China AI Model Monitor (Aug): Upgrade wave, Usage surge, Cloud upside

Industry Overview

Rapid model upgrade, Pricing hike, Cloud accelerates

Introducing the China AI Model Monitor: a visual, data-driven monthly recap of model intelligence, token usage & pricing trends. In Aug, we observed: 1) rapid model upgrade continued for both cloud incumbents (Tencent Hy4 preview, Alibaba's Qwen 3.8 Max/Flash) and independent AI labs (Zhipu's GLM5.3 Flash); 2) selected companies (e.g. DeepSeek) raised token price, which we attribute more to compute tightness rather than pivoting from high-performance-low-cost strategy; 3) more companies launched low-pricing small size models (GLM5.3 Flash), offering more high-performance-low-cost options; 3) a sharp increase of cloud platforms capex and cloud revenue acceleration in China – 2Q Capex: RMB53bn for Tencent, RMB68bn for BABA; 2Q cloud revenue: low-20% YoY for Tencent, 45% YoY for BABA.

Key trends in numbers

Intelligence level: According to Artificial Analysis, Anthropic currently leads the Intelligence Index with Claude Opus 5, followed by Claude Fable 5 & OpenAI's GPT-5.6 Sol. Kimi K3 is the highest ranked Chinese model/open-source model globally (95% of Opus 5), followed by GLM-5.3 & Qwen 3.8 Max. **Model usage:** OpenRouter data suggests Chinese models are leading for usage, with DeepSeek V4 Flash #1 followed by Tencent's Hy3. OpenCode data suggests strong adoption of lower cost open Chinese models, with DeepSeek V4 Flash dominating usage. Vercel data suggests DeepSeek gained usage share m/m and leads at 30% share, while Anthropic usage share was second (25%) and was first for spend (65%). **Token pricing:** As of August 26, LLM Token Expenditure Index was down 26% m/m but stabilizing at $1.07. DeepSeek raised price in August which we attribute to likely tight compute on hand. **Capex:** listed Chinese cloud incumbents saw major capex increase in 2Q26 – we now expect China top 4 cloud platforms combined capex up 98%/44% YoY in 26E/27E. Forward-12m consensus capex for Tencent/BABA was revised up by 25%/29% in the past 30 days.

Table 1: We expect a series of key events to impact major China AI companies
Key sector/company events ahead and implications for major AI companies in China

Time	Event
Aug 31	**Zhipu:** 1H26 results
Sep	**Alibaba**: potential model upgrade during Sep Alibaba Cloud Summit (Sep 22-24)
Sep-Oct	**MiniMax:** potential launch of M3.1, followed by M3 Pro
2H26	**Zhipu**: potential launch of GLM5.5
2H26	**DeepSeek:** potential launch of V5
4Q26	**Tencent:** potential launch of Hy4 official/Hy 5 preview
4Q26	**Bytedance:** potential launch of Seed model series upgrade
7-Jan-27	**Zhipu**: 12-month lock-up expiry. 40% shares unlocked
8-Jan-27	**MiniMax:** 12-month lock-up expiry. 90% shares unlocked post this event

Source: Companies, BofA Global Research estimates

Who's who in the China AI model market

Incumbents

Major internet companies such as ByteDance (private), Baidu, Alibaba, and Tencent—many of which have established cloud businesses—have developed proprietary AI foundation models in-house.

Independent AI labs

DeepSeek (private), MiniMax, Zhipu are independent AI labs comparable to OpenAI and Anthropic, distinct from established internet platforms.

Exhibit 1: AI model landscape is intensely competitive in China
AI model details, benchmark scores, and pricing comparison table

	DeepSeek	Alibaba	Minimax	Zhipu AI	Tencent	Moonshot AI	ByteDance	Google	OpenAI	Anthropic
Flagship model	**DeepSeek V4.0**	**Qwen-3.8 Max**	**Minimax-M3.0**	**GLM-5.3**	**Hy4 preview**	**Kimi K3**	**Seed2.1 Pro**	**Gemini 3.7**	**GPT-5.6 Sol**	**Claude Opus 5**
Country	China	China	China	China	China	China	China	US	US	US
Release date	8/13/2026	8/3/2026	6/1/2026	8/19/2026	8/28/2026	7/17/2026	6/23/2026	8/13/2026	7/9/2026	7/24/2026
Open/Proprietary	Open	Open weights	Open	Open	Open	Open	Proprietary	Proprietary	Proprietary	Proprietary
Model Architecture	MoE+hybrid	MoE+hybrid	MoE+hybrid	MoE+hybrid	MoE+hybrid	MoE+hybrid	MoE	Transformer	Transformer	Dense transformer
Context window	1mn	1mn	1mn	1mn	1mn	1mn	256k	1mn	1mn	1mn
Benchmark	**DeepSeek V4.0 Pro 0813**	**Qwen-3.8 Max**	**Minimax-M3.0**	**GLM-5.3 (max)**	**Hy4 preview**	**Kimi K3 (max)**	**Seed2.1 Pro**	**Gemini 3.7 Flash (high)**	**GPT-5.6 Sol (max)**	**Claude Opus 5 (max)**
AA Intelligence Index	53	58	45	60	NA	60	NA	56	61	63
AA Agentic Index	50	58	36	59	NA	54	NA	45	58	59
AA Coding Index	43	61	NA	NA	NA	63	NA	60	65	67
Reasoning										
GPQA Diamond	93%	93%	93%	92%	92%	94%	NA	95%	94%	93%
HLE	41%	43%	39%	42%	43%	47%	43%	48%	49%	55%
τ^3-Banking	40%	51%	15%	50%	NA	46%	NA	33%	44%	42%
AA-LCR	75%	74%	80%	76%	NA	83%	NA	80%	78%	76%
GDPval-AA v2	54%	62%	44%	63%	NA	59%	NA	51%	61%	67%
CritPt	18%	20%	4%	19%	17%	23%	NA	14%	32%	29%
Coding										
SciCode	49%	53%	45%	56%	NA	59%	60%	57%	56%	56%
Terminal-Bench v2.1	79%	81%	65%	84%	85%	85%	71%	86%	88%	89%
Knowledge										
AA-Omniscience Accuracy	49%	32%	17%	34%	NA	48%	NA	55%	59%	61%
AA-Omniscience Hallucination Rate	95%	42%	18%	30%	NA	53%	NA	65%	92%	61%
Pricing										
Input ($/1M)	0.7	2.0	0.3	1.4	0.9	3.0	0.9	0.8	4.0	5.0
Output ($/1M)	2.0	6.0	1.2	4.4	2.6	15.0	4.4	3.8	20.0	25.0

Source: OpenRouter, Artificial Analysis, Vals AI, company websites, BofA Global Research (Data as of August 28, 2026)
*All prices are after discounts and input price refers to input price (cache miss). Cell shading indicates relative performance across models for each benchmark. Darker green represents relatively stronger performance, while lighter green represents relatively weaker performance.



Model Usage / Adoption

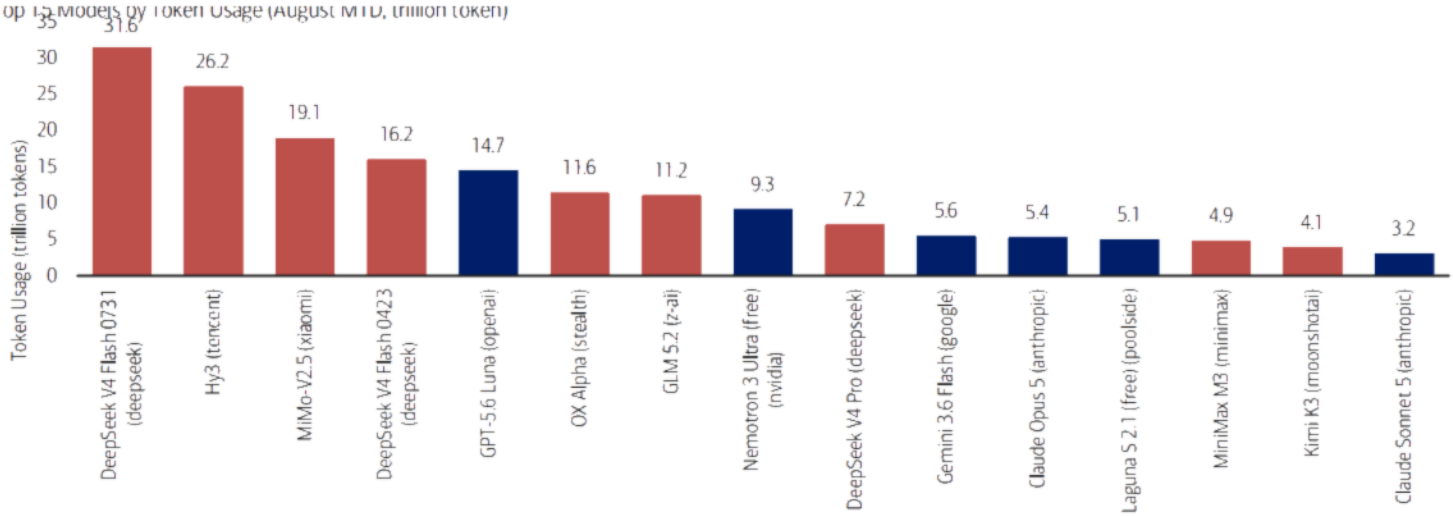
In this section, we track global frontier models' usage and adoption on OpenRouter, OpenCode and Vercel AI and highlight Chinese frontier models' performance. For more analysis on US frontier model capability and pricing trends, please refer to our US team's report: BofA Frontier AI Data Tracker: Model rankings, usage data, and token/GPU costs.

Token Usage - OpenRouter

Model usage trends from OpenRouter suggest strong adoption of lower-cost Chinese open models, with DeepSeek V4 Flash 0731 leading at 31.6tn tokens MTD, followed by

Tencent's Hy3 at 26.2tn and Xiaomi's MiMo-V2.5 at 19.1tn, as of August 17.

Exhibit 2: DeepSeek V4 Flash and Tencent Hy3 lead in token usage



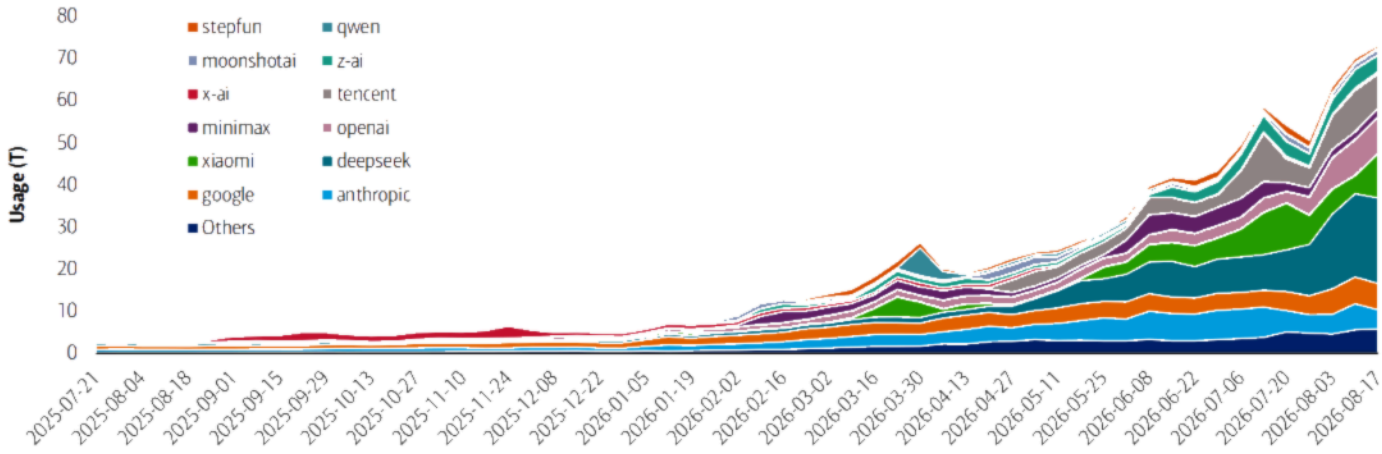
Source: Openrouter.ai data as of August 17

By providers, DeepSeek led the weekly token usage on OpenRouter at 20.3tn tokens, followed by Xiaomi's 10.5tn tokens, OpenAI's 8.8tn tokens, and Tencent's 8.2tn tokens in the week of August 17.

Among the top 50 models, DeepSeek accounts for the largest token usage share at 22%, followed by Xiaomi at 11% and OpenAI's 9% in the week of August 17. Compared to the same week in July, the most notable share gains were DeepSeek (+4.7pt) and OpenAI (+4.6pt), while the largest declines were Xiaomi (down 7.8pt), Anthropic (down 3.9pt).

Exhibit 3: DeepSeek led the token usage on OpenRouter in the week of August 17

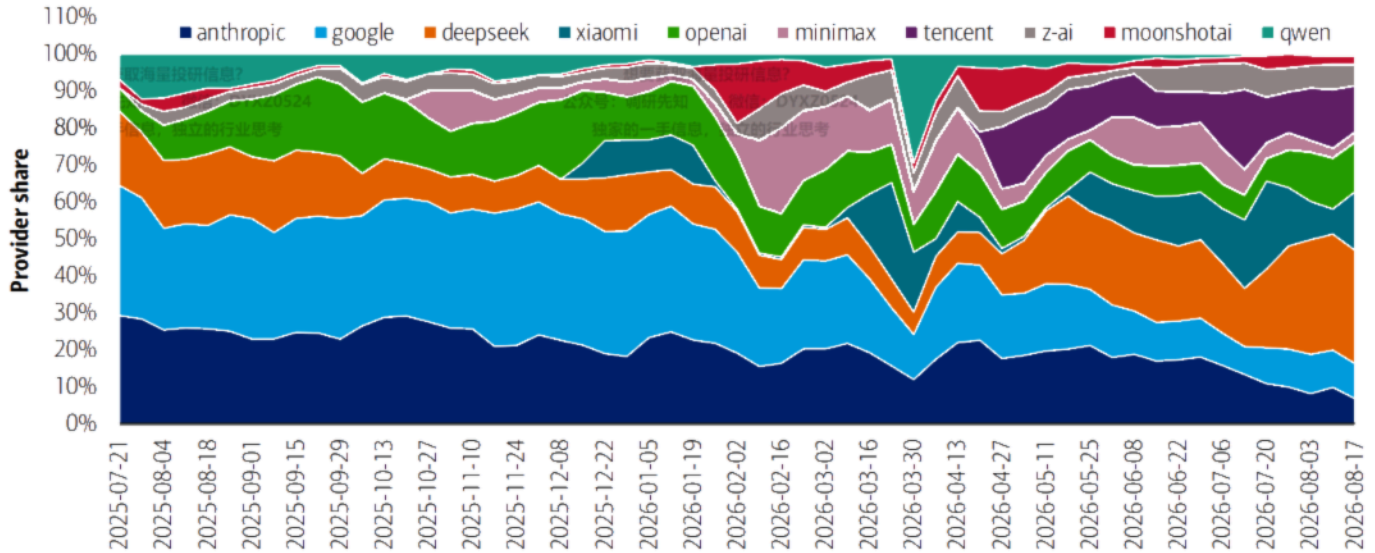
Weekly token volume by provides on OpenRouter



Source: Openrouter.ai data as of August 17

Exhibit 4: DeepSeek led token usage share on OpenRouter, accounting for 22% weekly volume share in the week of August 17

Weekly token volume share by provider among OpenRouter's top 50 models

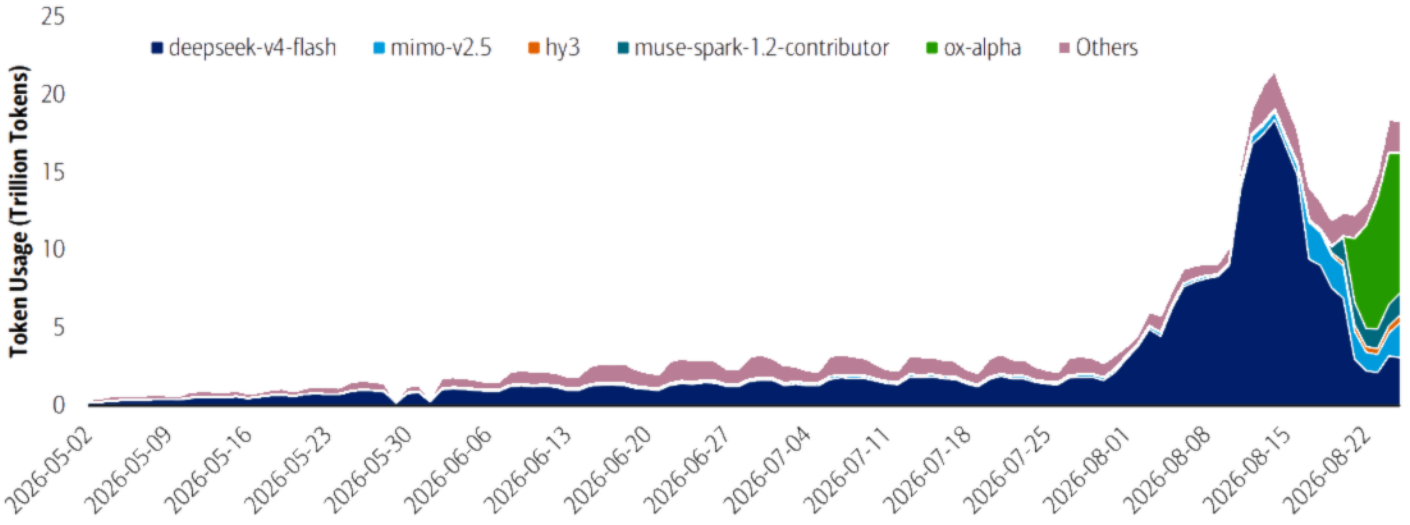


Source: Openrouter.ai data as of August 17

Token Usage – OpenCode

Model usage trends from OpenCode suggest strong adoption of lower cost open Chinese models, with DeepSeek dominating token usage especially after DeepSeek V4 Flash official launch in late July. Ox-alpha, which is reported to be Z.ai's GLM-5.3-Flash, saw a sharp surge in usage after launch in late August.

Exhibit 5: Ox-alpha (reported to be GLM-5.3-Flash) saw a sharp surge in usage since August 20
Daily token volume by model on OpenCode



Source: OpenCode data as of August 25

How to read OpenRouter/Opencode model usage data

OpenRouter is an AI model aggregation and routing platform that provides developers with access to multiple LLMs through a single API. OpenCode is an open-source AI coding platform that provides developers with access to multiple LLMs through a unified interface. OpenRouter/Opencode token usage rankings offer a real-time view of model adoption based on token volume routed through these platforms, making the data a useful indicator of developer mindshare, cost-performance trade-offs and commercial traction within the frontier AI ecosystem.

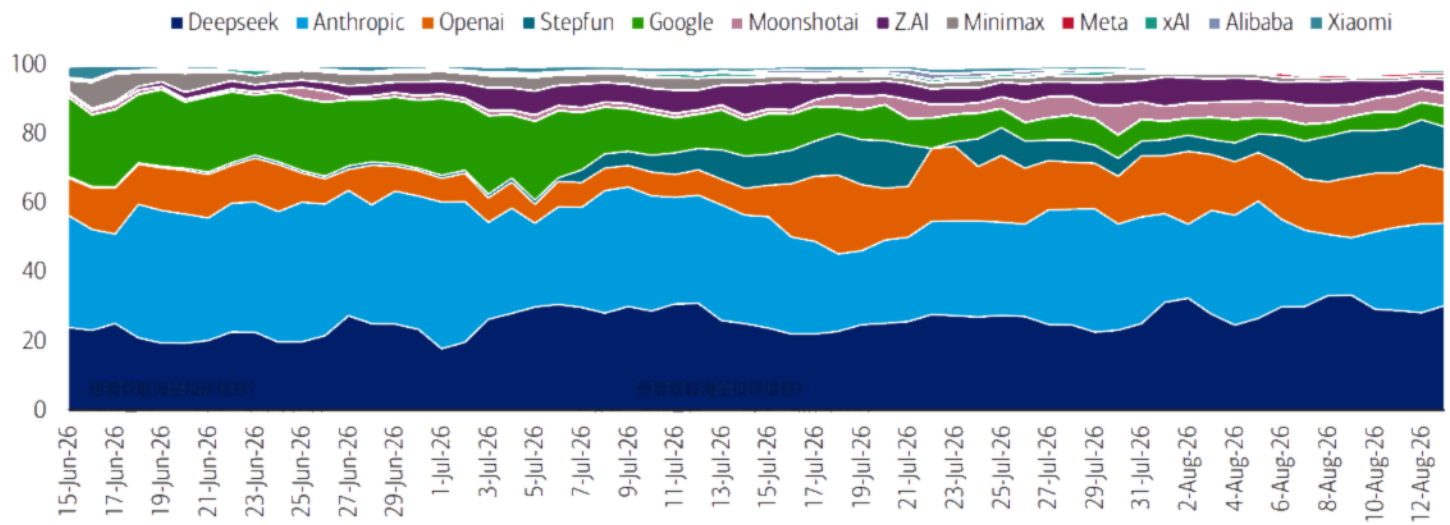
However, we note OpenRouter/OpenCode data should not be interpreted as total model usage across the broader AI market. The dataset captures usage from developers and applications that route requests through OpenRouter/OpenCode but does not include workloads that go directly through provider-specific APIs or cloud platforms such as OpenAI API, Anthropic API, Google Vertex AI/Gemini API, AWS Bedrock or Azure OpenAI. It also excludes usage from enterprise private deployments, self-hosted inference stacks, hyperscaler internal workloads and company-owned data centers unless those workloads are routed through OpenRouter.

Token Volume Share on Vercel AI Platform

DeepSeek led the average MTD token usage share on the Vercel AI platform at 29.7% (vs 26.0% in July), followed by Anthropic with average MTD token market share at 24.7% (vs. 30.1% in July), OpenAI's 16.3% (vs. 12.3% in July) and Google 5.1% (vs. 11.5% in July).

The most notable share gains were OpenAI (+4.1pt) and DeepSeek (+3.7pt), followed by StepFun (+3.3pt), while the largest declines were Google (down 6.4pt), Anthropic (down 5.4pt), followed by MiniMax (down 1.3pt).

Exhibit 6: DeepSeek led AI Token Usage on Vercel, capturing approximately 25% share MTD in August (vs 30.1% in July)
Daily Token Volume Share on Vercel AI Platform (%)



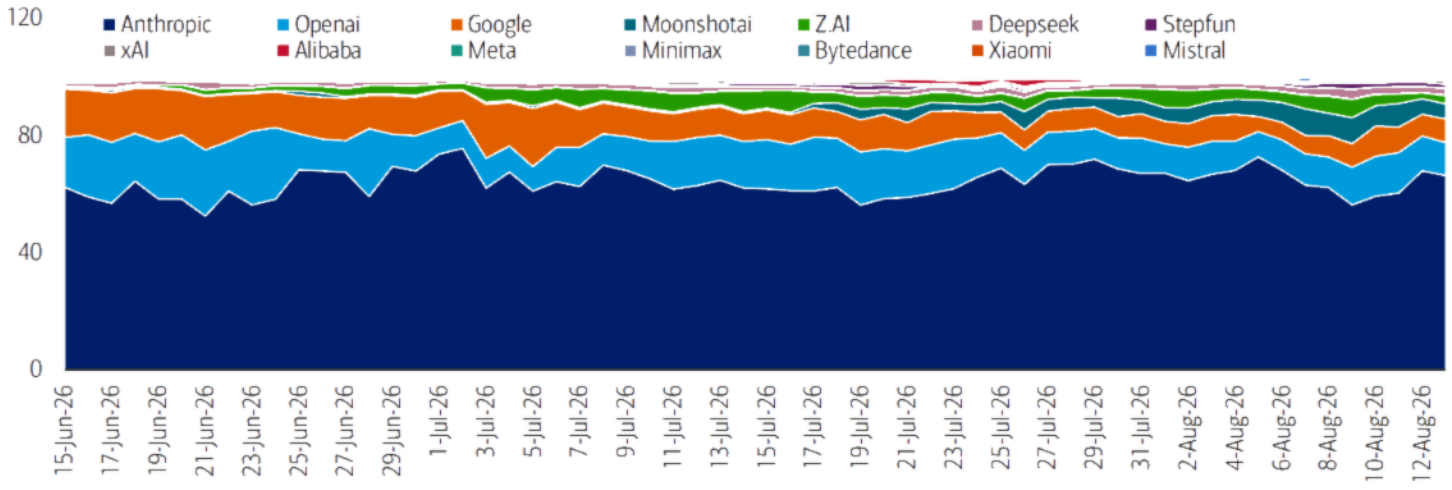
Source: Vercel.com

Spend Share on Vercel AI Platform

Anthropic led the average MTD spend share on the Vercel AI platform at 64.8% (same as July), followed by OpenAI at 11.2% (vs. 13.4% in July), Google 7.8% (vs. 10.5% in July) and Moonshot.ai at 6.4% (vs 2.1% in July).

The most notable spend-share gains were Moonshot AI (+4.3pt), DeepSeek (+0.5pt), and StepFun (+0.4pt), while the largest declines were Google (-2.8pt), OpenAI (-2.1pt), and Alibaba (-0.2pt). Overall, Opensource models gained some share m/m.

Exhibit 7: Anthropic led AI model spend on Vercel, capturing approximately 65% share
Daily Spend Share on Vercel AI Platform (%)



Source: Vercel.com

How to read Vercel AI data

Vercel's daily token share data captures each AI lab's share of the total tokens routed through the Vercel AI Gateway, regardless of the price paid per token. Similar to OpenRouter, the data reflects only activity routed through the Vercel gateway and excludes AI inference activity handled directly through the respective labs' APIs or through third-party cloud providers.

Vercel's daily spend share represents each AI lab's share of total AI Gateway spend. Vercel calculates spend using market-rate pricing based on providers' published list prices, providing a normalized measure across teams, including those using their own API keys. The metric therefore reflects both the volume of tokens routed to each lab and the market-rate price of those tokens. The data covers only activity routed through Vercel AI Gateway and excludes AI inference handled directly through the respective labs ' APIs or through other cloud providers.

Model Intelligence Ranking

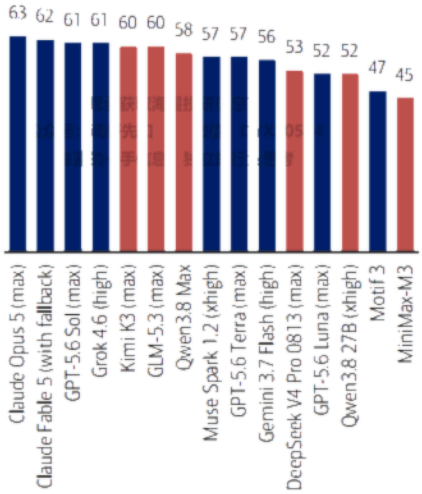
Intelligence Index

Artificial Analysis Intelligence Index tracks progress toward artificial general intelligence across mathematics, science, coding and reasoning by incorporating nine key benchmarks. Claude Opus 5 (max) was ranked the highest at 63, followed by Claude Fable 5 (with fallback) at 62, and GPT-5.6 Sol (max) at 61. Kimi's K3 was ranked 5th at 60 and Zhipu's GLM-5.3 (max) was ranked 6th at 60.

Artificial Analysis Agentic Index represents the weighted average of models' agentic capabilities benchmarks. Claude Opus 5 (max), GLM-5.3 (max) and Grok 4.6 (high) were ranked the highest at 59, followed by GPT-5.6 Sol (max) and Qwen 3.8 Max at 58.

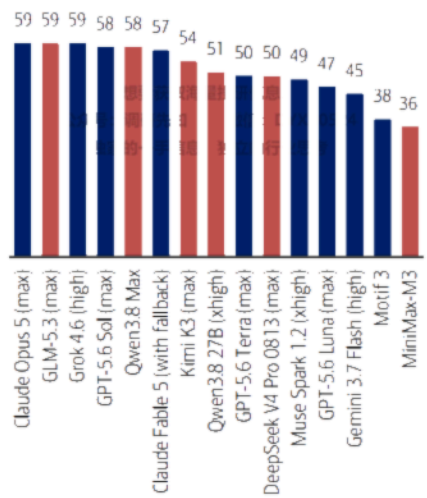
Artificial Analysis Coding Agent Index assesses models' performance, cost, and execution time for leading coding agents on end-to-end software engineering tasks. Claude Opus 5 (xhigh) was ranked the highest at 68, followed by Claude Fable 5 (with fallback) and Claude Opus 5 (max) at 67. Kimi's K3 was ranked 6th at 63 and Alibaba's Qwen 3.8 Max was ranked 8th at 61.

Exhibit 8: Claude Opus 5 (max) was ranked highest in Intelligence index
Intelligence Index - Top 15 Models



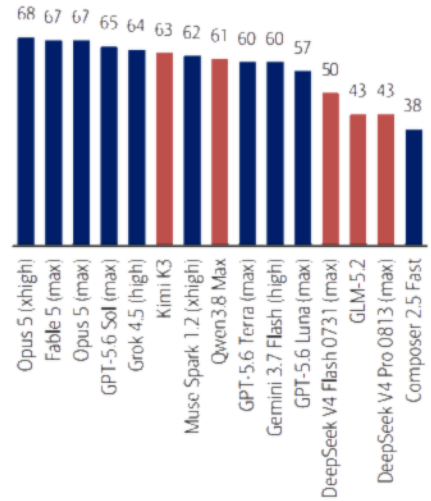
Source: Artificial Analysis as of Aug 26
*Ranking based on models selected b

Exhibit 9: Claude Opus 5 (max) was ranked highest in Agentic index
Agentic Index - Top 15 Models



Source: Artificial Analysis as of Aug 26
*Ranking based on models selected b

Exhibit 10: Claude Opus 5 (xhigh) was ranked highest in Coding Agent index
Coding Agent Index - Top 15 Models

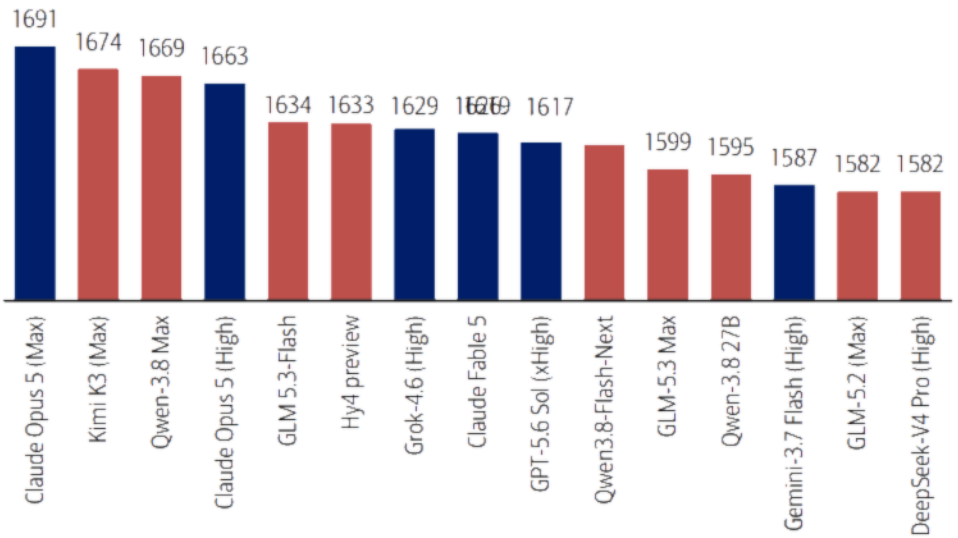


Source: Artificial Analysis as of Aug 26
*Ranking based on models selected b

Code Arena – Web Development

Code Arena – WebDev ranks models on front-end web development tasks, including agentic coding workflows that require multi-step reasoning and tool use. Claude Opus 5 (max) was ranked the highest at 1691, followed by Kimi K3 (Max) at 1674, and Qwen-3.8 Max at 1669. Tencent's Hy4 preview was ranked 6th at 1633.

Exhibit 11: Claude Opus 5 was ranked highest, followed by Kimi K3 (Max) and Qwen-3.8 Max
Arena AI Leaderboard Code Arena: WebDev



Source: Code Arena

Top Models by Intelligence Score by BenchLM

According to BenchLM, Anthropic's Claude Mythos 5, Claude Opus 5 and Claude Fable 5 lead the overall intelligence rankings, each with a score of ~83, followed by OpenAI's GPT-5.6 Sol at ~82 and Kimi K3 at ~80. Meta's Muse Spark 1.2 ranked 8th with a score of ~77, while Google's Gemini 3.6 Flash ranked 10th with a score of ~75.

Exhibit 13: Anthropic Claude models lead in the intelligence score
Top 20 Models by Intelligence by BenchLM

	Model			Pricing		Intelligence Score					
Rank	Model	Creator	Type	Input Price	Output Price	Overall Score	Coding	Agentic	Reasoning	Knowledge	Math
1	Claude Mythos 5	Anthropic	Proprietary	$10	$50	83	80	76	-	94	-
2	Claude Opus 5	Anthropic	Proprietary	$5	$25	83	78	81	91	93	-
3	Claude Fable 5	Anthropic	Proprietary	$10	$50	83	80	75	-	67	-
4	GPT-5.6 Sol	OpenAI	Proprietary	$5	$30	82	79	68	93	83	97
5	**Kimi K3**	**Moonshot AI**	**Pending**	**$3**	**$15**	**80**	**78**	**74**	**-**	**85**	**-**
6	**Qwen3.8 Max**	**Alibaba**	**Proprietary**	**-**	**-**	**80**	**67**	**79**	**96**	**67**	**-**
7	Claude Opus 4.8	Anthropic	Proprietary	$5	$25	77	72	61	76	88	67
8	Muse Spark 1.1	Meta	Proprietary	-	-	77	64	65	-	93	-
9	Grok 4.5	xAI	Proprietary	$2	$6	76	58	62	60	67	-
10	Gemini 3.6 Flash	Google	Proprietary	$2	$8	75	65	46	-	68	-
11	GPT-5.4	OpenAI	Proprietary	$3	$15	73	59	57	78	77	66
12	GPT-5.5	OpenAI	Proprietary	$5	$30	73	71	64	87	79	71
13	GPT-5.6 Terra	OpenAI	Proprietary	$2	$12	72	66	63	86	82	97
14	Claude Opus 4.7 (Adaptive)	Anthropic	Proprietary	$5	$25	72	65	64	79	83	-
15	Claude Opus 4.7	Anthropic	Proprietary	$5	$25	72	68	57	-	65	62
16	**Qwen3.7 Max**	**Alibaba**	**Proprietary**	**-**	**-**	**72**	**54**	**39**	**87**	**72**	**82**
17	Muse Spark	Meta	Proprietary	-	-	71	63	65	52	75	56
18	**MiMo-V2.5-Pro**	**Xiaomi**	**Proprietary**	**-**	**-**	**69**	**55**	**42**	**-**	**71**	**-**
19	**MiniMax M3**	**MiniMax**	**Open Weight**	**$0.3**	**$1.2**	**69**	**48**	**40**	**-**	**64**	**-**
20	**Hy3**	**Tencent**	**Open Weight**	**$0.0**	**$0.0**	**68**	**63**	**61**	**-**	**-**	**-**

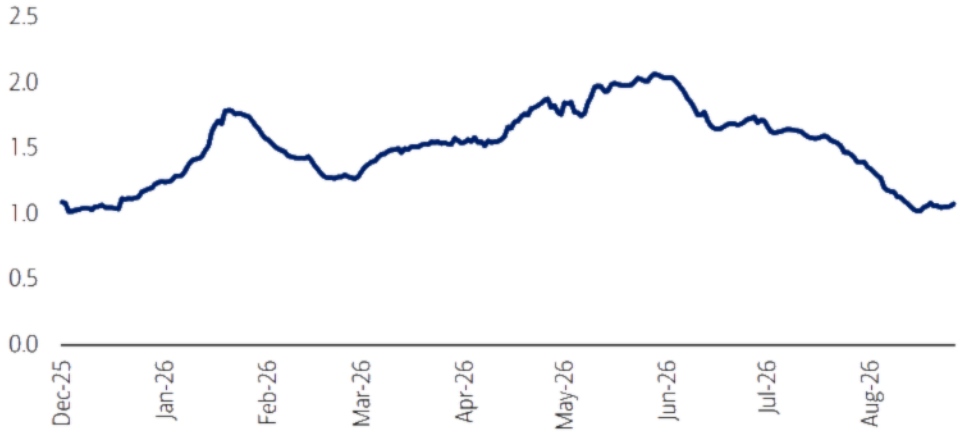
Source: benchlm.ai

Token Pricing

LLM Token Expenditure Index (Silicon Data)

In August, LLM Token Expenditure Index declined 26% m/m to $1.07 (vs $1.45 in July).

Exhibit 13: LLM token expenditure index declined to c.US$1.07 in August
Silicon Data LLM Token Expenditure Index



Source: Bloomberg, Silicon Data LLM Token Expenditure Index, as of August 26

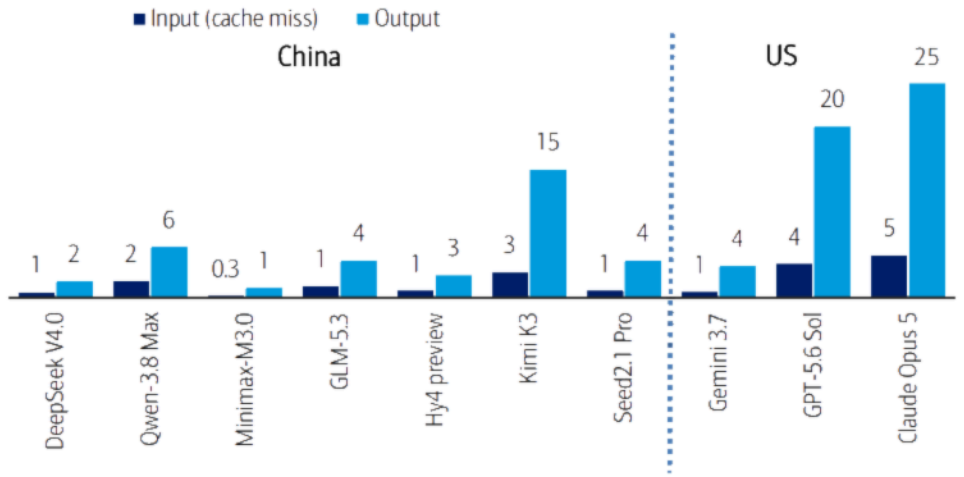
How to read LLM Token Expenditure Index

Silicon Data's LLM Token Expenditure Index provides a useful benchmark for tracking token costs. The index, created by Silicon Data—a data aggregator specializing in GPU and model usage—shows the average price paid per million LLM tokens, weighted by expenditure and independent of model type.

Model input and output pricing

US frontier models continue to maintain a notable pricing premium over Chinese peers. Claude Opus 5 has the highest token prices among the models in our comparison, at US$5/mn input tokens and US$25/mn output tokens, followed by GPT-5.6 Sol at US$4/mn and US$20/mn. Within Chinese frontier models, Kimi K3 leads with input and output token prices of US$3/mn and US$15/mn, respectively, followed by Qwen-3.8 Max at US$2/mn and US$6/mn.

Exhibit 14: China frontier models have relatively lower token pricing compared to US peers
Key models' input and output prices comparison (in US$/mn tokens)

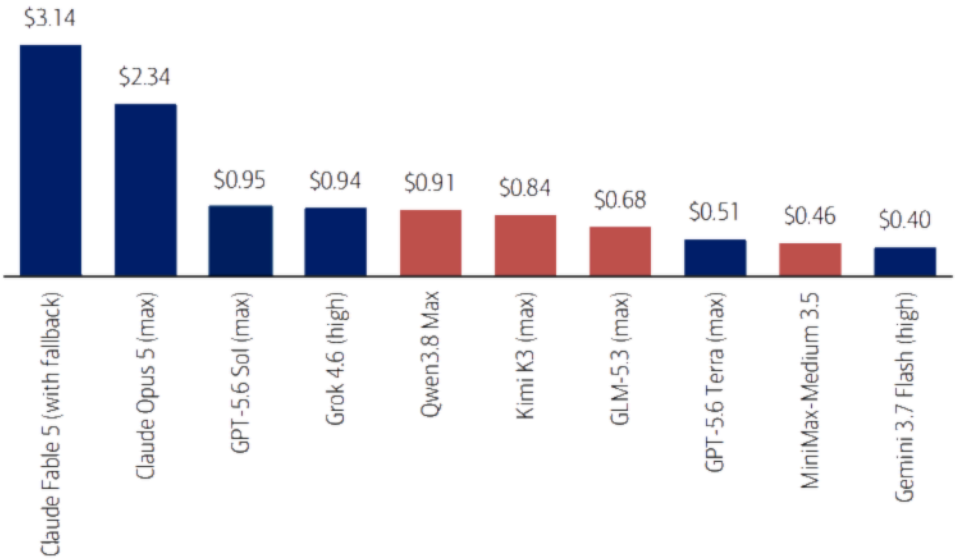


Source: DeepSeek, Alibaba Cloud, Minimax, Z.ai, Tencent, Kimi, Volcano Engine, OpenAI, Claude platform

Cost per Task (Artificial Analysis)

According to Artificial Analysis, Claude Fable 5 (w/fallback) has the highest cost per task among leading frontier AI models at $3.14 per Intelligence Index task, followed by Claude Opus 5 at $2.34 and GPT-5.6 Sol at $0.95. Qwen 3.8 Max was 5th at $0.91 and Kimi K3 was ranked 6th at $0.84.

Exhibit 15: Claude Fable 5 (w/fallback) has the highest cost per task among leading AI models
Weighted average Dollar Cost per Intelligence Index task, by model



Source: Artificial Analysis

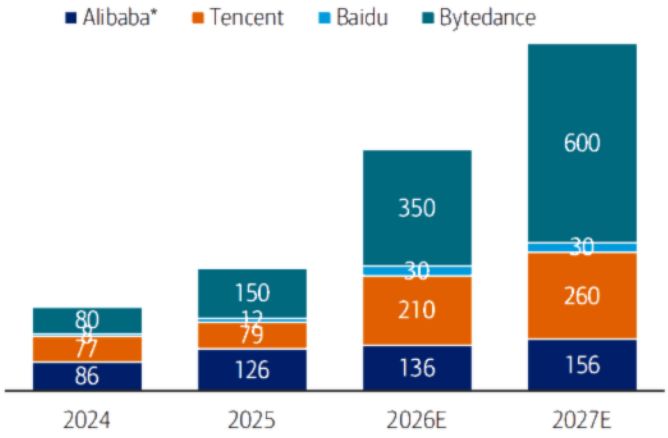
Artificial Analysis' Cost per Intelligence Index Task measures the effective cost of delivering a unit of AI capability, adjusting for model quality as captured by the Artificial Analysis Intelligence Index. The metric estimates the weighted average cost (USD) of completing a standardized AI task by incorporating all relevant inference costs, including input, cache, reasoning, and output tokens, and weighting each evaluation according to its contribution to the Intelligence Index. Lower values indicate greater cost efficiency, meaning a model can achieve a given level of intelligence at a lower inference cost.

China cloud capex trend

China hyperscalers capex

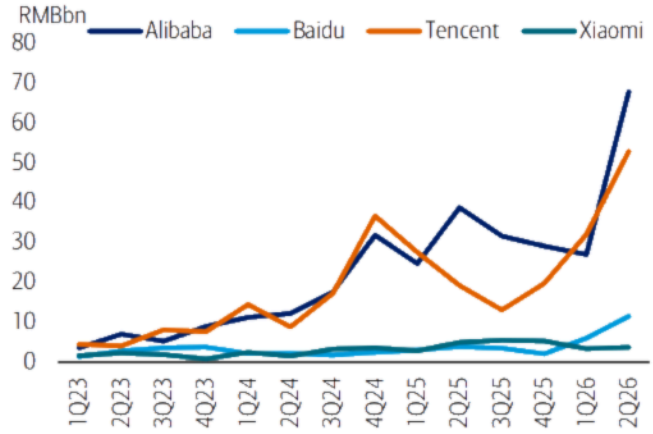
Within China, leading internet platforms ramp up investment in AI infrastructure since 2024 – the top 4 China capex spenders combined capex is close to RMB400bn in 2025 and is set to reach RMB726bn in 2026E and RMB1.0tn in 2027E.

Exhibit 16: Capex of major China internet companies is estimated to be up by 98% YoY in 2026E
Annual capex trend from 2024 to 2027E (RMBbn)



Source: company reports, SCMP, Reuters, Financial Times, BofA Global Research
*Alibaba 2024/2025E/2026E/2027E used FY2025/FY2026E/FY2027E/FY

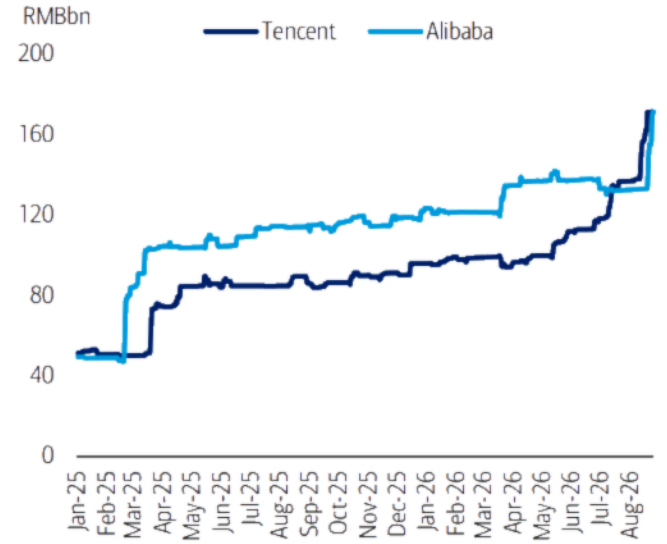
Exhibit 17: China Internet companies have increased capex significantly over the past two years
Quarterly capex trend from 1Q23 to 2Q26 (RMBbn)



Source: company reports, BofA Global Research

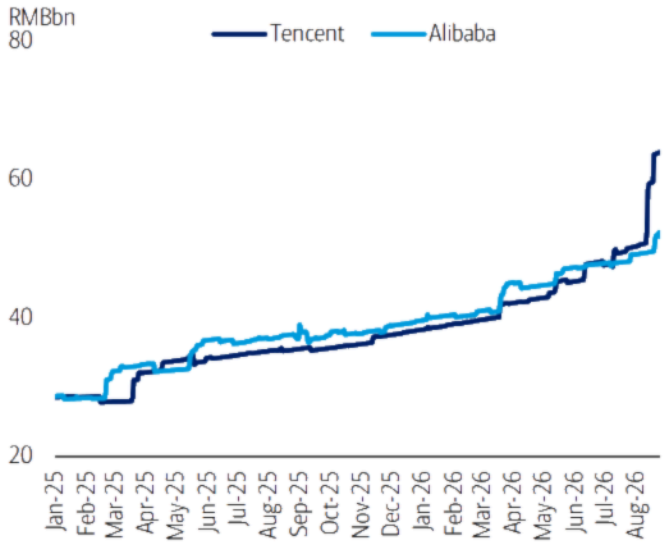
Capex/depreciation revision of China hyperscalers

Exhibit 18: Alibaba and Tencent's capex estimates trend higher amid AI investment cycle
Forward 12 months depreciation revision trend



Source: Bloomberg, BofA Global Research

Exhibit 19: Alibaba and Tencent's depreciation estimates trend higher following higher capex estimates
Forward 12 months depreciation revision trend



Source: Bloomberg, BofA Global Research

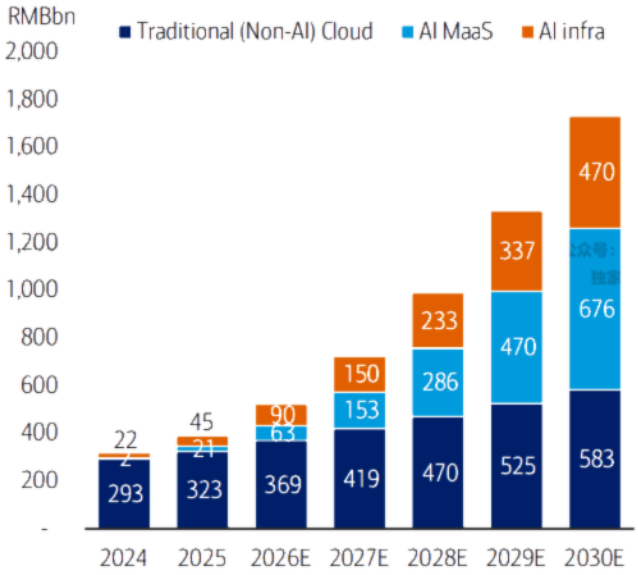
China cloud market forecasts

We summarize our key forecasts for China enterprise AI (B2B) monetization in this section. **Cloud, including AI MaaS/AI compute/non-AI cloud, remains the largest and most visible enterprise AI service revenue model** in China, with strong demand of compute power (for training and inference) and of AI model capability ("Model as a service"). *See Appendix for our China Cloud Market Model.*

We expect that **major internet incumbents** (e.g. Alibaba, Tencent, Bytedance) with substantial cloud operations will likely gain more from the rise in AI-driven enterprise spending. Meanwhile, **independent AI labs** are projected to experience faster growth between 2026E and 2030E due to their smaller revenue base and greater presence in international markets.

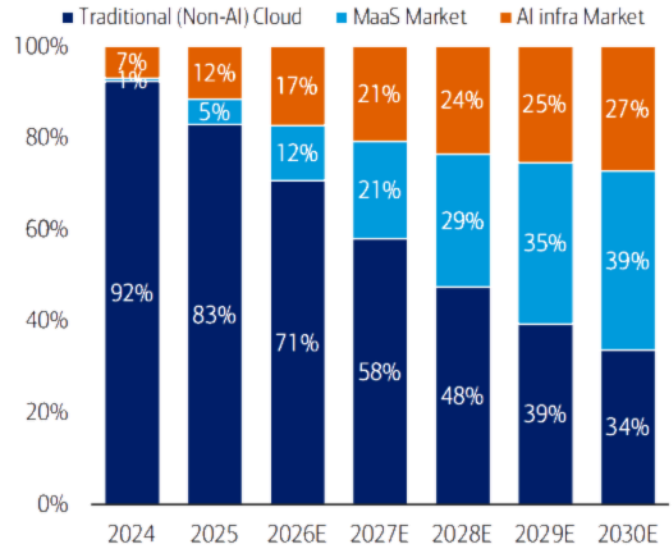
Overall we expect the China cloud market to grow from RMB389bn in 2025 to RMB1,729bn in 2030E, driven predominately by AI-related cloud business (MaaS and infra).

Exhibit 21: We expect China cloud market to reach RMB16tn by 2030E
China cloud market forecasts, 2024-2030E



Source: company, IDC, Gartners, Omdia, BofA Global Research

Exhibit 21: We expect AI-driven cloud (MaaS + infrastructure) to become the primary growth driver
China cloud market forecast breakdown by types, 2024-2030E



Source: company, IDC, Gartners, Omdia, BofA Global Research

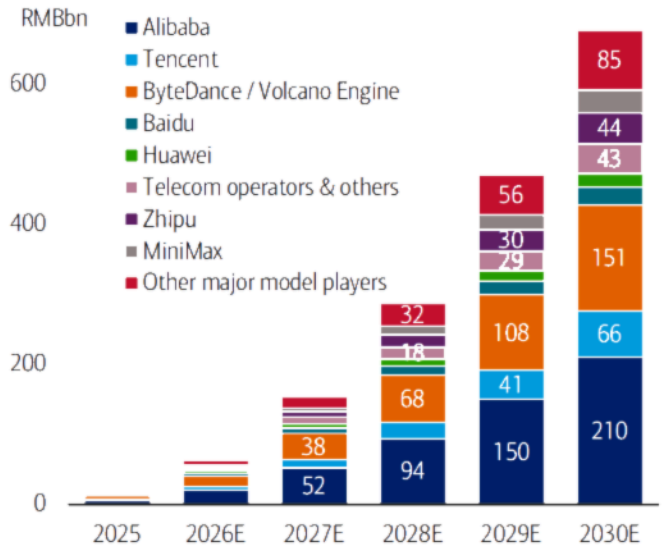
AI MaaS (AI Model as a service)

MaaS is still early but structurally critical — it converts compute + models into API-based revenue. With sector token usage exploding YTD 2026, this segment is forecasted to be the fastest growing subsegment within AI cloud stack.

Overall we expect China AI MaaS market to grow from RMB21bn in 2025 to RMB676bn in 2030E. Within the MaaS market, we expect incumbents. platforms (Alibaba, Bytedance, Tencent) to capture much of the incremental growth.

Exhibit 23: We expect China AI MaaS market to reach RMB676bn in 2030E

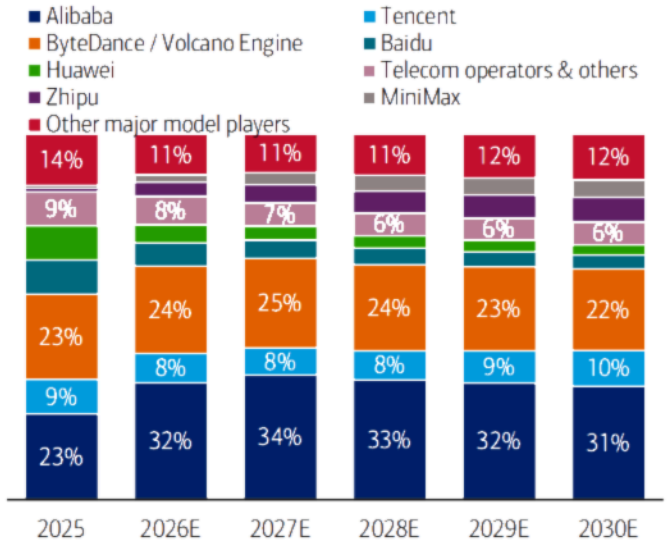
China AI MaaS cloud market forecasts, 2025-2030E



Source: company, IDC, Gartners, Omdia, BofA Global Research

Exhibit 24: We expect incumbents to capture much of incremental growth

China AI MaaS cloud market share forecasts by players



Source: company, IDC, Gartners, Omdia, BofA Global Research

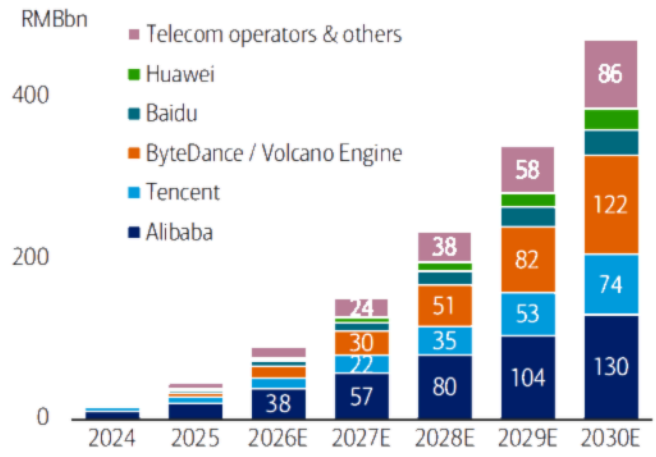
AI infrastructure cloud

AI infrastructure—including GPU compute, training and inference systems, and AI Platform-as-a-Service—is currently the most rapidly expanding large-scale revenue segment, primarily due to generative AI adoption.

Overall we expect China AI infra cloud market to grow from RMB45bn in 2025 to RMB470bn in 2030E.

Exhibit 25: We expect China AI infra cloud market to reach RMB470bn by 2030E

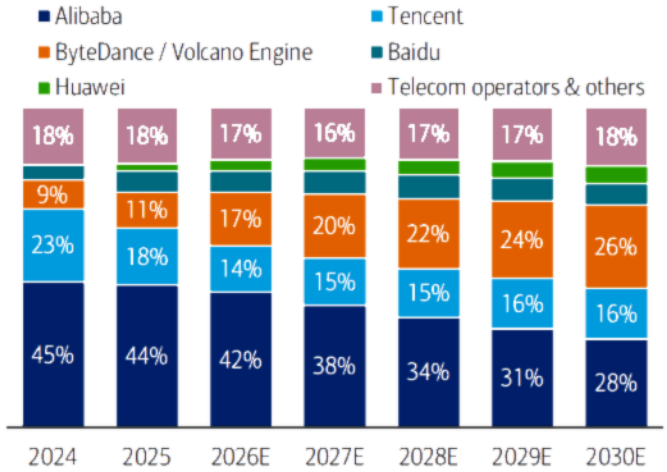
China AI infra cloud market forecasts, 2024-2030E



Source: company, IDC, Gartners, Omdia, BofA Global Research

Exhibit 26: We expect incumbents to capture much of the incremental growth

China AI infra cloud market share forecasts by company



Source: company, IDC, Gartners, Omdia, BofA Global Research

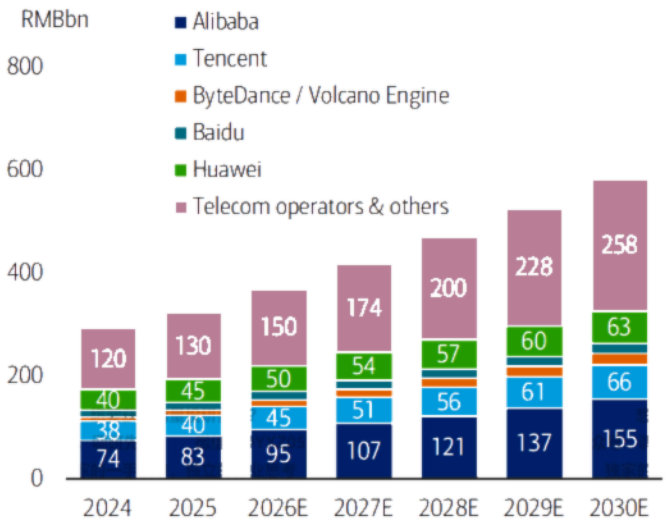
Non-AI cloud

China's non-AI cloud market remains the "base layer" and largest revenue pool for key names, but is now in a more mature scaling phase. Growth is increasingly concentrated in IaaS (compute/storage), which is the fastest-growing segment as enterprises migrate workloads and prepare for AI adoption.

Overall we expect China non-AI cloud market to grow from RMB323bn in 2025 to RMB583bn in 2030E.

Exhibit 27: We expect China non-AI cloud market to reach RMB471bn by 2030E

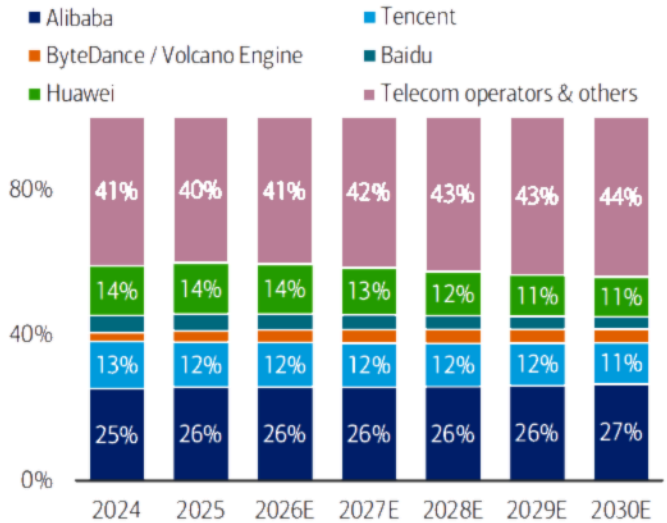
China cloud market forecasts, 2024-2030E



Source: company, IDC, Gartners, Omdia, BofA Global Research

Exhibit 28: We expect incumbents and telecom operators to capture much of the incremental growth

China cloud market share forecasts by company

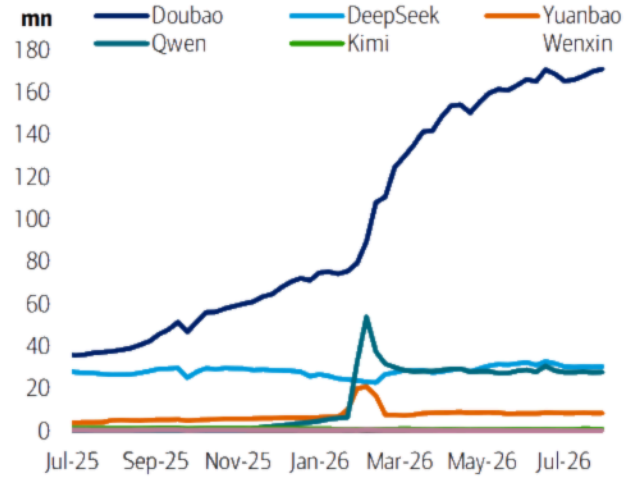


Source: company, IDC, Gartners, Omdia, BofA Global Research

China AI chatbot

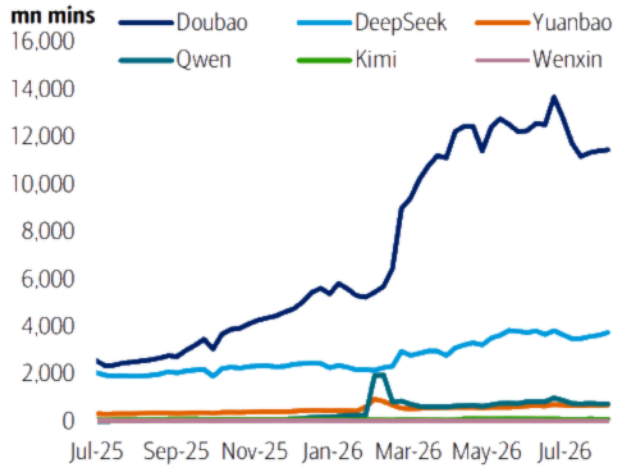
ByteDance's Doubao weekly DAU/total time spent grew to 171.3mn/11.5bn mins in the week of August 3, followed by DeepSeek at 30.7mn/3.8bn mins, Qwen at 28.0mn/752mn mins, and Yuanbao at 8.7mn/695mn mins. On a per-user basis, engagement trends were mixed. Doubao's daily time spent/number of sessions per user down notably in June/July, while DeepSeek's edged up to 18 mins/8 times in the week of August 3.

Exhibit 28: Doubao DAU surged to 171mn, widening its lead over peers
AI apps weekly DAU trend



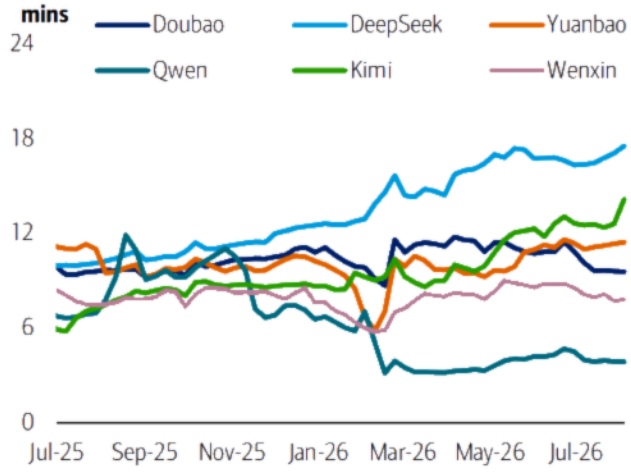
Source: QuestMobile, BofA Global Research

Exhibit 29: Doubao retained a commanding lead in weekly time spent despite a recent moderation
AI apps weekly time spent trend



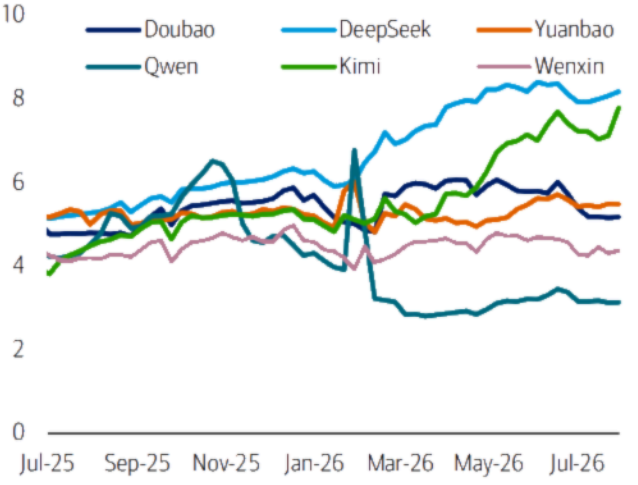
Source: QuestMobile, BofA Global Research

Exhibit 30: Average daily time spent per user remained broadly stable amid diverging app trends in the week of Aug 3 '26
AI apps daily time spent per DAU



Source: QuestMobile, BofA Global Research

Exhibit 31: Sessions per user diverged across AI apps in the week of Aug 3 '26
AI apps daily session per user



Source: QuestMobile, BofA Global Research

Appendix: China Cloud Market Model

Based on our company revenue forecasts and various industry consultants (IDC, Gartner, Omdia), we expect cloud to remain the largest and most visible monetization model for China enterprise AI services, spanning AI MaaS, AI compute infrastructure, and non-AI cloud. Demand should be driven by both rising compute requirements for AI training and inference, as well as growing adoption of model capabilities delivered as "Model as a Service."

Exhibit 33: We estimate China Cloud market to reach RMB1.7tn by 2030E
China AI Cloud market size forecasts

RMBbn	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
Total China Cloud Market	**317**	**389**	**522**	**722**	**989**	**1,332**	**1,729**	**38%**
Total Cloud YoY		23%	34%	38%	37%	35%	30%	
Within Cloud:								
Traditional (Non-AI) Cloud	**293**	**323**	**369**	**419**	**470**	**525**	**583**	**12%**
Non-AI Cloud Share within total cloud	92%	83%	71%	58%	48%	39%	34%	
Non-AI Cloud YoY		10%	14%	13%	12%	12%	11%	
AI Cloud Market	**24**	**66**	**153**	**303**	**519**	**807**	**1,146**	**66%**
AI Cloud Share within total cloud	8%	17%	29%	42%	52%	61%	66%	
AI Cloud YoY		175%	130%	98%	71%	56%	42%	
Within AI Cloud:								
MaaS Market	**2**	**21**	**63**	**153**	**286**	**470**	**676**	**81%**
MaaS Share within AI Cloud	9%	32%	41%	51%	55%	58%	59%	
MaaS Share within total Cloud	1%	5%	12%	21%	29%	35%	39%	
MaaS YoY		914%	193%	144%	87%	64%	44%	
AI infra Market	**22**	**45**	**90**	**150**	**233**	**337**	**470**	**51%**
AI infra Share within AI Cloud	91%	68%	59%	49%	45%	42%	41%	
AI infra Share within total Cloud	7%	12%	17%	21%	24%	25%	27%	
AI infra YoY		105%	100%	67%	55%	45%	39%	

Company Cloud revenue

Company	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
Alibaba	85	108	153	216	295	391	494	34%
Tencent	43	50	63	84	114	155	206	35%
ByteDance / Volcano Engine	10	20	43	83	136	210	296	62%
Baidu	15	20	26	35	48	62	76	31%
Huawei	40	48	56	66	77	91	107	18%
Telecom operators & others	124	140	170	208	257	315	387	23%
Zhipu	0	0	2	8	17	30	44	109%
MiniMax	0	0	1	5	13	22	33	124%
Other major model players	0	3	7	16	32	56	85	86%
Total	317	389	522	722	989	1332	1729	35%

Traditional non-AI cloud revenue

Company	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
Alibaba	74	83	95	107	121	137	155	13%
Tencent	38	40	45	51	56	61	66	10%
ByteDance / Volcano Engine	7	10	13	16	18	20	22	14%
Baidu	14	15	16	17	18	19	20	5%
Huawei	40	45	50	54	57	60	63	6%
Telecom operators & others	120	130	150	174	200	228	258	14%
Zhipu	0	0	0	0	0	0	0	
MiniMax	0	0	0	0	0	0	0	
Other major model players	0	0	0	0	0	0	0	
Total	293	323	369	419	470	525	583	12%

MaaS revenue

Company	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
Alibaba	1	5	20	52	94	150	210	80%
Tencent	0	2	5	12	23	41	66	91%

Exhibit 33: We estimate China Cloud market to reach RMB1.7tn by 2030E
China AI Cloud market size forecasts

RMBbn	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
ByteDance / Volcano Engine	1	5	15	38	68	108	151	78%
Baidu	0	2	4	8	13	19	25	58%
Huawei	0	2	3	6	10	15	19	58%
Telecom operators & others	0	2	5	10	18	29	43	71%
Zhipu	0	0	2	8	17	30	44	109%
MiniMax	0	0	1	5	13	22	33	124%
Other major model players	0	3	7	16	32	56	85	86%
Total	2	21	63	153	286	470	676	81%

AI infra revenue (ex-MaaS)

Company	2024	2025	2026E	2027E	2028E	2029E	2030E	CAGR 26'-30'
Alibaba	10	20	38	57	80	104	130	36%
Tencent	5	8	13	22	35	53	74	55%
ByteDance / Volcano Engine	2	5	15	30	51	82	122	69%
Baidu	1	3	6	11	17	24	31	51%
Huawei	0	1	3	6	11	17	26	71%
Telecom operators & others	4	8	15	24	38	58	86	55%
Zhipu	0	0	0	0	0	0	0	
MiniMax	0	0	0	0	0	0	0	
Other major model players	0	0	0	0	0	0	0	
Total	22	45	90	150	233	337	470	51%

Source: company, IDC, Gartners, Omdia, BofA Global Research

Exhibit 34: Glossary of abbreviations
Abbreviations and definitions

Term	Definition
API	Application Programming Interface
B2B	Business-to-Business
DAU	Daily Active Users
GPU	Graphics Processing Unit
IaaS	Infrastructure-as-a-Service
LLM	Large Language Model
MaaS	Model as a service
MTD	Month to date
YTD	Year to date

Source: BofA Global Research