

NAVIGATING CHINA AI MODELS

LLM primer: Critical point of intelligence for global proliferation

From DeepSeek's moment last year (on cost efficiency) to Zhipu's GLM moment this year (on model intelligence)

China's AI open-source/open-weight models are reaching a critical point of intelligence performance vs. global proprietary models, with a significant ramp up in domestic enterprise & global SME adoption that will enable a positive data flywheel of further model improvement, in our view. We evaluate *How* these models achieve such performance at low costs/tight computing resources; *Why* they pursue an open-source/open-weight approach and how they monetize; *What* the key addressable markets are, as global enterprises shift from 'token-maxxing' to ROI-first, where we highlight two favored ARR quadrants for Chinese models; and *Who* the potential long-term winners will be under our Competitive Positioning framework. We note any stricter access to China's future most frontier models for overseas markets, Western markets' policies on China models, and access to high-end computing in model training as three key swing factors/risks to our China AI model token/revenue growth trajectory.

We introduce our Chinese AI model Competitive Positioning framework based on pricing power, cost advantage and financial strength, overlaying with token scale and market share progressions, and identify Knowledge Atlas (Zhipu, initiation) and DeepSeek (private) as the strongest positioned in foundation models, and ByteDance (private) in multi-modal.

Ronald Keung, CFA

+852-2978-0856
ronald.keung@gs.com
Goldman Sachs (Asia) L.L.C.

Eric Sheridan

+1(917)343-8683
eric.sheridan@gs.com
Goldman Sachs & Co. LLC

Allen Chang

+852-2978-2930
allen.k.chang@gs.com
Goldman Sachs (Asia) L.L.C.

Lincoln Kong, CFA

+852-2978-6603
lincoln.kong@gs.com
Goldman Sachs (Asia) L.L.C.

Timothy Zhao

+852-2978-2673
timothy.zhao@gs.com
Goldman Sachs (Asia) L.L.C.

Damian Xie

+852-2978-1398
damian.xie@gs.com
Goldman Sachs (Asia) L.L.C.

Iris Xiao

+852-2978-0202
iris.xiao@gs.com
Goldman Sachs (Asia) L.L.C.

Steve Qiu

+852-2978-2672
steve.qiu@gs.com
Goldman Sachs (Asia) L.L.C.

Luqing Zhou

+852-3465-4207
luqing.zhou@gs.com
Goldman Sachs (Asia) L.L.C.

Eunice Liu

+852-2978-7472
eunice.liu@gs.com
Goldman Sachs (Asia) L.L.C.

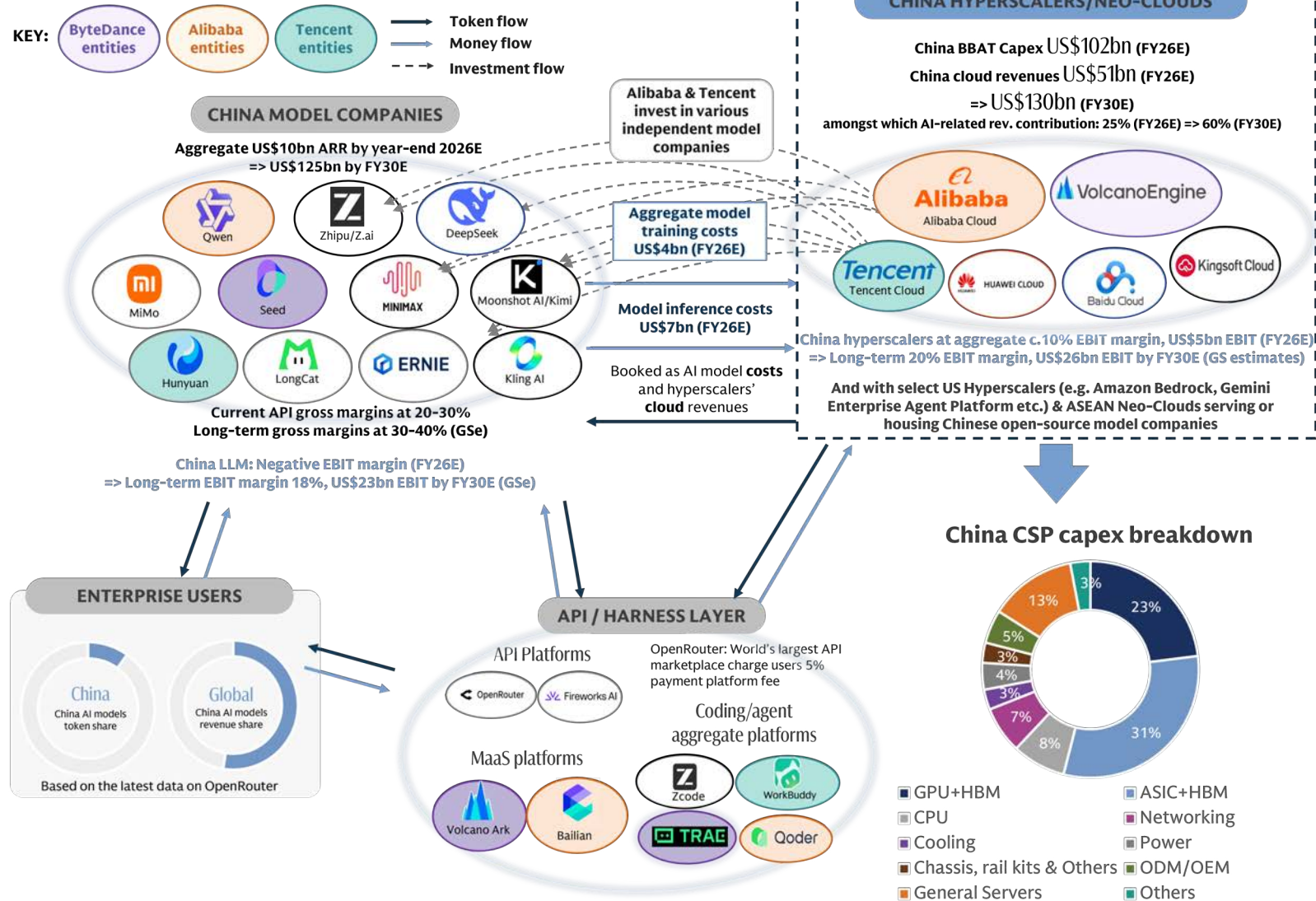
Goldman Sachs does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision. For Reg AC certification and other important disclosures, see the Disclosure Appendix, or go to www.gs.com/research/hedge.html. Analysts employed by non-US affiliates are not registered/qualified as research analysts with FINRA in the U.S.

Table of Contents

Dissecting China open-source models unit economics today and path to profitability	4
PM Summary	5
Comparing China's key LLM players	11
Decoding China's AI model tokens & our forecasts on token/revenue share	12
How do Chinese models achieve competitive performance at low costs/tight computing resources?	13
Why are Chinese models pursuing an open source/open weight approach, and ways to monetize?	20
What are the key addressable markets, domestically and internationally, and key risks?	23
Who are best positioned to be the long term winners? Introducing our Competitive Positioning framework	26
Foundation models	26
Multi-modal generation models	32
Key upside and downside risks	37
China's key AI players at a glance	40
Price Target Methodology & Risks	47
Disclosure Appendix	49

CHINA AI MODEL & HYPERSCALERS ECOSYSTEM

Our mapping of key players, revenue/profit pools & interdependencies



1) The calculation of China cloud revenues includes Alibaba, Tencent, ByteDance, Baidu, and Kingsoft Cloud. 2) China CSP capex breakdown is based on GS Greater China Tech team estimates; 3) Companies included are for illustrative purposes and may not be an exhaustive list.

Source: Company data, Goldman Sachs Global Investment Research

Dissecting China open-source models unit economics today and path to profitability

Exhibit 1: Unit economics for China's value-for-money agentic model and top performing coding model, vs. hypothetical economics for global SOTA model

UNIT ECONOMICS OF TWO-TIERED CHINA AI MODELS (PER 1M BLENDED TOKENS)

	Current (Year-end 26E)			Long-term	
	China's value-for-money agentic model	China's best performing coding model	Hypothetical economics for 'most frontier' SOTA model	China's value-for-money agentic model	China's best performing coding model
API Pricing	\$0.2	\$1	\$3.9	\$0.2	\$2
Inference cost as % of ARR	-0.15 -74%	-0.70 -70%	-0.98 -26%	-0.11 -57%	-1.02 -51%
Gross Profit (ARR - inference cost)	0.03	0.18	2.70	0.07	0.80
% Gross margin	10-15%	15-20%	70%	30-40%	30-40%
Training as % of ARR	-0.06 -30%	-0.40 -40%	-1.89 -49%	-0.03 -13%	-0.25 -13%
EBIT (Gross profit - Training cost & OpEx)	-0.06	-0.39	0.00	0.03	0.44
% EBIT margin	-30%	-39%	0%	14%	22%

CHINA'S VALUE FOR MONEY MODELS BLENDED PRICE: US\$0.2 per blended 1M tokens



(C) **CHINA'S TOP PERFORMING CODING MODELS BLENDED PRICE: US\$1 per blended 1M tokens**



IN ABSOLUTE US\$ BILLIONS

	Current (Year-end 26E)			Long-term	
	China's value-for-money agentic model	China's best performing coding model	Hypothetical economics for 'most frontier' SOTA model	China's value-for-money agentic model	China's best performing coding model
ARR, US\$	\$0.8bn	\$1.5bn	\$55bn	\$10.9bn	\$30.6bn
Inference cost as % of ARR	-\$0.6bn -74%	-\$1.1bn -70%	-\$14bn -26%	-\$6.2bn -57%	-\$15.6bn -51%
Gross Profit (ARR - inference cost)	\$0.1bn	\$0.3bn	\$39bn	\$3.6bn	\$12.2bn
% Gross margin	10-15%	15-20%	70%	30-40%	30-40%
Training as % of ARR	-\$0.2bn -30%	-\$0.6bn -40%	-\$27bn -49%	-\$1.6bn -13%	-\$4.4bn -13%
EBIT (Gross profit - Training cost & OpEx)	-\$0.2bn	-\$0.6bn	-\$0.1bn	\$1.5bn	\$6.7bn
% EBIT margin	-30%	-39%	0%	14%	22%

← We estimate aggregate of US\$10bn ARR for China AI model industry by year-end FY26E; we estimate to reach US\$125bn for FY30E

← US\$7bn inference costs for China AI model industry in FY26E, we estimate US\$55bn by FY30E, which translates into hyperscalers'/neo-clouds' revenues

← US\$4bn training costs for China AI model industry in FY26E, US\$20bn by FY30E, as model sizes continue to scale up

← Negative EBIT margin for China AI model companies in FY26E; we estimate to reach 18% EBIT margin/US\$23bn profit pool by FY30E

Source: Company data, Goldman Sachs Global Investment Research

PM Summary

From DeepSeek's moment last year (on cost efficiency) to Zhipu's GLM moment this year (on model intelligence)

China's AI open-source/open-weight models are reaching a critical point of intelligence performance vs. global proprietary models, with a significant ramp up in domestic enterprise adoption and a proliferation of global consumer and SME demand. In this report, we evaluate How these models achieve such performance at low costs/tight computing resources; Why they pursue an open source/open weight approach and how they monetize; What the key addressable markets are, where we highlight two favoured ARR quadrants and risk factors; and Who the potential long-term winners will be under our Competitive Positioning framework.

We believe Chinese models are reaching a critical 'good enough' stage for agentic tasks/specific coding scenarios, and rising fragmentation in China's AI model landscape (but the strong will get stronger). While pricing power remains strong for models with frontier performance/multi-modal, the lower-end segment is in a price war; nevertheless, Agentic AI is driving explosive demand for these value-for-money models at the lower-end. Access to computing will be a swing factor, where US/China regulations, balance sheet and inference efficiency are key.

Accordingly, we introduce our Chinese AI model Competitive Positioning framework based on pricing power, cost advantage and financial strength, overlaying with token scale and market share progressions, and identify Knowledge Atlas (Zhipu, initiation) and DeepSeek (private) as the strongest positioned in foundation models, and Bytedance (private) in multi-modal.

- **How do Chinese models achieve competitive performance at low costs/tight computing resources?** By leveraging smaller-sized parameter models for equivalent benchmark performance, and Mixture-of-Expert and new architectural innovations. Chinese AI models are bifurcating into a two-tiered market where performance and time to market are key to pricing power, thereby leading to two 'ARR maximising' quadrants based on token adoption and pricing. A positive flywheel is taking effect for top Chinese AI models driven by increasing actual real world coding adoption, and reducing their reliance on model distillation practices.
- **Why are Chinese models pursuing an open source/open weight approach, and ways to monetize?** Open source allows for greater flexibility in model training/deployment, and allows for the widest adoption and an open community. We believe open source models' disclosed ARRs are likely understating total deployment and revenue potentials, and we expect more shifts to open weight (with Community License, i.e., commercial terms if for commercial use) among Chinese AI models down the road.
- **What are the key addressable markets, domestically and internationally, and key risks?** We highlight two favoured ARR quadrants and estimate China/China AI models market to see token growth of 25X by 2030E; the coding landscape

to consolidate while the agentic/low-end segment could remain fragmented. International (going global) should be a key upside, with the potential for higher pricing and global proliferation, especially in non-US markets as global enterprises increasingly pivot from a token-maxxing to a ROI-first model that prioritizes clear task boundaries, number of agents per day, back-end process automation and actual output over pure computational token volume. Key risk factors are market access/anti-distillation and regulations, including any stricter access to China's future most frontier models for overseas markets, access to highest-end leased computing equipment (used for training), market access to western markets, further restriction lists/entity list designations (but this could be positive for China's path to further AI self-sufficiency across software/CPU/ASICs), and competition from SLM/threat from AI architectures.

- **Who are best positioned to be the long term winners within China's AI model companies?** We expect players with the largest ARR scale with a gross margin advantage + financial strength to be the long-term winners. We highlight independent AI model companies mostly stand out in our Competitive Positioning framework in pricing power + cost advantage (in aggregate represent over US\$200bn in implied valuations, based on latest market cap/funding rounds), where Zhipu and DeepSeek are the most strongly positioned in text based foundation models, while ByteDance leads in multi-modal capabilities based on our framework.

Signposts we are looking for into 2H 2026

- **Harness/agentic applications:** We expect China AI model companies to increasingly focus on positioning their harness/agentic applications as key entry points, especially in coding (e.g. Zhipu's ZCode, Tencent's Workbuddy, and Alibaba's Qoder which are aggregate platforms that support the full suite of AI models) as model companies attempt to close the loop in capturing more real-life coding and agentic data for scaling. Co-work and industry expert agentic products could be the next priorities. Enterprises will increasingly focus on overall cost per task instead of headline pricing per token, with an openness to using different models (multiple-models approach).
- **Multiple large parameter high-end coding model launches and potentially stricter access to the most advanced Chinese AI models outside of China:** We anticipate multiple new Chinese AI model launches over 2H26 with significantly larger total parameter sizes of 2-5 trillion across Chinese AI model players. Coding/programming segment competition will intensify as Chinese models try to challenge Zhipu GLM's leadership via training on high-quality real-life coding data (where available) and scaling to larger parameter model sizes. We also expect a potential shift from an open-source to an open-weight approach for best-performing models (i.e. from free-for-all use cases to requiring revenue sharing/a take rate for commercial use). The addition of multi-modal/visual understanding will be the next upgrades amongst Chinese AI foundation models (e.g. for GLM, DeepSeek), while MiniMax M3 already excels in these given the multi-modal focus of MiniMax from the start. That said, we note [press reports](#) on potential future restrictions on overseas access to China's most advanced AI models (both closed and open source if at the frontier level) as cutting-edge artificial intelligence is increasingly being seen as a critical national asset.

- **Continued suppressed API pricing for the lower end agentic-focused segment:** While DeepSeek recently announced an increase in peak-hour pricing from mid-July, we expect API pricing and therefore gross margins to remain under pressure at the lower-end pricing segment (around US\$0.1-0.2 per 1M tokens) into the second half, as China's AI model players have significant cash buffers post-fund raising to subsidize competitive pricing at zero/negative gross margins in the near term. As a result, we see financial strength as one of the three most important metrics (alongside pricing power and cost efficiencies) in assessing a Chinese AI model company, where cash on hand, net cash as % of assets and valuation multiples will be the critical financial strength metrics for long-term success. This said, we are Buy-rated on MiniMax as the company stands out on cost efficiency/cost advantage metrics under our Competitive Positioning framework. With its M3 model well positioned in the favored ARR maximizing quadrant (attractive pricing + high token volumes), alongside its discounted valuation at 13X P/2026E year-end ARR (vs. China/global peers which command multiples several times higher at similar ARR stage), we believe risk-reward is skewed to the upside and reiterate our Buy rating. Key swing factors for MiniMax to improve its overall competitive positioning will hinge on its pricing power and financial strength. We expect its H3 video generation model performance (imminent launch in a more favorable industry landscape vs. text models) and time-to-market for its next M3 updates (focusing on further coding intelligence level uplift from post-training/reinforcement learning, we estimate over July-Aug, and a larger parameter size M3 model later in 2H 2026) will be the next key drivers.
- **Multi-modal/video-generation models to see further ARR ramp-up from global adoption:** We expect continued healthy industry pricing and gross margins within video generation (unlike foundation text) where key players ByteDance's SeeDance, Kuaishou's (Buy) Kling and MiniMax's Hailuo/upcoming H3 models to enjoy healthy growth over 2H 2026 amid new functionality breakthroughs (combination of video-generation with LLM) and tight computing resources where demand significantly outpaces capacity. According to China [news reports](#) like LatePost and 36Kr, ByteDance's Seedance gross margins have been at a healthy 70% at its latest US\$2bn+ ARR run-rate.
- We also expect increasing domestic ASICs supply, tighter access to overseas computing resources and potential market access limitations (could mirror TikTok's trajectory where rapid expansion in western markets was followed by more regulations/focus on ensuring data security where computing will have to be conducted within local jurisdictions).

Our assessment of AI model companies: ARR scale x gross margin advantage + financial strength

- Largest ARR scale (Token scale x pricing power)
- Gross margin advantage (Training & Inference efficiency, Technology)
- Financial strength (Balance sheet, Access to computing)

Accordingly, we lay out our Competitive Positioning framework for AI model players based on quantifiable metrics, **1) Pricing Power** (amongst which, based on Time-to-market of model launch, Arena score based on actual usage cases and pricing

level), **2) Cost advantage** (based on token volume scale, throughput/cache hit rate, parameter sizes/activation ratio and our estimate of inference gross margins), and **3) Financial strength** (based on cash on-hand, net cash as % of assets, and valuation multiples).

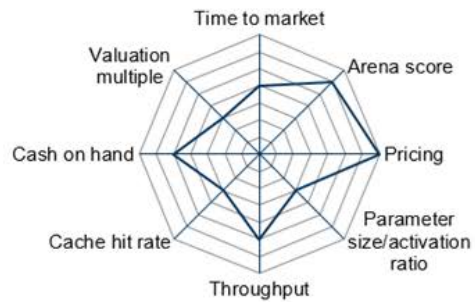
Exhibit 2: Competitive positioning framework for AI model players

Key themes	Chinese models reaching critical stage of intelligence for wide adoption across agentic tasks/specific coding scenarios			Fragmentation between China AI models but strong to get stronger; pricing power to come from frontier performance/multi-modal			Agentic AI driving higher demand for value-for-money models/SLMs			Access to computing, where balance sheet and inference efficiency are key		
Drivers of corporate performance	Competitive positioning											
CP category	Pricing Power				Cost advantage				Financial strength			
CP metrics	Time to market	Arena score	Pricing	Token volume	Throughput /Cache hit rate	Parameter size/ activation ratio	Inference GPM	Cash on hand	Net cash as % of asset	Valuation multiples		

Source: Goldman Sachs Global Investment Research

Exhibit 3: Overview of competitive analysis for key LLM labs' positioning

Alibaba Qwen: Full-stack AI player with multi-year AI/cloud upside, leading position of most areas



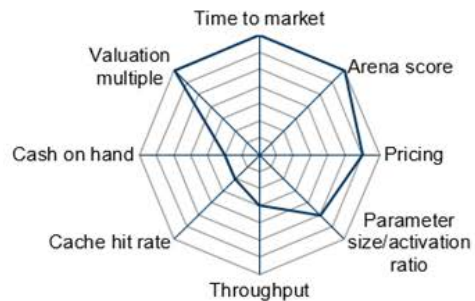
Tencent Hunyuan: Strong contender after AI strategic revamp; potential in Weixin AI agent



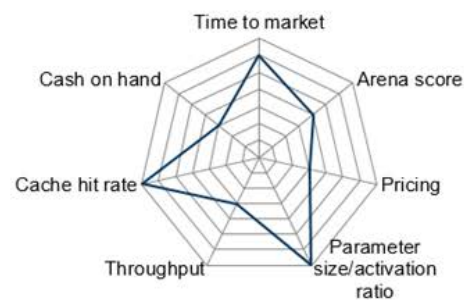
MiniMax M-series: Unique full-modal capabilities among independent players; cost-efficient computing architectures



Zhipu GLM: Leading model capabilities in coding/agentic scenarios



DeepSeek: Extreme cost efficiency to sustain scalable token usage on competitive pricing

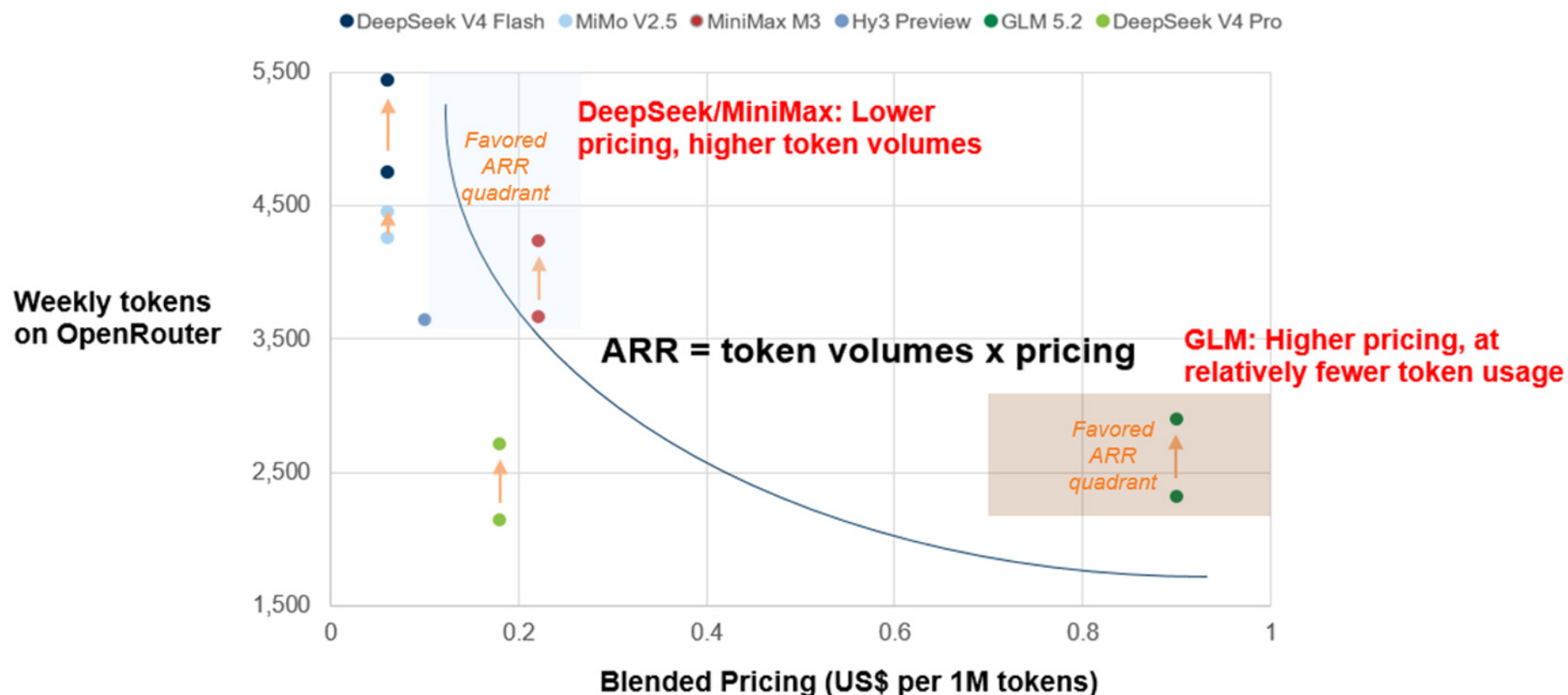


ByteDance Seedance: leading in model performance with strong financial/ecosystem back-up for quick adoption & commercialization



N/A for DeepSeek/ByteDance valuation multiple

Source: Goldman Sachs Global Investment Research

Exhibit 4: We identify two favorable ARR quadrants (in maximizing ARR) which is a combination of maximizing token volumes and/or pricing level

Source: Company data, OpenRouter, Goldman Sachs Global Investment Research

Related research

Americas Technology: Decoding the Agentic Economy - The Coming Inflection in AI Usage and Margins, 5 May 2026, by James Schneider, Eric Sheridan and team

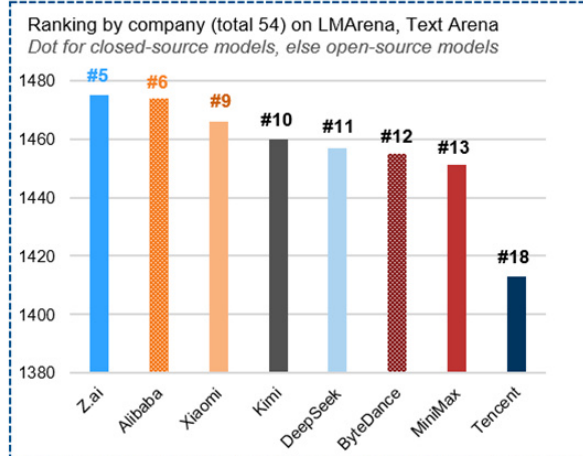
Global Server: Raising both AI and General servers; better mix and higher memory costs drive CSP spending, 24 June 2026, by Allen Chang and team

Knowledge Atlas Technology (2513.HK): Initiate with Neutral at US\$110bn valuation; China's top enterprise/coding AI model, July 10, 2026

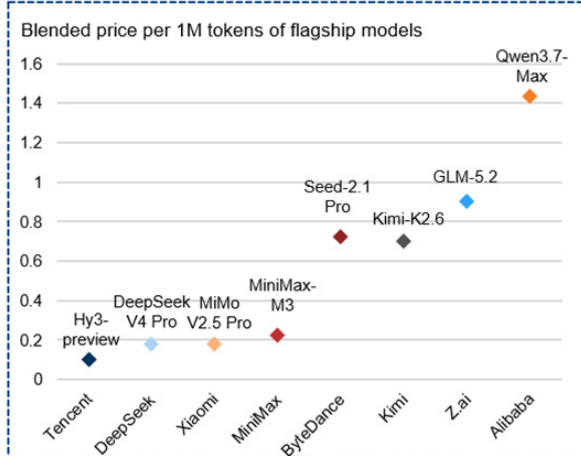
MiniMax Group (0100.HK): Key focuses post-share unlock; upside skew to risk-reward; Buy, July 10, 2026

Comparing China's key LLM players

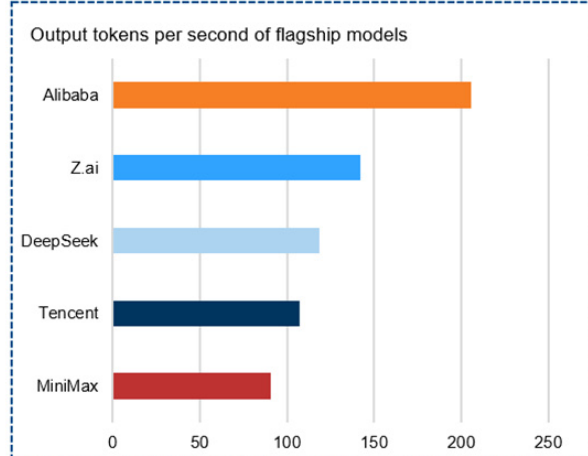
Performance: China LLM labs keep climbing up the Text Arena ranking, with flagship models hitting new high



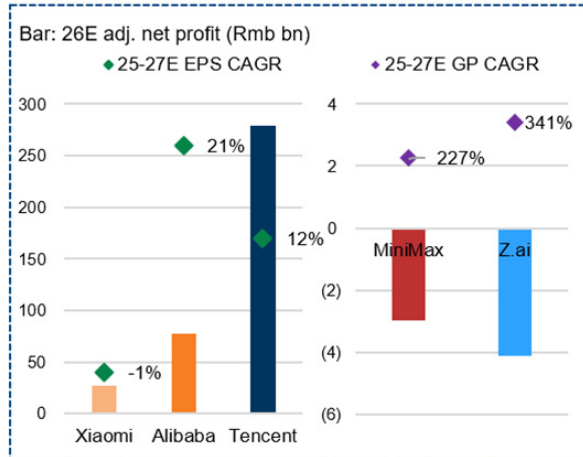
Two-tier price: premium pricing on model intelligence vs. low cost for price-sensitive demands



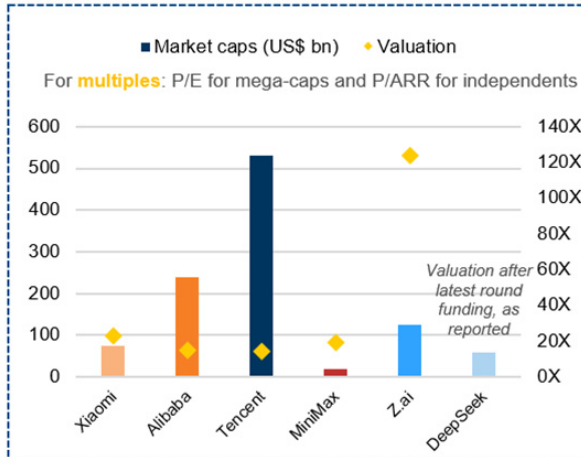
Speed: first-party API performance to reflect inference efficiency and infra capabilities



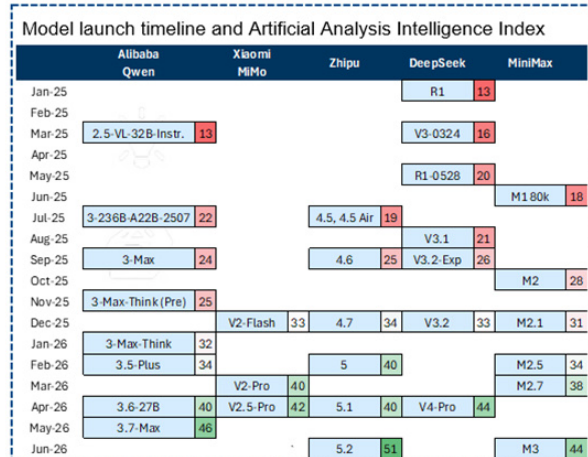
P&L: mega-caps AI funding from stable profits while independents still scaling in commercialization



Valuation: latest market caps and reported private valuation for LLM players



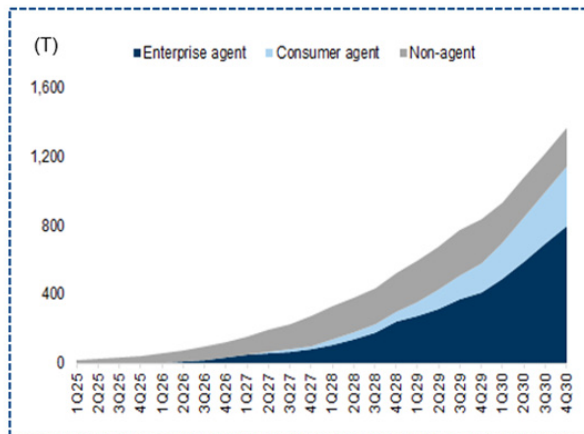
Model updates into 2026: faster iteration with rising intelligence level



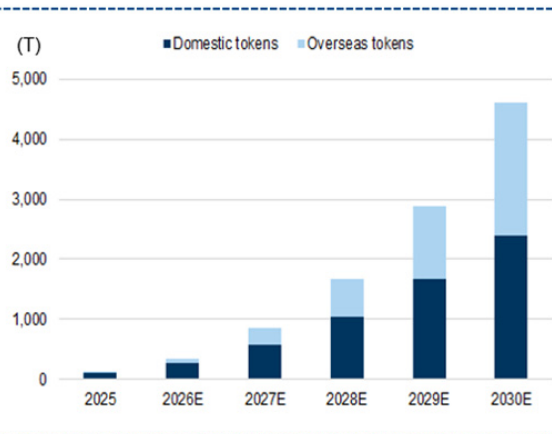
Source: Company data, Goldman Sachs Global Investment Research, WSJ, Artificial Analysis

Decoding China's AI model tokens & our forecasts on token/revenue share

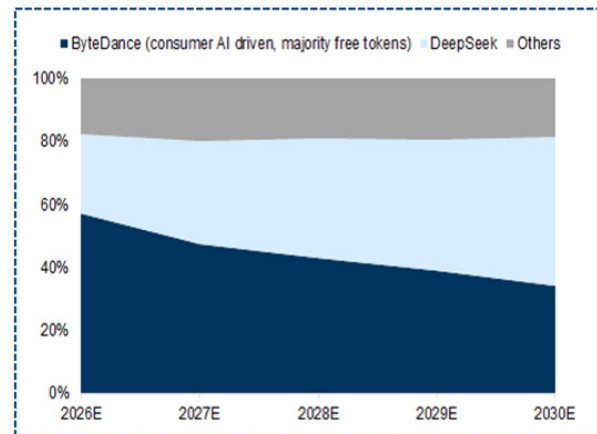
By 2030, we estimate tokens generated by China AI model companies will be 25X higher than today



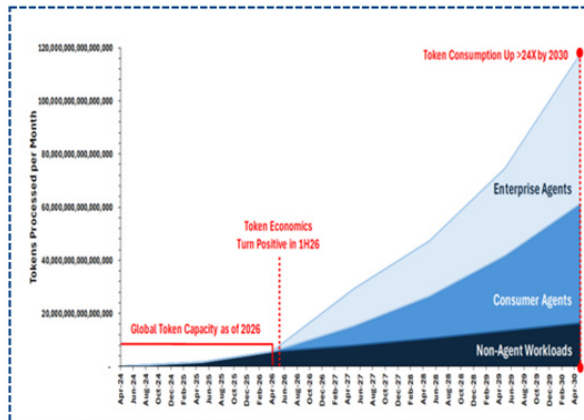
We estimate overseas tokens generated by China AI model companies will reach 55% of global tokens by 2030E



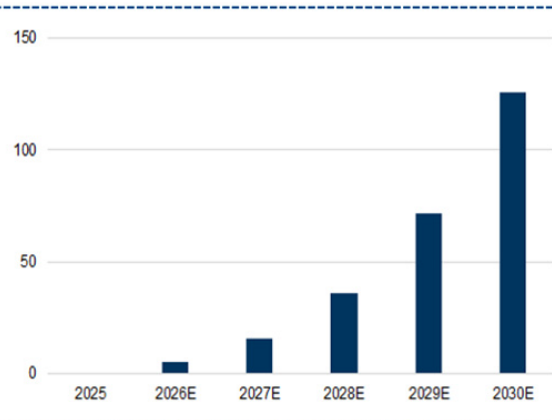
We expect ByteDance and DeepSeek to account for 80% of tokens generated by China AI model companies by 2030E



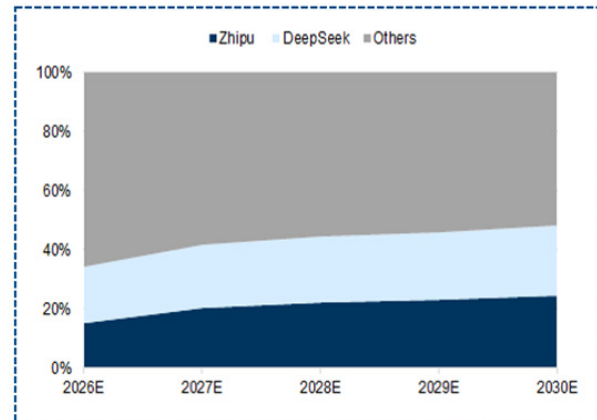
Our US research team estimates consumer and enterprise agents could push token consumption 24X above today's global capacity (ex-China)



We estimate China AI API + subscription revenue pool to reach US\$125bn in 2030E



We expect Zhipu and DeepSeek to capture most of the incremental revenue pool



DeepSeek/ByteDance/Zhipu's (Not Covered) relevant metrics (e.g. token consumption/revenue) are calculated by analyzing the industry and companies that we cover, and then extrapolating them to DeepSeek, ByteDance and Zhipu.

Source: Goldman Sachs Global Investment Research

How do Chinese models achieve competitive performance at low costs/tight computing resources?

Smaller-sized parameters models for equivalent benchmark performances, Mixture-of-Expert and new architectural innovations

Compared with a year ago, Chinese models' coding and agentic capabilities have reached a critical level in being able to complete more coding and autonomous agentic tasks with higher success rates as context windows have been expanded to 1 million tokens. The smaller parameter model sizes of Chinese models (spanning from 200bn to 1.6T parameters, at 2-10% of leading SOTA models, due to constrained access to high-end computing), and highly efficient architectures (MoE, Sparse Attention, OCR etc., at 3-5% activated parameters only vs. total parameter sizes) all contribute to the much lower training and inference costs for Chinese models vs. leading US models. We would attribute the recent step improvement of Chinese models in coding according to Arena.ai to data curation, distillation techniques and reinforcement learning post training, despite their relatively small parameter model sizes (1.6T for DeepSeek V4 Pro, 0.7T for Zhipu's GLM5.2 and 0.4T for MiniMax's M3). On June 27, DeepSeek introduced DSpark, a speculative decoding framework that makes existing DeepSeek-V4 models serve faster. DSpark has already been deployed in DeepSeek-V4 Flash / Pro online serving, improving per-user DeepSeek-V4 generation 60-85% faster on V4-Flash and 57-78% faster on V4 Pro without changing the model's weights or output quality.

Chinese AI models bifurcating into two-tiered market (where performance and time to market are key to pricing power), with two 'ARR maximising' quadrants based on token adoption and pricing

We are seeing pricing power for the highest performing Chinese AI models, e.g. Zhipu's GLM5.2 model and Alibaba's Qwen3.7 Max models at around US\$1 per blended 1M tokens, at 5X that of low-end Chinese AI models. The reported tighter US processes in allowing most SOTA model access has also opened new arenas for China's top performing coding models for enterprises and SMEs. Smaller parameter and activated ratios allow China's top performing models to be priced at US\$1, 10-25% vs. US SOTA models at US\$4-8 per blended 1M tokens, while generating double digit 10-20% gross margins (GSe) that are lower than global SOTA models due to relatively lower pricing power.

In the lower end segment, agentic focused models are priced at US\$0.06-0.2 per blended 1M tokens, which are enabling these models to tap into new global TAMs for price sensitive SMEs and one-man companies. MiniMax generates 60-70% revenues from overseas.

We also note DeepSeek announced that its V4 official version is set for launch in mid-July, alongside the introduction of peak/off-peak API pricing to better allocate resources and improve service stability. V4 Pro/Flash non-peak pricing remains unchanged, while peak hour (9am-12pm/2pm-6pm China time) will be charged at 2X non-peak rates, implying blended

pricing of US\$0.35/US\$0.12 per 1M tokens due to strong Chinese AI model demand that is increasingly causing significant compute tightness in work/productivity scenarios.

Positive flywheel is taking effect for top Chinese AI models from increasing actual real world coding adoption, with less reliance on model distillation ahead

Compared with learning and distillation tactics from global SOTA models in the past, we believe top Chinese models like GLM5 are reaching a critical stage of adoption by China's major enterprises and global users. As per LatePost, AI-generated code has increased to as high as 90% at some China mega-cap companies, up from 20-30% in 2H25, which we believe will enable a positive data flywheel effect of further improvements via reinforcement learning with actual user data (both successful and unsuccessful coding cases) and post training. We think these have underpinned the step improvements in GLM5.2 from GLM5.1 in just over the course of a few months in 2026, and we expect to see further step improvements to Chinese AI models over the next 6-12 months.

Exhibit 5: China's model count is rising in top rank bands across usage (text, coding and agent)

LMarena ranking of Text, WebDev (Coding) and Agent Arena

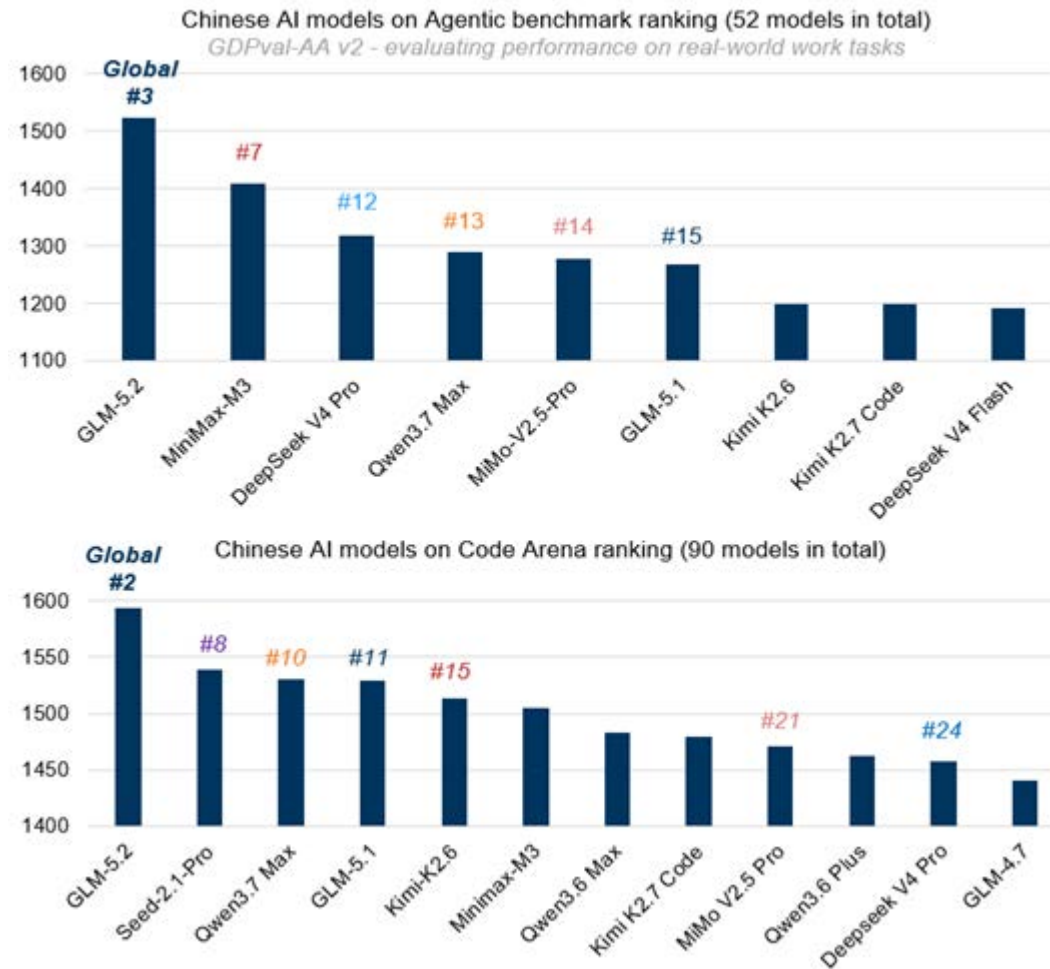
Text Arena			Code Arena			Agent Arena		
	3 months ago	Current		3 months ago	Current		early-June	Current
Top 10% # of China model ranked within the top	7	10 ▲ 3	Top 10% # of China model ranked within the top	0	2 ▲ 2	Top 30% # of China model ranked within the top	0	1 GLM 5.2 at #7
Top 30% # of China model ranked within the top	43	49 ▲ 9	Top 30% # of China model ranked within the top	6	11 ▲ 5	Top 50% # of China model ranked within the top	1	3 ▲ 2
Top 50% # of China model ranked within the top	70	83 ▲ 13	Top 50% # of China model ranked within the top	9	20 ▲ 11	By agentic domain 		
Leading China AI labs 			Best China model rank GLM-5 #9/56 → ZHIPU-AI → GLM-5.2 #2/90			Legal LegalBench Finance Finance Agent v2		

Baseline date for Text/WebDev (Coding) Arena ranking = 3/18/2026, and current date refers to latest date available, i.e. 6/16/2026 & 6/20/2026. Agent Arena ranking started at June 5, 2026, with current date = 6/30/2026

Source: LMArena, Data compiled by Goldman Sachs Global Investment Research

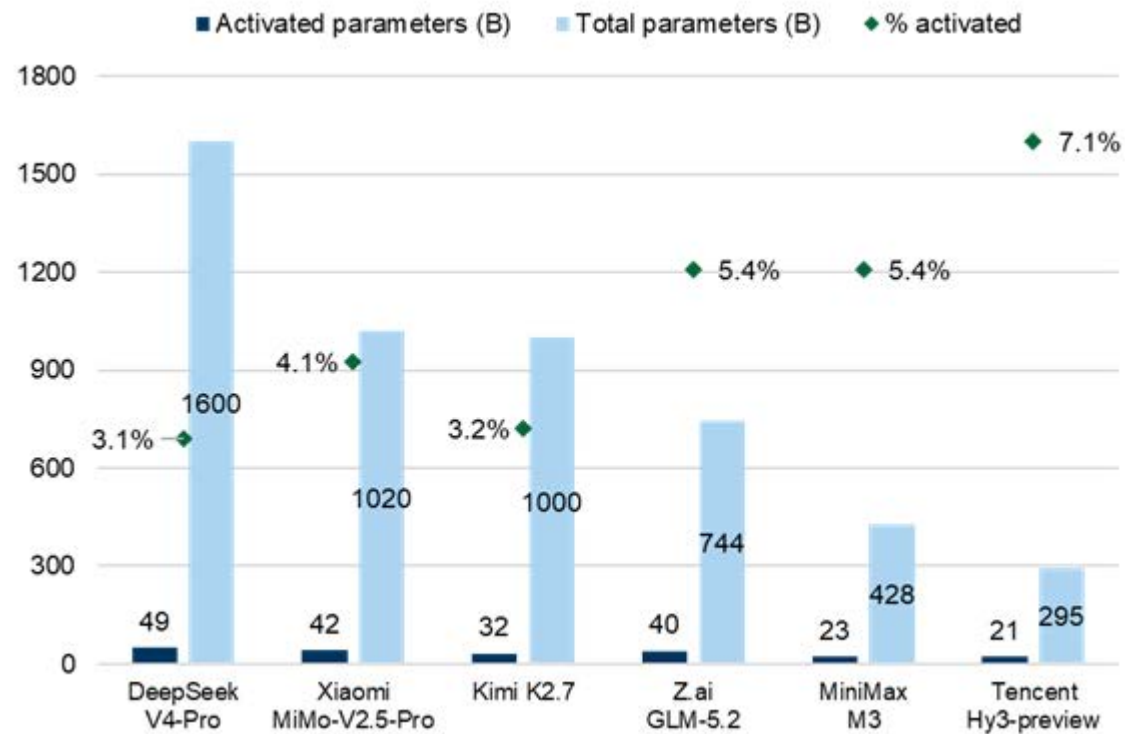
Exhibit 6: Notably in coding/agent tasks, Chinese players are reaching global top-tier positions

As of Jun 20, 2026



Source: GDPval-AA, LMArena, Data compiled by Goldman Sachs Global Investment Research

Exhibit 7: Compared with global SOTA leading models of several tens of trillions, China open source models are mostly around or below 1 trillion parameter in size, and adopt a MoE structure, with low activated to total parameter ratios for higher inference efficiency



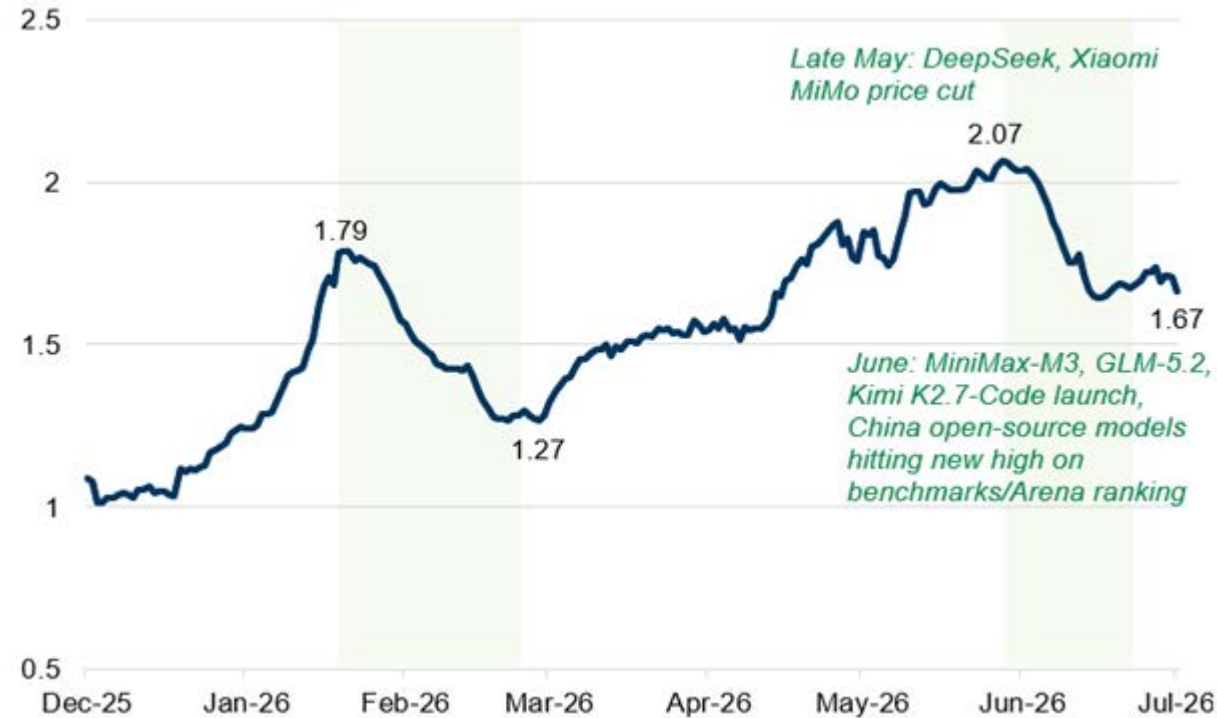
A Mixture-of-Experts (MoE) structure is a sparsely activated neural architecture, where it dynamically selects a small subset of expert networks to process each input, allowing the model to increase its total capacity without activating all parameters during inference.

Source: Company data, Data compiled by Goldman Sachs Global Investment Research

Exhibit 8: Blended LLM token price (SDLLMTK) has been declining since early June, potentially driven by rising adoption of cost effective China models

US\$ per 1M tokens

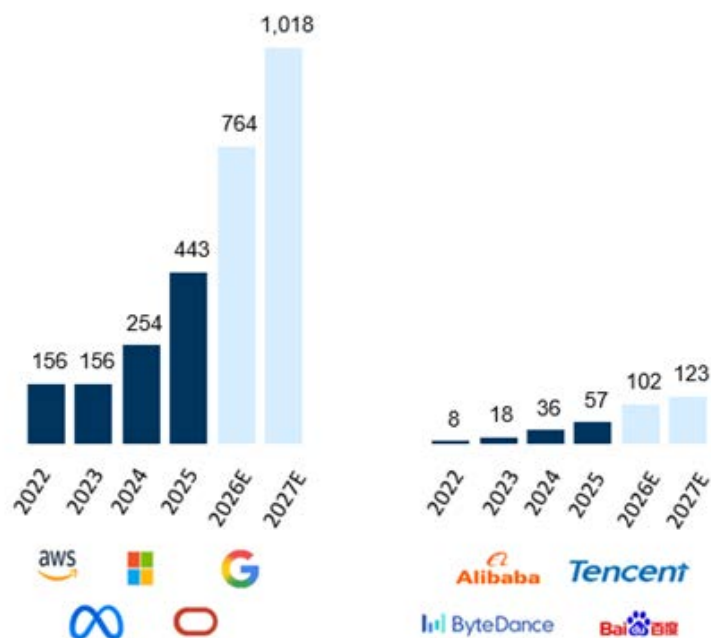
US\$ per 1M tokens



The Silicon Data LLM Token Expenditure Index (SDLLMTK) is a daily blended measure of what the market actually pays to run LLM inference across frontier and open-weight models.

Source: Silicon Data, Bloomberg, Data compiled by Goldman Sachs Global Investment Research

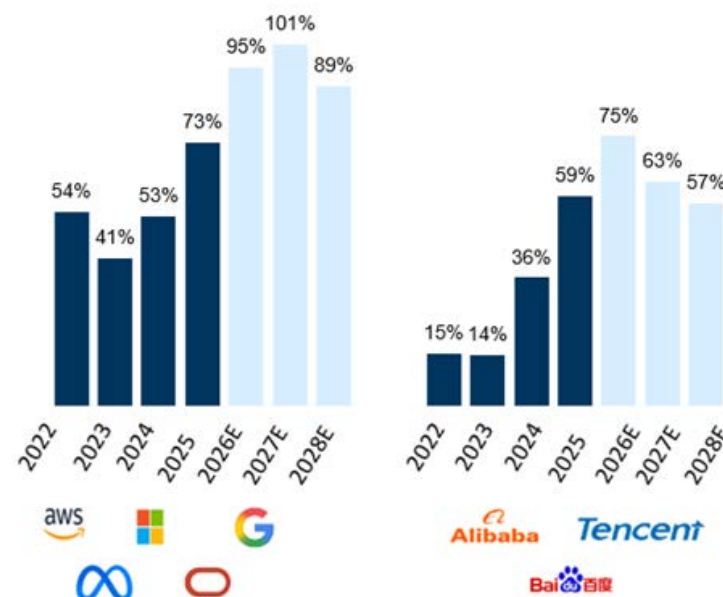
Exhibit 9: Historical and projected capex for major US & China cloud service providers
US\$bn



We calculate ByteDance's (Not Covered) relevant operating metrics by analyzing the industry and companies that we cover, and then extrapolating them to ByteDance.

Source: Company data, Goldman Sachs Global Investment Research

Exhibit 10: Capex to operating cash flow ratio is still healthy for China hyperscalers



Source: Company data, Goldman Sachs Global Investment Research

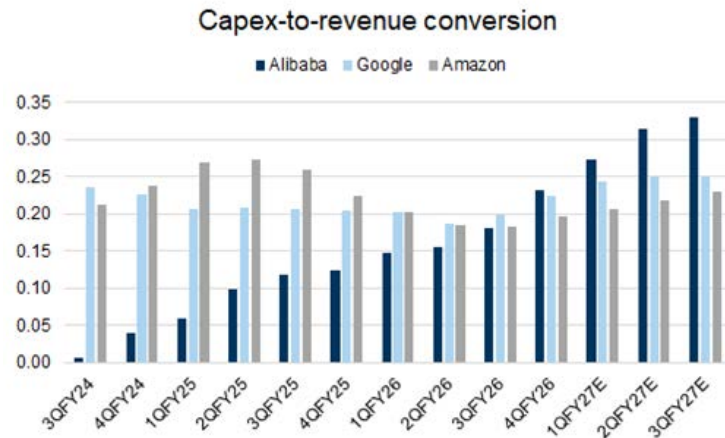
Case study on Meituan's LongCat 2.0: A milestone for China's domestic AI infrastructure

Released on June 30, 2026, we believe Meituan's LongCat 2.0 marks a major milestone as China's first official 1.6 trillion-parameter open-source Mixture-of-Experts (MoE) model trained and deployed entirely on a 50,000-card domestic compute cluster. Purpose-built for agentic coding and complex software engineering workflows, the model features a native 1-million-token context window enabled by LongCat Sparse Attention (LSA) and dynamically activates an average of 48 billion parameters per token to optimize inference costs. **Implications for a more self-sustainable China AI model outlook that is less reliant on foreign high-end chips for model training:** The successful end-to-end pre-training and inference of a trillion-parameter class model on Chinese silicon (reportedly utilizing Huawei Atlas-950 SuperPods) fundamentally drives a more sustainable China's AI model development outlook, in our view. While previous Chinese flagship models, such as DeepSeek V4-pro, mentioned domestic chips for inference, LongCat 2.0's ability to overcome critical memory bottlenecks

and distributed stability challenges during the compute-heavy pre-training phase proves the viability of a wider and more localized hardware stack for AI model training in the future.

Exhibit 11: Alibaba capex-to-revenue conversion ratio vs. Amazon and Google

Rolling 12m incremental cloud revenue/rolling 12m capex (lagged by 1 quarter)

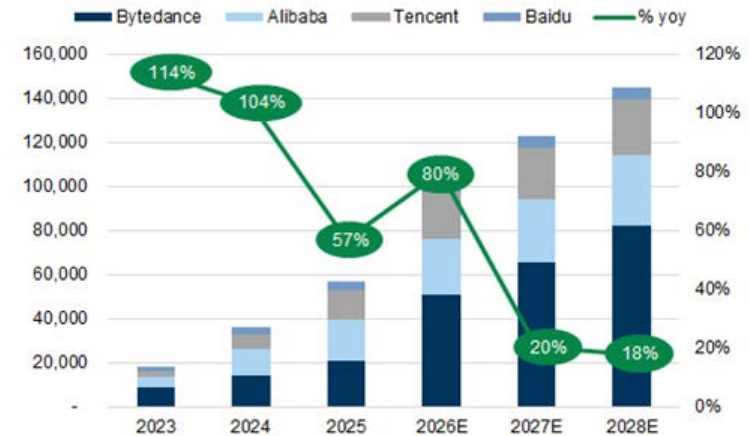


We benchmark Alibaba external cloud revenue growth to Google and Amazon Cloud (lagged c.2 years)

Source: Company data, Goldman Sachs Global Investment Research

Exhibit 12: China Internet yearly capex forecast

US\$mn



For ByteDance (Not Covered), we impute capex based on industry trends.

Source: Company data, Goldman Sachs Global Investment Research, Bloomberg

Why are Chinese models pursuing an open source/open weight approach, and ways to monetize?

Open source allows for higher flexibility in model training/deployment, and allows for the widest adoption and an open community

Alibaba's Qwen model family has long pursued an open source approach (before adopting a closed source for its largest and highest performing Qwen-Max models for better monetization), while other key Chinese AI model players have mostly pursued an open source/open weight approach including DeepSeek, Zhipu's GLM and MiniMax M3 series models with the exception of ByteDance's full closed proprietary approach for its Seed model. We believe the open source approach allows for the highest flexibility in terms of the locations of where models are trained (and thus where the models can be deployed both inside and outside of Mainland China after training). An open source approach also allows for the highest adoption amongst the AI community with full transparency of the model parameters/architecture for trust, and an open community in driving more user feedback and thus model iterations/improvements. The open source approach vs. world's leading closed/proprietary approach also provides an alternative choice for worldwide AI users when the best performing closed models have stricter user access and higher costs from their premium pricing, especially at a time when 'token-maxxing' has become a key cost consideration for many corporates.

Open source models' disclosed ARR are likely understating total deployment and revenue potentials

While open source model companies offer their own coding plans and a chargeable open platform API channel (where the model companies conduct their own model inference), the majority of open source models also allow individuals/third-party hyperscalers/neocloud providers to deploy the models without a charge even for commercial use (e.g. Alibaba Cloud's Bailian MaaS platform can house GLM5.2 open source model without needing to pay a fee/take rate to Zhipu). As a result, while Zhipu's last stated ARR target for year-end 2026 is at US\$1bn, we believe the actual deployment of GLM5.2 model worldwide is and will be multiple-fold higher vs. Zhipu's own API channel token volumes and revenue. There is also increasing post training of Chinese AI models that are re-branded by global companies (e.g. Composer 2 etc.) where the Chinese AI models do not necessarily receive any revenues given the open source spirit.

We expect more shifts to open weight (with Community License) approach among Chinese AI models down the road

While Zhipu's open source GLM model's MIT license allows for free for all use cases (regardless of revenue), MiniMax M series models have pursued a restricted license (where the industry terms it as an open weight with Community License model), which requires MiniMax's agreement and commercial terms (e.g. revenue sharing/a take rate) on commercial use. We believe this will be the likely next path for other open source models in the Chinese AI model industry, in order for eventual gross profits of inference tokens to cover training costs and for each AI model company to achieve a sustainable path to returns.

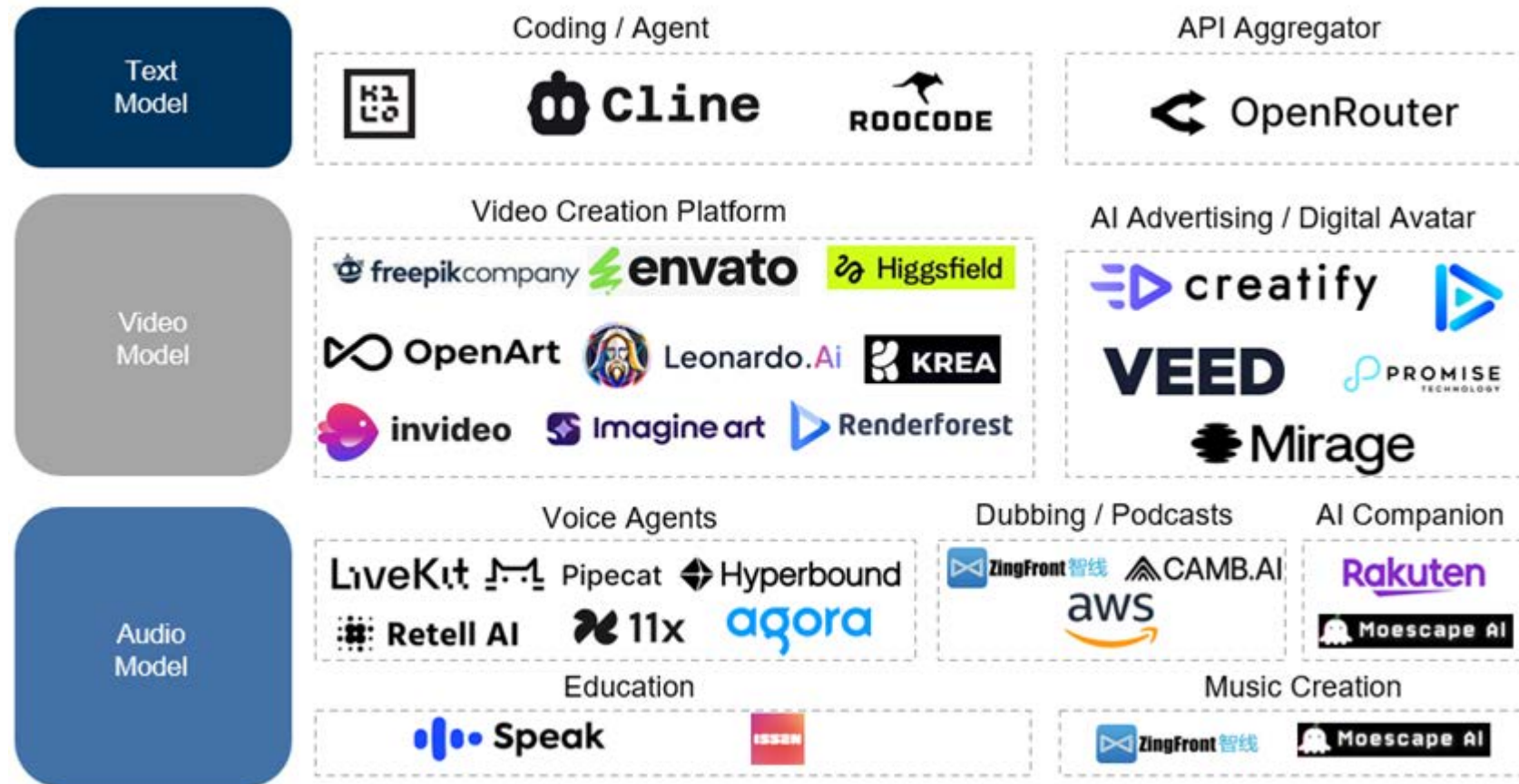
Exhibit 13: Comparison of different open source licenses

Feature	Zhipu GLM 5.2	MiniMax M3
License Type	MIT License (Permissive)	MiniMax Community License (Restricted)
Commercial Use	Unrestricted. Free for all users regardless of revenue	Restricted. Free only for entities with <US\$20M annual revenue
Approval Process	None. No permission needed to deploy commercially	Required. Must email api@minimax.io . <US\$20M revenue needs notice; >\$20M requires formal written license
Attribution	Standard copyright notice retention (MIT terms)	Mandatory "Built with MiniMax" label on products/UI
Prohibitions	Standard legal limitations apply	Explicit restrictions on military use, hate speech, and illegal content

Source: Company data, Data compiled by Goldman Sachs Global Investment Research

What are API platform services?

Chinese AI models typically monetize via subscription plans or their in-house API platform services. API access is typically offered on third party platforms like OpenRouter, Cline or coding agent platforms. As we expect more open source models could shift to open weight, we believe the unit economics of Chinese AI model companies could improve as not only can official API channels be monetized, but also by revenue sharing agreements when these open weight models are deployed and sold on third-party neocloud/hyperscalers like AWS Bedrock, Alibaba Cloud Bailian. Such arrangements could become highly margin accretive as AI model companies will not need to conduct inference/bear computing costs themselves.

Exhibit 14: API platform international enterprise customers through which customers can access Chinese AI models

Source: Company data, Data compiled by Goldman Sachs Global Investment Research

What are the key addressable markets, domestically and internationally, and key risks?

We estimate China AI models' aggregate API+subscription revenue to increase from Rmb35bn in 2026E to Rmb879bn by 2030E from rising model intelligence, in particular with recent models like GLM5.2 reaching a critical point for global adoption and attractive pricing. Our revenue pool estimates for China AI models imply total Chinese AI model daily token consumption of 350T in 2026E to increase to 4,600T by 2030E.

We estimate domestic market to see token growth of 25X by 2030E; coding landscape to consolidate while agentic/low-end segment could remain fragmented

At 140tn daily token volumes for the country in March 2026 and several hundred trillions by June 2026 as per the National Bureau of Statistics, we estimate open source/open weight models have roughly a 30% token share (vs. 70% token market share by ByteDance alone, which is closed source and mainly driven by its Doubao app enterprise and individual user base, as the #1 used AI chatbot in China). We expect similar to the US, the coding segment (at a premium pricing level) will continue to be dominated by SOTA best performing models. Meanwhile, we expect the lower-end segment focused on agentic AI will remain fragmented with multiple players due to the financial strength of AI model companies/mega-caps which would sustain the lower-end segment price war for longer.

International (going global) to be the key upside; with potential for higher pricing and global proliferation, especially in non-US markets

Our US research team estimates ([link](#)) agentic AI will drive 24X growth in token consumption by 2030 (from 2026) to 120 quadrillion tokens per month (or 4 quadrillion tokens per day, from their estimate of 170 trillion daily tokens today), with the biggest driver at 55X from enterprise agents and 12X from consumer agents. The global (ex. China) landscape has seen significant token share gains from Chinese AI models, as rising model intelligence and attractive token costs have driven higher adoption across 24/7 Hermes/Claw/Productivity agents, and shifting global SME mindset on using Chinese models in managing token costs given Chinese models have reached a 'good enough' stage in terms of intelligence/performance.

Pivoting from 'token-maxxing' to ROI-focused metrics, e.g. Daily Active Agents/Agentic Work Units

We believe the AI token proliferation is undergoing a paradigm shift from 'token-maxxing' (an initial focus from late 2025 to early 2026 where enterprises equated high AI token consumption directly with organization productivity) towards an 'ROI-first' model that prioritizes clear task boundaries and output over raw computational volume.

- **'Token-maxxing' has been attributed to corporate inefficiencies and cost overruns:** Data from a Jellyfish AI Engineering trends [study](#) indicated heavy AI users at enterprises consumed 10X more tokens but only had a 2X increase in

output. Meanwhile, we note multiple US Internet companies have commented back in April 2026 that either their engineering teams utilized a full year's AI budget in just four months using agentic products (promoting stricter monthly caps per tool) or changed their internal token utilization leaderboards that previously wrongly incentivized staff to launch inefficient/low value autonomous agent tasks.

- **Enterprise framework is pivoting towards an ROI-first model that prioritizes clear task boundaries, number of agents per day, backend process automation and actual output over pure computational token volume.** Besides a marked increase in adoption of Chinese AI models, global enterprises have been downgrading default models to cheaper flash models for more typical tasks (while reserving SOTA models for only the top most value creating tasks like coding/programming). We see a transition from just token tracking to metrics like Daily Active Agents (DAA) and Agentic Work Units (AWU), and believe the overall cost per task will become more relevant than price per token metrics.

Open source/open weight approach allows for the option for U.S. hyperscalers to host Chinese models, operated within U.S. cloud ecosystem

Alphabet and Amazon's (both covered by Eric Sheridan) respective cloud services Gemini Enterprise Agent Platform (via. its Model Garden) and AWS Bedrock already offer a broad selection of Chinese AI models including DeepSeek, MiniMax, Moonshot, GLM and Qwen which are fully managed by the US hyperscalers. Besides the model layer, into applications, worthy of note is Microsoft (covered by Gabriela Borges) CEO's recent remarks at a Wall Street Journal interview ([link](#)) where he noted Microsoft is considering hosting versions of DeepSeek on Copilot as an optional, cost-effective model which could give its customers access to cheaper choices alongside US proprietary models. Microsoft indicated that if it hosts DeepSeek, the model would operate within its cloud ecosystem, ensuring customer data stays inside Azure.

Noting key risks around any potential tighter geopolitical policies given Chinese AI models inroads into western markets.

Key risks to the 'going global' opportunity will hinge on end market access (especially in western countries, and with a focus on where computing is done/data is stored), access to high-end computing for AI model training (that could impact iteration pace and cost structure of Chinese models), and restrictions on Chinese AI model companies could impact supplier relationships/access to US companies and/or access to US capital. More in the key risks section.

Exhibit 15: China AI models now only account for limited share of spend across different tasks, per OpenRouter

As of week starting Jul 1, 2026



Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

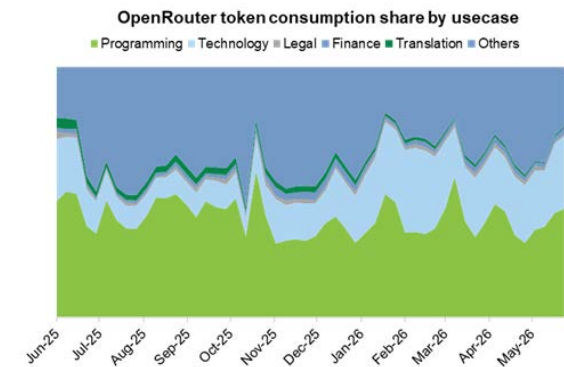
Exhibit 16: While China AI models already capture majority of token usage on price competitiveness, esp. in long duration coding/agentic tasks

As of week starting Jul 1, 2026



Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

Exhibit 17: Token consumption is concentrating in complex, high value scenarios like programming/technology



Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

Who are best positioned to be the long term winners? Introducing our Competitive Positioning framework

We expect players with the largest ARR scale with a gross margin advantage + financial strength to be the long-term winners.

ARR scale x gross margin advantage + financial strength

- Largest ARR scale (Token scale x pricing power)
- Gross margin advantage (Training & Inference efficiency, Technology)
- Financial strength (Balance sheet, Access to computing)

Accordingly, we introduce our Chinese AI model Competitive Positioning framework based on pricing power, cost advantage and finance strength, overlaying with token scale and market share progressions, and identify Knowledge Atlas (Zhipu initiation) and DeepSeek (private) as the most strongly positioned in foundation models, and Bytedance (private) in multi-modal capabilities.

Foundation models

We assess each key player's competitive positioning from its flagship foundation model, scored across 3 aspects: pricing power, cost advantage and financial strength (of the company), with each aspect further built upon granular quantitative metrics.

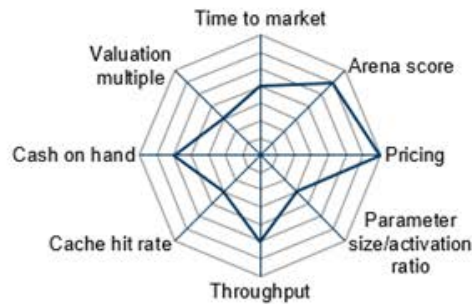
Exhibit 18: Competitive positioning framework for AI model players

Key themes	Chinese models reaching critical stage of intelligence for wide adoption across agentic tasks/specific coding scenarios			Fragmentation between China AI models but strong to get stronger; pricing power to come from frontier performance/multi-modal			Agentic AI driving higher demand for value-for-money models/SLMs		Access to computing, where balance sheet and inference efficiency are key	
Drivers of corporate performance	Competitive positioning									
CP category	Pricing Power			Cost advantage				Financial strength		
CP metrics	Time to market	Arena score	Pricing	Token volume	Throughput /Cache hit rate	Parameter size/ activation ratio	Inference GPM	Cash on hand	Net cash as % of asset	Valuation multiples

Source: Goldman Sachs Global Investment Research

Exhibit 19: Overview of competitive analysis for key LLM labs' positioning

Alibaba Qwen: Full-stack AI player with multi-year AI/cloud upside, leading position of most areas



Tencent Hunyuan: Strong contender after AI strategic revamp; potential in Weixin AI agent



MiniMax M-series: Unique full-modal capabilities among independent players; cost-efficient computing architectures



Zhipu GLM: Leading model capabilities in coding/agent scenarios



DeepSeek: Extreme cost efficiency to sustain scalable token usage on competitive pricing



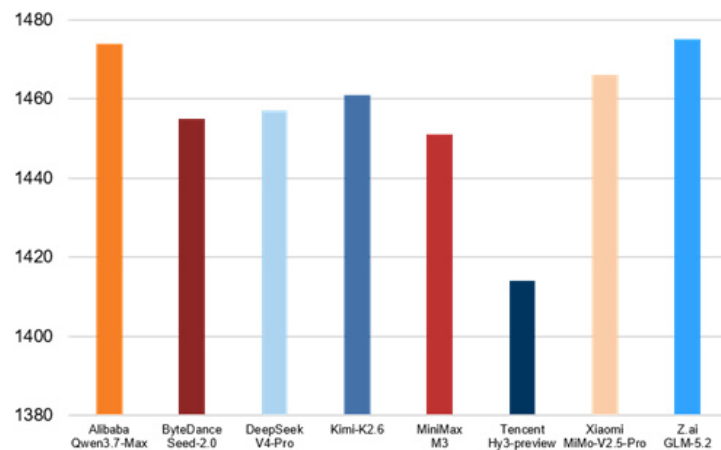
N/A for DeepSeek valuation multiple

Source: Goldman Sachs Global Investment Research

Pricing Power

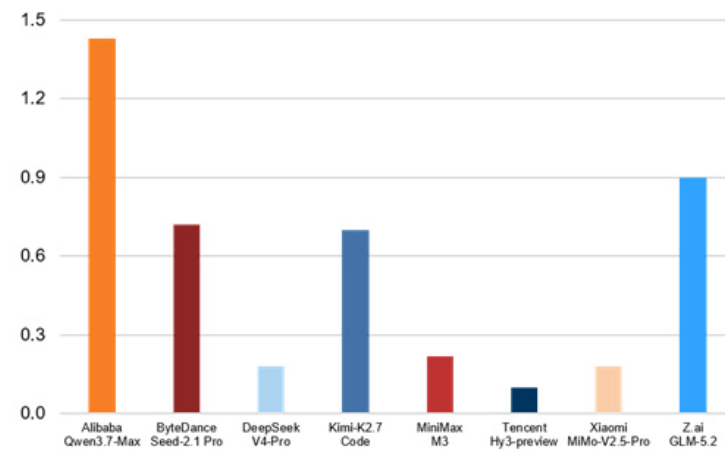
- **Time to market:** We assess this in terms of how quickly and effectively a player delivers frontier competitive models. From comparing the launch date of the company's flagship models with its prior generation & other models at similar performance level, we aim to assess whether the release delivers a meaningful capability step-up over its past generation, and whether it narrows the gap to the current global SOTA models.

- **Arena score (overall text):** We view model intelligence as the primary determinant of pricing power, since models capable of handling higher-value tasks can deliver clearer ROI and therefore sustain their pricing premium. We use LMArena's score instead of any static benchmark because it reflects a large scale of blind user reviews, making it a more objective read on the real-world capability.
- **Pricing (blended, US\$ per 1M tokens):** We also refer to the realized headline price (across input/output/cached input) for flagship models, where sustained higher pricing with iteration signals stronger pricing power, whereas pricing cuts may indicate a more volume-prioritized strategy.

Exhibit 20: Text Arena score of key domestic AI models

As of June 2026

Source: LMArena, Data compiled by Goldman Sachs Global Investment Research

Exhibit 21: Weighted pricing of key player models
US\$ per 1M tokens

As of June 2026

Source: Artificial Analysis, Goldman Sachs Global Investment Research

Cost advantage: the structural cost-to-serve (i.e. inference costs) that sets the floor for price and margin

We refer to the below metrics as proxies for inference efficiency:

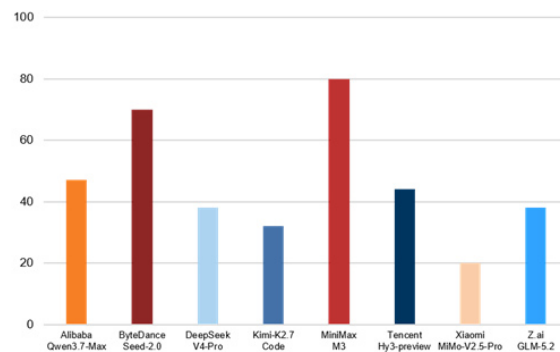
- **Throughput (tokens/second):** Measured by the number of tokens a single GPU generates per second. The more tokens a GPU outputs per second, the more fixed hourly compute cost is spread out. Therefore, throughput is a strong indicator of inference efficiency, which depends on model architecture (sparsity & attention design) and serving efficiency (batching & utilization).
- **Cache hit rate (%):** The share of input tokens served from cache rather than recomputed. In conversations and agentic workflows, much of the input repeats across calls, so models can read those tokens from cache memory instead of

recomputing from new inputs. A higher rate cuts compute, and since cached tokens are near-free to serve despite being billed at a discount (10X-100X cheaper than input cost), it is also margin accretive.

- **Parameter size/activation ratio:** The share of total parameters activated per token (available for open-source MoE models), where fewer active parameters mean fewer FLOPs per token and therefore lower inference cost, at any given level of performance. We also read the absolute parameter size since calculation scales with this degree of freedom.
- **Inference GPM (where disclosed, or GS estimates based on total parameter size/activated parameters):** The realized gross margin on model API, where ~90% of COGS is inference costs, as a direct indication on cost efficiency.

Exhibit 22: Throughput of key player models (tokens per second)

Processing capacity per unit time, representing how fast the model writes

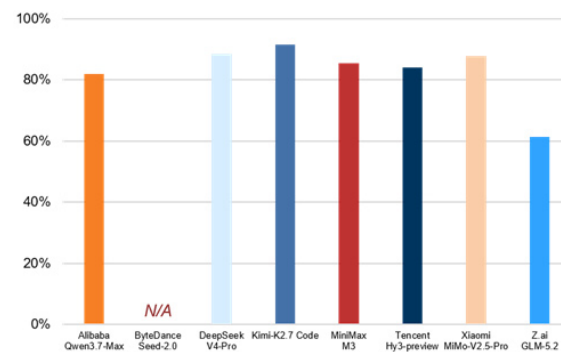


As of June 2026

Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

Exhibit 23: Cache hit rate of key player models

% of request served from cache memory instead of new inputs, higher = better for cost/speed

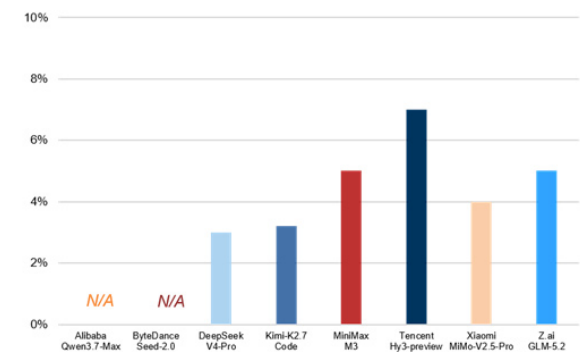


As of June 2026

Source: OpenRouter, Data compiled by Goldman Sachs Global Investment Research

Exhibit 24: Active to total parameter ratio of key player models

Share of parameters activated per token; lower means cheaper inference costs given the same performance



Parameter data unavailable for closed-source models (Alibaba's Qwen and ByteDance's Seed); As of June 2026

Source: Company data, Data compiled by Goldman Sachs Global Investment Research

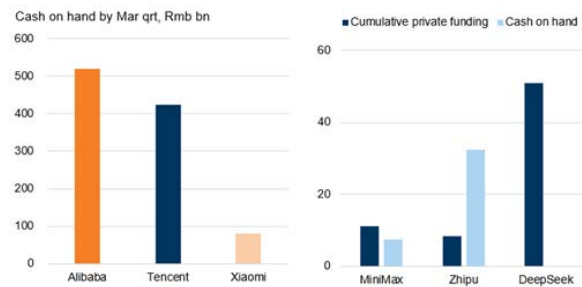
Financial strength: capacity to keep funding frontier R&D and training before a profitability turnaround

Cash on hand and net cash/debt as % of assets: We use total cash on hand as a measure of balance sheet strength, with net cash as % of assets to normalize across players of different scale. We see mega-caps (compared with individual players) as having greater resources to accumulate compute, fund multi-modal exploration and push faster product distribution.

Valuation multiples: For independents, we use P/ARR 2026E multiples as they are not yet profitable and P/ARR is a more comparable measure of business scale and future monetization potential, and for mega-caps we use P/E 2026E.

Exhibit 25: Cash on hand for key players

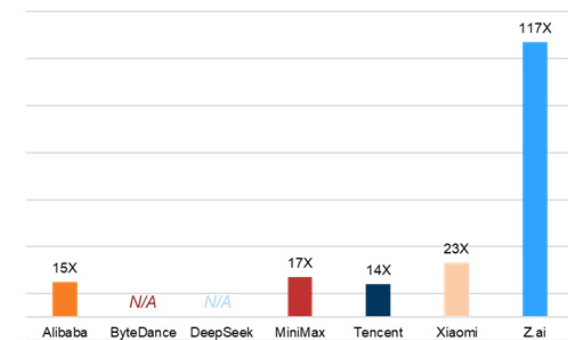
Rmb bn



Source: Company data, The Information, Data compiled by Goldman Sachs Global Investment Research

Exhibit 28: Valuation multiples for key players

Mega-caps valuation on P/E, independent players valuation on P/ARR, 2026E

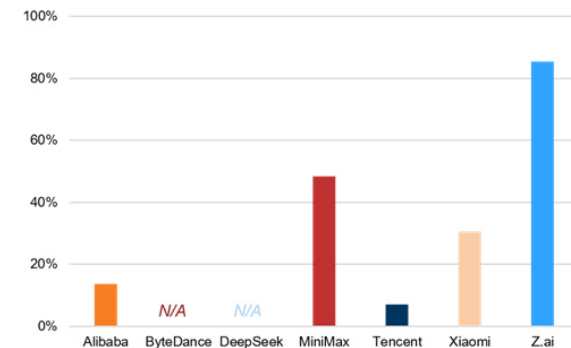


P/E for Alibaba, Tencent, and Xiaomi; P/ARR for MiniMax and Z.ai; Data as of Jun 26, 2026

Source: Company data, Goldman Sachs Global Investment Research

Exhibit 26: Net cash as % of total asset for key players

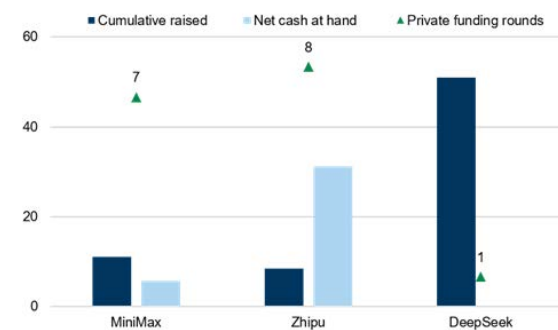
As of July 2026



Source: Company data, Goldman Sachs Global Investment Research

Exhibit 29: Independent AI players' cumulative external private funding (actual) and GSe for net cash as of end-2026E

Rmb bn

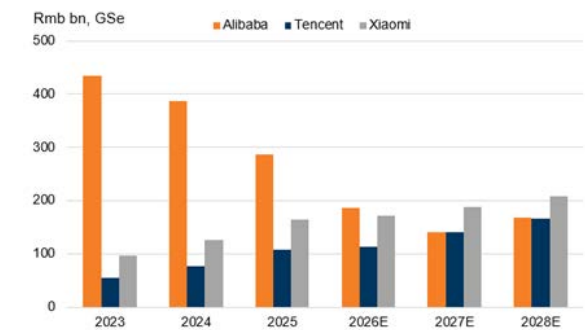


N/A for DeepSeek's net cash at hand

Source: Company data, Goldman Sachs Global Investment Research

Exhibit 27: Net cash estimates for Alibaba, Tencent and Xiaomi

Rmb bn, GSe

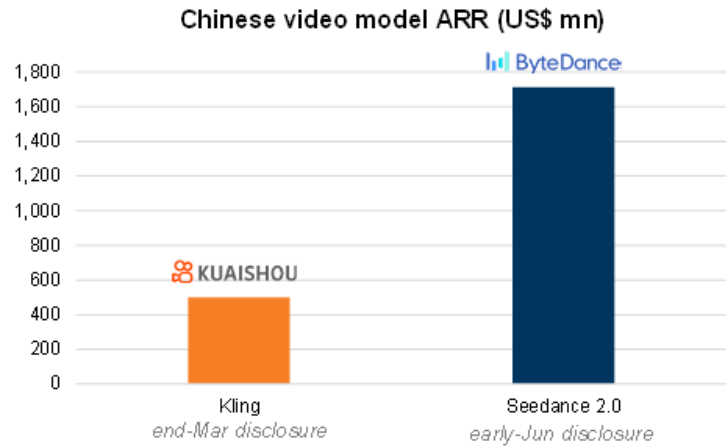


Source: Company data, Goldman Sachs Global Investment Research

Multi-modal generation models

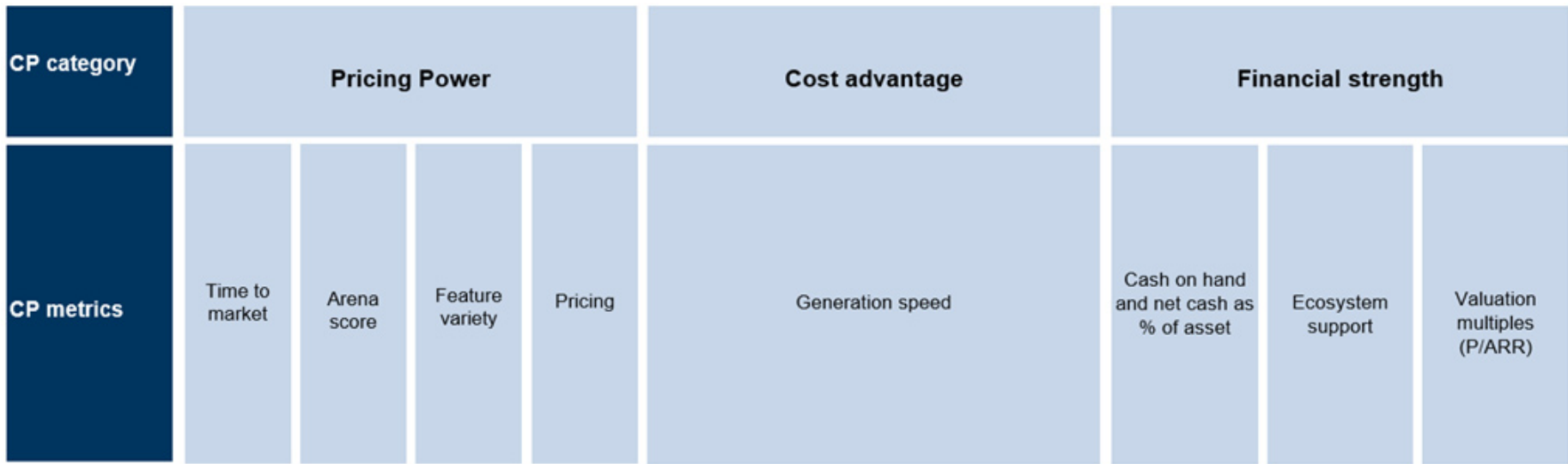
We assess the competitive positioning of video generation players (ByteDance, Alibaba, Kuaishou and MiniMax) under a similar framework across 3 dimensions: pricing power, cost advantage, and financial strength, with each built upon more detailed metrics.

Exhibit 30: Latest disclosure of Chinese AI video model ARR (US\$ mn)



Source: Company data, 36kr, Data compiled by Goldman Sachs Global Investment Research

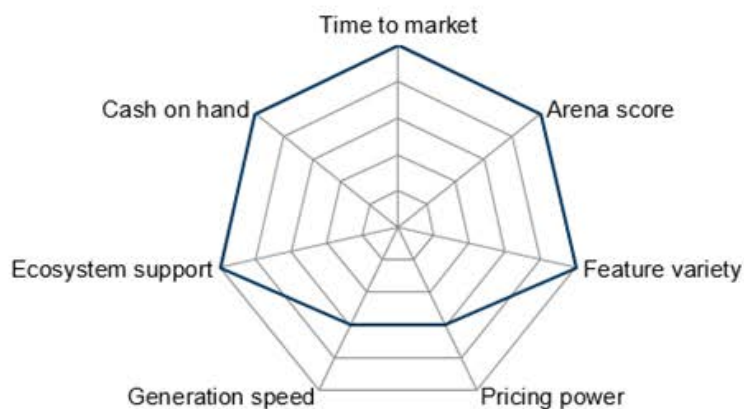
Exhibit 31: Competitive positioning framework for multi-modal generation players



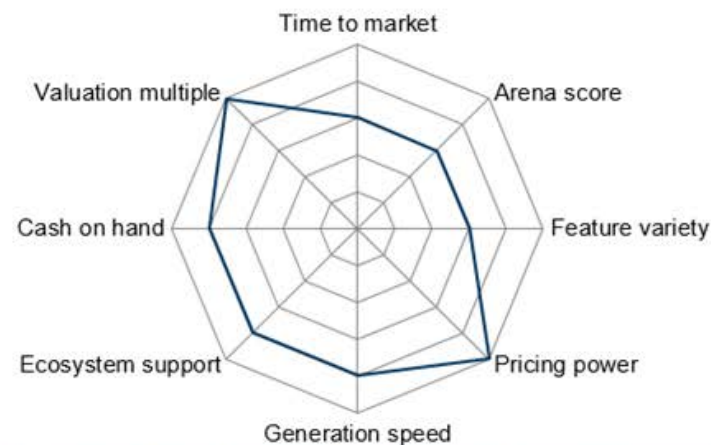
Source: Goldman Sachs Global Investment Research

Exhibit 32: Overview of competitive analysis for key multi-modal labs' positioning

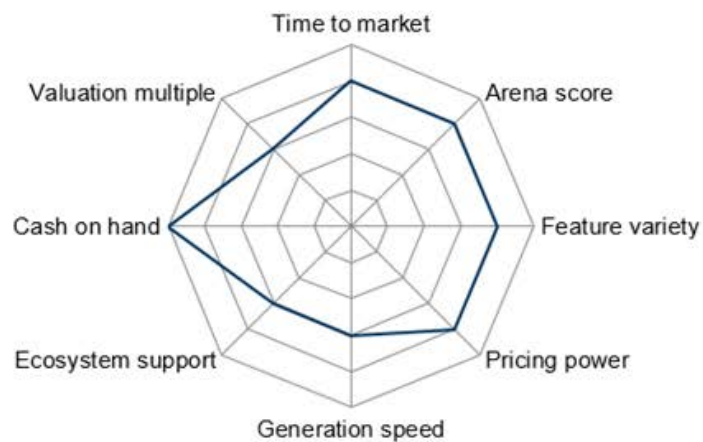
ByteDance: leading in model performance with strong financial/ecosystem back-up for quick adoption & commercialization



Kuaishou: prioritizing enterprise & professional users with strong pricing power and clearer monetization path, supporting a more manageable ROI



Alibaba: abundant resources to support rapid iteration for catching up in performance / feature offerings



MiniMax: native multi-modal training (text+video) to offer synergy on cost efficiency & capability, with anticipated H3 launch as key catalyst



ByteDance's (Not Covered) relevant metrics are calculated by analyzing the industry and companies that we cover, and then extrapolating them to ByteDance. N/A for ByteDance's valuation multiple

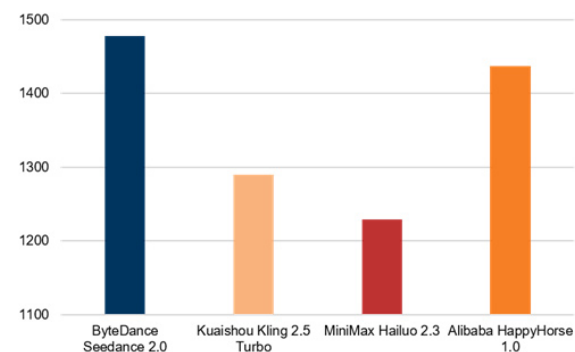
Source: Goldman Sachs Global Investment Research

Pricing Power

- **Time to market:** Measured by how quickly and effectively a player delivers frontier competitive models. From comparing the launch date of the company's flagship models with its prior generation & other models at a similar performance level, we aim to assess whether the release delivers a meaningful capability step-up over its past generation, and whether it narrows the gap to the current global SOTA models.
- **Arena score (average from text-to-video generation and image-to-video generation):** We view model intelligence as the primary determinant of pricing power, since stronger generation models handle higher-value tasks/commercial work and deliver clearer ROI. We use the LMArena score instead of any static benchmark because it reflects a large scale of blind user reviews, making it a more objective read on the real-world capability.
- **Feature variety:** Reflects the generation quality and accessibility of the model, where we further look at **1)** native audio (Y/N): whether the model directly outputs audio matched to the visuals, rather than requiring a separate voice/audio model layered on; **2)** max duration: how long a single clip the model can generate; **3)** max # of storyboards; how many distinct shots the model can generate within one clip, implying its ability to sustain narrative coherence; and **4)** max # of multi-modal reference inputs: how many reference materials beyond text prompt (images or video clips) the user can feed in to guide generation.
- **Pricing (US\$ per min of 1080p video from text-to-video generation, without audio):** We also refer to the realized headline price for flagship models, where sustained higher pricing with iteration signals stronger pricing power.

Exhibit 33: Video Arena score of key players

Average score of text-to-video generation and image-to-video generation

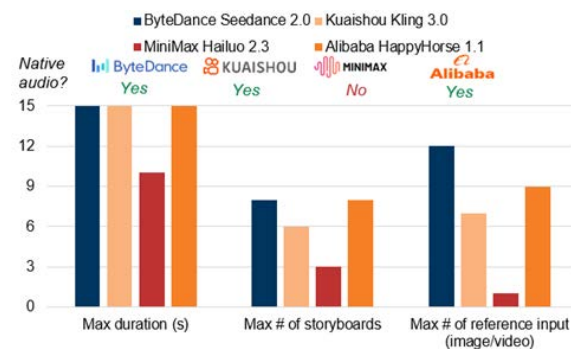


As of Jun 2026

Source: LMArena, Data compiled by Goldman Sachs Global Investment Research

Exhibit 34: Feature variety of key players

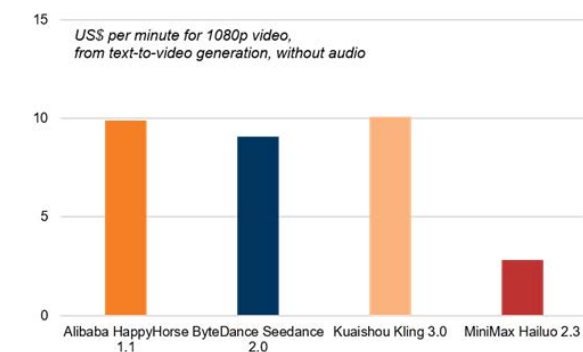
Native audio (Y/N), max duration (s), max # of storyboards and max # of reference inputs (as image/video, beyond text prompt)



Source: Company data, Data compiled by Goldman Sachs Global Investment Research

Exhibit 35: API pricing of key players at the same settings

US\$ per minute



ByteDance Seedance 2.0 datapoint is for 720p video generation (as default setting), all else equal, as per Artificial Analysis

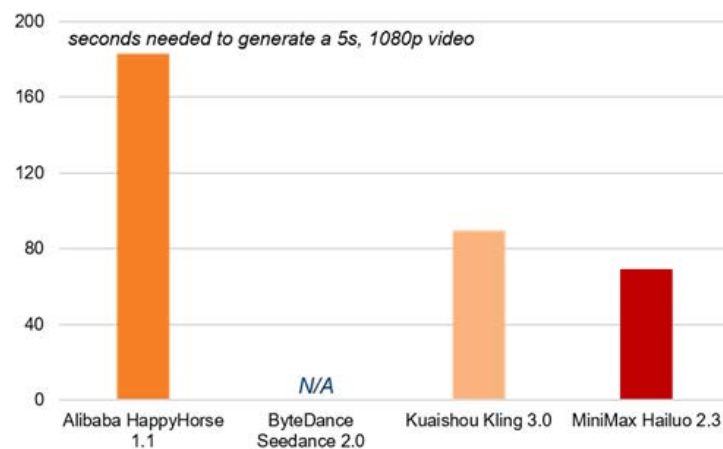
Source: Artificial Analysis, Data compiled by Goldman Sachs Global Investment Research

Cost advantage

- **Generation speed:** Seconds to generate a 5s 1080p clip, where faster speeds indicate higher unit compute utilization and lower inference cost per video.

Exhibit 36: Generation speed of key players

seconds, lower = better



As of early-July, 2026

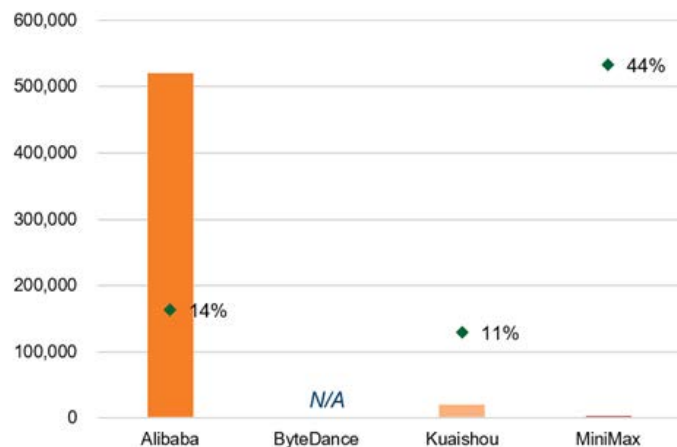
Source: Artificial Analysis, Data compiled by Goldman Sachs Global Investment Research

Financial strength

- **Cash on hand and net cash/debt as % of assets:** We use total cash on hand as a measure of balance sheet strength, with net cash as % of assets to normalize across players of different scale. We see mega-caps (compared with individual players) as having more resources to accumulate compute, fund multi-modal exploration and push faster product distribution.
- **Ecosystem support:** Customer acquisition and distribution channels within the player's ecosystem (including content platforms, editing tools and MaaS), where platform companies such as ByteDance and Kuaishou have an edge in strengthening enterprise commercial adoption and consumer traffic acquisition, thereby powering their data flywheels.
- **Valuation multiples:** For both MiniMax and Kuaishou (with Kling AI valuation at US\$18bn post-money implied valuation, see our team's [analysis](#)), we use P/ARR 2026E multiple.

Exhibit 37: Cash on hand and net cash as % of total assets by 2025 for key players

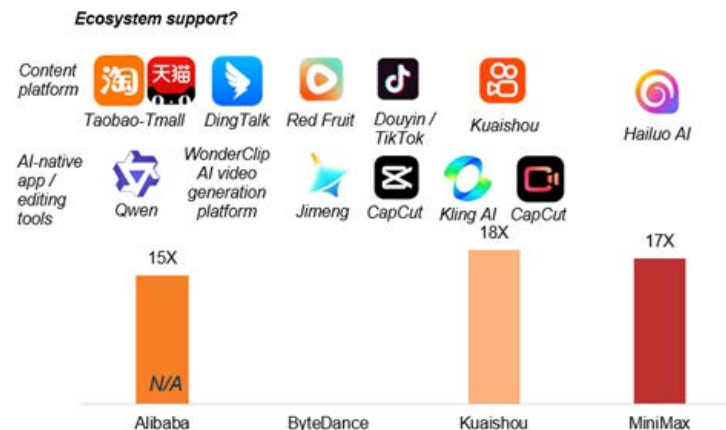
Rmb mn



Source: Company data, Goldman Sachs Global Investment Research

Exhibit 38: Ecosystem support and valuation multiples of key players

P/ARR multiples, as of end-2026E



Source: LatePost, Company data, Goldman Sachs Global Investment Research

Key upside and downside risks

Upside risks:

- **Stronger-than-expected model intelligence uplift from self-sustainable positive flywheel:** Breakthroughs in coding and multi-modal performances, offering seamless integration and understanding across text, image, and audio, would unlock new applications, physical AI and AGI potentials.
- **Faster-than-expected path to profit visibility:** If Chinese AI model companies shift faster toward being open-weight (from pure open source) and demonstrate superior operational efficiency, such as optimizing model inference costs and training infrastructure, or diversify monetization channels to achieve higher revenue growth or better unit economics sooner, it would accelerate the timeline to positive cash flow and net income. Early profitability would boost investor confidence on the AI model industry, potentially reducing the need for future external funding.
- **Additional non-inference take rate revenue streams:** If Chinese AI model companies are able to train and achieve SOTA model performance, an open-weight approach charging take rates/revenue sharing on neoclouds/hyperscalers that house such models and sell to customers may transform AI model company business models to a high gross margin business (without needing to conduct inference for each token sold, but earning via revenue sharing with platforms that

sell top performing Chinese models).

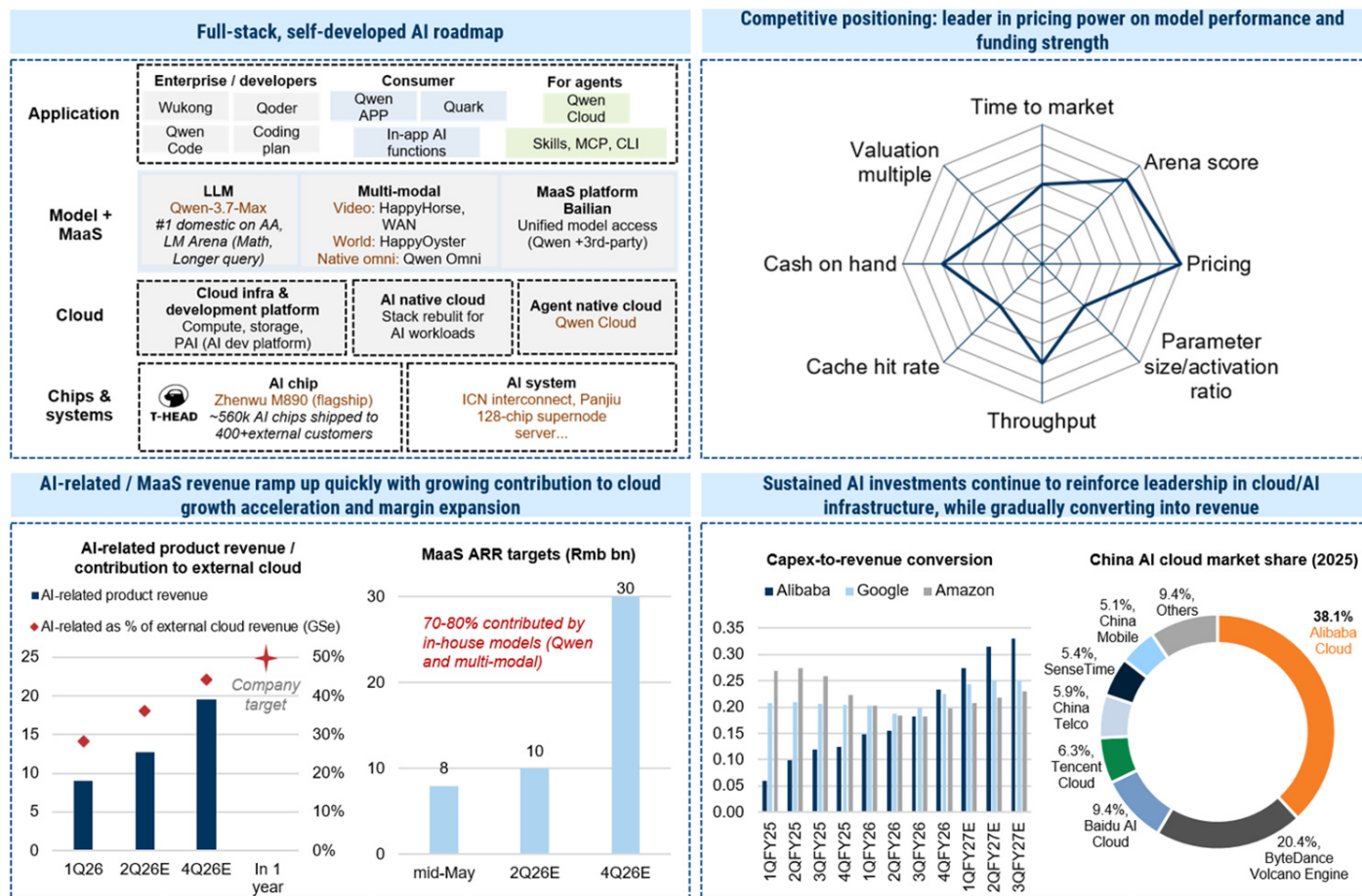
- **Stronger-than-expected commercialization capability:** Growth of new applications, Code and Co-Worker platforms, coupled with ‘digital worker’ agent models could provide multiple channels for revenue generation. If these channels achieve higher-than-expected adoption rates, conversion efficiencies, or ARPU, Chinese AI models could outperform current revenue forecasts.

Downside risks:

- **Competition in global foundation model industry, vs. US/China mega-caps:** Chinese AI models operate within the intensely competitive and rapidly evolving global foundation model industry, where the risks include the challenge of keeping pace with swift technological advancements, frequent new product introductions, and dynamic customer demand. Failure to adapt products and invest significantly in R&D in a cost-effective and timely way, could lead to diminished competitiveness. In addition, the early-stage nature of the market opportunities makes customer demand unpredictable and an inability to meet user requirements would risk losing existing users and adversely impact revenue.
- **Limited near-term profit visibility due to high R&D expenses:** Achieving and sustaining profitability of Chinese AI model companies in the near future could be a challenge, as success depends on continuous technological development, user experience improvement, and effective commercialization. Failure to manage escalating costs and generate adequate revenue could materially impact the company’s financial position and profitability.
- **Risks around IP/content generation:** Chinese AI models, especially multi-modal generation models, face risks from the generation of inappropriate content. Similar to challenges encountered by other AI applications, such as lawsuits against Character AI’s chatbot for its alleged role in teens’ safety, image/video-generation models could be exposed to reputational damage and potential lawsuits. Robust content moderation and ethical AI development are crucial to prevent adverse output and ensure a safe user experience. Chinese companies could also face risks from third party intellectual property/distillation method infringement claims, from both the data used to train their models and the content they generate.
- **Cash burn/self-funding capability:** The intense capital requirements for advanced AI model training, coupled with rapid technological advancements by competitors, could pose financing risks to the relatively fragmented Chinese AI model landscape. In addition, any models that fail to keep pace with leading peers could run into difficulties in generating sufficient revenue (ARR x inference gross margins) to cover training costs.
- **Risk around geopolitics under an intensified tech race between the US and China:** Any US restrictions on the use of Chinese AI models could limit market access and adoption, and Chinese AI model companies could also face increased costs if barred from training its models on US cloud service providers or be included into any list which limits procurement of US supplies from hardware to software. The evolving global chip availability dynamics also present a critical supply

chain risk.

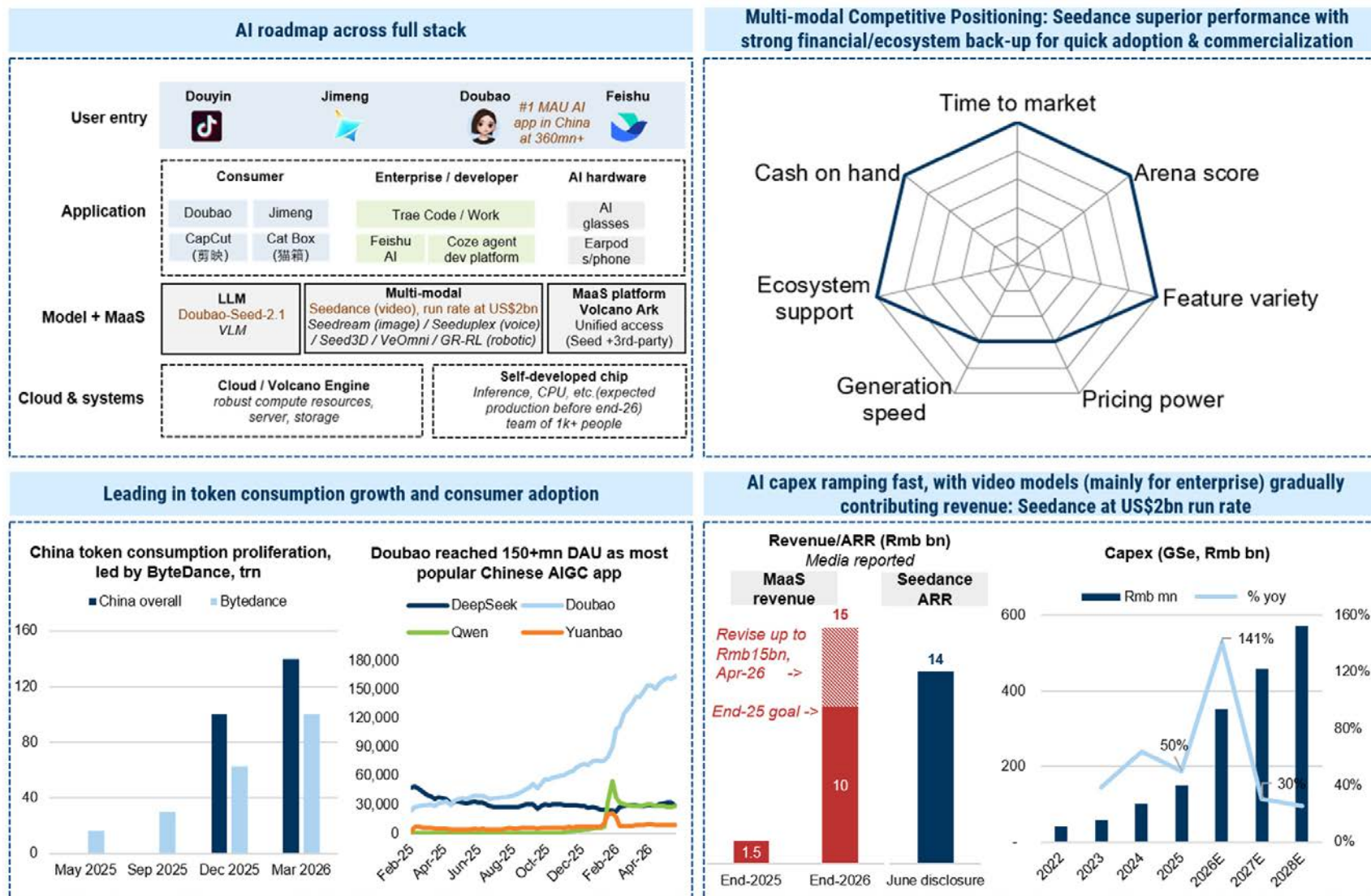
China's key AI players at a glance

Mega-caps**Exhibit 39: Alibaba****Alibaba: Full-stack AI player with multi-year AI/cloud upside, leading position of most areas**

Source: Company data, Goldman Sachs Global Investment Research, Omdia

Exhibit 40: ByteDance

ByteDance: Deep resources for broad AI exploration with strong consumer base and product portfolio

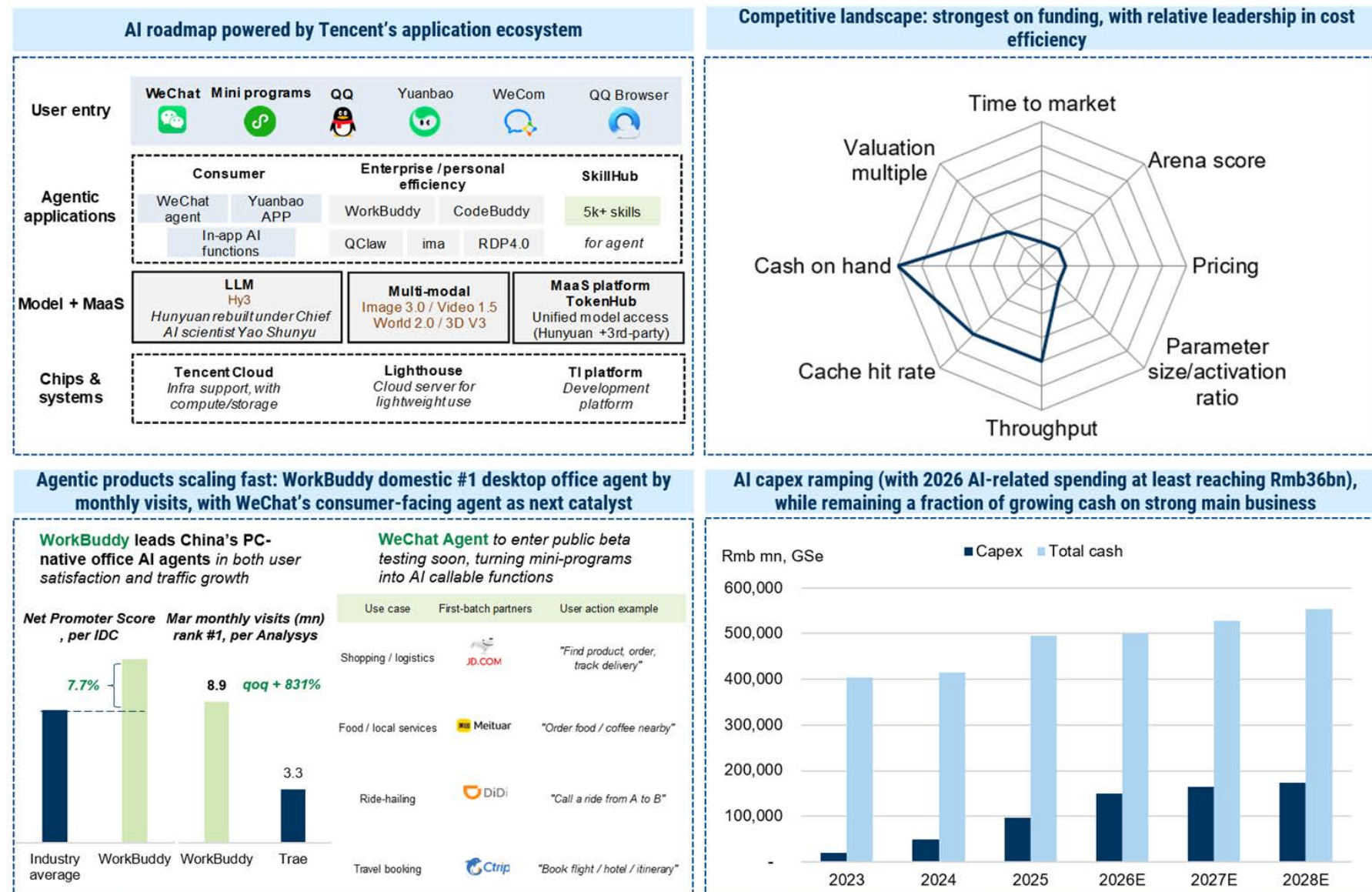


ByteDance's (Not Covered) competitive positioning metrics are calculated by analyzing the industry and companies that we cover, and then extrapolating them to ByteDance. N/A for the valuation multiple in the competitive positioning chart

Source: Company data, Goldman Sachs Global Investment Research, LatePost, QuestMobile

Exhibit 41: Tencent

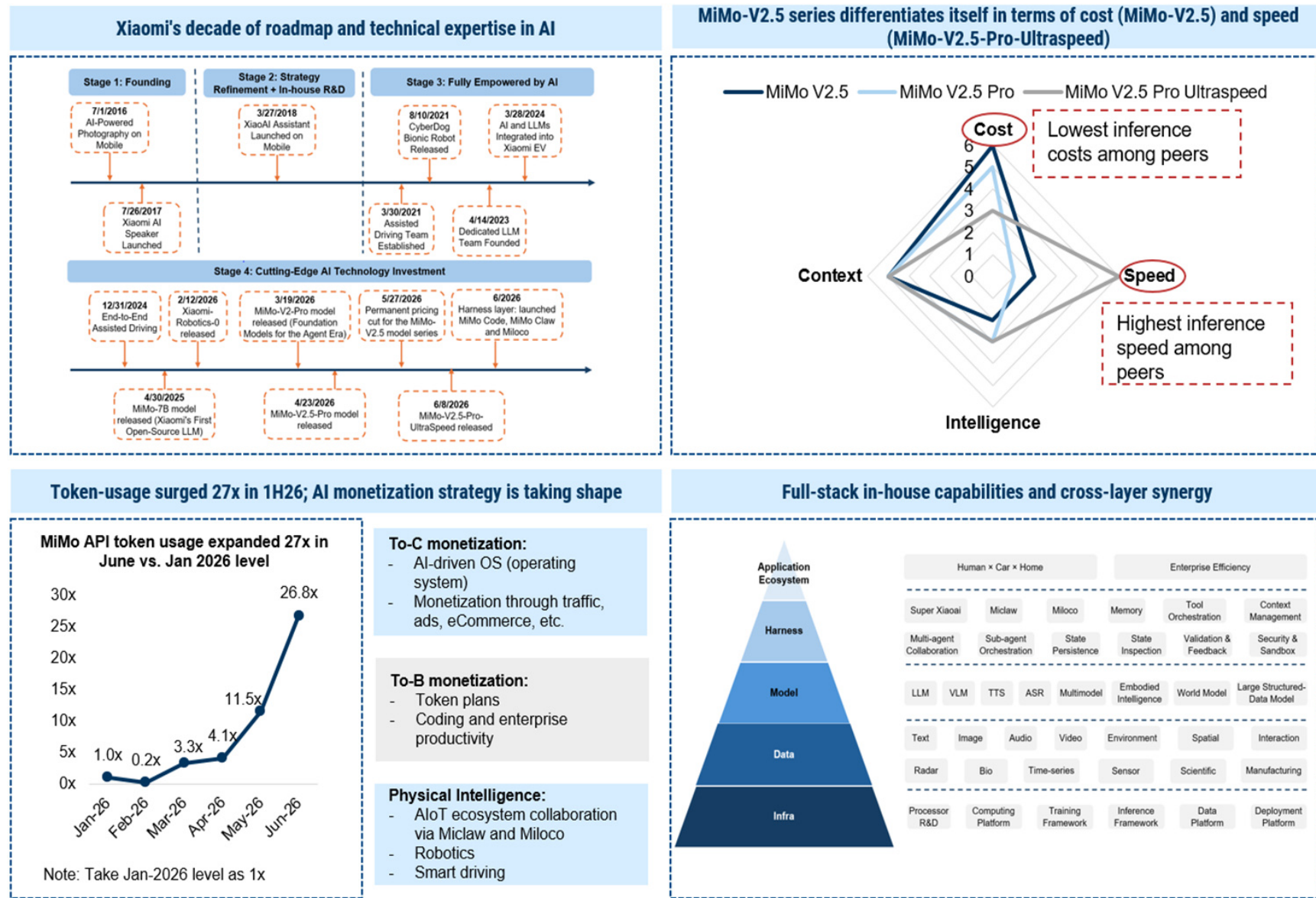
Tencent: Strong contender after AI strategic revamp; potential in Weixin AI agent



Source: Company data, Goldman Sachs Global Investment Research, Analysys, IDC

Exhibit 42: Xiaomi

Xiaomi: Pushes efficiency frontiers of LLMs; Rich applications scenarios in to-C/to-B and physical intelligence

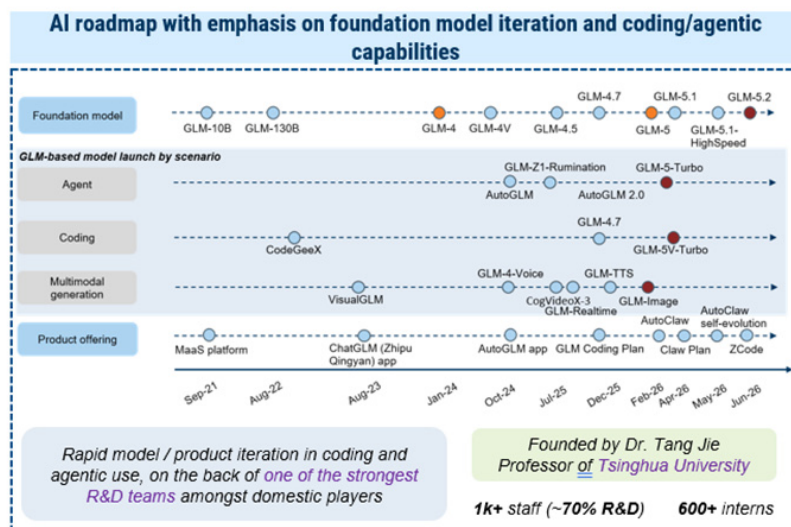


Source: Company data, Goldman Sachs Global Investment Research

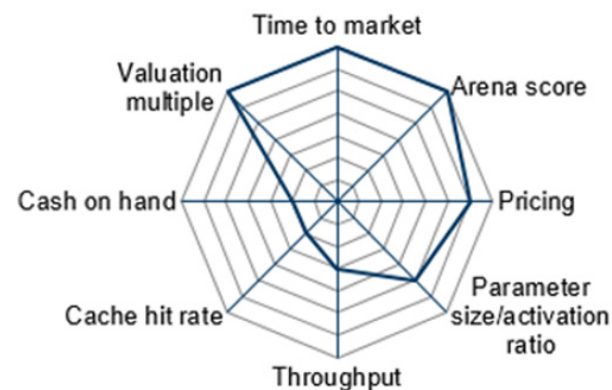
Key independent AI players

Exhibit 43: Zhipu

Zhipu: Leading model capabilities in coding/agentic scenarios

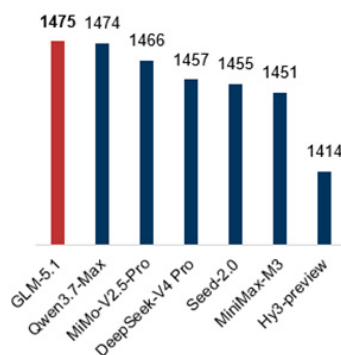


Competitive Positioning: leading in operational efficiency and coding model intelligence, which Zhipu believes to determine pricing power



Top coding/agentic model performance support continuous pricing uplift

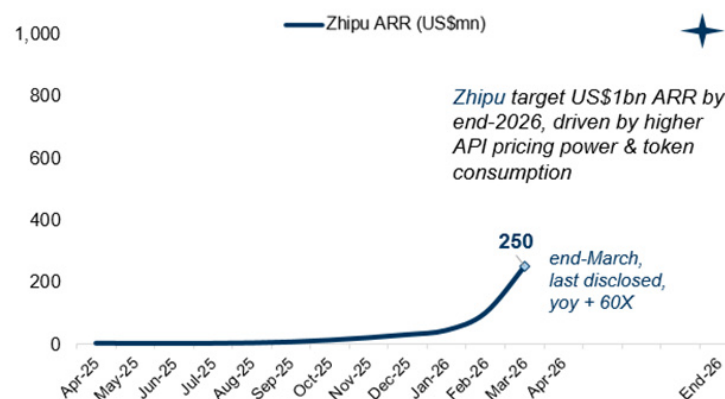
LMarena text-to-text task leaderboard across math, coding, creative writing, etc.



GLM-series API & subscription pricing trend higher YTD with new model launch

Date	Product	Pricing action
API pricing		
Feb-26	API	GLM-5 API price up by 50%
Mar-26	API	GLM-5-Turbo API price up by 20%
Apr-26	API	GLM-5.1 API price up by 10%
May-26	API	GLM-5.1 High-Speed for enterprise users at higher price
Accumulated price uplift YTD		98%+
Coding plan		
Feb-26	GLM coding plan	Subscription price up by 30%
Apr-26	Overseas GLM coding plan	Subscription price up by 80-140%

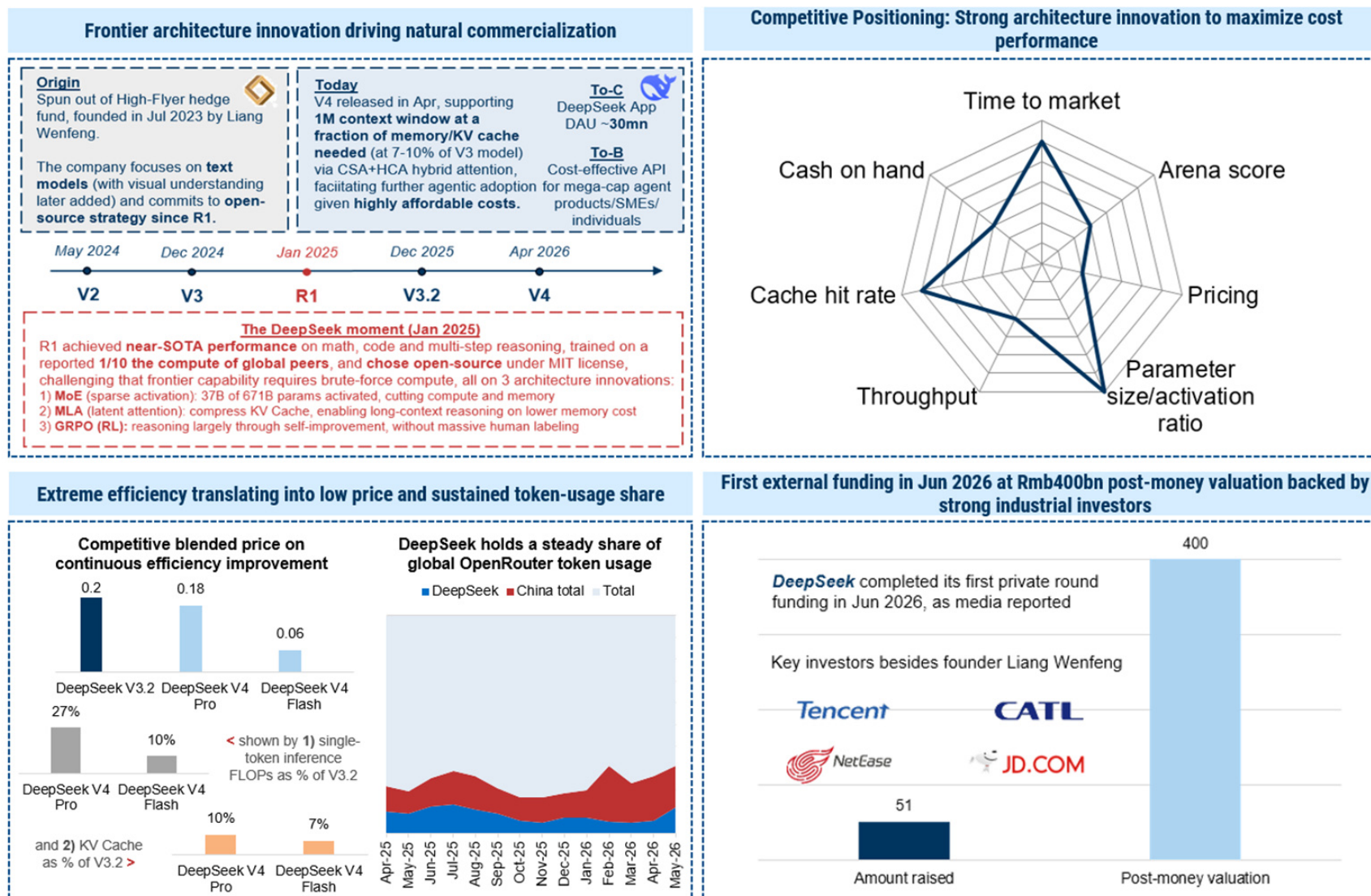
End-March ARR run-rate at US\$250mn (+60X yoy), targeting US\$1bn by end-2026 on higher pricing and token consumption



Source: Company data, Goldman Sachs Global Investment Research, LMarena

Exhibit 44: DeepSeek

DeepSeek: Extreme cost efficiency to sustain scalable token usage on competitive pricing

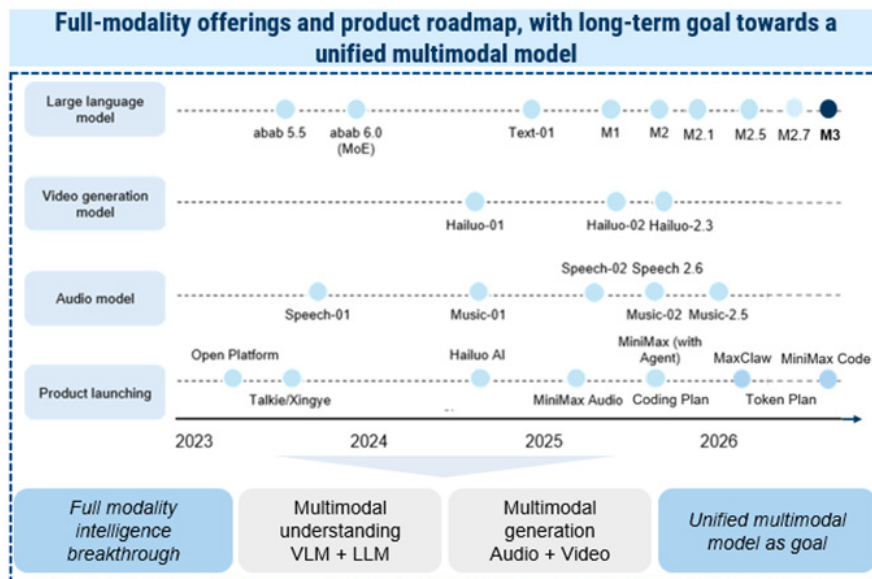


DeepSeek's (Not Covered) competitive positioning metrics are calculated by analyzing the industry and companies that we cover, and then extrapolating them to DeepSeek. DeepSeek's latest funding round based on Reuters.

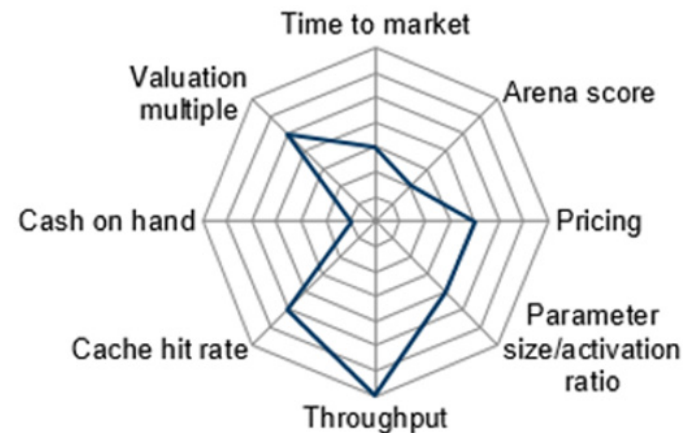
Source: Company data, Goldman Sachs Global Investment Research, OpenRouter, Reuters

Exhibit 45: MiniMax

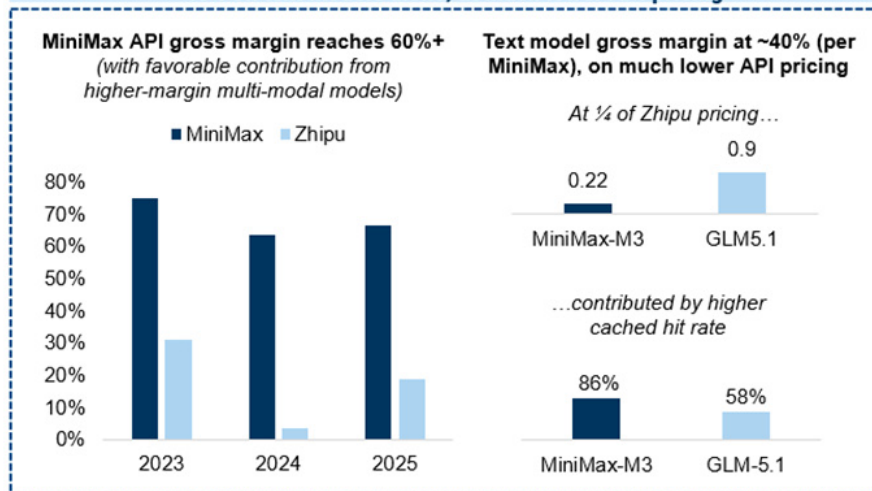
MiniMax: Unique full-modal capabilities among independent players; cost-efficient computing architectures



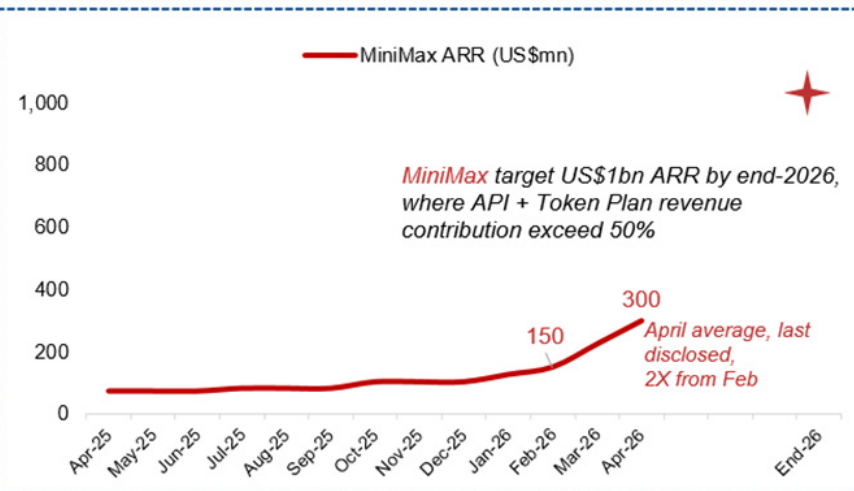
Competitive positioning: strong on multi-modality offerings and cost efficiency among independent players



API gross margin above 60%, with higher-margin multi-modal contribution and ~40% on text models alone, at much lower API pricing



April ARR run rate at US\$300mn (2X from Feb), targeting US\$1bn by end-2026 with API + Token Plan contribution above 50%



Source: Company data, Goldman Sachs Global Investment Research

Price Target Methodology & Risks

Price Target Risks and Methodology – Alibaba Group

Valuation methodology: We are Buy rated on BABA/9988.HK, with SOTP-based 12-month target prices of US\$186/HK\$180 per ADS/share (FY27E-based).

Key risks: (1) Lower-than-expected GMV growth due to macro/competition; (2) Slower-than-expected monetization in China retail; (3) Weaker-than-expected execution in key strategic investments; and (4) Cloud revenue growth deceleration.

Price Target Risks and Methodology – Tencent Holdings

Valuation methodology: We are Buy rated on Tencent (0700.HK), with a 12-m SOTP-based TP of **HK\$700**.

Key risks: More intense industry competition in performance-based advertising, unexpected delays in game launches/Banhao approvals, moderated growth in FinTech and Cloud businesses, slower-than-expected AI development progress, reinvestment risk.

Price Target Risks and Methodology – Xiaomi Corp.

We are Buy rated on Xiaomi with a 12-month target price of HK\$40, based on a SOTP valuation methodology: 1) 16x target 12m-fwd EV/NOPAT on Xiaomi core; 2) DCF-based valuation for Xiaomi EV of US\$42bn (12% WACC and 3% TGR); 3) a 10% holdco discount. Key downside risks: 1) More intense competition and weaker market share gains within the global smartphone industry. 2) Higher GPM pressure on the smartphone/EV business. 3) Below-expected execution on Xiaomi brand premiumization and the EV business. 4) Intensifying geopolitical risks and regulatory uncertainties. 5) Softer macro environment and weaker smartphone/IoT demand. 6) FX fluctuation.

Price Target Risks and Methodology – MiniMax Group

We are Buy rated on MiniMax with a 12-month target price of HK\$860. Our valuation is based on DCF valuation, with 12% WACC and 2% terminal growth rate.

Key risks: Weaker-than-expected model performance amid the competition in the global foundation model industry; Slower-than-expected path to profit visibility; Weaker-than-expected commercialization capability; risks around IP/content generation; cash burn/self-funding capability; risks around geopolitics under an intensified tech race between the US and China.

Price Target Risks and Methodology - Kuaishou Technology

Valuation methodology: We are Buy rated on Kuaishou with a 12m target price of **HK\$68**, based on P/E valuation (target P/E multiple of 15X applied to our avg. 2026E-27E EPS).

Key downside risks: 1) Slower-than-expected ad budget recovery; 2) Weaker-than-expected Kling monetization; 3) Slower-than-expected growth of its aggregate user engagement base (DAU x time spent per DAU); 4) Lower-than-expected profitability; 5) Weaker-than-expected AI progress.