

AI Series #44:AI Lab Included in Tech Index; DS Keep Discounted Price Permanent

May 24, 2026

Highlights (1) HSI announced inclusion of MiniMax and Zhipu to Hang Seng Tech Index effective on 8 Jun; (2) MiniMax: upcoming release of M3.0 and Hailuo 3.0 are to be anticipated; (3) Zhipu: breakthrough in GLM-5.1-highspeed model and innovations in network efficiencies (ZCube); (4) DeepSeek-V4-Pro API price will keep at 25% of original price after promotions ended; (5) Artificial Analysis: Qwen3.7-Max ranked no 5 in intelligence index globally

In this report, we provide investors with over 50 data points on token consumption in OpenRouter, pricing for different models, model intelligence in Artificial Analysis, user trends by sub-sector and other industry data for analysis.

Hang Seng Indexes announced MiniMax and Zhipu are to be included in Hang Seng Tech Index after HK market close last Friday. The inclusion will be effective on 8 Jun. For **MiniMax**, we expect catalyst is upcoming release of M3.0 and Hailuo 3.0. ARR is not a linear but step up function. MiniMax is a tier 1 player in coding and agentic AI with differentiations in model multimodality and pursues model tiering strategies. For **Zhipu**, it launched GLM-5.1-highspeed version with output tokens (400 tokens/s max) surpasses overseas peers. The integration of flagship model intelligence with low latency time is positive to long horizon tasks, thanks to enhancement in throughput, orchestration and computing resources utilization. Post recent updates on infra-engineering to enhance efficiencies amid supply constraints ([link](#)), it launched ZCube which achieves network efficiencies via fully flattened network topology. It achieves cost optimization (switch and optical module capex reduced by 33%), throughput improvement (up 15%) and latency improvement (TTFT P99 reduced by 40.6%)

DeepSeek -V4-Pro announced API will keep at 25% of original price after promotions ended on 31 May. In addition, media (e.g. Bloomberg, 22 May) reports emphasis of DeepSeek is about technology innovations / GI as priority vs commercialization. Amidst surging demand and supply constraints, modeling industry is expected to have fast iterations, tech innovations and inference efficiencies down the road. Model companies have respective strengths across different used case scenario and pursues model tiering strategies (i.e flagship models priced at premium on intelligence)

Qwen3.7-Max ranked no 5 globally per Artificial Analysis. The model has competitive moats in long-horizon autonomous execution, frontier coding agent, productivity and cross-harness generalization. It can handle long horizon tasks for 35 hours continuously using 1000+ tool calls with zero hand-holding. Please refer to report ([link](#)) for details. According to Artificial Analysis, it is ranked no 5 globally after GPT-5.5, Opus 4.7, Gemini 3.1 and GPT-5.4 (xhigh). It ranked first among China models. On Terminal-Bench 2.0 (coding) and SWE-bench Pro (agentic), Qwen3.7-Max ranked first and ahead of DS-V4-Pro Max/GLM-5.1/Kimi K2.6/Opus-4.6 Max.

BAT, KC, AI labs and Kling are Beneficiaries. We expect surge in token consumption on rising demand for AI models and agents to benefit CSPs like **Baidu, Alibaba, Tencent (BAT), Kingsoft Cloud (KC), AI labs and Kling**. According to third party SemiAnalysis on 1 May, Anthropic ARR is forecast to reach USD44bn. According to analysis from Artificial Analysis, (1) there is a narrowing intelligence gap between US and China models; (2) API cost is just a fraction vs US benefiting from MoE, linear attention, and MFU ([Exhibit 8-14](#)); (3) Coding, Workspace scenario and mid- to long-form content should see significant opportunities ahead ([Exhibit 6/-/1](#)).

Exhibit 1 - Weekly token consumption over the past 12 months as of 18 May (tn)

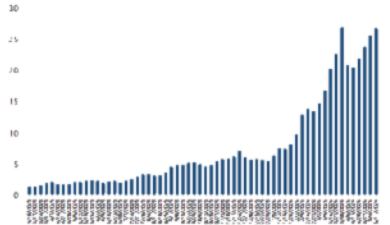


Exhibit 2 - Top 10 ranking in terms of token consumption by model the latest week

No	Company	Model	Token
1	Tencent	Hy3 gptmax	2 667
2	Deepseek	Deepseek v4 labh	2 067
3	Anthropic	Claude Sonnet 4.5	1 537
4	Anthropic	Claude 3.5 Sonnet 4.7	1 547
5	Google	Gemini 1.5 Flash Preview	1 157
6	MooshuAI	Katana K2.6	1 037
7	Deepseek	Deepseek v3.2	1 037
8	OpenRouter	Qwen Alpha	8958
9	Deepseek	Deepseek V4 Pro	8918
10	MiniMax	MiniMax M2.7	7508

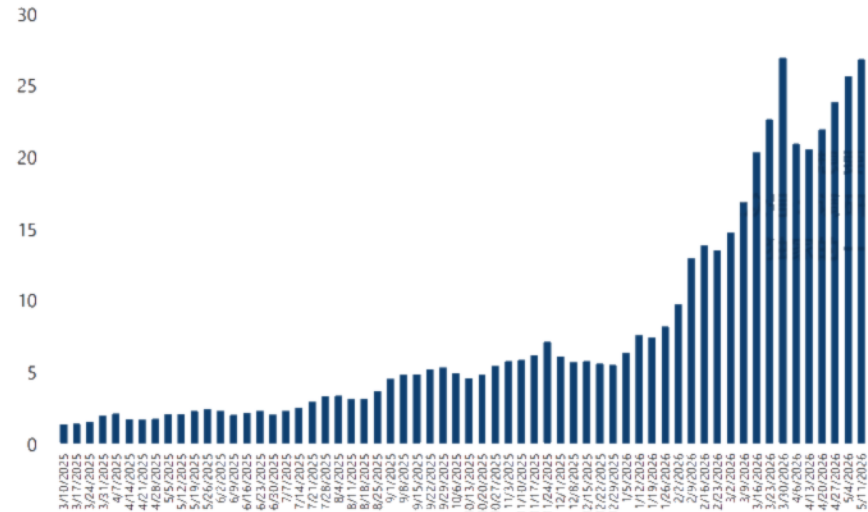
China vs US comparison

AI Models

According to OpenRouter, token consumption went up significantly in Feb-26 due to (a) a **surge in adoption of coding / agents**, which allows people using AI from "to chat" to "to work". It connects with large models which can be deployed locally or on servers which allows you to automate workflow through natural language chat; (b) **release of a number of high performance and cost-efficient models** which include Kimi K2.5 (Jan-26), M2.5 (Feb-26) and GLM-5 (Feb-26).

According to OpenRouter, rising adoption of Coding / Agents and release of high-performance and cost-efficient China models are driving token consumption

Exhibit 3 - Weekly token consumption over the past 12 months as of 18 May (tn)

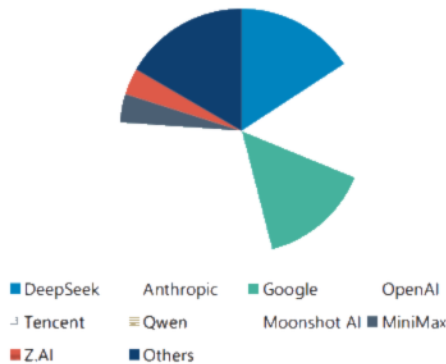


Source: OpenRouter, Jefferies

According to OpenRouter (between 11 May and 17 May), China models token consumption declined by 3% compared to prior week at 7.7tn after the release of Hy3 preview (free) with promotions till 8 May. This compares to US models at 4.2tn token consumption.

In OpenRouter, China models surpassed US models in terms of token consumption the week of 11 May post the release of Hy3 preview (free)

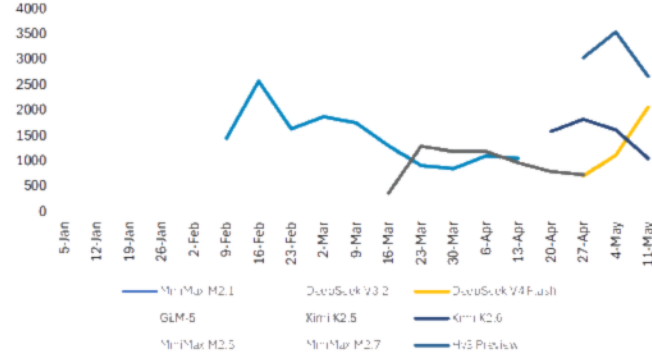
Exhibit 4 - Market share: weekly token consumption by company as of 18 May



Source: OpenRouter, Jefferies

On token consumptions by company, DeepSeek (15.8%) comes first, followed by Anthropic (15.4%), Google (14.8%), OpenAI (9.9%) and Tencent (9.9%). The top 5 weekly token consumption by model include Hy3 preview (2.66tn), DeepSeek V4 Flash (2.06tn), Claude Sonnet 4.6 (1.55tn), followed by Claude Opus 4.7 (1.54tn) and Gemini 3 Flash Preview (1.15tn).

Exhibit 5 - Weekly token consumption since Jan-26 (bn)



Source: OpenRouter, Jefferies

Hy3 preview ranked first in terms of token consumption the week of 11 May

Exhibit 6 - Top 10 ranking in terms of token consumption by model the latest week

No	Company	Model	Tokens
1	Tencent	Hy3 preview	2.66T
2	DeepSeek	DeepSeek V4 Flash	2.06T
3	Anthropic	Claude Sonnet 4.6	1.55T
4	Anthropic	Claude Opus 4.7	1.54T
5	Google	Gemini 3 Flash Preview	1.15T
6	MoonshotAI	Kimi K2.6	1.05T
7	DeepSeek	DeepSeek V3.2	1.03T
8	OpenRouter	Owl Alpha	895B
9	DeepSeek	DeepSeek V4 Pro	893B
10	MiniMax	MiniMax M2.7	750B

Source: Data from Artificial Analysis

Based on Artificial Analysis, China models DeepSeek, Qwen, Kimi, MiniMax and Zhipu are closing the gap with US models in terms of intelligence. China models are highly cost-efficient on Mixture of Experts (MoE) architecture, linear attention, and MLU (Model Flog Utilization). According to 2026 Stanford AI Index Report (Apr-26), the top US model leads Chinese models by 2.7%. Separately, Anthropic (1503), xAI (1495), Google (1494), OpenAI (1481), Alibaba (1449) and DeepSeek (1424) occupy the top tier of Arena Elo ratings as of Mar-26.

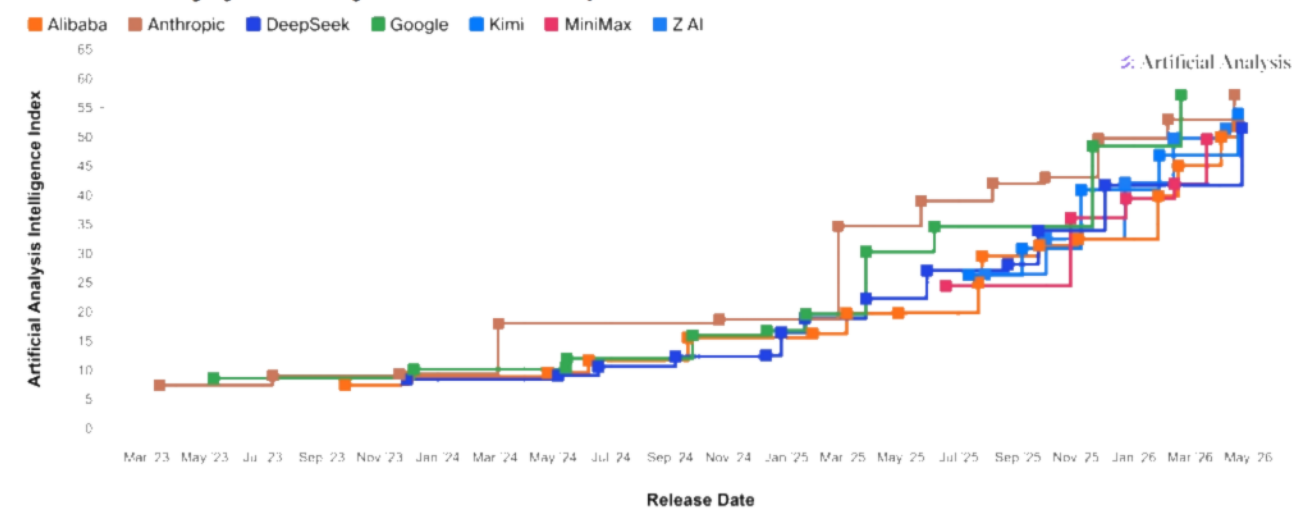
Exhibit 7 - No 11-20 ranking in terms of token consumption by model in latest week

No	Company	Model	Tokens
11	StepFun	Step3.5 Flash	709B
12	NVIDIA	Nemotron 3 Super (free)	616B
13	Anthropic	Claude Opus 4.6	574B
14	Google	Gemini 2.5 Flash Lite	568B
15	Google	Gemini 2.5 Flash	565B
16	OpenAI	GPT-5.5	460B
17	InclusionAI	Ling-2.6-1T (free)	435B
18	OpenAI	gpt-oss-120b	421B
19	Google	Gemini 3.1 Pro Preview	414B
20	Qwen	Qwen3.6 Plus	378B

Source: Data from Artificial Analysis

China models are narrowing the gap with US models in terms of intelligence. On Input/Output API, China large models are highly efficient vs US models.

Exhibit 8 - Frontier Language Model Intelligence over time as of 18 May



Source: Data from Artificial Analysis

In terms of Input/Output API price, China models are just a fraction vs US models.

Exhibit 9 - Intelligence index for top 20 models as of 18 May

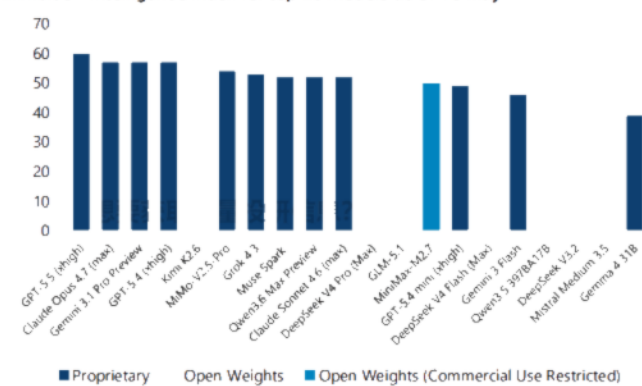
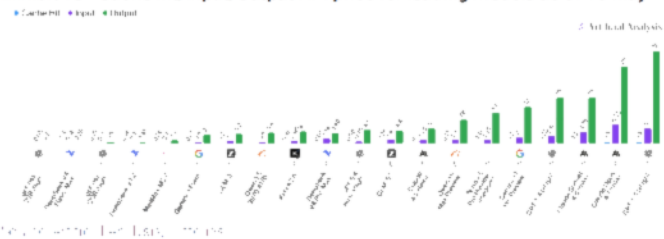


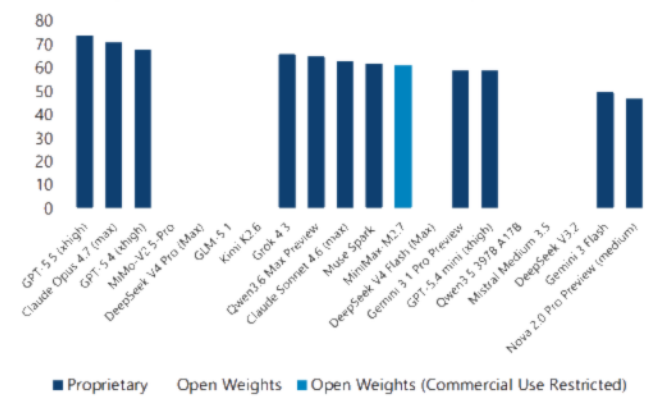
Exhibit 10 - Cache hit/Input/Output API price for leading models as of 18 May



Source: Data from Artificial Analysis

els, there are 9 which are China models and open weighted as of 18 May.

Exhibit 11 - Agentic index for top 20 models as of 18 May



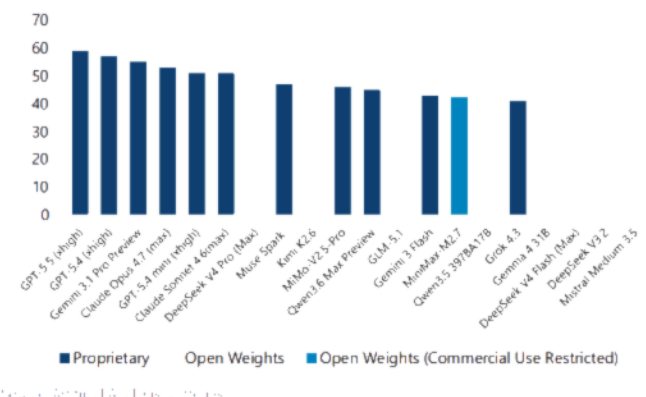
Source: OpenAI, Anthropic, Google, etc.

Exhibit 13 - Cost to run Artificial Intelligence Index as of 18 May



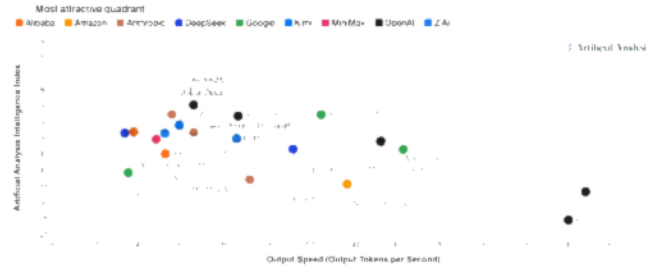
Source: OpenAI, Anthropic, Google, etc.

Exhibit 12 - Coding index for top 20 models as of 18 May



Source: OpenAI, Anthropic, Google, etc.

Exhibit 14 - Output speed to run Artificial Intelligence Index as of 18 May



Source: OpenAI, Anthropic, Google, etc.

Input/Output Price for Different Models / Pricing for Coding Plan and Agents Offerings

Red color denotes price hike / new price during the month

Exhibit 15 - Pricing plans for Alibaba, Tencent, Baidu and ByteDance

Models Provider	Model	Context	May-26	
			Input (RMB) (million tokens)	Output (RMB) (million tokens)
Alibaba	Qwen3.5-Max-Preview	(0, 128k]	9	54
		(128k, 256k]	15	90
	Qwen3.5-Plus	(0, 256k]	2	12
		(256k, 1m]	8	48
		(0, 1276k]	0.8	4.8
Tencent	Qwen3.5-Plus	(128k, 256k]	2	12
		(256k, 1m]	4	24
	HY 3.0 Preview	(0, 16k]	1.2	4
		(16k, 32k]	1.6	6.4
		(32k, 256k]	2	8
	HY 2.0 Think	(0, 32k]	3.975	15.9
		(32k, 128k]	5.8	23.2
	HY 2.0 Instruct	(0, 32k]	3.18	7.95
		(32k, 128k]	4.505	11.13
	Hunyuan-T1		1	4
Baidu	Hunyuan-Turbo5		0.8	2
	Tencent HY Vision 1.5 Instruct		3	9
	Hunyuan-translation		1.2	3.6
	ERNIE-S.1	(0, 32k]	4	16
		(32k, 128k]	6	22
ByteDance	ERNIE-S.0	(0, 32k]	6	24
		(32k, 128k]	10	40
	ERNIE-4.5-Turbo		0.8	3.2
			3	9
	ERNIE-4.5-Turbo-VL	[0, 37k]	1.7	16
		(37k, 128k]	4.8	24
	Doubao-seed-2.0-pro	(128k, 256k]	9.6	48
		(0, 32k]	0.6	3.6
	Doubao-seed-2.0-lite	(32k, 128k]	0.9	5.4
		(128k, 256k]	1.8	10.8
ByteDance	Doubao-seed-2.0-mini	(0, 32k]	0.2	2
		(32k, 128k]	0.4	4
	Doubao-seed-2.0-code	(128k, 256k]	0.8	8
		(0, 37k]	3.7	16
		(32k, 128k]	4.8	24
		(128k, 256k]	9.6	48

Source: AI Lab, Zhipu, Kimi

Exhibit 16 - Pricing plans for AI Labs MiniMax, Zhipu, Kimi

Models Provider	Model	Context	May-26	
			Input (RMB) (million tokens)	Output (RMB) (million tokens)
MiniMax	MiniMax-M2.7		2.1	8.4
	MiniMax-M2.7-highspeed		4.2	16.8
	MiniMax-M2.5		2.1	8.4
	MiniMax-M2.5-highspeed		4.2	16.8
	GLM-5.1	[0, 37k] (32k-i)	6 8	24 28
Zhipu AI	GLM-5-Turbo	[0, 32k] (32k-i)	5 7	22 26
	GLM-5	[0, 32k] (32k-i)	4 6	18 22
	GLM-4.7	Input length: [0, 37k] Output length: [0, 0.2k] Input length: [0, 37k] Output length: [0.2k-i] (32k, 200k)	2 3 4 6.5	8 14 16 27
	Kimi K2.6		4	16
	Kimi K2.5		4	21
Kimi	kimi-k2-thinking		4	16
	kimi-k2-thinking-turbo		8	58
	kimi-k2-turbo-preview			
	kimi-k2-Q905-preview		4	16
	kimi-k2-Q711-preview			
	moonshot-v1-8k		2	10
	moonshot-v1-8k-vision-preview			
	moonshot-v1-32k		5	20
	moonshot-v1-32k-vision-preview			
	moonshot-v1-128k		10	30
	moonshot-v1-128k-vision-preview			

Exhibit 17 - Pricing plans for AI Labs StepFun, Xiaomi, and DeepSeek

Models Provider	Model	Context	May-26	
			Input (RMB) (million tokens)	Output (RMB) (million tokens)
StepFun	step-3.5-flash		0.7	2.1
	step-1.5-mini		2.5	8
	step-3	inputs 4k & outputs 4k	1.5	4
		inputs 4k & output > 4k	1.5	6
		4k < input ≤ 64k	4	10
	step-2-mini		1	2
	step-2-16k			
	step-2-16k-exp		38	120
	step-1-8k		5	20
	step-1-32k		15	70
Xiaomi	MiMo-V2.5-Pro	(0, 256k]	7	21
		(256k, 1m]	14	42
	MiMo-V2.5	(0, 256k]	2.8	14
		(256k, 1m]	5.6	28
	MiMo-V2-Pro	(0, 256k]	7	21
		(256k, 1m]	14	42
	MiMo-V2-Omini		2.8	14
	MiMo-V2-TTS		limited-time free	
	MiMo-V2-Flash		0.7	2.1
	DeepSeek V4 Flash		1	2
DeepSeek	DeepSeek V4 Pro		3	6
	DeepSeek V3.2		0.7	2

Source: Anthropic, OpenAI, DeepSeek

Exhibit 18 - Pricing plans for AI Labs StepFun, Xiaomi, and DeepSeek

Models Provider	Model	Context	May-26	
			Input (RMB) (million tokens)	Output (RMB) (million tokens)
StepFun	step-3.5-flash		0.7	2.1
	step-1.5-mini		2.5	8
	step-3	inputs 4k & outputs 4k	1.5	4
		inputs 4k & output > 4k	1.5	6
		4k < input ≤ 64k	4	10
	step-2-mini		1	2
	step-2-16k			
	step-2-16k-exp		38	120
	step-1-8k		5	20
	step-1-32k		15	70
Xiaomi	MiMo-V2.5-Pro	(0, 256k]	7	21
		(256k, 1m]	14	42
	MiMo-V2.5	(0, 256k]	2.8	14
		(256k, 1m]	5.6	28
	MiMo-V2-Pro	(0, 256k]	7	21
		(256k, 1m]	14	42
	MiMo-V2-Omini		2.8	14
	MiMo-V2-TTS		limited-time free	
	MiMo-V2-Flash		0.7	2.1
	DeepSeek V4 Flash		1	2
DeepSeek	DeepSeek V4 Pro		3 (original 12, promo till 31 May)	6 (original 24, promo till 31 May)
	DeepSeek V3.2		0.7	2

Source: Anthropic, OpenAI, DeepSeek

Exhibit 19 - Pricing plans for overseas players Anthropic, OpenAI and Google

LLM Provider	Model	Context	May-26	
			Input (US\$) (million tokens)	Output (US\$) (million tokens)
Anthropic	Opus 4.7/4.6/4.5		5	25
	Sonnet 4.6/4.5		3	15
	Haiku 4.5		1	5
	GPT-5.5		5	30
	GPT-5.4		2.5	15
	GPT-5.4-mini		0.75	4.5
OpenAI	GPT-5.4-runs		0.7	1.75
	GPT-5.4-pro		30	180
	GPT-5.2		1.75	14
	GPT-5.2-pro		21	168
	GPT-5.1		1.25	10
	GPT-5		1.25	10
Google	Gemini 3.1 Pro Preview	(0, 200k]	2	12
		(200k, ∞)	4	18
	Gemini 3.1 Flash-Lite Preview	text/image/video	0.25	1.5
		audio	0.5	1.5
	Gemini 3 Flash Preview	text/image/video	0.5	3
		audio	1	3

Source: Anthropic, OpenAI, Google

Video generation model offerings

Red color denotes price hike / new price during the month

Pricing Plans for different video generation models

Exhibit 20 - Video generation model pricing in RMB

App	Model	Monthly standard plan (RMB)	Monthly credit quota	Credit per video	Length per video	Cost (RMB) per video
HappyHorse AI	HappyHorse-1.0	73	8.7k	16000	4.75s	1.52-76
	Video 30/30 Omni - 720p	58	660	27.11s	3.75s	227.1186
	Video 3.0/3.0 Omni - 1080p	58	660	36.80	3.75s	316.1582
Kling AI	Video 01 - 720p	58	660	30-60	5.70s	2.68-5.77
	Video 01 - 1080p	58	660	30-80	5.70s	1.50-7.91
	Video 2.0	58	660	60-90	5.70s	4.99-8.19
	Video 2.5 Turbo - 720p	58	660	15-30	5.70s	7.32-26.4
	Video 2.5 Turbo - 1080p	58	660	15-30	5.70s	22.4-39
Jimeng AI	Seedance 2.0 Fast	99	72k	2677	4.75s	190.79
	Seedance 2.0	99	72k	32770	4.75s	107.1142
	Seedance 1.5 Pro	67	72k	30720	5.73s	476.1142
	Seedance 1.0	99	72k	50700	5.70s	4.76-9.52
	Huakuo 2.3 - 768p	98	1000	50700	6.70s	3.8-6.8
	Huakuo 2.3 - 1080p	98	1000	180	6s	10.88
	Huakuo 2.3 Fast - 768p	98	1000	10-60	6.70s	2.04-4.08
Haibao AI	Huakuo 2.3 Fast - 1080p	98	1000	180	6s	5.8
	Huakuo 2.0 - 720p	98	1000	23570	6.70s	14.3-3.4
	Huakuo 2.0 - 768p	98	1000	60700	6.70s	1.48-6.8
	Huakuo 2.0 - 1080p	98	1000	180	6s	10.88
	Video 03 Pro - 540p	59	800	10760	1.75s	0.74-1.18
Vidu	Video 03 Pro - 720p	59	800	75400	1.75s	7.88-79.1
	Video 03 Pro - 1080p	59	800	80480	1.75s	2.71-71.8
	Video 03 Turbo - 540p	59	800	8178	1.75s	0.98-9.44
	Video 03 Turbo - 720p	59	800	12762	1.75s	0.89-14.6
	Video 03 Turbo - 1080p	59	800	14224	1.75s	1.03-16.32
	Video 1.1 reference - 540p	59	800	80740	4.75s	2.71-71.8
	Video 1.1 reference - 720p	59	800	80770	5.75s	4.44-21.8
	Video 03 reference - 1080p	59	800	75400	3.75s	5.58-29.5

Source: Official website of the app

Exhibit 21 - Video generation model pricing in USD

App	Model	Monthly standard plan (USD)	Monthly credit quota	Credit per video	Length per video	Cost (USD) per video
HappyHorse AI	Video 30/30 Omni - 720p	8.8	660	27.11s	3.75s	0.86-1.8
	Video 3.0/3.0 Omni - 1080p	8.8	660	36.80	3.75s	0.88-2.4
	Video 01 - 720p	8.8	660	30-60	5.70s	0.4-0.8
	Video 01 - 1080p	8.8	660	30-80	5.70s	0.98-1.87
	Video 2.0	8.8	660	60-90	5.70s	0.67-1.31
Kling AI	Video 2.5 Turbo - 720p	8.8	660	15-30	5.70s	0.87-0.4
	Video 2.5 Turbo - 1080p	8.8	660	15-30	5.70s	0.33-0.67
	Seedance 2.0 Fast	15	1575	56770	4.75s	0.93-2
	Seedance 2.0	15	1575	68225	4.75s	0.66-2.14
	Seedance 1.5 Pro	15	1575	100740	5.73s	0.96-22.9
Jimeng AI	Seedance 1.0	15	1575	23750	5.70s	0.28-0.48
	Huakuo 2.3 - 1080p	9.99	1000	180	6s	1.6
	Huakuo 2.3 Fast - 768p	9.99	1000	10-60	6.70s	0.6-0.6
	Huakuo 2.3 Fast - 1080p	9.99	1000	180	6s	1
	Huakuo 2.0 - 768p	9.99	1000	30700	6.70s	0.57
	Huakuo 2.0 - 1080p	9.99	1000	180	6s	1.6
Google Gemini	Video 3.1	With Google AI Plus (\$19.99 per month with 200 credits) and Pro (\$55.99 per month with 1000 credits) plans				4.8s
	Video 3.1	With Google AI Ultra (\$49.99 per month with 25,000 credits)				4.8s
PixVerse	PixVerse V1 - 720p	10	1200	10710	1.75s	0.08-0.17
	PixVerse V1 - 1080p	10	1200	20700	1.75s	0.17-0.3
	PixVerse V1 - 720p	10	1200	10710	1.75s	0.08-0.17
	PixVerse V1 - 1080p	10	1200	20700	1.75s	0.17-0.3
	PixVerse V5.6 - 720p	10	1200	10700	1.70s	0.08-0.83
	PixVerse V5.6 - 1080p	10	1200	20740	1.8s	0.17-0.31
	PixVerse V5 - 720p	10	1200	8780	1.70s	0.07-0.67
	PixVerse V5 - 1080p	10	1200	16728	1.8s	0.13-1.07
Runway	Gen-2.5	15	625	24720	2.70s	0.98-2.88
	Gen-2 Video	15	625	40770	5.70s	7.88-2.88
	Gen-2 Video Turbo	15	625	25750	5.70s	0.67-2

Source: Official website of the app

Exhibit 22 - Video generation membership comparison (original price)

App	Membership	Monthly price (RMB)	Cost per credit (RMB)	Cost per second (720P)	Cost per second (1080P)
HappyHorse AI	Standard	87.5	0.10	RMB 0.93s	RMB 1.6s
	Pro	350	0.10	RMB 0.93s	RMB 1.6s
	Gold	66	0.10	RMB 0.93s	RMB 1.2s
Kling AI - Video 3.0	Platinum	268	0.09	RMB 0.63s	RMB 1.07s
	Diamond	966	0.08	RMB 0.73s	RMB 1.0s
	Black Gold	1514	0.05	RMB 0.45s	RMB 0.63s
Jimeng AI - Seedance 2.0	Basic	79	0.11		
	Standard	299	0.11	RMB 0.67s	RMB 0.63s
	Pro	649	0.11		

Source: Official website of the app

Exhibit 23 - Video generation membership comparison (after promotion)

App	Membership	Monthly price (RMB)	Cost per credit (RMB)	Cost per second (720P)	Cost per second (1080P)
HappyHorse AI	Standard	70	0.08	RMB 0.51s	RMB 0.9s
	Pro	245	0.07	RMB 0.44s	RMB 0.70s
	Gold	56	0.06	RMB 0.79s	RMB 1.05s
Kling AI - Video 3.0	Platinum	244	0.08	RMB 0.70s	RMB 0.94s
	Diamond	526	0.07	RMB 0.66s	RMB 0.80s
	Black Gold	1214	0.05	RMB 0.43s	RMB 0.63s
Jimeng AI - Seedance 2.0	Basic	65	0.10		
	Standard	199	0.09	RMB 0.72s	
	Pro	499	0.08	RMB 0.65s	

Source: Official website of the app

Coding Plans and Agents offerings

Pricing Plans for different Coding and Agent offerings by leading players

Red color denotes price hike during the month



Exhibit 24 - Volcano Engine Ark Agent Plan

Program	Price	Limit
Small	RMB 40/month	Monthly AFP: 20,000
		Weekly AFP: 7,000
		5 hrs AFP: 2,000
		VLM daily AFP: 10,000
Medium	RMB 200/month	Monthly AFP: 100,000
		Weekly AFP: 35,000
		5 hrs AFP: 10,000
		VLM daily AFP: 50,000
Large	RMB 500/month	Monthly AFP: 250,000
		Weekly AFP: 87,500
		5 hrs AFP: 25,000
		VLM daily AFP: 125,000
Max	RMB 1000/month	Monthly AFP: 500,000
		Weekly AFP: 175,000
		5 hrs AFP: 50,000
		VLM daily AFP: 250,000

Source: Tencent AI agent plan

Exhibit 25 - Tencent AI products pricing

Product	Version	Pricing
QClaw	Internal testing	TBD
	Individual Trial	Limited-time free, 500 credits per month
	Individual Pro	RMB58 per month, 2,000 credits per month (incl. 500 credits for trial)
WorkBuddy	Enterprise Flagship	RMB78 per month, 2,000 credits per month
	Enterprise Pro	RMB58 per month, 2,000 credits per month
	Starter	RMB45 per month, 2 vCPUs, 2GB MEM, 40GB SSD, 3Mbps Bandwidth, 200GB monthly data package
	Starter (Basic)	RMB50 per month, 2 vCPUs, 2GB MEM, 50GB SSD, 4Mbps Bandwidth, 300GB monthly data package
Lighthouse	General (Advanced)	RMB700 per month, 2 vCPUs, 4GB MEM, 100GB SSD, 7Mbps Bandwidth, 7,000Gb monthly data package
	General (Pro)	RMB245 per month, 4 vCPUs, 8GB MEM, 180GB SSD, 12Mbps Bandwidth, 2,000Gb monthly data package
	Razor Speed	RMB45 per month, 2 vCPUs, 2GB MEM, 40GB SSD, 200Mbps Bandwidth, unlimited data package
	Razor Speed	RMB230 per month, 4 vCPUs, 8GB MEM, 120GB SSD, 200Mbps Bandwidth, unlimited data package

Source: Tencent AI agent plan

Exhibit 26 - Hy token plan

Program	Price	Unit price (RMB/Mtok)	Tokens amount (m)
Lite	RMB 28/month	0.8	35
Standard	RMB 78/month	0.78	100
Pro	RMB 238/month	0.74	320
Max	RMB 468/month	0.72	650

Source: Hy agent plan

Exhibit 27 - Tencent token plan

Program	Price	Tokens amount (m)
Lite	RMB 39/month	35
Standard	RMB 99/month	100
Pro	RMB 299/month	320
Max	RMB 599/month	650

Source: Tencent AI agent plan

Exhibit 28 - BABA Coding Plan

Program	Price	Usage limitations
Coding Plan Pro	RMB 200/month	Every 5 hours: Maximum 6000 requests
		Every week: Maximum 45000 requests
		Every subscription month: Maximum 90000 requests

Source: Alibaba AI agent plan

Exhibit 29 - BABA token plan

Region	Program	Price	Usage limitations
China (Team Subscription)	Standard Seat	RMB 198/month	25,000 credits/month
	Pro Seat	RMB 698/month	100,000 credits/month
	Max Seat	RMB 1,398/month	250,000 credits/month
Overseas	Pro Version	USD 50/month	Maximum of 90,000 requests/month

Source: Alibaba AI agent plan

Exhibit 30 - Baidu coding plan

Program	Price	Usage limitations
Coding Plan Lite	RMB 40/month	Every 5 hours: Maximum 1200 requests
		Every week: Maximum 9000 requests
		Every subscription month: Maximum 18000 requests
Coding Plan Pro	RMB 200/month	Every 5 hours: Maximum 6000 requests
		Every week: Maximum 45000 requests
		Every subscription month: Maximum 90000 requests

Source: Baidu, 2020, 2021, 2022

Exhibit 31 - MaxClaw subscription prices

MaxClaw subscription prices					
Region	Program	Free	Basic	Pro	Ultra
China	Standard price	R1190	R18133 per month	R15483 per month	
	Monthly program	R1190	R17939 per month	R15313 per month	
	Privileges	200 credits per day 1,000 free credits available for 3 days	5,000 credits per month About 45 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MacClaw	16,000 credits per month About 110 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MaxClaw	N/A
		Standard price	US230 per month	USD119 per month	USD239 per month
		Monthly program	US230	USD69 per month	USD129 per month
		Yearly program	US210	USD56 per month	USD109 per month
Overseas	Privileges	200 credits per day 1,000 free credits available for 3 days	5,000 credits per month About 40 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MacClaw	20,000 credits per month About 120 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MaxClaw	40,000 credits per month About 240 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MaxClaw
		Standard price	US230 per month	USD119 per month	USD239 per month
		Monthly program	US230	USD69 per month	USD129 per month
		Yearly program	US210	USD56 per month	USD109 per month
	Privileges	200 credits per day 1,000 free credits available for 3 days	5,000 credits per month About 40 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MacClaw	20,000 credits per month About 120 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MaxClaw	40,000 credits per month About 240 Pro mode tasks available 200 daily credits Peak hour priority Remove watermark Custom domain Deploy 24/7 Cloud MaxClaw
		Standard price	US230 per month	USD119 per month	USD239 per month

Source: MaxClaw, 2020, 2021, 2022

Exhibit 32 - MaxClaw add-on prices

Region	Starter pack	Booster pack	Mega pack
China	RMB59 for 5,000 credits (valid for 1 year)	RMB109 for 10,000 credits (valid for 1 year)	RMB149 for 15,000 credits (valid for 1 year)
Overseas	USD39 for 5,000 credits (valid for 1 year)	USD78 for 10,000 credits (valid for 1 year)	USD156 for 20,000 credits (valid for 1 year)

Source: MaxClaw, 2020, 2021, 2022

Exhibit 33 - MiniMax token plan

Program	China	Overseas	Usage limitations
Starter	RMB 29/month	USD 10/month	M2 T: 400 request / 5 hrs Speech 2 B: 1500 request / 5 hrs Speech 2 B: 4,000 character/day Image-01: 50 /day
Plus	RMB 49/month	USD 20/month	M2 T: 1500 request / 5 hrs Speech 2 B: 9,000 character/day Image-01: 100 /day
Plus-Highspeed	RMB 98/month	USD 40/month	M2 T: 4,500 request / 5 hrs Speech 2 B: 11,000 character/day Image-01: 200 /day
Max	RMB 179/month	USD 50/month	M2 T: 4,500 request / 5 hrs Speech 2 B: 11,000 character/day Image-01: 200 /day Halluc2 3-Fast 768P 6s: 2/day Halluc2 3 768P 6s: 7/day
Max-Highspeed	RMB 199/month	USD 80/month	M2 T: 4,500 request / 5 hrs Speech 2 B: 19,000 character/day Image-01: 200/day Halluc2 3-Fast 768P 6s: 3/day Halluc2 3 768P 6s: 3/day

Source: MiniMax, 2020, 2021, 2022

Exhibit 34 - Xiaomi MiMo token plan

Region	Program	Price	Content
China	Lite	RMB 39/month, RMB 34 32/month for the first purchase	60 million credits
	Standard	RMB 99/month, RMB 87.12/month for the first purchase	200 million credits
	Program	RMB 329/month, RMB 289.52/month for the first purchase	700 million credits
	Max	RMB 659/month, RMB 579.92/month, for the first purchase	
Overseas	Lite	USD 6/month, USD 5.28/month for the first purchase	60 million credits
	Standard	USD 16/month, USD 14.08/month for the first purchase	200 million credits
	Program	USD 50/month, USD 44/month for the first purchase	700 million credits
	Max	USD 100/month, USD 88/month for the first purchase	1600 million credits

Source: Xiaomi, 2020, 2021, 2022

Exhibit 35 - GLM coding plan

Region	Program	Pricing	Content
China	Lite	RMB 49/month	5-hour limit: up to approx. 80 prompts, weekly limit: up to approx. 400 prompts
		RMB 132.3/quarter	
		RMB 470.4/year	
	Pro	RMB 149/month	5-hour limit: up to approx. 400 prompts, weekly limit: up to approx. 2000 prompts
		RMB 402.3/quarter	
		RMB 1430.4/year	
Overseas	Max	RMB 469/month	5-hour limit: up to approx. 1600 prompts, weekly limit: up to approx. 8000 prompts
		RMB 1266.3/quarter	
		RMB 4502.4/year	
	Lite	USD 18/month	5-hour limit: up to approx. 80 prompts, weekly limit: up to approx. 400 prompts
		USD 48.6/quarter	
		USD 172.8/year	
	Pro	USD 72/month	5-hour limit: up to approx. 400 prompts, weekly limit: up to approx. 2000 prompts
		USD 194.4/quarter	
		USD 691.2/year	
	Max	USD 160/month	5-hour limit: up to approx. 1600 prompts, weekly limit: up to approx. 8000 prompts
		USD 432/quarter	
		USD 1536/year	

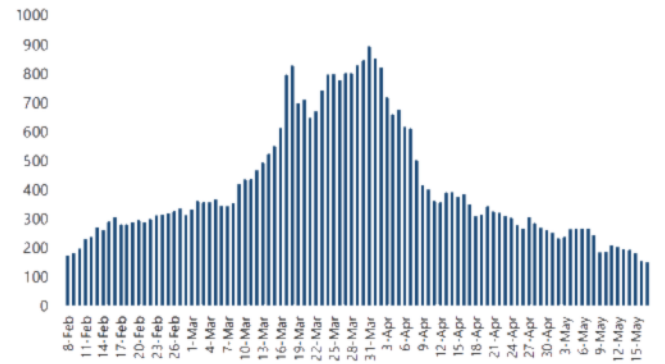
Notation: China, Overseas

OpenClaw

According to OpenRouter, top 3 models for OpenClaw based on token consumption come from China, namely, GLM 5 Turbo, MiniMax M2.7, and MiniMax M2.5. Other models include DeepSeek V3.2 and Qwen3.6 Plus.

According to OpenRouter, GLM 5 Turbo, MiniMax M2.7, and MiniMax M2.5 are ranked top 3 in terms of token consumption on OpenClaw

Exhibit 36 - Token consumption for OpenClaw (bn)



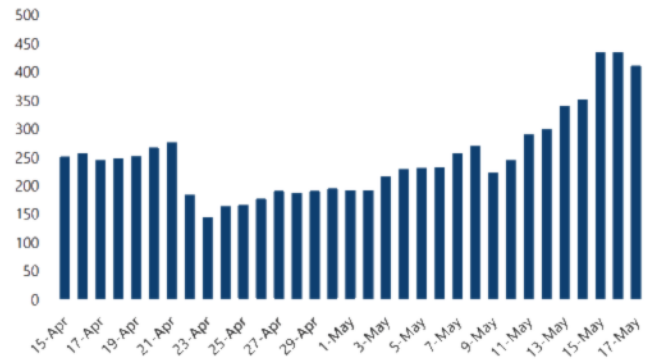
Source: OpenRouter, OpenClaw

Exhibit 38 - Top 15 models used by OpenClaw over the last 30 days (as of 18 May)

No	Model	Tokens
1	Claude Sonnet 4.6	618B
2	Nemotron 3 Super	564B
3	GLM 5 Turbo	524B
4	MiniMax M2.7	514B
5	Gemini 2.5 Flash Lite	453B
6	Claude Opus 4.7	374B
7	Claude Opus 4.6	321B
8	MiniMax M2.5	304B
9	Gemini 3 Flash Preview	213B
10	DeepSeek V4 Flash	209B
11	gpt-oss-120b	169B
12	DeepSeek V3.2	168B
13	Hy3 preview	148B
14	Kimi K2.6	144B
15	Gemini 2.5 Flash	134B

Source: OpenRouter, OpenClaw

Exhibit 37 - Token consumption for Hermes Agent (bn)



Source: OpenRouter, Hermes Agent

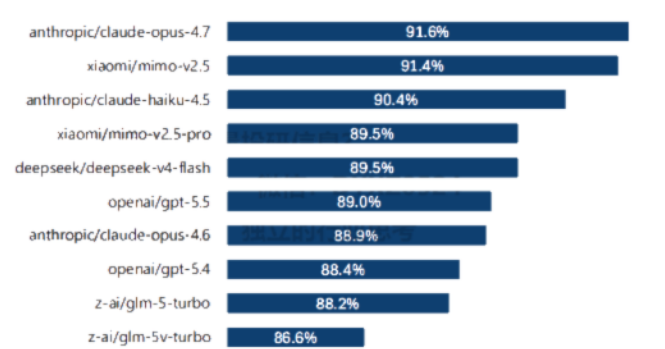
Exhibit 39 - Top 15 models used by Hermes Agent over the last 30 days (as of 18 May)

No	Model	Tokens
1	DeepSeek V4 Flash	917B
2	MiMo-V2-Pro	662B
3	MiniMax M2.7	581B
4	Owl Alpha	563B
5	Step 3.5 Flash	473B
6	Nemotron 3 Super	445B
7	DeepSeek V4 Pro	366B
8	Qwen3.6 Plus	352B
9	Hy3 preview	324B
10	Kimi K2.6	310B
11	Claude Opus 4.7	216B
12	Claude Sonnet 4.6	194B
13	Gemini 3 Flash Preview	153B
14	GLM 5.1	137B
15	GPT-5.5	134B

Source: OpenRouter, Hermes Agent

According to PinchBench, Anthropic’s Claude Opus 4.7 ranked first in terms of success rate (91.6%), followed by Xiaomi MiMo V2.5 Pro (91.4%) and Anthropic Claude Haiku 4.5 (90.4%).

Exhibit 40 - Global: Top 10 models by success rate as of 18 May



Video generation models

According to Artificial Analysis (as of 11 May), HappyHorse-1.0 ranked top 5 in terms of text to video and image to video. Others include Seedance 2.0, Veo 3.1, Kling 3.0 Omni 1080p, Kling 3.0 1080p, Veo 3.1 Fast, and grok-imagine-video.

Exhibit 46 - Top 15 image to video leaderboard (with audio)

No	Model	ELO
1	Dreamina Seedance 2.0 720p	1,179
2	HappyHorse-1.0	1,163
3	grok-imagine-video	1,087
4	Veo 3.1 Fast	1,086
5	Veo 3.1	1,083
6	Kling 3.0 1080p(Pro)	1,078
7	PixVerse V6	1,076
8	Kling 3.0 720p(Standard)	1,069
9	Kling 3.0 Omni1080p (Pro)	1,062
10	Veo 3.1 Lite	1,057
11	Vidu Q3 Pro	1,051
12	Kling 3.0 Omni 720p (Standard)	1,048
13	Kling 2.6 Pro(January)	1,003
14	Seedance 1.5 pro	1,000
15	PixVerse V5.6	958

Source: Artificial Analysis, 11 May 2025

Exhibit 47 - Top 15 text to video leaderboard (with audio)

No	Model	ELO
1	Dreamina Seedance 2.0 720p	1,222
2		1,214
3	Kling 3.0 Omni1080p (Pro)	1,105
4	Kling 3.0 1080p (Pro)	1,103
5	Veo 3.1	1,102
6	Veo 3.1 Fast	1,099
7	Kling 3.0 720p (Standard)	1,090
8	Sora 2 (December)	1,088
9	Vidu Q3 Pro	1,086
10	Kling 3.0 Omni 720p (Standard)	1,086
11	Veo 3.1 Lite	1,083
12	grok-imagine-video	1,083
13	Sora 2 Pro	1,082
14	PixVerse V6	1,071
15	Wan 2.6	1,033

Source: Artificial Analysis, 11 May 2025

Global and China AI Mobile App MAU Ranking

Exhibit 48 - Global AI mobile app MAU ranking (Apr-26)

Rank	App	MAU (m)	MoM
1	ChatGPT	957.8	-0.36%
2	Doubao	336.04	1.51%
3	Qwen	220.35	
4	Quark	162.3	-5.77%
5	Gemini	139.37	9.29%
6	DeepSeek	138.98	1.79%
7	Yuanbao	111.13	2.96%
8	Baidu Drive	103.8	-5.24%
9	Jimeng AI	92.49	15.70%
10	Grok	75.97	-5.88%
11	Dola	70.78	-3.18%
12	Nova	69.63	-1.20%
13	Meta view	60.12	31.34%
14	AI Mirror	34.91	0.37%
15	Perplexity	31.54	-5.95%

Source: AICPB, Statista

Exhibit 49 - China AI Mobile App MAU Ranking (Apr-26)

Rank	App	MAU (m)	MoM
1	Doubao	336.04	1.51%
2	Qwen	220.35	-1.02%
3	Quark	162.3	-5.77%
4	DeepSeek	138.98	1.79%
5	Yuanbao	111.13	2.96%
6	Baidu Drive	103.8	-5.24%
7	Jimeng AI	92.49	15.70%
8	Kimi AI	25.33	-1.69%
9	Nano AI Search	8.91	-4.61%
10	Baidu Wenku	8.51	-6.52%
11	Kling	7.85	6.78%
12	ERNIE Bot	7.7	-6.51%
13	Zhipu Qingyan	5.29	-4.37%
14	Lovekey Keyboard	4.57	-5.49%
15	Cat Box	4.05	-3.38%

According to aicpb.com, MAU for China AI apps, which include Doubao, Qwen, Quark, and Yuanbao, experienced strong MoM growth vs ChatGPT due to Chinese New Year in Feb

In domestic market, MAU growth for Kling and Jimeng AI experienced fast MoM growth in March

Mar 26: Monthly App Tracker from QuestMobile Analysis

Online shopping

- MoM trend: DAU for Taobao/PDD decreased by 1%/2% respectively and VIPS increased by 3% in Apr. DAU for JD increased by 2% MoM in Apr.
- YoY trend: DAU for JD increased 9% YoY and Taobao/PDD increased by 6%/10% YoY respectively.

PDD maintained fast DAU growth YoY basis in Apr

Exhibit 50 - MAU for Taobao / PDD / JD in Apr-26

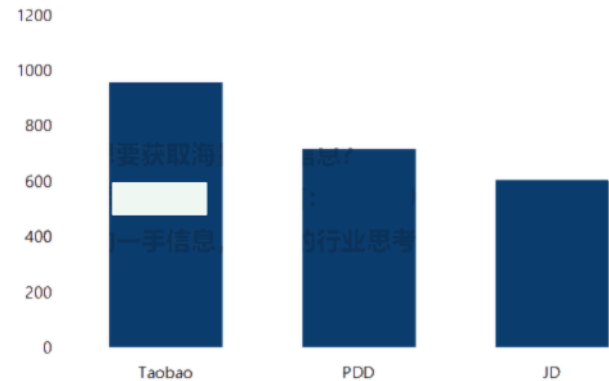
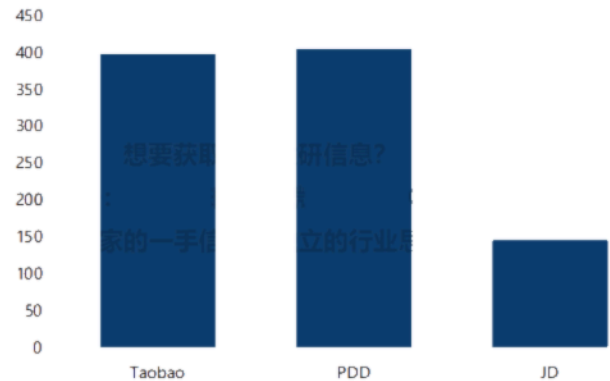


Exhibit 51 - DAU for Taobao / PDD / JD in Apr-26

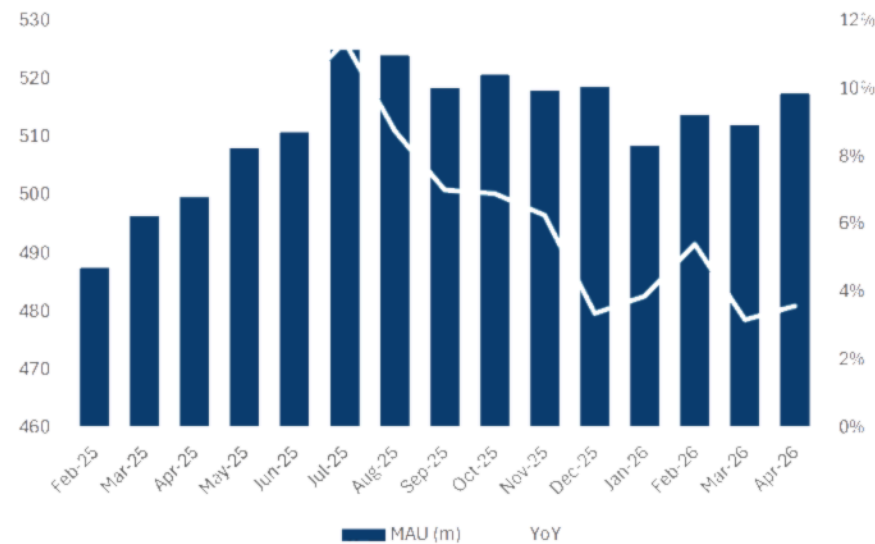


Food delivery

- MoM trend: Meituan MAU increased by 1% MoM during the month.
- YoY trend: Meituan MAU increased by 4% (vs 3% YoY in Mar) in Apr.

Meituan MAU was largely stable MoM in Apr

Exhibit 52 - MAU for Meituan app in past 12 months



Source: QuestMobile, QuestMobile

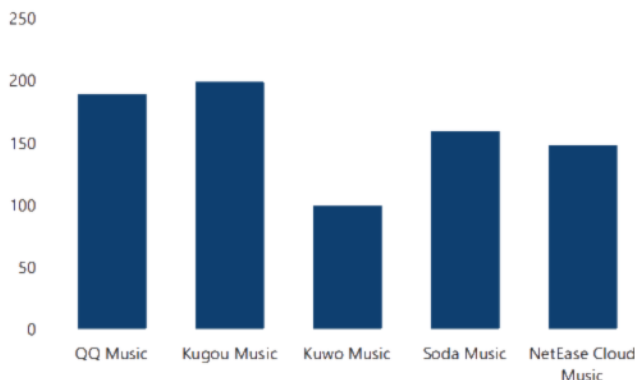
Online Music

- MoM trend: MAU for Soda Music increased by 2% MoM to 159.9m in Apr. For TME, MAU for QQ Music/Kugou/Kuwo declined by 6%/3%/3% MoM respectively in Apr. MAU for NetEase Cloud Music increased 1% MoM during the month.

Soda Music maintains fast MAU YoY growth while largely stable MoM in Apr

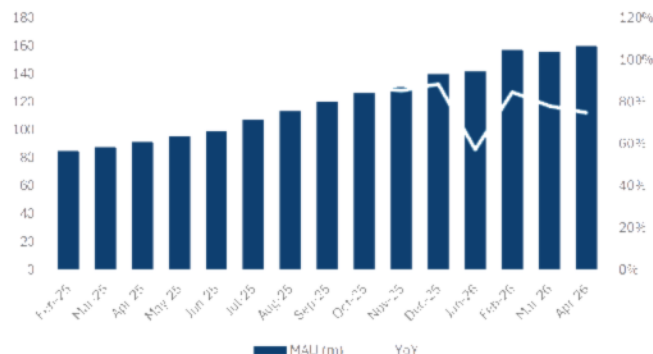
- YoY trend: MAU for Soda Music increased by 75% YoY while NetEase Cloud Music MAU increased by 2% YoY in Apr. For TME, MAU for QQ Music increased by 1% while that of Kugou/ Kuwo declined by 8%/16% YoY respectively.

Exhibit 53 - MAU for TME, Soda Music and NetEase Cloud Music in Apr-26



Source: Company announcements

Exhibit 54 - MAU for Soda Music in past 12 months

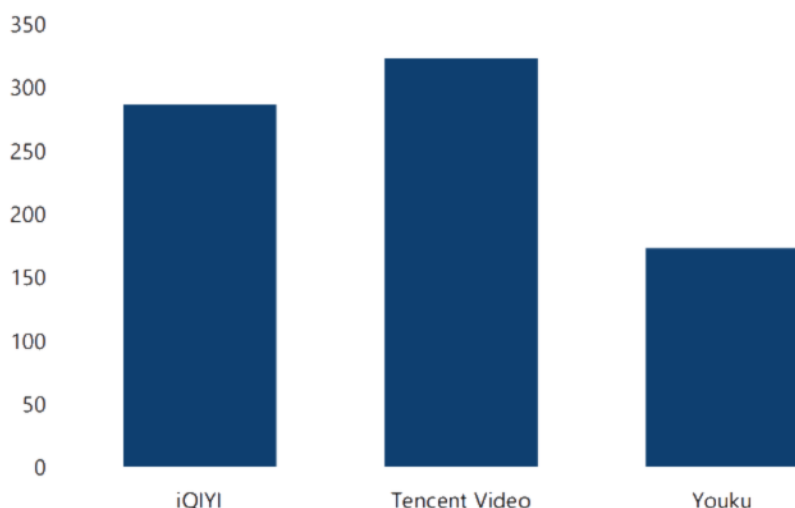


Source: Company announcements

Online video

- MoM trend: MAU for Tencent Video/iQ declined by 3%/7% and Youku increased by 8% MoM. Mango TV maintained stable MoM growth in MAU.
- YoY trend: MAU for leading long form video players declined YoY during the month.

Exhibit 55 - MAU for iQ, Tencent Video and Youku in Apr-26



Source: Company announcements

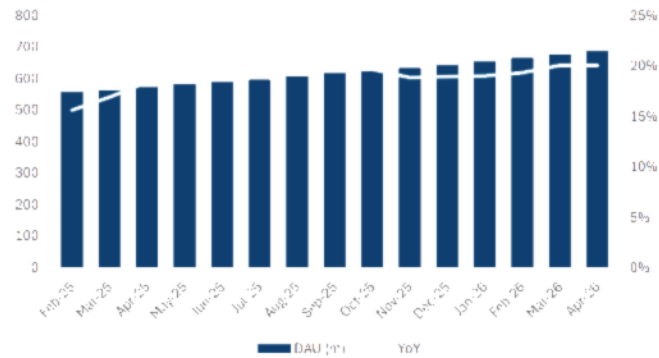
Live-streaming and short video

- MoM trend: DAU for Douyin main app increased 2% MoM in Apr. For game broadcasting, MAU for Huya increased by 3% and Douyu decreased by 8% MoM in Apr. For BILI, MAU maintained stable MoM during the month.
- YoY trend: DAU for Douyin main app maintains solid growth at 20% YoY. For game broadcasting, both Huya and Douyu experienced 4% and 15% YoY decline in MAU respectively. For BILI, MAU increased by about 4% YoY during the month.

MAU for leading long form video platform declined YoY in Apr

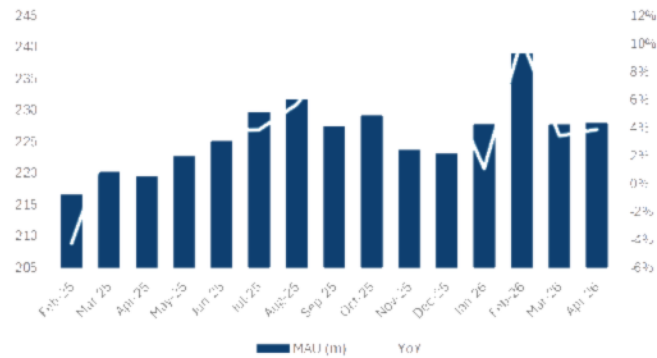
Douyin DAU maintained solid YoY growth in Apr

Exhibit 56 - DAU for Douyin over the past 12 months



Sources: Statista, Company Reports

Exhibit 57 - MAU for BILI over the past 12 months



Sources: Statista, Company Reports

Online travel

- MoM trend: MAU for TCOM / Qunar increased by 7%/8% MoM respectively in Apr.
- YoY trend: MAU for TCOM/Qunar declined by 5%/10% YoY in Apr.

Exhibit 58 - TCOM's MAU over the past 12 months



Sources: Statista, Company Reports

News app

- MoM: MAU for Weibo declined by 1% while MAU for Toutiao decreased by 3% in Apr.
- YoY: MAU for Weibo decreased by 2% YoY while Toutiao decreased by 3% in Apr.

TCOM MAU in domestic market increased MoM in Apr after CNY in Feb

MAU for Weibo experienced negative YoY growth in Apr

Exhibit 59 - Weibo MAU over the past 12 months



Source: Statista, Weibo App

Baidu app

- MoM: MAU remained stable and DAU increased by 2% in Apr.
- YoY: MAU declined by 7% YoY and DAU declined by 20% YoY in Apr.

Baidu app MAU experienced stable MoM and declined YoY in Apr

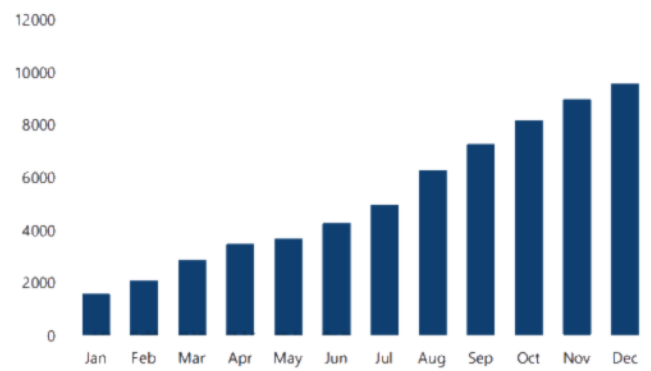
Exhibit 60 - Baidu MAU over the past 12 months



Source: Statista, Baidu App

IDC estimates on tokens consumptions on public cloud and outlook

Exhibit 61 - China's MaaS market monthly average daily token consumption in 2025



Source: IDC, © IDC

Exhibit 62 - China's MaaS market share by token consumption in 2025

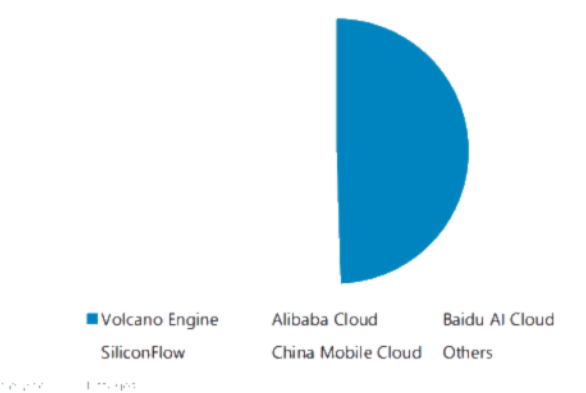
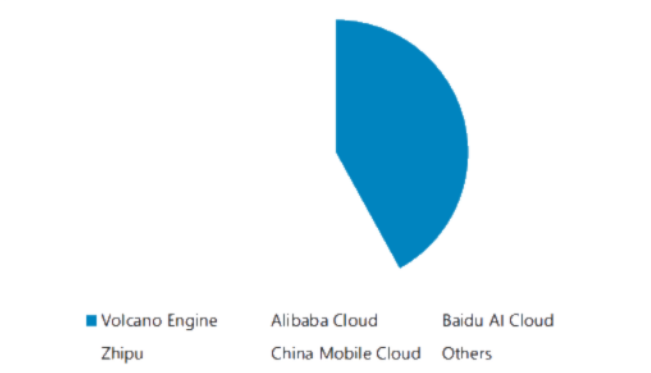
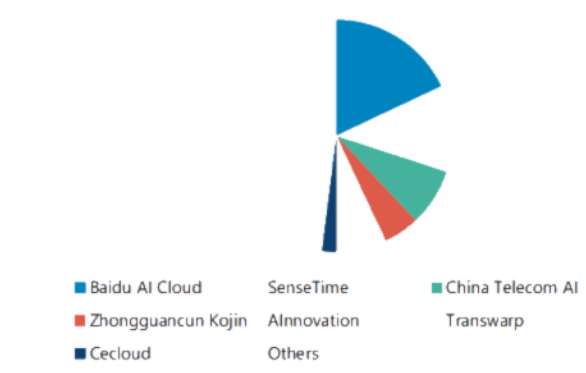


Exhibit 63 - Market share of major vendors in China's MaaS market in 2025 (revenue basis)



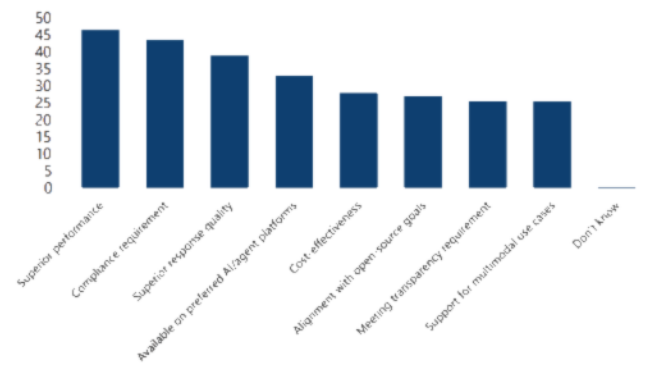
Source: IDC, © IDC

Exhibit 64 - Market share of major vendors in China's private large model platform market in 2025 (revenue basis)



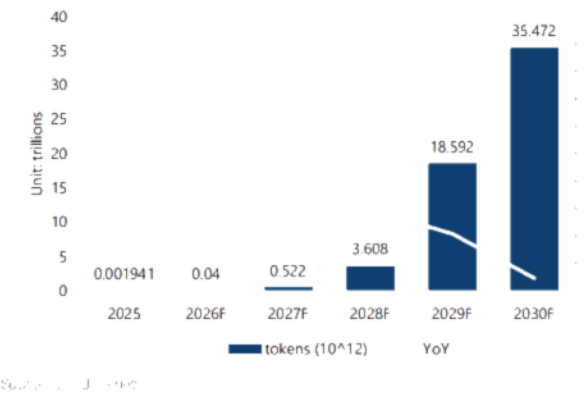
Source: IDC, © IDC

Exhibit 65 - The most important factors that drive people's decision to deploy model widely



Source: IDC, © IDC

Exhibit 66 - China's MaaS market scale and forcaset 2024-2030 (token basis)



Source: IDC, © IDC

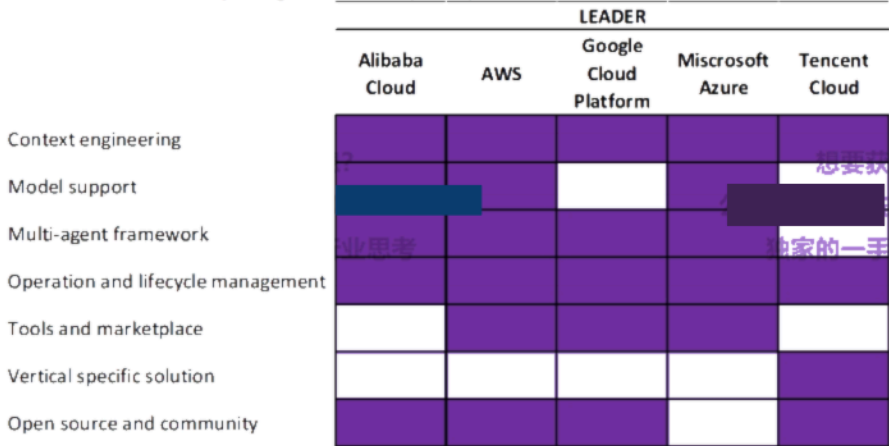
Appendix

Omdia heatmap for agentic AI development platform in Asia and Oceania

According to Omdia, agentic AI software market is expected to increase at 105% CAGR from USD271m in 2025 to USD9.7bn by 2030. Alibaba Cloud captures the highest rankings in 5 areas: (1) context engineering, (2) model support, (3) multi-agent framework, (4) operation and lifecycle management, and (5) open source and community.

According to Omdia, Alibaba Cloud is ranked as one of the leaders and top choice for building and deploying AI agents. It captures the highest ranking for 5 out of 7 metrics

Exhibit 67 - Omdia heatmap for agentic AI development platform in Asia and Oceania



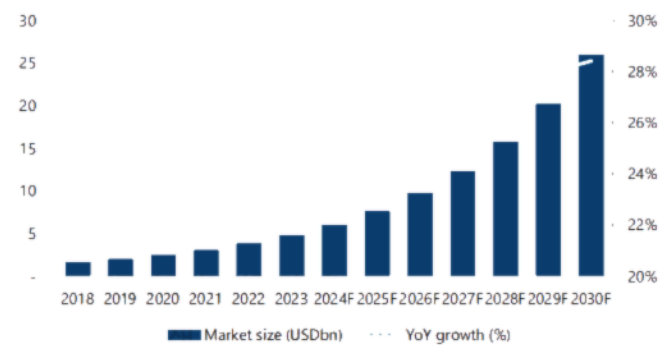
Source: Omdia, 2024

Market size: Coding, Workspace, and Content Creation

Global AI code tools market is expected to be 3x the scale in 2030 vs 2025

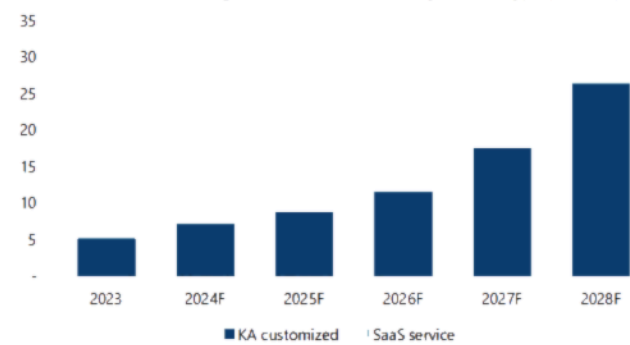
According to Grand View Research, the global AI code tools market is expected to increase by 3x to USD26bn from 2025 to 2030. On the other hand, F&S predicts that the China AI code generation market will reach about USD4.7bn in 2028. KA customized solutions are expected to be a key revenue contributor.

Exhibit 68 - Global AI code tools market size (USDbn)



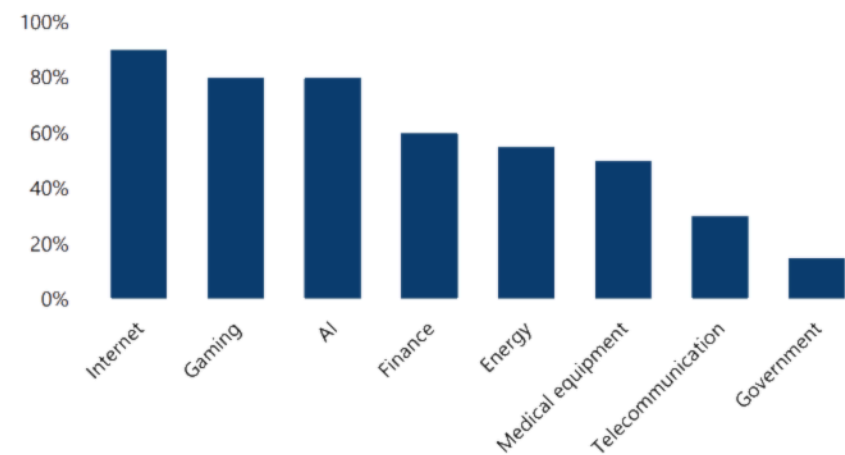
Source: Grand View Research, 2024

Exhibit 69 - China AI code generation market size by service type (RMBbn)



Source: Frost & Sullivan, 2024

Exhibit 70 - China AI code penetration in different industries

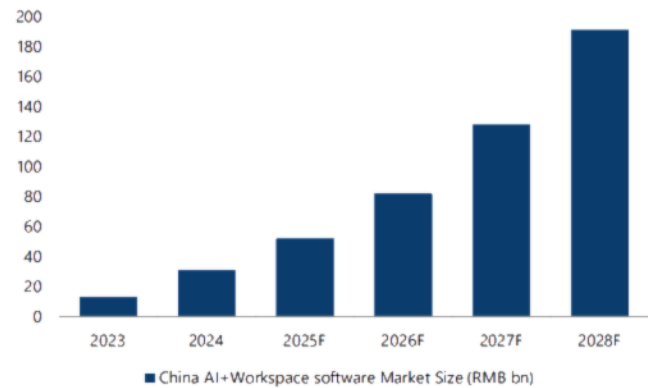


In terms of AI code penetration, internet, gaming, AI, finance, and energy are top 5 key sectors

Source: Gartner, Statista, etc.

According to Gonyn Industry Research (2024), market size of AI+Workspace software is expected to reach RMB191bn in 2028, 6x the scale vs 2024.

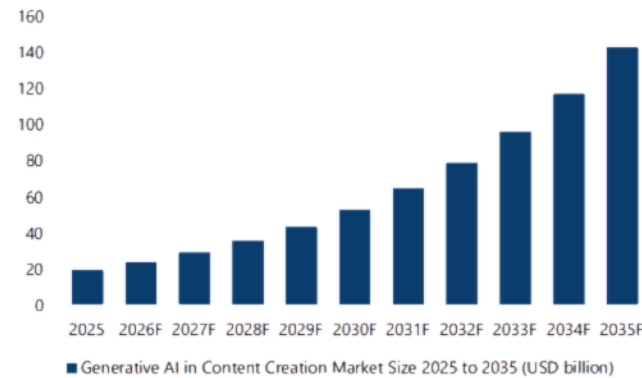
Exhibit 71 - China AI + Workspace Market Size



Source: Gartner, Statista, etc.

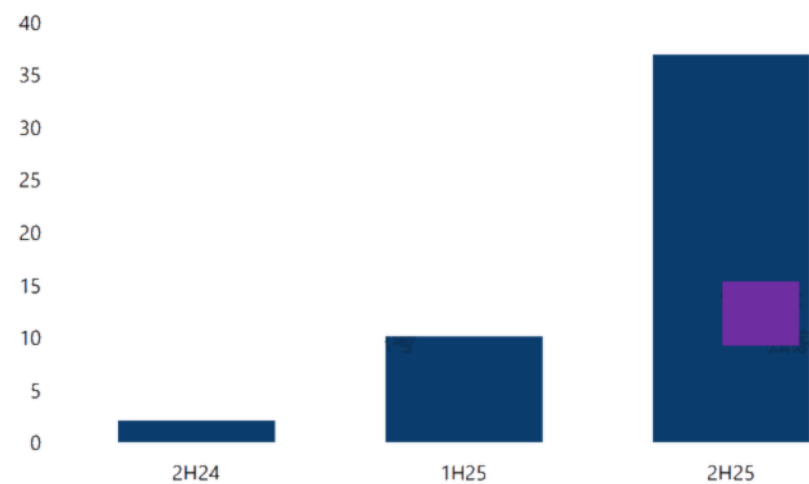
According to Precedence Research, global generative AI content creation market is expected to increase by 7x surpassing USD140bn in 2035 vs 2025.

Exhibit 72 - Generative AI in Content Creation Market Size 2025 to 2035 (USD bn)



Source: Precedence Research, Statista, etc.

Exhibit 73 - Daily token consumption in China by enterprises in 2H24 vs 1H25 (tn)



According to F&S, enterprise daily token consumption increased by 263% HoH to 37tn, among which market share for Alibaba Cloud increased the most and captures no 1 position reaching 32% in 2H25 (1H25:17.7%), followed by 21.3% for Doubao (1H25: 14.1%) and 18.4% for DeepSeek (1H25: 10.3%).

Alibaba's Qwen ranked first in terms of token consumption among enterprises in 2H25

Exhibit 74 - Market share in terms of tokens used among enterprises in 1H25



Source: F&S, 1H25, 1H25, 1H25, 1H25

Exhibit 75 - Market share in terms of tokens used among enterprises in 2H25



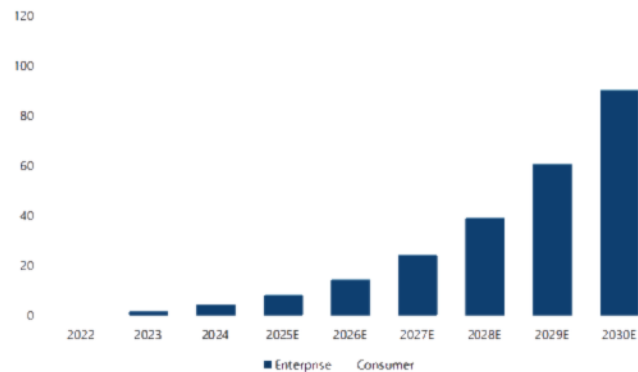
Source: F&S, 2H25, 2H25, 2H25, 2H25

China Large Language Models

According to F&S / CAICT, China LLM market is expected to increase by about 64% CAGR between 2024 and 2030 and surpass RMB100bn in 2030. By segment, consumer revenue is expected to grow at a slower rate vs enterprise as willingness to pay remains low.

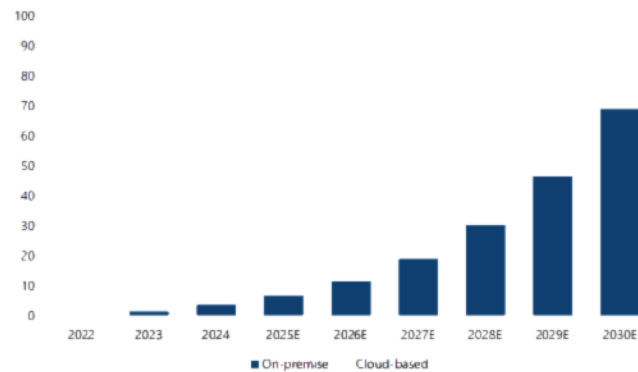
China large language models are expected to increase by 64% CAGR between 2024 to 2030. By segment, on-premise is key revenue contributor

Exhibit 76 - Market size of China LLM (RMB billions)



Source: F&S / CAICT, Bloomberg, 30 Apr.

Exhibit 77 - Market size of China Enterprise LLM (RMB billions)

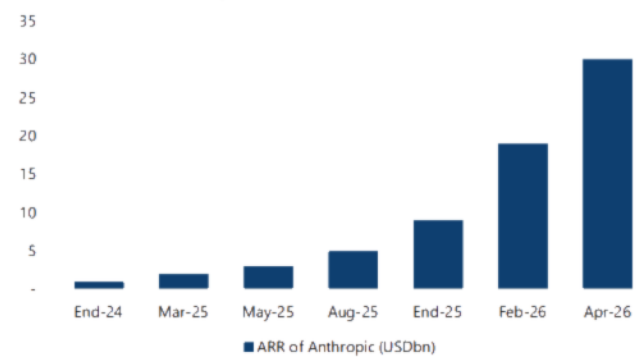


Source: F&S / CAICT, Bloomberg, 30 Apr.

Anthropic and OpenAI

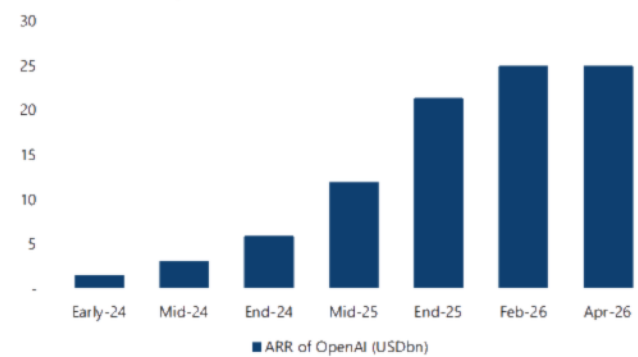
- Latest news flow on Anthropic is about USD900bn valuation (Bloomberg, 30 Apr), vs USD852bn valuation for OpenAI
- According to The Information, OpenAI revenue is expected to reach USD200bn in 2030, over 6x the scale in 2026. Rising Agent revenue contribution is expected over the next few years.

Exhibit 78 - ARR of Anthropic



Source: The information, Bloomberg, 30 Apr.

Exhibit 79 - ARR of OpenAI



Source: The information, Bloomberg, 30 Apr.

Exhibit 80 - OpenAI revenue between 2023 and 2030F (as of 3Q25)

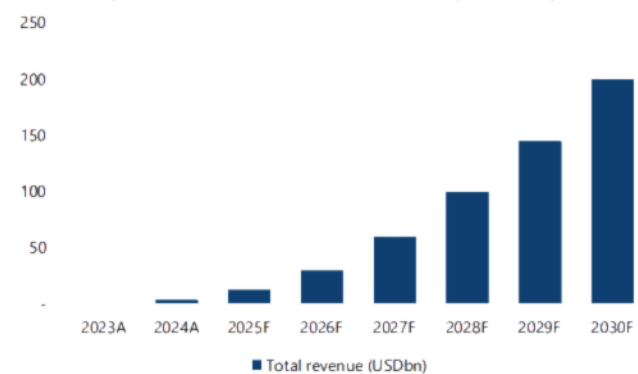


Exhibit 81 - Revenue mix of OpenAI between 2024 and 2030F (as of 3Q25)

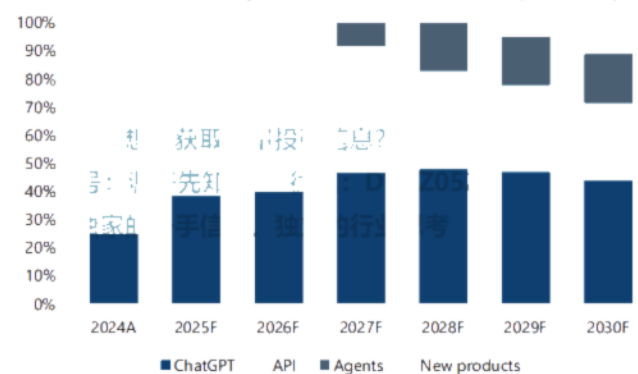
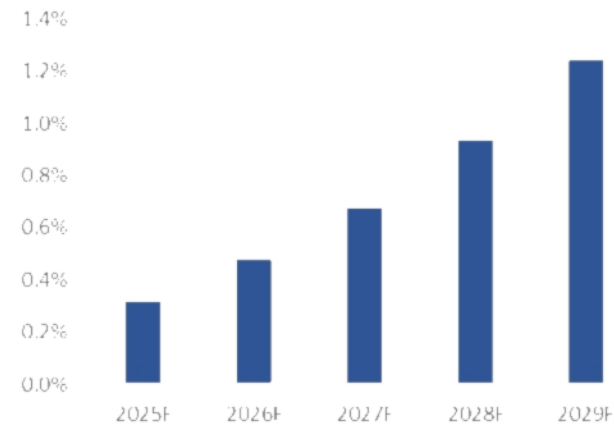


Exhibit 82 - MiniMax: % of OpenAI revenue over the next few years



SOURCE: OPENAI, ANALYST ESTIMATES

Models of Leading Players in China

Exhibit 83 - Alibaba model list

Release Date	Model	Core Functions
Jan-24	Qwen VL Plus / Max	Enhanced visual reasoning, high-resolution OCR, and detailed image understanding via API
Feb-24	Qwen1.5	First major open-source iteration with improved multilingual support and 0.5B to 72B sizes.
Jun-24	Qwen2	Optimized for coding, mathematics, and long context processing (up to 128K tokens).
Aug-24	Qwen2-AudioIn	Native audio-to-text and audio-to-audio dialogue capabilities
Aug-24	Qwen2-VL	Introduced dynamic resolution for visual inputs and long video understanding
Sep-24	Qwen2.5	Significant upgrades in instruction following, structured data (tables) processing, and coding
Nov-24	QwQ-32B-Preview	Reasoning-focused model using Reinforcement Learning for complex logic and chain-of-thought
Jan-25	Qwen2.5-Max	Large-scale MoE (Mixture of Experts) flagship matches or exceeds GPT-4o benchmarks
Jan-25	Qwen2.5-VL	"Agentic" vision model capable of identifying and interacting with mobile/PC UI screens.
Mar-25	Qwen2.5-Omnis	End-to-end full-modality model supporting seamless real-time text, audio, and visual interaction
Apr-25	Qwen3 (Series)	New reasoning architecture with hybrid inference modes to reduce hallucinations in complex tasks
Sep-25	Qwen3-Max-Preview	Trillion parameter preview version optimized for massive enterprise scale applications.
Dec-25	Wan2.6-Image	Supports image editing and mixed image-text output
Jan-26	Qwen3-Max	Frontier reasoning, 11+ params, agentic workflow mastery.
Feb-26	Qwen3.5	Latest flagship series featuring a 997B MoE model with superior cross-lingual reasoning
Feb-26	Wan2.6-22w-flash	Generates multi-shot videos with automatic dubbing, using a character's appearance from a reference video or image
Feb-26	Qwen3.5-Plus	Outstanding performance comparable to the very latest state-of-the-art models with significant leaps in both pure-text and multimodal capabilities compared to the 3 series
Feb-26	Qwen3-Coder-Next	Multi-turn tool interactions, future-ready development support
Mar-26	Qwen3.5-Omnis	Support for 256K long-text input and can process more than 10 hours of audio input and over 400 seconds of 720P audio-visual input at 1 FPS.
Apr-26	Wan2.7-Image	The Wan 2.7 image generation and editing model supports text-to-image, text-to-image-set, image-to-image-set, image editing, multi-image reference generation, and interactive editing, with improved performance in text rendering, subject consistency, and complex instruction following. The Pro series supports 4K output, the accelerated version balances quality and response speed.
Apr-26	Qwen3.5-Plus	Native multimodal, 1M context, agentic coding
Apr-26	Harpy Oyster	AI model that creates interactive 3D scenes with realistic physics for game, film, and TV
Apr-26	Qwen3.6-Max-Preview	It brings stronger world knowledge and instruction following, along with significant agentic coding improvements across a wide range of benchmarks
Apr-26	Qwen-Image2.0	Delivers a significant leap in visual quality — especially texture detail, lighting, and materials. It supports multilingual in-image text generation and delivers more balanced artistic style performance
Apr-26	HappyHorse-1.0 (Test)	It supports all four video generation modalities: Text to Video and Image to Video, each with and without native audio

Source: Alibaba Group

Exhibit 84 - Tencent model list

Release Date	Model	Core Functions
May-24	Hunyuan-DiT	The industry's first Chinese-native Diffusion Transformer (DiT) for high-quality text-to-image generation.
Sep-24	Hunyuan-Turbo	An upgraded LLM based on the Mixture-of-Experts (MoE) architecture, doubling training efficiency and reducing inference costs
Nov-24	Hunyuan3D-1.0	First open-source 3D generation model supporting text-to-3D and image-to-3D workflows
Dec-24	HunyuanVideo	A large-scale video generation framework capable of producing high-definition videos with complex motion.
Jan-25	Hunyuan3D-2.0	Significant upgrades in geometry accuracy and texture representation for 3D asset creation.
Feb-25	Hunyuan-Turbo-S	A "fast-thinking" model using Hybrid-Mamba-Transformer architecture, optimized for near-instant responses and competitive reasoning
Mar-25	HunyuanVideo-12V	A specialized image-to-video model built on the foundation of HunyuanVideo
Jun-25	Hunyuan3D-2.1	Production-ready 3D asset generation with a new PBR (Physically Based Rendering) material model
Sep-25	HunyuanImage3.0	Flux-to-image generation model (80B parameters) topping global rankings for semantic understanding
Nov-25	HunyuanVideo-1.5	A highly efficient and powerful update to the foundation video generation model
Dec-25	Hunyuan-2.0	Next-generation LLM featuring 406B parameters and a 256K context window for complex reasoning tasks.
Jan-26	HunyuanImage3D-Instruct	An image-to-image editing model supporting precise single-sentence editing and multi-image fusion
Mar-26	HY-WU	A scalable framework for on-the-fly conditional generation of low-rank(LoRA) updates.
Apr-26	HY-World 2.0	A multi-modal world model framework for world generation and world reconstruction. It accepts diverse input modalities — text, single-view images, multi-view images, and videos — and produces 3D world representations
Apr-26	Hy3 preview	A 295B-parameter Mixture-of-Experts (MoE) model with 21B active parameters and 3.8B MTP layer parameters. It improves significantly on complex reasoning, instruction following, context learning, coding, and agent tasks.

Source: Tencent Group

Exhibit 85 - Baidu model list

Release Date	Model	Core Functions
Mar-24	ERNIE Speed / Lite / Tiny	Lightweight models optimized for low latency and high cost-efficiency
Jun-24	ERNIE 4.0 Turbo	Efficiency upgrade for the 4.0 architecture with faster response speeds
Mar-25	ERNIE 4.5	Baidu's first native multimodal foundation model (text, image, audio, video)
Mar-25	ERNIE X1	Deep-thinking reasoning model designed for complex logic and multi-step tasks
Apr-25	ERNIE 4.5 Turbo / X1 Turbo	Performance-optimized version of X1 with significantly lower inference costs
Sep-25	ERNIE X1.1	Major upgrade in factuality, instruction following, and agentic capabilities
Nov-25	ERNIE 5.0 (Preview)	Preview of next-gen omni-modal model with advanced real-time reasoning
Jan-26	ERNIE 5.0 (Stable)	Current flagship 7.4T parameter model, top-ranked Chinese model on LMArena
Apr-26	ERNIE-Image	A text-to-image generation model built on a single-stream Diffusion Transformer (DiT) with 8B DiT parameters. Achieving leading performance among open weights models on several challenging benchmarks.
May-26	ERNIE 5.1	Significantly reduces pretraining cost while compressing total parameters to ~1/3 and activated parameters to ~1/2 — using only ~6% of the pretraining cost compared to models at similar scale, while achieving leading performance in its class.

Exhibit 86 - MiniMax model list

Release Date	Model	Core Function
Jan-25	MiniMax-Text-01/MiniMax-V1-01	New generation n° text and vision language model
Feb-25	T2V-01-Director/2V-01-Director	Director-level video generation model, better adherence to camera movement/description instructions
Apr-25	Image-01	Support generating images of various sizes from text descriptions
Jun-25	MiniMax-M1	80K mind chain and 7M rhout
Jun-25	MiniMax-Hailuo 02	Native 1080p, SOTA instruction following, extreme physics mastery
Jun-25	Musai-1.5 (Beta)	Support music generation by routing musical inspiration and lyrics
Aug-25	Speech-2.5	Significantly enhanced multilingual performance, more realistic and accurate voice cloning, over 40 languages
Sep-25	Music-1.5	Extending song generation to a full 4 minutes, one person is the band
Oct-25	MiniMax M2	Open-source MoE, strong code & agent
Oct-25	MiniMax-Hailuo-2.3	Complex video performance & merka agent
Oct-25	MiniMax-Speech-2.6	Real-time response, intelligent parsing, fluent LoRA voice
Oct-25	Music-2.0	A lead forward in model's understanding and expression of music
Dec-25	MiniMax M2.1	Exceptional multilingual capabilities to elevate code engineering
Jan-26	Music-2.5	Creation more precise and music more real
Jan-26	MiniMax M2-Hier	Multi-character roleplay, attractive long-horizon interaction
Feb-26	MiniMax M2.5	Industry-leading coding and reasoning capabilities at a fraction of the cost
Mar-26	Music-2.5+	Enhanced creation experience & efficiency, smarter melodies, purer sound quality
Mar-26	MiniMax M2.7	M2.7 is capable of building complex agent harnesses and completing highly elaborate productivity tasks, leveraging capabilities such as Agent Teams, complex Skills, and dynamic tool search.
Apr-26	Music-2.6	A text-to-audio music generation model, the output supports up to 256kbits bitrate and 44.1kHz sample rate

Source: MiniMax AI website

Exhibit 87 - Moonshot AI model list

Release Date	Model	Core Function
Jan-25	Kimi K1.5	An o1-level multi-modal model
Apr-25	Kimi-VL	Offers advanced multimodal reasoning, long-context understanding, and strong agent capabilities
Apr-25	Kimera-Prover Preview	Can reason in a human-like way and prove mathematical theorems in the Lean 4 language
Apr-25	Kimi-Audio	Open-source audio foundation model excelling in audio understanding, generation, and conversation
Jun-25	Kimi-Dev	A strong and open-source coding LLM for issue resolution
Jun-25	Kimi-Reasoner	End-to-end RL training for emerging agentic capabilities
Jul-25	Kimi K2	A state-of-the-art MoE language model, featuring 32 billion activated parameters and a total of 1 trillion parameters
Sep-25	Kimi-K2-Instruct-0905	The post-trained model best for drop-in, general-purpose chat and agentic experience
Nov-25	Kimi K2 Thinking	Exceed up to 200-300 sequential tool calls without human interference, reasoning coherently across hundreds of steps to solve complex problems
Jan-26	Kimi K2.5	A multi-modal model, state-of-the-art coding, vision capabilities, self-directed agent swarm paradigm
Apr-26	Kimi K2.6	Kimi's latest and most intelligent model, possessing stronger and more stable long-term code writing capabilities, significantly improved instruction compliance and self-correction capabilities, and the ability to handle more complex software engineering tasks

Source: Moonshot AI website

Exhibit 88 - Z.ai model list

Release Date	Model	Core Function
Jul-25	GLM-4.5-Air	Adopts a more streamlined design with 706B total parameters and 1.2B active parameters.
Jul-25	GLM-4.5	Total parameter count of 355B with 32B active parameters per forward pass
Aug-25	GLM-4.5V	Achieved SOTA performance among open-source VLMs of the same level with 106B parameters and 1.2B activation parameters
Sep-25	GLM-4.6	Advanced agentic, reasoning and coding capabilities
Dec-25	AutoG-M	Autonomous foundation agents for GUIs
Dec-25	GLM-4.6V	Open source multimodal models with native tool use
Dec-25	GLM-ASR-Nano	A robust, open-source speech recognition model with 1.5B parameters
Dec-25	GLM-TTS	Zero-shot TTS with multi-reward reinforcement learning
Dec-25	GLM-4.7	Complex system's engineering and long-horizon agentic tasks
Jan-26	GLM-4.7-Flash	Open source and free
Feb-26	GLM-OCR	SOTA performance, mastering complex document recognition
Feb-26	GLM-5	MoE flagship, complex engineering & long-horizon agent
Mar-26	GLM-5-Turbo	GLM-5-Turbo is a foundation model deeply optimized for the OpenCLAW scenario. It has been specifically optimized for the core requirements of OpenCLAW tasks since the training phase, enhancing key capabilities such as tool invocation, command following, timed and persistent tasks, and long-chain execution.
Apr-26	GLM-5V-Turbo	It strengthens its multimodal tool use and GUI task execution, allowing the model to better interpret interfaces, design mockups, documents, and charts, plan the next action, and complete end-to-end Agent workflows while maintaining strong pure-text coding and reasoning performance
Apr-26	GLM-5.1	Built for long-horizon tasks, it runs autonomously for 8+ hours, refining strategies through thousands of iterations, and it delivers significant leaps in coding and agentic performance with an performance boost. In high-performance database optimization, it reached 21.3k OPS over 600+ iterations and 6,000+ tool calls, which is 6x the performance of a standard 50-turn session

Source: Z.ai website

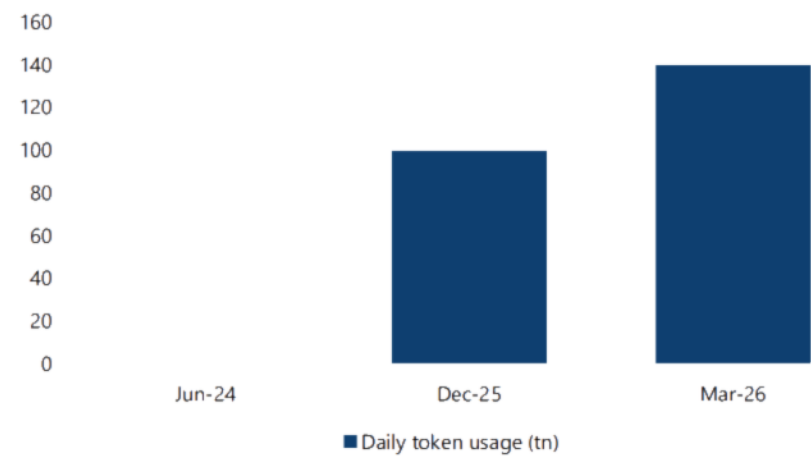
Exhibit 89 - Benchmark glossary

Category	Benchmark	Measure	Definition
Knowledge & Reasoning	MMLU-Pro	General knowledge	Measuring Massive Multi-Task Language Understanding Professional
	SimpleQA-Venheo	Fact accuracy	Simple Question Answering for factuality evaluation
	Chinese-SimpleQA	Chinese facts	Chinese-specific Simple Question Answering benchmark
	GRQA Diamond	Advanced science	Graduate-level Google-Proof Question Answering
	HLE	Hard reasoning	Humanity's Last Exam for expert-level reasoning
	LiveCodeBench	Live coding	Contamination-free Live Code Benchmarking
	Codeforces	Algorithm solving	Benchmarking Competition-level Code Generation of LLMs
	HMNT 2026 Feb	Math Competition	Feb 2026 Harvard-MIT Mathematics Tournament
	IMOAnswerBench	Olympiad Math	International Mathematical Olympiad Answer-based Benchmark
	Apex	Advanced logic	The AI Productivity Index, a benchmark for assessing whether frontier AI models can perform knowledge work with high economic value
Long Context	Apex Shortlist	Advanced logic	The MathArena Apex Shortlist
	MRCR 1M	Long retrieval	Multi-Round Conference Relevance Retrieval at 1 Million tokens
	CorpusQA 1M	Long document QA	A long-context question answering benchmark designed to evaluate models at approximately 1 million token contexts
	Terminal Bench 2.0	Terminal tasks	A command-line benchmark suite designed to evaluate AI agents on 89 real-world multi-step terminal tasks
Agentic	SWE Verified	Code repair	A human-filtered subset of 500 instances from SWE-bench
	SWE Pro	Code repair	A substantially more challenging benchmark that builds upon the best practices of SWE-Bench
	SWE Multilingual	Code repair	A benchmark designed to evaluate the software engineering capabilities of LLMs across multiple programming languages
	BrowseComp	Web browsing	A simple and challenging benchmark that measures the ability of AI agents to locate hard-to-find information
	HLE w/look	Tool use	Humanity's Last Exam with external tools
	GDPval-AA	Task Planning	A benchmark designed to evaluate AI models on real-world, economically valuable tasks across 44 occupations
	MCPAtlas Public	Multi-tool use	A large-scale benchmark for tool-use competency with real MCP servers
	Toolathlon	Tool Proficiency	A comprehensive benchmark for evaluating language agents on diverse, realistic, and long-horizon tasks

Sourced from OpenAI

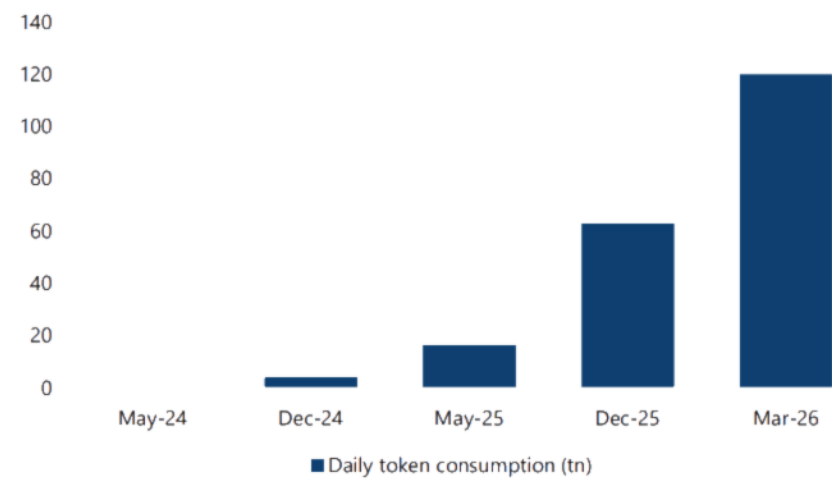
We see different third party analysis shows China has meaningful contribution to global token consumption. According to Hugging Face, China models accounted for 41% of global downloads in the past 12 months ended Feb-26.

Exhibit 90 - China Daily token usage



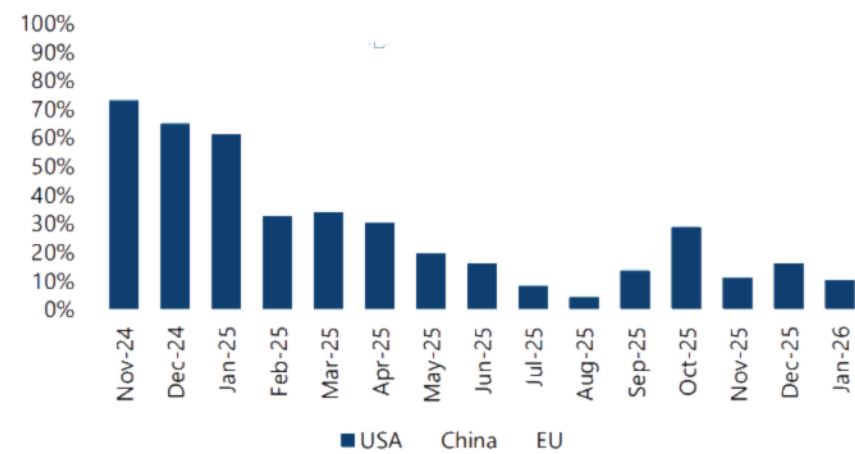
Sourced from Hugging Face and OpenAI

Exhibit 91 - Doubao daily token consumption



Source: ChatGPT, Bloomberg Intelligence

Exhibit 92 - Token share by region



Source: ChatGPT, Bloomberg Intelligence