment formats
- Tightened regulations on antitrust and amid US/China tension

Alibaba Group Holding (BABA.N)

Base case, discounted cash flow model. Key assumptions include a 10% WACC and 3% terminal growth rate, in line with our Chinese Internet coverage range.

Risks to Upside

- Better core e-commerce monetization, driving earnings growth upside
- Faster enterprise digitalization, re-accelerating cloud revenue growth
- Stronger AI demand to drive cloud revenue

Risks to Downside

- More intense competition
- Higher-than-expected reinvestment costs
- Weaker consumption amid a slower post-Covid recovery
- Slower pace of enterprise digitalization pace
- Regulation - additional scrutiny of Internet platforms

Risk Reward Reference links

1. View explanation of Options Probabilities methodology - Options_Probabilities_Exhibit_Link.pdf

2. View descriptions of Risk Rewards Themes - RR_Themes_Exhibit_Link.pdf

3. View explanation of regional hierarchies - GEG_Exhibit_Link.pdf

4. View explanation of Theme/Exposure methodology - ESG_Sustainable_Solutions_External_Link.pdf

5. View explanation of HERS methodology - ESG_HERS_External_Link.pdf