China AI value chain

The expanding pie and race for value: Initiate broader coverage; BABA, Tencent and Z.AI are our preferred exposures

Internet

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

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

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

J.P. Morgan Securities (China) Company Limited

We initiate broader coverage of China AI with a constructive sector view. We expect exponential intelligence consumption growth to support growth across China's AI value chain, generating **US$544bn of cumulative operating profit over 2026E-30E**, with models + applications contributing **one-third of industry profits and nearly two-thirds of the increase in annual earnings between 2026E and 2030E** vs **c.14% of the China-AI-attributable listed market's estimated value today**. We believe the market is pricing the current revenue distribution, rather than anticipating the shift in revenue and earnings that our model implies. Our **Opportunity × Scarcity × Value Capture** framework identifies **Alibaba, Tencent and Z.AI (Zhipu)** as our preferred exposures to this shift, alongside nine initiations/ assumptions.

- **Expanding intelligence consumption supports substantial growth across every layer.** We expect China's broader adoption and agent workloads to **lift token demand to c.60x its 2026E level by 2030E**, supporting hardware investment, cloud utilization and paid model/application usage. Across these stages, we expect subsector revenue to grow **1.8x for hardware, 12.5x for cloud and 26.9x for models + applications from 2026E to 2030E**. As revenue scales and margins improve, we forecast a rise in China's AI industry operating profit from US$27bn in 2026E to US$243bn in 2030E.

- **We expect China's AI profit pools to shift structurally from infrastructure to downstream monetization.** Hardware remains the largest near-term beneficiary, but we expect its profit share in the value chain to fall from 93% in 2026E to 30% by 2030E. Cloud will become a more durable earnings layer, while models + applications move from losses to 54% of industry profit by 2030E. For investors, this implies a changing opportunity set: Hardware offers near-term earnings visibility, but the largest medium-term earnings and valuation upside should increasingly migrate toward cloud and model + application companies that can scale AI adoption into profitable growth.

- **Our framework favors companies with scarce assets and a clear route to monetization.** We prefer Alibaba for integrated chip, cloud and model capabilities; Tencent for scarce AI application assets that we believe are largely unpriced; and Zhipu for differentiated intelligence driving paid usage. Alongside these preferred exposures, we initiate coverage on SenseTime, Phancy and Xunce at Overweight, and on Meitu at Neutral. We assume coverage of Kingsoft at Overweight, on Kingdee, JST and Manycore at Neutral, and on Yonyou at Underweight. Key risks include slower-than-expected intelligence-consumption growth, an erosion of company-specific scarcity, and a weaker conversion of adoption into gross profit and cash flow.

See page 42 for analyst certification and important disclosures, including non-US analyst disclosures.

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

J.P.Morgan

Table of contents

Executive summary 5

Exponential intelligence consumption growth creates a larger market and growing earnings across the AI value chain 5

Downstream contributes one-third of five-year profits and nearly two-thirds of earnings growth, creating room for valuations to rise 7

Opportunity × Scarcity × Value Capture favors Alibaba, Tencent and Zhipu alongside nine initiations/ assumptions 9

Three preferred exposures and nine initiations/ assumptions apply these judgments 10

AI value chain: A growing end-demand pool with earnings shifting toward models + applications..... 13

Intelligence consumption: About 60x token growth over 2026E-30E, roughly two-thirds by enterprise 14

>US$1.4tn provides the end-demand anchor, with further upside from AI-driven value creation..... 15

Value is concentrated upstream today, and that matches where revenue currently sits 17

How we build the forecast: One dollar of investment, earned three times..... 18

AI value chain layer analysis: Where each layer earns, and what remains scarce 24

Investment framework: Opportunity × Scarcity × Value Capture 30

Opportunity: Start with the economic pool around the workload 30

Scarcity: Identify what remains difficult to replicate or substitute..... 31

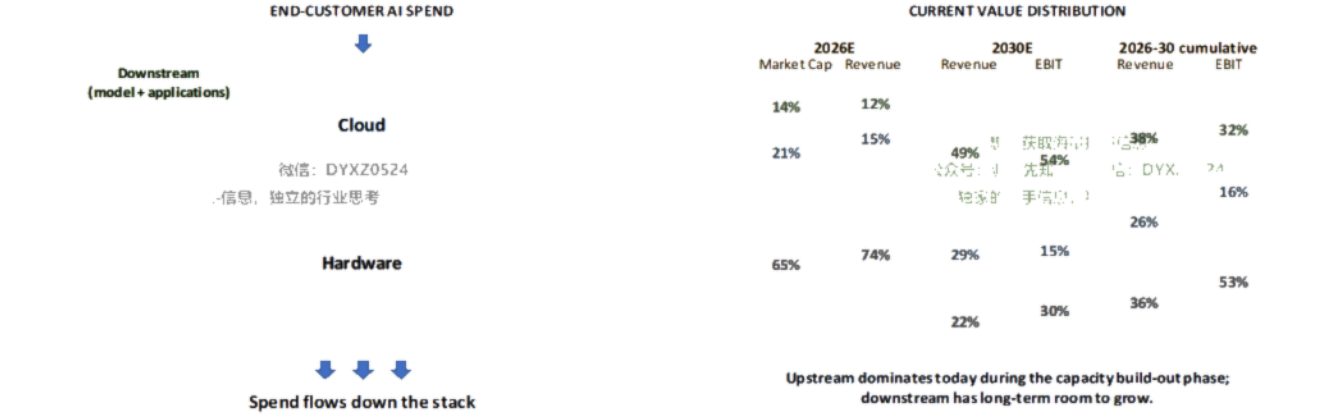
Value Capture: Scarcity has to translate into company economics..... 36

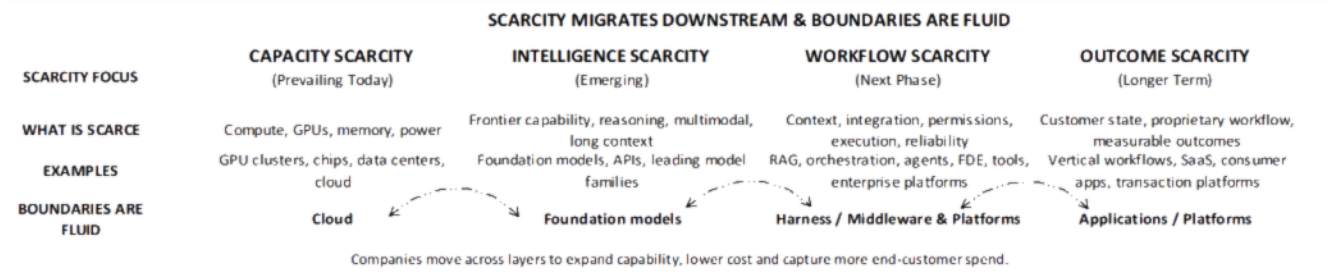
Valuation: Attractive economics still need to offer stock upside 38

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

J.P.Morgan

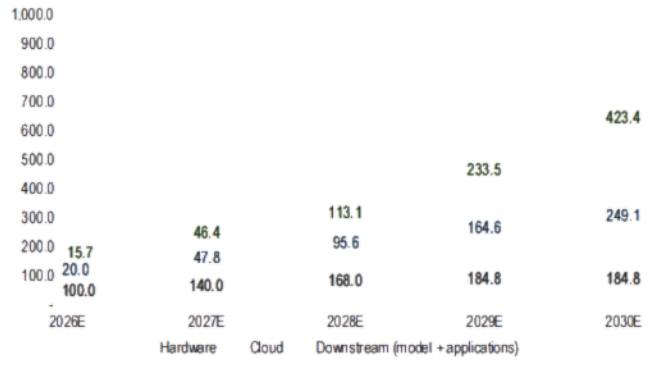
Figure 1: China AI value chain: Growing TAM, migrating scarcity, fluid boundaries





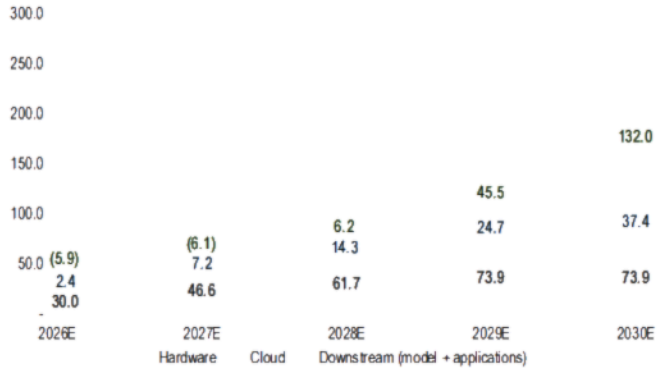
Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 2: 2026-30E China AI value chain revenue pool (US$ bn)



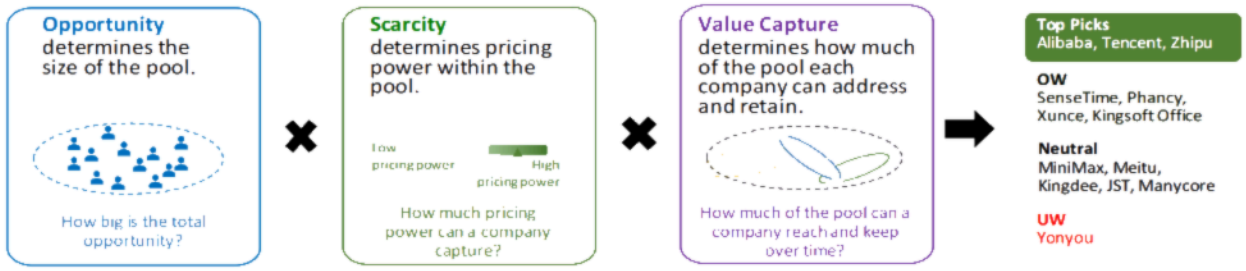
Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 3: 2026-30E China AI value chain operating profit pool (US$ bn)



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 4: Our Opportunity × Scarcity × Value Capture framework identifies Alibaba, Tencent and Z.AI (Zhipu) as our preferred exposures to this value chain shift



Source: J.P. Morgan estimates.

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

Table 1: Stock summary of China AI value chain (midstream cloud and downstream)

Company	Ticker	Price	Rating	Price Target	Pot. up/ downside	Position	Framework mapping	Investment case	Financial conversion & valuation
Alibaba	BABA / 9988.HK	**$116/ HK$115**	OW	$210 / HK$205	**81% / 79%**	Hardware / cloud / model + application – integrated	Cross-layer scarcity + capacity returns	T-Head, cloud and the Qwen franchise give integrated participation across three layers	FY27E-29E total revenue 23% CAGR and cloud revenue 62% CAGR; adjusted NP to Rmb193bn by FY2028E; PT at 16x FY28E P/E
Tencent	700.HK	**HK$452**	OW	HK$690	**53%**	Cloud / model + application	Unpriced application scarcity + execution capture	Distribution, proprietary context and transaction execution provide direct access to existing consumer and commercial activity; such AI application assets are largely unpriced now	2026E-28E revenue 12% CAGR; adjusted NP to Rmb279bn by 2028E; PT at 20x 2027E P/E
Zhipu	2513.HK	**HK$742**	OW	HK$2,000	**170%**	Foundation model	Intelligence scarcity + downstream capture	Competitive model capability and commercial traction support intelligence scarcity; proprietary harnesses can extend the charging point	2026E-28E revenue 173% CAGR; adjusted NP to Rmb6.7bn by 2028E; PT at 20x 2030E P/E discounting back at 15% CAGR
SenseTime	20.HK	**HK$1.3**	OW	HK$2.0	**58%**	Cloud / model + application / enterprise deployment	Optimization scarcity + deployment capture	Participate in AI infrastructure demand, model & application assets provide additional capture opportunities as value shifts toward utilization and intelligence efficiency	2026E-28E revenue 28% CAGR; adjusted NP to Rmb1.1bn by 2028E; PT at 15x 2028E EV/EBITDA
Phancy	6682.HK	**HK$28**	OW	HK$50	**78%**	Cloud / application / enterprise deployment	Heterogeneous-compute optimization + deployment capture	China's fragmented chip environment increases the value of scheduling, optimization and deployment	2026E-28E revenue 36% CAGR; adjusted NP to Rmb898mn by 2028E; PT at 15x 2028E EV/EBITDA
Xunce	3317.HK	**HK$104**	OW	HK$160	**54%**	Enterprise deployment	Enterprise context + productization	Monetizes the implementation gap around context, governance and production integration; Token/TokenOS and module reuse provide a path toward more scalable software and usage-based economics	2026E-28E revenue 91% CAGR; adjusted NP to Rmb2.3bn by 2028E; PT at 20x 2028E P/E
Kingsoft Office	688111.CH	**Rmb231**	OW	Rmb325	**40%**	Productivity application	Distribution + workflow capture	WPS combines large-scale distribution, document state and workflow ownership; sourcing increasingly interchangeable intelligence can expand AI monetization while falling model costs support application economics	2026E-28E revenue 21% CAGR; adjusted NP to Rmb3.4bn by 2028E; PT at 45x 2028E P/E
MiniMax	100.HK	**HK$286**	Neutral	HK$300	**5%**	Foundation model	Large opportunity + capture still developing	Strong intelligence demand and usage growth support the opportunity, while durable economics depend on sustained relative capability, pricing power and conversion of usage into gross profit	2026E-28E revenue 127% CAGR; adjusted net loss at Rmb960mn in 2028E, turning profitable in 2030E; PT at 20x 2030E P/E discounting back at 15% CAGR
JST	6687.HK	**HK$13**	Neutral	HK$14	**8%**	Application	Strong workflow assets offset by e-commerce cycle and valuation	Live commerce state, platform connectivity and execution remain valuable as agents perform more operating work; subdued e-commerce conditions, growth expectations and valuation offset the strength of those assets	2026E-28E revenue 14% CAGR; adjusted NP to Rmb377mn by 2028E; PT at 15x 2028E P/E
Meitu	1357.HK	**HK$4.1**	Neutral	HK$4.5	**10%**	Application	TAM expansion + rising model-level substitution	AI materially expands visual creation and marketing workloads; improving general models make individual features easier to replicate, creating uncertainty over incremental growth and monetization	2026E-28E revenue 14% CAGR; adjusted NP to Rmb1.4bn by 2028E; PT at 15x 2028E P/E
Kingdee	268.HK	**HK$7.9**	Neutral	HK$8.3	**5%**	Application	System-of-record scarcity + capture still developing	Enterprise state, permissions and workflow provide residual scarcity; AI attach, ACV, NRR and usage need to translate into clearer revenue acceleration and incremental gross profit	2026E-28E revenue 12% CAGR; adjusted NP to Rmb1.3bn by 2028E; PT at 20x 2028E P/E
Manycore	68.HK	**HK$8.2**	Neutral	HK$7	**-14%**	Model + application	Workflow + proprietary spatial context	Specialized design workflows and proprietary 3D/spatial assets support residual scarcity; Spatial Intelligence monetization remain earlier	2026E-28E revenue 7% CAGR; adjusted NP to Rmb139mn by 2028E; PT at SOTP: 15x 2028E P/E on traditional SaaS, and US$2.1bn on Spatial Intelligence, which is 1/3 of the average private-market valuations for global Spatial Intelligence/World Model peers (AMI Labs at US$3.5B pre-money and World Labs at ~US$5B, per Bloomberg)
Yonyou	600588.CH	**Rmb9.9**	UW	Rmb6	**-39%**	Application	Strategic scarcity + weak financial conversion	Valuable enterprise state, permissions and workflow have yet to translate into comparable consolidated growth, profitability and cash generation	2026E-28E revenue 6% CAGR; adjusted NP to Rmb470mn by 2028E; PT at 45x 2028E P/E

Source: Bloomberg Finance L.P. (as of Sep 22, 2026), J.P. Morgan estimates.

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

Executive summary

Exponential intelligence consumption growth creates a larger market and growing earnings across the AI value chain

We expect China's token consumption to grow c.60x from 2026E to 2030E as better intelligence expands both the amount and complexity of work AI can perform. Frontier models make harder tasks feasible, while falling inference costs bring established capabilities to a broader customer base. As customers adopt AI across more activities, longer reasoning, richer context and multi-step execution also raise token consumption within each workflow. Together, these two effects support sustained demand across China's AI value chain.

We expect enterprises to account for approximately 69% of China's token consumption in 2030, with consumers contributing the remaining 31%. Enterprise usage scales as AI becomes embedded in recurring work such as coding, research, customer service and operations. On the consumer side, easier access and lower costs expand creation, productivity and personalized services. This mix gives the industry two complementary sources of demand: repeated business workflows that support sustained consumption, and broad consumer adoption that expands the range of use cases.

Growing intelligence consumption translates into a large spending opportunity as AI creates measurable value for end customers. Enterprises can compare AI spending with the labor or outsourcing cost of completed work, while consumers pay for useful services and lower creation costs. On that basis, our market estimates extend beyond traditional software budgets to **US$1.4tn of global AI end customer spending in 2030E, including approximately one-quarter of demand in China**. We expect the revenue evidence to appear in API and subscription receipts, enterprise contracts, usage fees and transaction revenue. The profit evidence should appear in gross margin after inference and delivery costs. All subsequent industry forecasts refer to China.

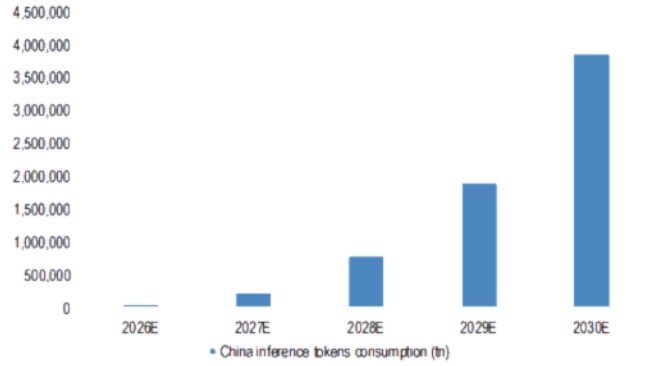
The capacity and services needed to meet this demand create revenue throughout the value chain. We model **US$778bn of AI investment by China's hyperscalers over 2026-30**. Each investment cohort generates hardware sales when capacity is procured. Once deployed, that capacity produces recurring cloud revenue, while models + applications charge for the intelligence, workflows and outcomes it supports. Successive investment cohorts and rising utilization allow cloud and downstream businesses to monetize an expanding installed base. We consequently forecast 2030 revenue at **1.8x its 2026E level for hardware, 12.5x for cloud and 26.9x for models + applications**.

We expect operating profit across China's AI value chain to rise from US$27bn in 2026E to US$243bn in 2030, with models + applications contributing the largest share of incremental profit, followed by cloud. Higher hardware procurement and margins raise upstream earnings. A larger installed base creates more billable compute, allowing cloud revenue to scale faster than operating costs. Downstream turns profitable as gross profit from models + applications covers training and other expenses. We therefore expect earnings to grow across all three layers, with hardware profit reaching 2.5x its starting level, cloud 15.6x and models +

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

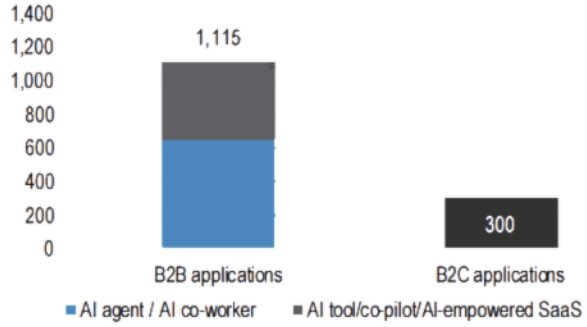
applications moving from losses to profit. The timing of these transitions determines where earnings accumulate over the investment cycle.

Figure 5: We expect China's inference token consumption to grow by c.60x in 2026-30



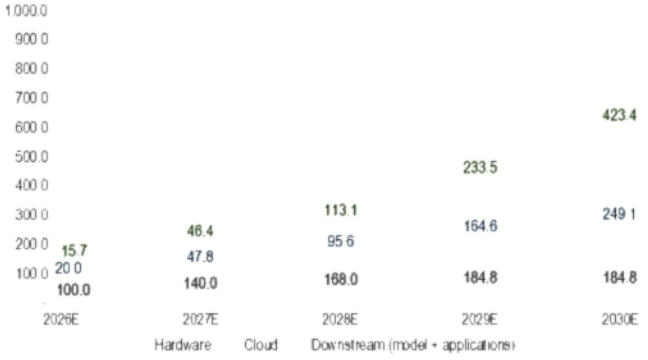
Source: National Data Administration (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 6: Our bottom-up analysis projects that the global AI TAM will reach US$1.4tn by 2030, of which China accounts for one-quarter



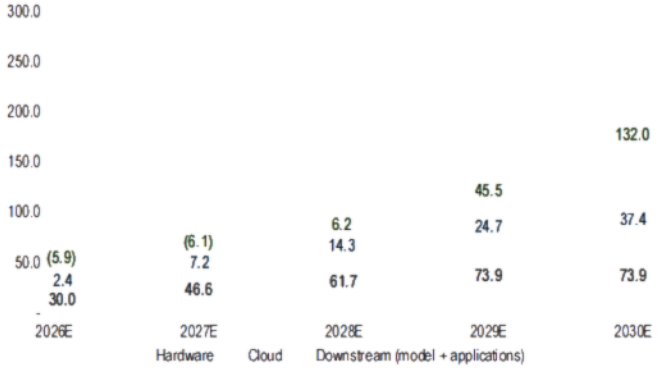
Source: International Labor Organization, National Bureau of Statistics of China, U.S. Bureau of Labor Statistics, Sensor Tower; (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 7: 2026E-30E China AI value chain revenue pool (US$ bn)



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 8: 2026E-30E China AI value chain operating profit pool (US$ bn)



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

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

Downstream contributes one-third of five-year profits and nearly two-thirds of earnings growth, creating room for valuations to rise

The earlier recognition of buildout earnings gives hardware approximately half of the US$544bn cumulative industry operating profit that we forecast for 2026-30. Once that capacity enters service, cloud earns recurring fees and models + applications charge for the intelligence and workflows it supports. We expect models + applications to consequently contribute **32% of five-year profit, including their early losses, while their subsequent expansion should supply 64% of the increase in annual industry profit from 2026E to 2030E**. In investment terms, downstream (AI foundation models + applications) combines a substantial five-year earnings contribution with the largest share of incremental growth.

The turning point is 2028E, when we expect downstream (AI foundation models + applications) operating profit to turn positive, at US$6.2bn. We expect the gross margin of **downstream players (AI foundation models + applications)** to rise to 40% by 2028, while training falls to approximately 25% of downstream revenue. We expect the gross margin of domestic AI model players to reach 50% by 2030 and training to represent 15% of compute power spending, allowing US$423bn of downstream revenue to generate US$132bn of operating profit after training and other expenses. The same workloads continue to require inference compute, memory, networking and cloud services, so downstream growth reinforces demand for upstream capacity.

We believe this expanding earnings contribution leaves room for greater market recognition, with models + applications representing c.14% of the estimated China-AI-attributable listed market value today. That share corresponds closely to where the subsector's revenue sits today: Models + applications accounts for c.12% of 2026E value chain revenue, rising to c.49% by 2030, on our forecasts. **The market is therefore pricing the current revenue distribution, which we regard as reasonable, rather than anticipating the shift in revenue and earnings that our model implies.** The valuation case strengthens when listed companies report rising paid usage, gross profit and cash generation from these businesses. Our market value estimate uses select companies and AI attribution assumptions, while the industry pool also includes unlisted competitors and platform businesses. Company selection therefore requires evidence of competitive strength and funding through the earnings ramp, particularly as approximately three-quarters of modeled downstream (AI models + applications) five-year profit arrives in 2030E.

Paid usage and gross profit progression should provide evidence ahead of that earnings inflection. We would look for higher cloud utilization, recurring API and subscription revenue, more paying application users, and gross margin after inference and delivery costs over the next 12-24 months. These metrics show whether consumption is becoming revenue and whether revenue is becoming profit. We favor incremental exposure to cloud and models + applications when companies report this progression in segment results and cash flow.

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

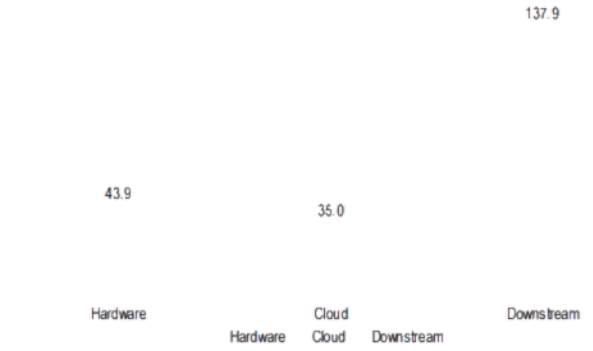
Table 2: 2026E-30E model output: Revenue grows across every layer; hardware earns first, downstream (AI models + applications) earns the growth

US$ bn	2026E	2027E	2028E	2029E	2030E	2026E-30E Total
Capex and installed base						
Cloud capex – headline	100.0	140.0	168.0	184.8	184.8	**777.6**
YoY growth		40%	20%	10%	0%	
Cumulative capex	100.0	240.0	408.0	592.8	777.6	
Monetized cloud capex	100.0	140.0	168.0	184.8	184.8	
Hardware						
Revenue	100.0	140.0	168.0	184.8	184.8	**777.6**
Gross profit	45.0	67.2	87.4	101.6	101.6	**402.8**
Gross margin	45%	48%	52%	55%	55%	
Operating profit (EBIT)	**30.0**	**46.6**	**61.7**	**73.9**	**73.9**	**286.1**
Operating margin	30%	33%	37%	40%	40%	
Cloud						
Opening pre-2026 installed base	20.0	16.6	12.3	7.8	3.1	
2026 capex cohort	0.0	31.2	37.9	45.0	46.8	
2027 capex cohort	0.0	0.0	45.4	55.1	65.5	
2028 capex cohort	0.0	0.0	0.0	56.7	68.8	
2029 capex cohort	0.0	0.0	0.0	0.0	64.9	
2030 capex cohort	0.0	0.0	0.0	0.0	0.0	
Total cloud revenue	**20.0**	**47.8**	**95.6**	**164.6**	**249.1**	**577.1**
Cloud EBITA margin	12%	15%	15%	15%	15%	
Cloud EBIT (EBITA)	**2.4**	**7.2**	**14.3**	**24.7**	**37.4**	**86.0**
Cloud EBIT margin	12%	15%	15%	15%	15%	
Cloud D&A	10.0	25.0	51.0	83.6	118.6	308.2
Cloud EBITDA	12.4	32.2	65.3	108.3	155.9	394.1
Cloud EBITDA margin	62%	67%	68%	66%	63%	
Downstream (LLM & application)						
Training / model R&D share	45%	37%	29%	22%	15%	
Inference share	55%	63%	71%	78%	85%	
Cloud training spend / model R&D	9.0	17.7	27.7	36.2	37.4	
Cloud inference spend / downstream COGS	11.0	30.1	67.9	128.4	211.7	
Downstream gross margin assumption	30%	35%	40%	45%	50%	
Downstream revenue	**15.7**	**46.4**	**113.1**	**233.5**	**423.4**	**832.1**
Gross profit after inference	4.7	16.2	45.2	105.1	211.7	
Downstream gross margin	30%	35%	40%	45%	50%	
Less: training R&D	(9.0)	(17.7)	(27.7)	(36.2)	(37.4)	
Training R&D as % of revenue	57%	38%	25%	16%	9%	
Less: Other opex	(1.6)	(4.6)	(11.3)	(23.3)	(42.3)	
Downstream EBIT	**(5.9)**	**(6.1)**	**6.2**	**45.5**	**132.0**	**171.8**
Downstream EBIT margin	-37%	-13%	5%	19%	31%	
Summary – CY revenue and profit pools						
Operating profit summary	**2026E**	**2027E**	**2028E**	**2029E**	**2030E**	**2026E-30E Total**
Revenue (overlapping)						
Hardware	100.0	140.0	168.0	184.8	184.8	777.6
Cloud	20.0	47.8	95.6	164.6	249.1	577.1
Downstream (model + app)	15.7	46.4	113.1	233.5	423.4	832.1
Operating profit (additive)						
Hardware	30.0	46.6	61.7	73.9	73.9	286.1
Cloud	2.4	7.2	14.3	24.7	37.4	86.0
Downstream (model + app)	(5.9)	(6.1)	6.2	45.5	132.0	171.8
Total industry operating profit	**26.5**	**47.7**	**82.2**	**144.1**	**243.3**	**543.8**
Hardware share of operating profit	93%	87%	75%	51%	30%	53%
Cloud share of operating profit	7%	13%	17%	17%	15%	16%
Downstream share of operating profit	n/m	n/m	8%	32%	54%	32%

Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

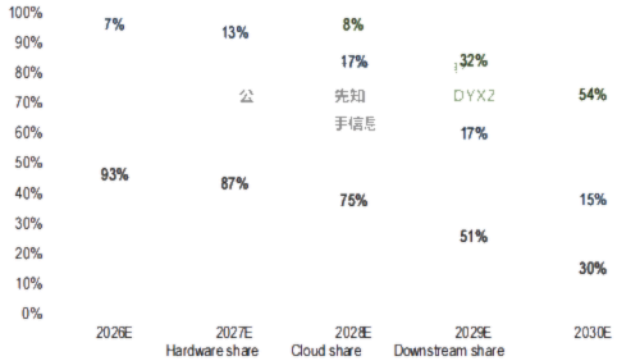
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Figure 9: Contribution to 2026E-30E annual OP increase (US$ bn): Downstream (AI models + applications) contributes US$138bn incremental profit in 2030E vs 2026E, or 64% of the increase in annual industry profit



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 10: 2026E-30E China AI value chain operating profit mix by layer



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Opportunity × Scarcity × Value Capture favors Alibaba, Tencent and Zhipu alongside nine initiations/ assumptions

Identifying who can earn the growing profit pool leads us to Opportunity × Scarcity × Value Capture. We start with the work and customer spending a company can address, then assess the capabilities customers find difficult to substitute. We test those advantages through reported pricing, retention, cost per inference, gross margin and cash returns. Valuation then determines the upside, connecting our preference for Alibaba, Tencent and Zhipu with the infrastructure, deployment and application businesses entering coverage.

Scarcity shifts as models improve and companies expand across the stack

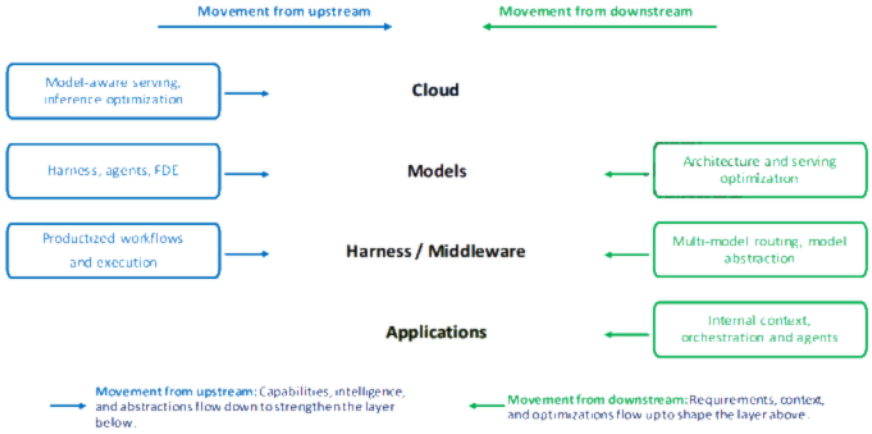
We expect differentiation to migrate from capacity and frontier intelligence toward bottlenecks in deployment and execution. A single workflow may require an advanced model, efficient serving, proprietary enterprise context and permission to act. Model providers can charge more when capability changes task outcomes; applications can retain more of the customer payment when their data, access and execution remain necessary after model prices fall.

The same model improvements that lower application costs also increase competition for application functionality. Generic content generation, analytics and configuration become easier to replicate, while live workflow state, permissions and transaction execution remain harder to reproduce. An agent may become the user entry point and still need these assets to complete a task. Applications therefore need to show that workflow access produces fees, transactions or higher retention.

Companies are responding by extending toward capabilities and customer relationships that show up in financial statements. Cloud providers can sell heterogeneous-compute optimization, model-aware serving and distribution; model providers can charge for infrastructure efficiency, proprietary harnesses, agents and deployment; middleware vendors can turn implementation work into reusable software; and applications can expose data and actions to agents. The relevant evidence is higher utilization, recurring revenue, gross profit per workload and stronger retention.

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

Figure 11: Scarcity migrates upstream and downstream



Source: J.P. Morgan estimates.

Three preferred exposures and nine initiations/ assumptions apply these judgments

Alibaba offers integrated chip, cloud and model exposure. T-Head broadens its hardware participation, while cloud and its model franchise connect infrastructure to customer demand. Our incremental-capacity reconstruction produces approximately a 26.6% after-tax project IRR vs 10% for the normalized industry cohort. This analyst scenario assumes stronger revenue yield and faster utilization. We expect these advantages to show up through higher cloud EBITA, lower serving cost per workload and growing paid model usage; cash conversion determines the shareholder benefit.

Tencent offers scarce AI application assets we believe are largely unpriced. Distribution, proprietary context and transaction execution give Tencent direct access to existing consumer and commercial activity. As model costs fall, these assets can support more workflows, transactions and advertising or merchant revenue. We focus on incremental commercial profit and paid AI adoption after development and inference costs, tying our valuation view to reported financial results within the group.

Zhipu is our preferred differentiated-intelligence exposure. Sustained relative capability can support API pricing, paid calls and customer retention, while serving optimization, proprietary harnesses and enterprise deployment can lower cost per token. We remain Overweight, focusing on paid API growth and gross profit sufficient to absorb model development. MiniMax remains Neutral: We need to see durable capability, pricing and gross profit conversion alongside its strong usage growth.

SenseTime and Phancy offer exposure to infrastructure utilization and optimization; we initiate both at Overweight. SenseTime's SenseCore connects capacity with production AI workloads, while Phancy addresses heterogeneous scheduling, adaptation and deployment. As installed compute expands, we expect value to accrue to tools that increase tokens or tasks processed per unit of capacity and reduce deployment cost. External adoption, reusable software revenue, utilization and returns on investment will show whether technical improvements scale.

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

Xunce addresses the implementation gap around enterprise context, governance and production integration; we initiate at Overweight. Token/TokenOS and module reuse can turn repeated customer projects into recurring software and usage revenue. We will look for shorter deployments, higher revenue per employee, more reusable modules and stronger cash collection. These metrics show whether enterprise demand is becoming a scalable software business.

Kingsoft Office combines distribution, document state and workflow ownership; we assume coverage at Overweight. Access to increasingly interchangeable intelligence can expand paid productivity use while lowering the cost of serving it. JST also controls valuable live orders, inventory, merchant rules and commerce execution. We assume coverage of JST at Neutral because subdued e-commerce conditions, growth expectations and valuation offset the strength of those workflow assets, in our view. For both companies, the evidence is incremental customer spending and gross profit after AI costs.

Meitu, Kingdee and Manycore offer meaningful application opportunities, with greater uncertainty over incremental growth and monetization; we initiate/ assume coverage at Neutral. Meitu benefits from expanding creation workloads, while improving general models make individual features easier to replicate. Kingdee's enterprise records and permissions provide durability, and Manycore's design workflows and spatial context support differentiation. We will track paid AI adoption, retention, incremental gross profit and cash conversion. We assume coverage of Yonyou at Underweight, as its enterprise assets require stronger consolidated growth, profitability and cash generation. Key sector risks include slower intelligence-consumption growth, the erosion of company-specific scarcity, and a weaker conversion of adoption into gross profit and cash flow.

For details, please refer to our company initiation/ assumption reports:

- SenseTime (20.HK): *Infrastructure at the core, deployment and multimodality as upside; initiate at Overweight*
- Phancy (6682.HK): *Heterogeneous compute meets rising intelligence demand; initiate at Overweight*
- Xunce Technology: *Enterprise AI deployment creates a new growth pool; initiate at Overweight*
- Kingsoft Office: *AI extends the productivity growth runway; assume coverage at Overweight*
- Meitu (1357.HK): *Expanding AI opportunity meets declining feature scarcity; initiate at Neutral*
- JST Group (6687.HK): *Workflow scarcity supports earnings durability, though growth expectations are running ahead; assume at Neutral*
- Kingdee International (0268.HK): *Cloud execution improving; AI upside and competitive impact remain to be proven; assume coverage at Neutral*
- Manycore Technology (0068.HK): *AI expands the opportunity; value capture remains to be proven; assume coverage at Neutral*
- Yonyou Network Technology (600588.SH): *Another transition, with value capture to prove; assume coverage at Underweight*

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

Our previous publications on China Artificial Intelligence:

- *China Artificial Intelligence: Global reach and model innovation drive next-gen leaders; initiate on Knowledge Atlas and MiniMax at OW* (February 2026)
- *China Artificial Intelligence: Addressing Ten Key Questions for Investors* (March 2026)
- *China Artificial Intelligence: After the rally – What's priced in, what could change and how to position?* (April 2026)
- *China Artificial Intelligence: DeepSeek V4 3-week check: first-party cost edge; compute-constrained market; SOTA is still the answer for Zhipu/MiniMax* (May 2026)
- *China Artificial Intelligence: When demand exceeds supply, a price cut is a verdict: remain OW Zhipu, downgrade MiniMax to Neutral* (June 2026)
- *China Artificial Intelligence: Open-weight commercialization playbook: winner takes more* (July 2026)
- *China Artificial Intelligence: Capital accelerates race; Raise Zhipu (OW) PT on higher ARR visibility; MiniMax (N) needs stronger model validation* (July 2026)
- *China Artificial Intelligence: Kimi K3 resets frontier-model valuation, but 50%+ sell-off overstates impact on Zhipu's ARR trajectory, retain OW* (July 2026)
- *China Artificial Intelligence: GLM-5.3 and DeepSeek's repricing shift the capability-cost frontier; raise Zhipu/MiniMax PT, remain OW/Neutral* (August 2026)

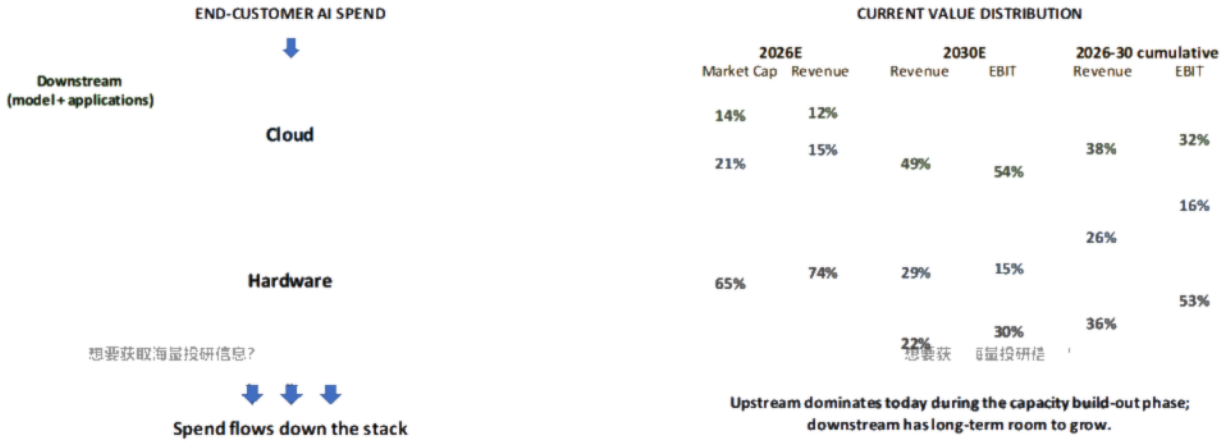
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

AI value chain: A growing end-demand pool with earnings shifting toward models + applications

China's AI development remains in a capacity-building phase, and this is where listed value currently sits: **Models + applications represents only c.14% of our China-AI-attributable listed market capitalization estimate**, with hardware and cloud accounting for the balance. At the same time, continued growth in intelligence consumption is expanding the end-demand pool across models, enterprise deployment and applications. Our industry model suggests that this downstream layer (AI models + applications) contributes **32% of cumulative 2026E-30E industry operating profit and 64% of the increase in annual profit** over the period. We first assess the size of the opportunity and how value is currently distributed before applying our **Opportunity × Scarcity × Value Capture** framework to determine which companies are best positioned to retain the resulting economics.

We organize the AI value chain into **three primary economic layers: hardware, cloud and models + applications**. Hardware supplies the physical capacity; cloud turns installed capacity into recurring compute and inference services; and models + applications monetizes the intelligence, workflows and outcomes running on top. Within the third layer, we further distinguish **foundation models, middleware / harness / enterprise deployment, and applications** because these sub-layers rely on different sources of scarcity and monetize at different points in the customer workflow.

Figure 12: China AI value chain: Growing TAM, migrating scarcity, fluid boundaries



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

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

Intelligence consumption: About 60x token growth over 2026E-30E, roughly two-thirds by enterprise

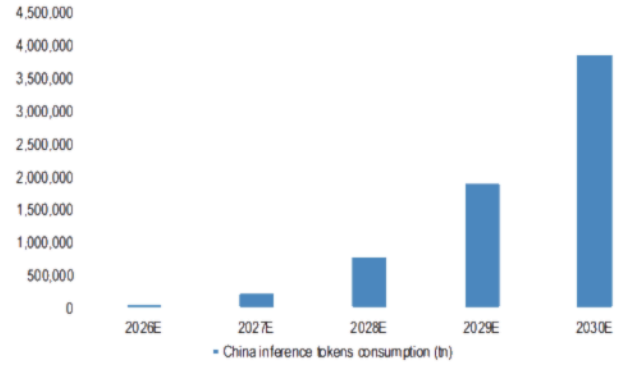
Token consumption is the physical measure of the expanding pie. We expect China inference-token consumption grow approximately **60x to 2030 vs its 2026E level**. Two mechanisms drive this simultaneously: Frontier capability makes harder tasks feasible, expanding the range of work AI can perform at all; falling unit costs bring established capability to a broader customer base and a longer tail of use cases. Within each adopted workflow, longer reasoning chains, richer context windows and multi-step agent execution raise consumption per task.

We expect enterprises to account for approximately 69% of 2030E consumption, with consumers contributing the remaining 31%. This split matters for how the downstream (AI models + applications) revenue pool behaves, not only its size:

- **Enterprise consumption is recurring and workflow-embedded.** Coding, research, customer service, document processing and operational workflows generate repeated, predictable token demand once deployed. This supports contracted revenue, usage-based pricing and higher retention, and it is the consumption base against which enterprises can benchmark labor and outsourcing costs.
- **Consumer consumption is broader, but more price-elastic.** Easier access and lower costs expand creation, productivity and personalized services across a much larger user base. Monetization runs through subscriptions, advertising and transaction revenue, rather than contracts, making distribution and cost per served user the critical variables.

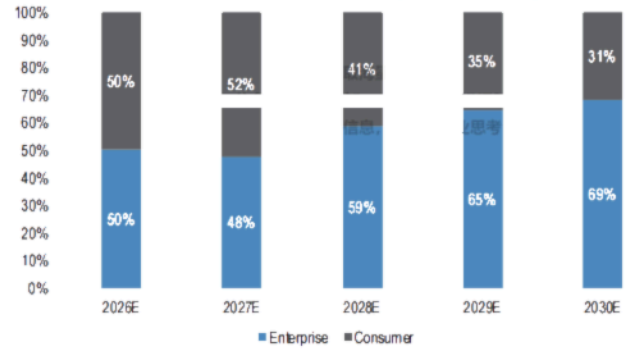
The two sources are complementary, rather than competing: Enterprise workloads provide the sustained consumption base, while consumer adoption expands the range of uses and drives scale economics in serving. **Falling unit costs can therefore coexist with rapidly rising aggregate industry revenue.** This **Jevons effect** is central to our sector view – reasoning, long context, multimodality and agents increase intelligence consumed per workflow, while lower prices make more workflows economically viable.

Figure 13: We expect China's inference token consumption to grow by c.60x in 2026-30...



Source: National Data Administration (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 14: ...enterprises to account for approximately 69% of 2030E consumption, with consumers contributing the remaining 31%



Source: National Data Administration (as of Sep 11, 2026), J.P. Morgan estimates.

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

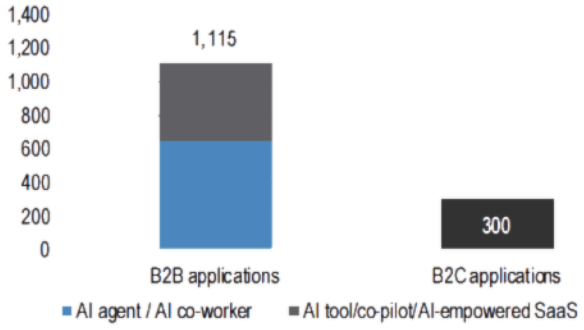
>US$1.4tn provides the end-demand anchor, with further upside from AI-driven value creation

Our Zhipu/MiniMax initiation estimated **>US$1.4tn of global AI end-customer spending by 2030**, including c.US$1.1tn of B2B demand and c.US$300bn of B2C demand, with **China representing roughly one-quarter**. Within China, we estimate a **c.US$250bn B2B opportunity**, comprising c.US$147bn from AI agents/co-workers and c.US$96bn from AI tools/co-pilots and AI-empowered SaaS, alongside c.US$100bn of consumer demand.

The estimate is anchored primarily to existing labor and service economics. Enterprises already pay for coding, customer service, research, design, marketing, sales support and operations; as AI takes on a larger share of these activities, customer willingness-to-pay can increasingly reference the economic value of the work performed. This is what allows the addressable pool to extend beyond traditional software and IT budgets.

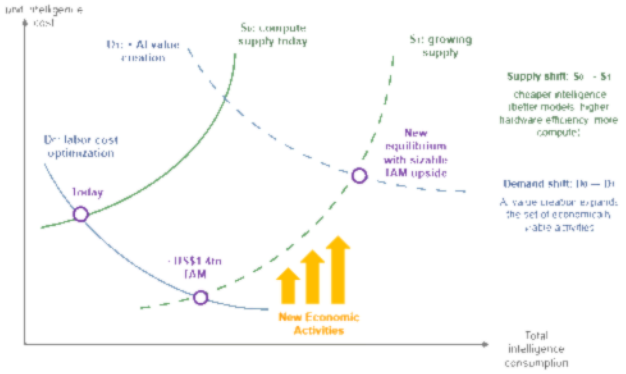
The longer-term opportunity can exceed the existing labor pool if cheaper intelligence also increases economic output. Lower creation costs can make additional software, content, design, analysis and personalized services economical to produce. We highlighted this mechanism in our previous initiation through the evolution of video: Falling creation and distribution costs supported a large increase in supply and, ultimately, new formats and business models. Recent developments in software provide early evidence of a similar response.

Figure 15: Our bottom-up analysis projects that the global AI TAM will reach US$1.4tn by 2030



Source: International Labor Organization, National Bureau of Statistics of China, U.S. Bureau of Labor Statistics, Sensor Tower; (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 16: We believe the US$1.4tn demand pool provides the anchor, while the significant upside lies in the value creation

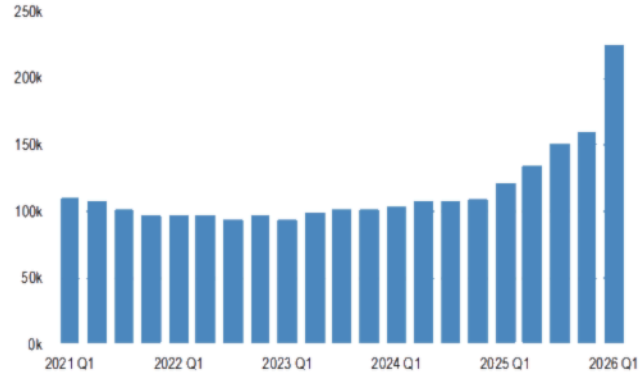


Source: J.P. Morgan estimates.

Software supply is accelerating. Sensor Tower data show that annual new App Store releases fell 48% between 2016 and 2024 before reversing to **+30% YoY in 2025, to nearly 600k, and +84% YoY in 1Q26, to 235.8k**. The data do not provide clean attribution to AI, although the timing is consistent with coding tools lowering the technical and labor requirements for software creation.

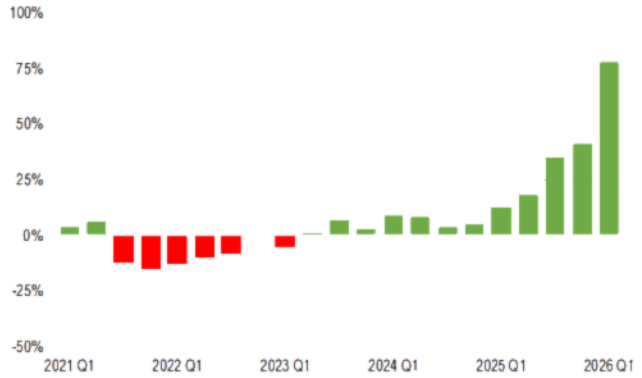
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Figure 17: The number of new apps in Apple's App Store has surged in recent quarters, likely due to a software boom sparked by new AI vibe coding tools



Source: Sensor Tower, Sina Finance (as of Sep 11, 2026)

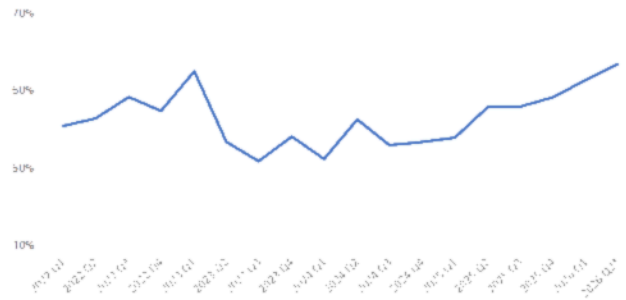
Figure 18: Quarterly global growth in new apps on Apple's App Store compared to the same period a year earlier



Source: Sensor Tower, Sina Finance (as of Sep 11, 2026)

Smaller teams are also reaching the market faster. Stripe Atlas reported that **63% of new Atlas C-corps were solo-founded** by 2Q26. By their second year, AI-native solo startups generated **nearly 2x the revenue of other solo-founded startups**, with the outperformance broad-based across roughly the 50th-95th revenue percentiles.

Figure 19: Solo founding is at an all-time high



Source: Stripe Atlas C corp incorporations (as of Sep 11, 2026)

Figure 20: Median revenue relative to other solo-founded startups



Source: Single-founder C corps incorporated through Stripe Atlas over 2022-23 (as of Sep 11, 2026)

We view these as early evidence of **AI value creation**, rather than a basis for mechanically increasing our >US$1.4tn estimate. Coding is one early example; similar economics could emerge across content, design, research, marketing and personalized services. Over a longer horizon, **AI for Science** could expand the opportunity into scientific discovery and the creation of new drugs, materials and technologies.

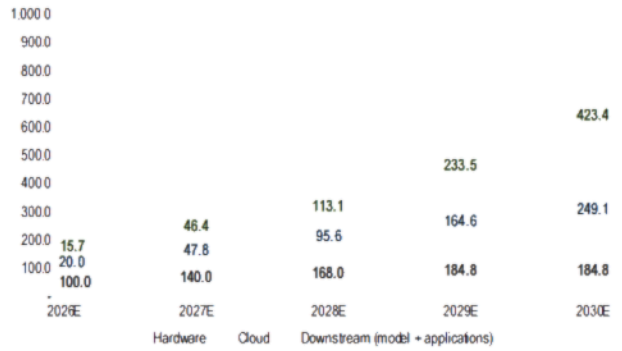
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Value is concentrated upstream today, and that matches where revenue currently sits

Models + applications represents c.14% of our estimated China-AI-attributable listed market capitalization, with hardware and cloud accounting for the balance. **Importantly, this c.14% corresponds closely to today's revenue distribution, rather than representing an obvious dislocation.** On our model, models + applications accounts for **c.12% of 2026E value-chain revenue** vs hardware at c.74% and cloud at c.15%. The market is therefore capitalizing roughly the revenue mix that exists today. We regard that as a reasonable starting point, not a mispricing, and it sharpens the actual investment question, which is whether the market is anticipating the shift ahead.

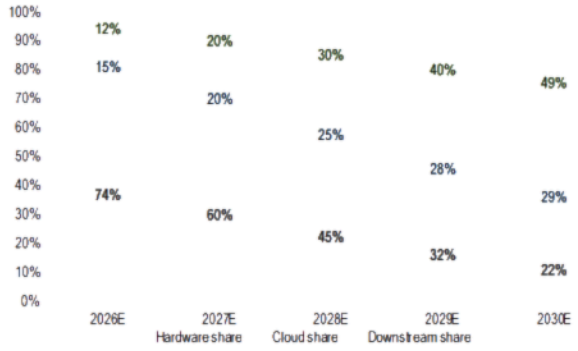
Our forecasts imply that the downstream (AI models + applications) revenue share will rise from c.12% to c.49% by 2030 and the downstream operating profit share will rise from negative to c.54%. The equity question is therefore not whether c.14% is wrong today, but whether the transition is sufficiently visible to be capitalized ahead of the reported earnings inflection. Foundation models and applications are earlier in monetization and, in many cases, continue to absorb training, inference and go-to-market costs before reaching mature margins, which is precisely why the revenue and market value shares currently agree.

Figure 21: 2026E-30E China AI value chain revenue pool
US$bn



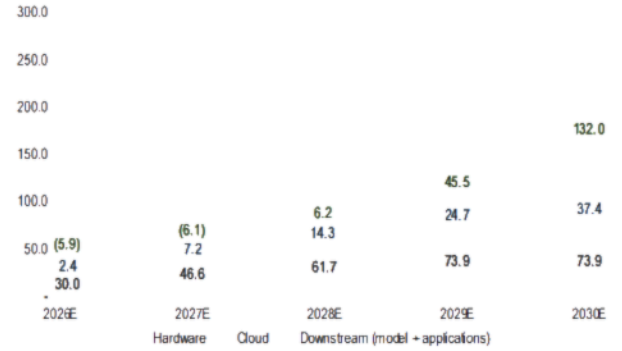
Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 22: 2026E-30E China AI value chain revenue mix



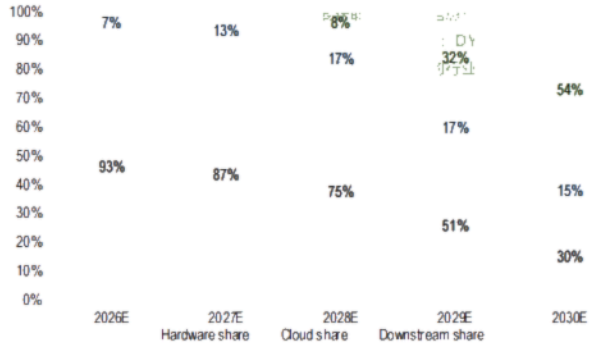
Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 23: 2026E-30E China AI value chain operating profit pool
US$ bn



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

Figure 24: 2026E-30E China AI value chain profit mix



Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

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

We do not read the upstream concentration as a mispricing on its own. Our model gives hardware approximately half of cumulative 2026E-30E industry operating profit, so a large upstream share of value is consistent with the earnings actually recognized over the forecast period. The narrower question is whether **32% of cumulative profit and 64% of profit growth** are adequately reflected in a **c.14%** share of attributable market value, given that the downstream (AI models + applications) earnings arrive later and with a wider dispersion of outcomes across companies.

How we build the forecast: One dollar of investment, earned three times

Our industry forecast is built upward from a single input, rather than allocated downward from a market size estimate. **We take one dollar of cloud investment and follow it until it has been earned three times.** Hardware books it the moment capacity is purchased. Cloud earns it back across the five years the asset stays in service. Models + applications then earns on top of the intelligence that capacity produces.

The three layers are therefore three claims on the same dollar, recognized on three different clocks. Deriving each layer from the one above it is what allows them to be compared consistently, and it imposes a constraint that a top-down TAM allocation does not: **no layer can be larger than what the layer above it can physically supply.** The cost of that constraint is that our conclusions concentrate in a handful of linked assumptions, which we set out in full below.

Layer revenues overlap and must not be summed as end-customer spending. Hardware sells to cloud; cloud sells inference to downstream (AI models + applications); downstream sells to end customers. Operating profit is additive and is the meaningful measure for cross-layer value distribution.

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

Table 3: 2026-30E model output: Revenue grows across every layer; hardware earns first, downstream (AI models + applications) earns the growth

US$ bn	2026E	2027E	2028E	2029E	2030E	2026E-30E Total
AI capex and installed base						
AI capex – headline	100.0	140.0	168.0	184.8	184.8	**777.6**
YoY growth		40%	20%	10%	0%	
Cumulative AI capex	100.0	240.0	408.0	592.8	777.6	
Monetized cloud capex	100.0	140.0	168.0	184.8	184.8	
Hardware						
Revenue	100.0	140.0	168.0	184.8	184.8	**777.6**
Gross profit	45.0	67.2	87.4	101.6	101.6	**402.8**
Gross margin	45%	48%	52%	55%	55%	
Operating profit (EBIT)	**30.0**	**46.6**	**61.7**	**73.9**	**73.9**	**286.1**
Operating margin	30%	33%	37%	40%	40%	
Cloud						
Opening pre-2026 installed base	20.0	16.6	12.3	7.8	3.1	
2026 capex cohort	0.0	31.2	37.9	45.0	46.8	
2027 capex cohort	0.0	0.0	45.4	55.1	65.5	
2028 capex cohort	0.0	0.0	0.0	56.7	68.8	
2029 capex cohort	0.0	0.0	0.0	0.0	64.9	
2030 capex cohort	0.0	0.0	0.0	0.0	0.0	
Total cloud revenue	**20.0**	**47.8**	**95.6**	**164.6**	**249.1**	**577.1**
Cloud EBITA margin	12%	15%	15%	15%	15%	
Cloud EBIT (EBITA)	**2.4**	**7.2**	**14.3**	**24.7**	**37.4**	**86.0**
Cloud EBIT margin	12%	15%	15%	15%	15%	
Cloud D&A	**10.0**	**25.0**	**51.0**	**83.6**	**118.6**	**308.2**
Cloud EBITDA	**12.4**	**32.2**	**65.3**	**108.3**	**155.9**	**394.1**
Cloud EBITDA margin	62%	67%	68%	66%	63%	
Downstream (LLM + application)						
Training / model R&D share	45%	37%	29%	22%	15%	
Inference share	55%	63%	71%	78%	85%	
Cloud training spend / model R&D	9.0	17.7	27.7	36.2	37.4	
Cloud inference spend / downstream COGS	11.0	30.1	67.9	128.4	211.7	
Downstream gross margin assumption	30%	35%	40%	45%	50%	
Downstream revenue	**15.7**	**46.4**	**113.1**	**233.5**	**423.4**	**832.1**
Gross profit after inference	4.7	16.2	45.2	105.1	211.7	
Downstream gross margin	30%	35%	40%	45%	50%	
Less: training R&D	(9.0)	(17.7)	(27.7)	(36.2)	(37.4)	
Training R&D as % of revenue	57%	38%	25%	16%	9%	
Less: Other opex	(1.6)	(4.6)	(11.3)	(23.3)	(42.3)	
Downstream EBIT	**(5.9)**	**(6.1)**	**6.2**	**45.5**	**132.0**	**171.8**
Downstream EBIT margin	-37%	-13%	5%	19%	31%	
Summary – CY revenue and profit pools						
Operating profit summary	**2026E**	**2027E**	**2028E**	**2029E**	**2030E**	**2026E-30E Total**
Revenue (overlapping)						
Hardware	100.0	140.0	168.0	184.8	184.8	777.6
Cloud	20.0	47.8	95.6	164.6	249.1	577.1
Downstream (model + app)	15.7	46.4	113.1	233.5	423.4	832.1
Operating profit (additive)						
Hardware	30.0	46.6	61.7	73.9	73.9	286.1
Cloud	2.4	7.2	14.3	24.7	37.4	86.0
Downstream (model + app)	(5.9)	(6.1)	6.2	45.5	132.0	171.8
Total industry operating profit	**26.5**	**47.7**	**82.2**	**144.1**	**243.3**	**543.8**
Hardware share of operating profit	93%	87%	75%	51%	30%	53%
Cloud share of operating profit	7%	13%	17%	17%	15%	16%
Downstream share of operating profit	n/m	n/m	8%	32%	54%	32%

Source: Bloomberg Finance L.P. (as of Sep 11, 2026), J.P. Morgan estimates.

The starting point: A US$778bn buildout that front-loads capacity

Everything begins with cloud investment. **We model US$100.0bn in 2026E, growing 40% in 2027E, 20% in 2028E and 10% in 2029E before flattening in 2030E at US$184.8bn, for a cumulative US$777.6bn over 2026E-30E.**

The deceleration is a judgment, rather than an extrapolation. Early-cycle investment is constrained by how much capacity exists; later-cycle investment is constrained by how intensively that capacity is being used. We assume the industry will cross that line around 2029E, after which utilization, rather than further procurement, becomes the binding constraint on serving growth. Because this path drives every other line in the model, **capex growth is our most consequential assumption** – and one where we would expect reasonable disagreement.

Hardware is paid at the point of purchase

Hardware revenue equals same-year cloud investment, running from **US$100.0bn to US$184.8bn** for the same US$777.6bn cumulative total. Operating margin rises from **30% in 2026E to 40% from 2029E** as the mix moves toward mature hardware economics.

The absence of any lag here is the reason hardware dominates early industry profit. Capacity generates revenue for its suppliers at the moment it is bought, before it has served a single workload, and before anyone downstream (AI models + applications) has earned anything from it. Hardware takes **93% of 2026E industry operating profit** for no reason more complicated than that it gets paid first.

Cloud earns the same dollar back over five years

Cloud operates on a slower clock. Each investment cohort has a **five-year revenue life beginning the year after procurement**, and its revenue depends on three things: **utilization**, which we ramp at 60%, 70% and 80% through Revenue Years 1, 2 and 3-5; **revenue yield**, at 0.50x of capex annually at full utilization in 2026E and improving 4% per year, to 0.585x by 2030E; and a **runoff contribution from pre-2026 capacity**, which accounts for the entire US$20.0bn 2026E cloud revenue base and declines from there.

The important consequence is that cohorts accumulate. By 2029E, four vintages are earning simultaneously, so **cloud revenue compounds far faster than the investment funding it – 12.5x its 2026E level by 2030E, against hardware's 1.8x**, despite both being driven by the same capex path. Cloud EBITA margin is 12% in 2026E and 15% thereafter, with cohort depreciation and the utilization ramp implying aggregate EBITDA margins of c.74% in 2027E and c.66% in 2030E.

The cloud dollar splits between training and inference, and that split matters twice. Within cloud revenue, the mix shifts from 45% training and model R&D / 55% inference in 2026E to 15%/85% by 2030E, on a straight line through 37.5%, 30% and 22.5%.

This is the assumption doing the most work in the model because it operates on the downstream layer (AI models + applications) from both sides at once. **The inference share sets how much paid downstream activity the installed base can support**, since inference spend is the compute those workloads actually consume. **The training share is simultaneously a cost borne by that same layer**, deducted below downstream gross profit. Early on, downstream is therefore squeezed from both directions: a small inference pool to sell against, and a large training bill to carry.

Downstream (AI models + applications) revenue is capped by the inference capacity it consumes

Because inference spend is downstream cost of goods sold, downstream revenue follows directly:

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

Downstream revenue = (inference share × cloud revenue) ÷ (1 − downstream gross margin)

With **gross margin improving from 30% in 2026E to 50% in 2030E** as inference economics improve, this gives downstream revenue of **US$15.7bn in 2026E, rising to US$423.4bn in 2030E, or 26.9x** – the fastest growth of any layer, from the smallest base.

Two effects compound to produce it. The inference pool itself expands as cloud revenue scales and tilts toward serving, and each dollar of inference supports progressively more revenue as gross margin improves. Neither is heroic on its own; together they are the difference between a modest downstream layer and one that ends the period larger than hardware and cloud combined.

The identity holds in every year, which is the discipline referred to above: **Downstream cannot sell more intelligence than the installed base can physically serve.** A top-down estimate of application spending carries no such check.

Training is what keeps downstream in loss until 2028E

The final step deducts the costs that a gross margin figure leaves out:

Downstream operating profit = (downstream revenue × gross margin) − training and model R&D − other operating expenses at 10% of downstream revenue

Training is the decisive line, and the arithmetic of its early years is unforgiving. In 2026E, it absorbs **US$9.0bn**, equal to 45% of all cloud revenue, against a downstream gross profit base of only **US$4.7bn**. The layer loses US$5.9bn before a single dollar of operating expense is counted.

By 2030E, training has reached US$37.4bn – barely higher in absolute terms – while downstream gross profit has grown to US$211.7bn. **The layer does not turn profitable because training spend falls. It turns profitable because the revenue base finally grows past it.** That distinction matters for stock selection: It implies that companies need to survive the crossing, rather than manage costs through it, which is why funding capacity appears repeatedly in our company assessments.

What the forecast actually rests on

We keep these assumptions visible because the model is an investor framework, rather than a set of point forecasts, and because a reader who disagrees with one of them should be able to see immediately what it changes. In rough order of impact on the downstream conclusion:

Table 4: Our forecast assumptions for China AI value chain model

Assumption	Base case	Why it matters
Downstream gross margin	30% in 2026E → 50% in 2030E	Sets how much revenue each dollar of inference supports, and the profit left after serving costs
Training / inference mix	45/55 in 2026E → 15/85 of cloud revenue in 2030E	Determines both the downstream revenue base and the largest downstream cost line
Cloud investment growth	+40% +20%/+10%/0% in 2027E/28E/29E/30E	Sets the size of the whole system; faster growth raises the upstream share of profit
Cloud revenue yield and utilization	0.50x rising 4% p.a.; 60% in Year 1, 70% in Year 2, 80% in Years 3-5	Determines whether capacity earns an acceptable return and can therefore keep growing
Hardware operating margin	30% in 2026E → 40% in 2030E	Determines the size of the earliest-recognized profit pool
Downstream other operating expenses	10% of downstream revenue	Modest individually, but applies to a very large 2030E revenue base

Source: J.P. Morgan estimates.

Revenue growth diverges sharply by layer. Hardware reaches 1.8x its 2026E level by 2030E, cloud 12.5x and downstream 26.9x. The divergence follows directly from the mechanics above: Hardware monetizes each year's investment once, cloud monetizes multiple active cohorts simultaneously, and downstream monetizes an inference pool that is itself growing while its gross margin improves.

Annual operating profit rises from US$26.5bn to US$243.3bn. Hardware rises from US$30.0bn to US$73.9bn (2.5x), cloud from US$2.4bn to US$37.4bn (15.6x), and downstream from a **US$5.9bn loss to a US$132.0bn profit**. The shift in the annual profit mix is stark: Hardware falls from 93% of the 2026E pool to 30% by 2030E, while downstream moves from loss-making to c.54%.

Where the five-year profit actually accumulates

Hardware earns approximately half of the US$543.8bn cumulative operating profit pool: c.53%. This is the most important correction to a purely "value migrates downstream" narrative. Because hardware recognizes economics at the point of procurement, it collects substantial profit throughout the buildout phase, while downstream absorbs training and inference costs before its revenue base is large enough to cover them. Cloud contributes c.16% and models + applications 32%.

The growth, however, is downstream. Annual industry operating profit increases by c.US$217bn between 2026E and 2030E. Models + applications supplies **64%** of that increase vs c.20% for hardware and c.16% for cloud. The investment implication is that hardware offers the more certain near-term earnings stream, while downstream offers the earnings trajectory – different propositions requiring different evidence.

2028E is the inflection; 2030E carries most of the downstream profit

The sequence matters more than the endpoints. **2026E-27E remains the capacity-to-utilization transition.** Hardware is already profitable and holds 93% and 87% of the industry profit pool in those two years, respectively. Cloud revenue begins to compound as the 2026 cohort enters service. Downstream remains loss-making – US$5.9bn in 2026E and US$6.1bn in 2027E – because training R&D is large relative to the initial revenue base.

2028E is the first financial inflection: Downstream operating profit turns positive, at US$6.2bn, or 8% of the industry pool. Gross margin reaches c.40% by that year as inference economics improve, while training falls to 30% of cloud revenue, equivalent to approximately 25% of downstream revenue. Both the margin improvement and the revenue scaling contribute; neither alone is sufficient.

2029E is where the mix changes materially: Downstream reaches 32% of the annual profit pool as revenue passes US$233bn, and hardware falls to 51% for the first time.

By 2030E, gross margin reaches 50% and training represents 15% of cloud spending, allowing US$423.4bn of downstream revenue to generate US$132.0bn of operating profit after training and other expenses.

Investors should note how concentrated that profit is in the final forecast year. Approximately **three-quarters of downstream five-year profit arrives in 2030E**, c.77% of the US$171.8bn cumulative total. This has two direct consequences for stock selection:

- **Funding matters as much as positioning.** A company with the right assets still needs to finance operations through 2026E-28E before the earnings ramp. Balance sheet, cash burn and access to capital are therefore selection criteria, not just quality checks.
- **Terminal-year sensitivity is high.** Because so much profit sits in one year, modest changes to gross margin, training intensity or the pace of paid adoption move the downstream pool materially. We treat the 2030E figure as a scenario outcome, rather than a forecast to be defended precisely.

A rising downstream share remains compatible with continued upstream growth. Higher model and application activity drives additional demand for inference compute, memory, networking and cloud capacity, since downstream revenue is derived from the inference capacity it consumes. The pie expands across several layers simultaneously even as the mix of incremental profit changes.

For downstream companies specifically, **gross profit is the critical bridge between usage and shareholder value.** Rapid token consumption or AI engagement can coexist with weak economics when model, inference or delivery costs absorb most of the incremental revenue. We therefore focus on whether rising customer spending translates into incremental gross profit, operating leverage and cash generation.

Cloud returns remain the key constraint on how the pool is shared

A larger downstream pool still requires economically sustainable upstream investment. Our cohort analysis follows one **US$100bn cloud investment cohort** through five revenue years. At a normalized 0.50x annual cloud revenue/capex yield, 15% steady-state EBITA margin and 55% EBITDA margin, the cohort generates c.**3.64 years of pretax cash payback**, a c.**11.6% pretax project IRR.**

This creates an important feedback loop. If cloud returns remain weak, investment growth eventually slows, utilization and pricing need to improve, or part of the economics must shift back toward infrastructure owners. If cloud returns are attractive, additional capacity can be financed and the downstream intelligence pool can continue expanding. **Our industry value chain analysis therefore describes a value chain with interdependent returns across upstream capacity, cloud utilization and downstream monetization.**

For investors, the key debate is therefore broader than whether AI investment is rising or falling. We focus on four linked questions: **(1) how fast intelligence consumption is growing; (2) whether installed compute earns an acceptable revenue yield and return; (3) whether falling inference costs expand downstream gross profit; and (4) which companies control the scarce capabilities and charging points within each layer.** This financial transmission provides the quantitative foundation for the Scarcity × Value Capture framework below.

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

AI value chain layer analysis: Where each layer earns, and what remains scarce

Hardware: Capacity remains the physical foundation and earns first

Hardware is the first layer of the value chain and the point at which new AI capacity enters the system. GPUs, memory, networking and related datacenter equipment convert cloud investment into deployable compute. In our model, hardware monetizes each new cohort at the time of procurement, which is why this layer captures economics earliest and collects **c.53% of cumulative 2026E-30E industry operating profit: US$286bn of the US$544bn pool**. Continued growth in training, inference and agent workloads can sustain a large absolute hardware profit pool even as more incremental value forms downstream.

The source of hardware scarcity evolves with workload mix. Frontier training remains sensitive to leading accelerators and high-bandwidth networking, while large-scale inference increases the importance of memory capacity, memory bandwidth, networking efficiency and power. As the enterprise share of token consumption rises toward our 69% 2030 assumption, and as agent workloads lengthen context and multiply calls per task, we expect the inference-weighted components of the hardware pool to grow faster than the training-weighted components. We therefore view hardware as a structurally important layer, with its share of total industry profit increasingly determined by the pace of new capacity additions relative to utilization and downstream monetization on the installed base.

Stock implications

- **Alibaba** participates in this layer through **T-Head**, which broadens its hardware exposure beyond the buyer side. In a fragmented domestic accelerator environment, in-house silicon can improve supply security, cost per unit of served capacity and the ability to co-design hardware with its own serving stack. The relevant evidence is deployment scale within Alibaba Cloud, cost per served workload and whether internal silicon improves cloud gross margin, rather than simply substituting external purchases.

Cloud: Converting installed capacity into recurring intelligence supply

Cloud is the second layer and the bridge between physical capacity and downstream intelligence consumption. CSPs aggregate hardware, operate the installed base and monetize it over multiple years through training, inference, model access, data, security and enterprise deployment environments. This creates a different financial clock from hardware: An investment cohort is monetized once by hardware suppliers, then repeatedly through cloud revenue as that capacity remains in service.

As installed capacity expands, cloud differentiation shifts from access to raw compute toward **utilization, scheduling, heterogeneous optimization, model-aware serving and intelligence output per unit of compute**. The progression can be summarized as:

Compute ownership → Resource scheduling → Heterogeneous optimization → Model-aware serving → Model/infra co-design

China's fragmented accelerator environment increases the value of this optimization layer. Different hardware stacks, software environments and performance characteristics create demand for portability, scheduling and serving efficiency across accelerators. This supports providers that can improve throughput, utilization and deployment economics on top of the physical capacity base.

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

Cloud returns provide the financial discipline for the whole system. Revenue yield, utilization and serving efficiency determine whether new investment earns an acceptable return and therefore whether capacity growth can continue. With the normalized cohort earning c.10% after tax, **the spread between average and best-in-class operators is economically significant, rather than marginal**. This is why our model treats cloud as a distinct layer from hardware, even though many companies participate across both activities.

Stock implications

- **Alibaba** is the clearest listed expression of cloud value capture in our coverage. Model/infra co-design is the end state of the progression above, and Alibaba is one of the few operators positioned to attempt it across silicon, cloud and models simultaneously. Our c.26.6% after-tax incremental IRR reconstruction reflects an assumption of stronger revenue yield and faster utilization than the industry cohort; **higher cloud EBITA margin, lower serving cost per workload and rising paid model usage are the reported metrics that would validate it**.
- **Phancy** is a direct coverage case for the optimization layer. The key question is whether HAMi, ModelHub and related capabilities can turn cross-chip scheduling, model adaptation and resource optimization into reusable software. Successful productization could shift part of the company's economics from capacity-heavy deployment toward optimization-heavy software, with evidence expected in gross margin, revenue per employee, deployment efficiency and reuse across customers.
- **SenseTime** represents another route through SenseCore. Running SenseNova alongside other AI workloads creates opportunities to optimize infrastructure against real production models. The longer-term value depends on how much of this capability can be generalized beyond internal workloads, with utilization, external adoption and margin progression among the relevant indicators.

Models + applications: The downstream pool where intelligence becomes customer spending

Models + applications forms the third economic layer. This is where cloud-delivered compute becomes paid intelligence, enterprise deployment, software usage and completed workflows. The layer is economically broad, so we retain three sub-layers for stock analysis: **foundation models**, **middleware / harness / enterprise deployment**, and **applications**. These sub-layers share the same downstream revenue and profit pool in our model, while competing for different parts of customer spending and relying on different forms of scarcity.

The distinction matters because value can move within the downstream pool even when the aggregate models + applications share rises. Foundation models capture value through differentiated intelligence and inference economics; middleware and harness providers monetize the gap between general intelligence and production deployment; applications monetize distribution, proprietary context, workflow ownership and execution. The relative share among these sub-layers depends on **intelligence scarcity, workflow ownership and control of monetization**.

Foundation models: Intelligence consumption remains the core growth driver

Foundation models turn compute into intelligence. Continued improvement in reasoning, coding, long context, multimodality and agentic capability expands the range of tasks models can perform, while declining inference costs increase the number of workflows that can economically use them.

This creates two simultaneous demand drivers. Frontier models move into more complex tasks where customers are willing to pay for higher capability; mature model capability becomes cheaper and expands into higher-volume use cases. Agents amplify both by making repeated model calls, consuming long contexts and operating across multi-step workflows.

Open-weight models can accelerate adoption through CSPs and inference platforms, private deployment and external capacity. Open distribution also increases price transparency and makes switching easier. Official endpoints can retain differentiation through faster model updates, serving optimization, routing, caching, enterprise support and proprietary high-speed or premium variants.

Model companies are also increasingly building proprietary coding and agent products. These products can deepen usage and create more direct customer relationships, while their economics remain tied to underlying model competitiveness. A harness becomes materially more valuable when the provider's intelligence is consistently strong enough to influence model choice.

The training cost line is what makes this sub-layer's profit path back-ended. In our model, training and model R&D absorb 45% of cloud revenue in 2026E – US$9.0bn – against downstream gross profit of only US$4.7bn, which is why the layer loses money before 2028E. Training reaches US$37.4bn by 2030E, barely higher in absolute terms, while downstream gross profit has grown to US$211.7bn: The layer turns profitable because the revenue base grows past the training line, not because training spend falls. For individual model companies, the equivalent question is whether gross profit grows fast enough to absorb continued model development without external funding.

For foundation models, the central operating questions remain **relative capability, capability-to-cost, token and ARR growth, pricing, official-endpoint economics, harness adoption and gross margin**.

Stock implications

- **Zhipu is our preferred differentiated-intelligence exposure, and we remain Overweight.** The key question is whether relative capability leadership can be sustained while commercial usage scales. Stronger model capability should support paid API growth and pricing, while proprietary harnesses, serving optimization and enterprise deployment can lower cost per token and deepen customer capture beyond the base API layer. **The specific test is gross profit sufficient to absorb model development.**
- **MiniMax remains at Neutral.** It offers similar exposure to intelligence-consumption growth, with a higher burden of proof on relative capability and pricing power. Its multimodal and agentic positioning supports strong usage growth, while durable economics depend on whether model performance remains competitive enough to sustain demand and attractive gross-profit conversion.
- **Alibaba** participates in this sub-layer through its model franchise, with the distinguishing feature being integration: models developed alongside owned silicon and cloud serving capacity, and distributed through an existing enterprise customer base.

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

Middleware / harness / enterprise deployment: Monetizing the production gap

General-purpose models do not automatically become production enterprise systems. Real-world deployment requires access to enterprise context, permissions, governance, integration, model routing, evaluation, reliability and operational systems. Middleware, harnesses and FDE monetize this implementation gap inside the broader models + applications layer.

The size of this sub-layer depends partly on how much complexity persists as models improve. Generic prompting and basic workflow construction can become easier to reproduce, while complex enterprise environments still require integration across proprietary information, permissions, multiple systems, regulatory requirements and operational execution.

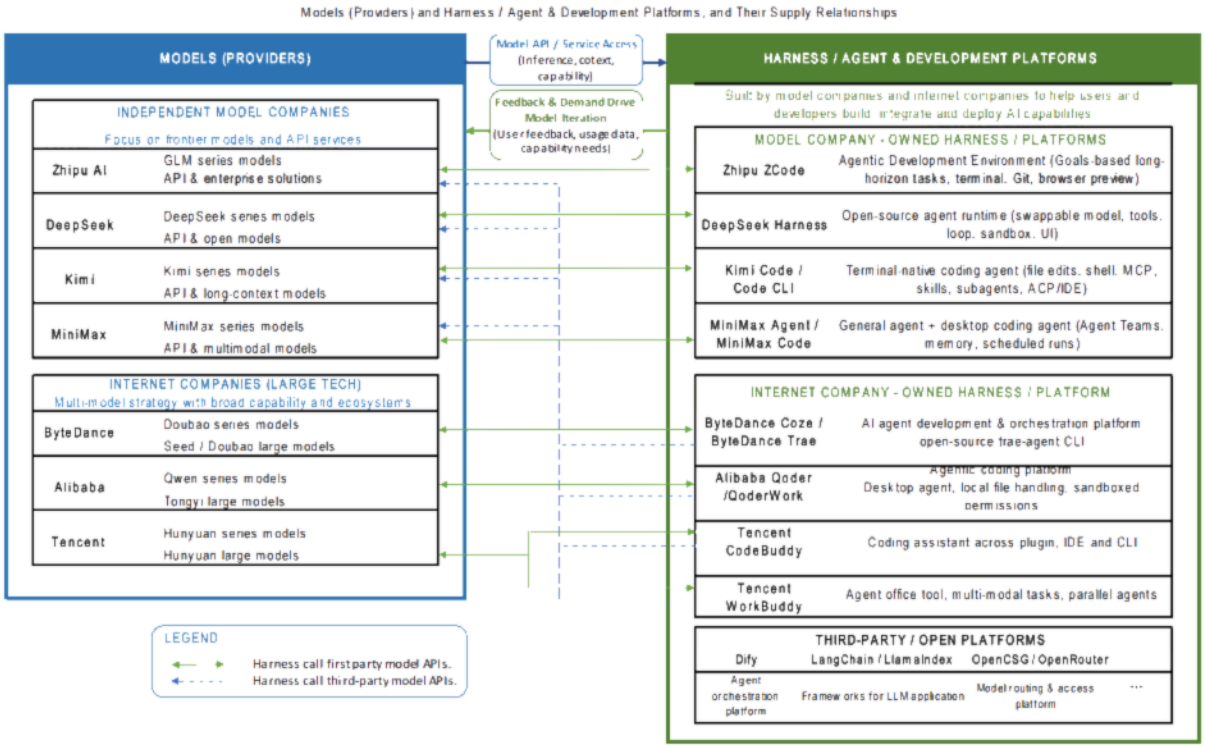
The economic opportunity improves when repeated implementation work becomes reusable software:

Custom project → Reusable module → Standardized deployment → Usage-based software

This is particularly relevant for **Xunce**, which we initiate on at Overweight. The customer owns the underlying enterprise data and context; Xunce's opportunity comes from monetizing the implementation gap around that context through governance, integration and production deployment. Token/TokenOS, reusable modules and cross-industry deployment can determine whether implementation-heavy demand evolves toward more scalable software economics. We therefore track **module reuse, deployment time, revenue per employee, cross-industry gross margin, usage/token revenue, NRR and CFO/FCF**.

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

Figure 25: China model and Internet companies harness ecosystem map



Source: Z.AI, DeepSeek, Kimi, MiniMax, Bytedance, Alibaba, Tencent, Dify, LangChain, LlamaIndex, OpenCSG, OpenRouter: company websites (as of Sep 11, 2026), J.P. Morgan estimates.

Applications: Workflow ownership determines how much downstream value is retained

Applications connect intelligence to end users, workflows and business outcomes. AI can materially increase the amount of work software addresses as agents move from assisting employees toward performing tasks directly. This expands potential spending from traditional IT budgets and software seats toward a broader labor and service pool.

The opportunity varies across application categories. **Systems of record and vertical workflow software** participate in operational processes where AI can extend software from recording and assisting work toward completing a larger share of the workflow. **Productivity and creation applications** can expand from individual software functions into broader knowledge-work, creation and design workloads. **Consumer and commerce platforms** combine distribution with transaction execution, allowing AI activity to convert into advertising, merchant or transaction revenue rather than software fees.

Within the broader models + applications pool, application economics improve when falling intelligence costs combine with durable distribution, proprietary context and control of workflow or execution. This is also where agent adoption creates the clearest tension between seat compression and machine-generated usage, making pricing model and gross-profit conversion critical.

The competitive pressure runs in both directions, which is central to several of our Neutral ratings. Improving general models expand the workloads applications can address while simultaneously making individual application features easier to replicate. An application benefits on net only where the assets an agent still requires – live state, permissions, proprietary context, execution – remain necessary after model capability improves.

Stock implications

- **Tencent** is our preferred application exposure. Distribution, proprietary context and transaction execution give it direct access to existing consumer and commercial activity at a scale few competitors can replicate. As model costs fall, these assets can support more workflows, transactions and advertising or merchant revenue. **We believe these AI application assets are largely unpriced**, and we focus on incremental commercial profit and paid AI adoption after development and inference costs, tying our valuation view to reported financial results within the group.
- **Kingsoft Office** can expand beyond document creation and editing into a broader knowledge-work pool spanning research, writing, presentation, collaboration and information processing. Its large user base gives it direct exposure to rising AI consumption in everyday productivity workflows, while access to increasingly interchangeable intelligence can expand paid use while lowering the cost of serving it. We assume coverage at Overweight.
- **JST** has exposure to a large pool of merchant operating work across orders, inventory, warehousing, logistics and platform operations. Agents can increase the amount of this work addressed through software and create additional usage around transaction and operational workflows. **We assume coverage at Neutral: The workflow assets are strong, but we believe subdued e-commerce conditions, growth expectations and valuation offset them.**
- **Meitu** has exposure to a large pool of image, video and commercial content creation work across photo editing, design, marketing and e-commerce visual production. AI lowers the cost of generating and editing visual content, expanding the share of this work addressed through software. **We initiate at Neutral: Expanding creation workloads are a genuine tailwind, but improving general models make individual features easier to replicate, creating uncertainty over incremental growth and monetization, in our view.**
- **Kingdee** can participate in broader enterprise automation across finance, supply chain, HR and other core operating processes. AI can extend ERP from employee software toward a larger share of work performed within enterprise workflows. Enterprise records and permissions provide durability; we assume coverage at Neutral pending clearer evidence of AI-driven revenue acceleration.
- **Manycore** can extend from professional design software into a broader pool of automated design, visualization, spatial content and design-to-production workflows as generative and spatial intelligence improve. Design workflows and spatial context support differentiation; we assume coverage at Neutral, as commercialization remains early.
- **Yonyou** has similar exposure across large-enterprise finance, supply chain, HR and operational workflows. The opportunity is sizeable if AI increases enterprise spending on automated work, but its enterprise assets require stronger consolidated growth, profitability and cash generation. We assume coverage at Underweight.

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

Investment framework: Opportunity × Scarcity × Value Capture

We assess AI stocks through **Opportunity × Scarcity × Value Capture**. **Opportunity determines how large the economic pool around a company's workloads can become; Scarcity determines which capabilities remain difficult to replicate or substitute; Value Capture determines how much of those economics ultimately convert into customer spending, gross profit, cash flow and returns.** Valuation is the final filter between attractive industry economics and attractive stock returns.

Table 5: Opportunity × Scarcity × Value Capture framework

	Core question	What we look for
Opportunity	How large can the economic pool around the company's workloads become?	Intelligence/compute consumption, addressable workflow, labor/service spend, agent work, new economic activity
Scarcity	What remains difficult to replicate or substitute?	Frontier capability, compute optimization, proprietary/live state, context access, workflow complexity, permissions, execution
Value Capture	How much of the opportunity ultimately accrues to the company?	Monetization, workflow/customer retention, ARPU/ACV, usage revenue, NRR, gross profit, operating leverage, CFO/FCF, incremental ROIC

Source: J.P. Morgan.

Opportunity: Start with the economic pool around the workload

The >US$1.4tn end-demand estimate discussed above provides the industry anchor. At the company level, Opportunity determines how much of this pool sits around the workloads each company serves and how much additional activity AI can create around them.

For **hardware and cloud**, Opportunity is primarily a function of aggregate compute demand, installed capacity, utilization and intelligence consumption. Within **models + applications**, the relevant pool varies by sub-layer: foundation models depend on paid intelligence consumption; middleware / harness / enterprise deployment depends on the scale and persistence of the production implementation gap; and applications depend on how much customer workflow, labor and service spending AI can address.

Agents materially increase the relevance of labor economics for applications. Traditional software monetization is largely tied to IT budgets and employee seats; when an agent performs work directly, customer willingness-to-pay can increasingly reference the cost or value of the task performed. Seat counts can also decline as automation increases output per employee, making the monetization model important alongside the size of the underlying workload.

Most of our coverage therefore screens positively on Opportunity. **Stock differentiation comes primarily from Scarcity and Value Capture.**

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

Scarcity: Identify what remains difficult to replicate or substitute

Scarcity determines where durable economic value can form. Today's AI stack contains several simultaneous sources of scarcity, and **we expect differentiation to migrate from capacity and frontier intelligence toward bottlenecks in deployment and execution** over the forecast period.

The first major scarcity pool formed around compute capacity. Frontier model progress then created scarcity around differentiated intelligence. As capability and supply improve, mature intelligence becomes cheaper and more interchangeable, while stronger models continue moving into harder tasks and establish a new frontier. **The relevant scarcity pool therefore evolves continuously, rather than settling.**

At the upstream level, advanced compute, memory, networking, power and deployable capacity remain relevant constraints, while efficient scheduling, heterogeneous optimization and serving create additional differentiation. At the model layer, **frontier intelligence** remains scarce when only a limited number of providers can complete the most demanding tasks at acceptable cost and reliability.

As capacity expands and intelligence costs decline, additional scarcity forms around **efficient serving, proprietary context, permissions, workflow and execution**. A model may understand what an enterprise wants to accomplish while still requiring access to live data, authorization, business rules, connected systems and execution capabilities to complete the task. **A single workflow may therefore require an advanced model, efficient serving, proprietary enterprise context and permission to act** – with different companies controlling each element.

This dynamic is especially important for the relative economics of models and applications. **Where model capability materially changes task outcomes, differentiated intelligence can retain pricing power.** As intelligence becomes easier to source, value increasingly depends on assets outside the model. Applications that control these assets can source intelligence as an input while retaining the customer relationship and a larger share of the economic spread.

At the same time, better models continuously absorb functionality that historically belonged to applications. Generic features, content generation, analytics and parts of software configuration become easier to replicate. **Application defensibility therefore depends increasingly on the portion of the workflow that remains difficult to reproduce through model capability alone**, and applications need to demonstrate that workflow access produces fees, transactions or higher retention.

Industry boundaries become fluid as companies extend toward adjacent scarcity pools

Because the scarcity pool moves, companies do not stay inside the layer they started in. Each of our three layers is currently extending toward the adjacent bottleneck and, more importantly, toward the adjacent charging point. This is the most important structural dynamic in the sector, and it is why layer labels are a starting point for analysis rather than a conclusion.

- **Cloud providers are moving from raw capacity toward heterogeneous optimization, model-aware serving and model distribution.** Selling compute by the hour is the least defensible position in the chain once capacity is abundant; selling throughput, portability across accelerators and access to models is considerably more so.

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

- **Model companies are extending in both directions** – upstream into serving efficiency and infrastructure optimization to lower cost per token, and downstream into proprietary harnesses, agents and enterprise deployment to hold the customer relationship rather than supplying an interchangeable input.
- **Middleware providers are productizing implementation expertise and moving closer to workflow execution**, converting project-based delivery into reusable software that can be sold repeatedly.
- **Applications are embedding and routing multiple models while exposing proprietary data, permissions and transaction capabilities to agents** – sourcing intelligence as a commodity input while making their own state and execution the necessary component.

The equity value of these moves depends entirely on whether they change what customers pay and how much economics the company retains. A proprietary harness can increase model usage and deepen the customer relationship. Infrastructure optimization can raise utilization and gross profit per unit of compute. Enterprise middleware can turn repeated implementation work into reusable software. An application can combine falling intelligence costs with proprietary workflow and distribution to expand monetization. Each of these is a genuine economic move; each is also frequently announced without producing any of these results.

The relevant evidence is therefore financial, not strategic: Higher utilization, recurring revenue, gross profit per workload and stronger retention – not the announcement of a capability. We map the movements across our coverage as follows:

Table 6: Company movements map

Movement	Coverage examples	What would confirm it
Cloud → optimization and serving	Alibaba, SenseTime, Phancy	Utilization, optimization attach, serving cost per workload, cloud EBITA margin
Cloud → silicon and co-design	Alibaba	Deployment scale of internal silicon, cost per served workload, cloud gross margin
Model → infrastructure efficiency	Zhipu, MiniMax, Alibaba	Cost per token, official-endpoint gross margin
Model → harness, agents and deployment	Zhipu, MiniMax	Harness adoption, paid API growth, enterprise contract revenue
Middleware → reusable software	Xunce, Phancy	Module reuse, deployment time, revenue per employee, usage revenue
Application → agent-callable state	Tencent, JST, Kingdee, Yonyou	Transaction/action volume, usage revenue, NRR, gross profit after AI costs
Application → multi-model routing	Kingsoft Office, Meitu, Manycore	AI paid conversion, ARPU, incremental gross margin as model prices fall

Source: Alibaba, SenseTime, Phancy, Z.AI, MiniMax, Xunce, Tencent, JST, Kingdee, Yonyou, Kingsoft Office, Meitu, Manycore; 2025 annual reports, company websites (as of Sep 11, 2026), J.P. Morgan estimates.

Agents reinforce this dynamic, rather than resolving it. A general-purpose agent can increasingly become the user entry point while continuing to consume the data, permissions, software functions and execution capabilities of underlying applications. Human interaction with individual applications can therefore decline at the same time as machine-generated software consumption rises. This creates potential charging points around usage, tasks, actions, transactions and outcomes, **with value distributed according to who controls the customer**

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

relationship, proprietary context and execution – not according to who did the work. We develop this through the Claudeforce case study below.

Assessing model attackability

For applications and middleware, we assess model attackability across six dimensions:

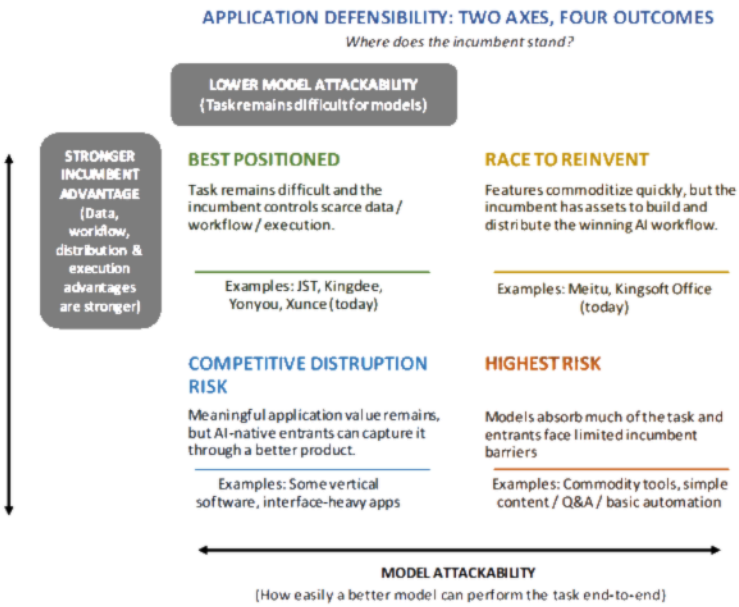
Table 7: Six dimensions of model attackability

Dimension	Higher attackability	Lower attackability / stronger residual scarcity
Outcome verifiability	Clear right/wrong outcome with rapid feedback	Subjective or delayed outcome
Workflow complexity	Single-step standardized digital task	Multi-step workflow spanning systems and continuous updates
Human / organizational interaction	Individual digital task	Sales, management, implementation and complex services
Strategic / game dynamics	Static problem with stable optimal answer	Negotiation, pricing, markets and competitive strategy
Proprietary data	Public or readily approximated information	Customer transactions, internal records, proprietary behavioral/industrial data
System / workflow / execution	User can take model output and leave	Completion requires live information, permissions, system access or physical execution

Source: J.P. Morgan estimates.

The same task can face a second source of disruption from AI-native competitors. An ERP workflow may remain difficult for a general model to replace end-to-end while an AI-native enterprise software vendor builds a better agent-based product around the existing system. **Model attackability and competitive attackability therefore need to be assessed separately.**

Figure 26: Application defensibility: Two axes, four outcomes



Source: Xunce, JST, Kingdee, Yonyou, Kingsoft Office, Meitu; 2025 annual reports, company websites (as of Sep 11, 2026), J.P. Morgan estimates.

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

Case study: Claudeforce – from user interface to proprietary context and execution

Recent developments in enterprise AI provide an early example of how application scarcity can change as agents perform more of the work. In August 2026, Salesforce and Anthropic expanded their partnership through Claudeforce. The first product, Salesforce in Claude, allows users to access Salesforce data and complete sales workflows from Claude through 37 pre-built skills, including meeting preparation, pipeline review, account planning and record updates. Salesforce exposes the underlying data and actions through Headless 360, MCP and APIs, with existing permissions, governance and business rules continuing to apply.

A salesperson can therefore ask Claude to review the pipeline, identify deals at risk and update next steps. Claude can perform the analysis and determine the required actions, while Salesforce provides the latest customer and opportunity data, determines which information the employee can access and records the resulting updates. **The interface through which the employee completes the task can change while the underlying application remains involved in providing proprietary context, permissions and execution.**

Similar developments are emerging across enterprise software. ServiceNow is opening workflow actions to third-party agents through Action Fabric; Workday allows external agents to access HR and finance capabilities through Agent-Ready Tools and MCP; Atlassian allows agents to access and act on Jira and Confluence; and Box allows external AI agents to work with enterprise content under existing access controls. These examples point to a broader shift toward making enterprise data and actions callable by both native and third-party agents.

Our first read-across is that agent adoption can change where application scarcity sits. As models absorb more search, analysis, reporting and user interaction, the interface itself becomes less differentiated. Scarcity increasingly resides in what an agent still needs to complete the work: proprietary and live enterprise state, permissions, business rules, integrations and execution. For **Tencent**, this includes consumer and merchant relationships, proprietary context and transaction execution; for **JST**, orders, inventory, platform connectivity and commerce execution; for **Kingdee and Yonyou**, financial and operational records, permissions, approval rules and transaction workflows; and for **Xunce**, access to enterprise context, governance and production systems. The relevant question increasingly becomes not whether employees continue to operate the application directly, but **which application assets remain necessary when an agent performs the task**.

This distinction also changes how we think about software usage. **Lower human interaction does not necessarily imply lower underlying software consumption.** One employee supported by agents can generate more data retrievals, record updates, workflow actions and transactions than several employees operating the same systems manually. Agent adoption can therefore create two opposing forces: fewer employees and potentially fewer traditional seats, but substantially more machine-generated activity through the underlying software.

The second read-across is that this can expand the application revenue pool while changing the unit of monetization. Traditional SaaS revenue is predominantly linked to human users through seats and licenses. Agentic workloads create additional charging points around **usage, actions, transactions, completed workflows and outcomes**. An application can therefore participate in a larger amount of work even if employee counts or direct human usage grow more slowly.

This is particularly relevant in China, where linking software monetization to work performed provides a potential path beyond traditional IT budgets toward the much larger labor and service pool being automated.

The transition is unlikely to be linear. **Seat compression can precede machine-usage monetization**, creating an initial period in which agent adoption increases underlying workload without generating a commensurate increase in application revenue. The outcome should differ by business model. Transaction- or usage-linked applications can participate more directly as machine activity rises, while predominantly seat-priced vendors need to establish AI premium SKUs, usage pricing or other charging mechanisms before higher system activity translates into revenue.

Revenue upside also does not translate mechanically into profit upside. Agentic workflows introduce model and inference costs that were much less relevant to traditional SaaS economics. When AI functionality is bundled into existing subscriptions or usage scales ahead of pricing, higher activity can dilute incremental gross margin. The financial test is therefore not only whether AI increases application usage or revenue, but whether incremental monetization exceeds model, inference and deployment costs.

These economics can improve as the model market develops. Applications can route simpler tasks toward cheaper models, reserve frontier intelligence for workflows where capability matters and benefit from declining inference costs as model supply expands. **Falling intelligence costs can therefore benefit applications twice: More workflows become economical to automate, while the cost of intelligence embedded in each workflow declines.** If customer pricing remains linked to the value of the workflow or outcome, the spread between customer willingness-to-pay and model cost can widen.

This creates a dynamic relationship between models and applications. **When differentiated intelligence remains scarce, model providers can capture a larger share of incremental AI economics**, particularly for workflows where model quality materially affects outcomes. As several models reach acceptable capability and become easier to substitute, enterprises and application vendors can route workloads across providers, putting pressure on model pricing and shifting more bargaining power toward proprietary context, workflow and execution.

The reverse is also possible. An external agent can control where users initiate work, decide which models and tools to call and become the destination for incremental AI spending. **The underlying application can therefore process substantial or even growing activity without capturing a proportional share of the economics.** Operational importance, usage growth and value capture can move differently.

The case therefore illustrates the distinction between Scarcity and Value Capture. Proprietary context, permissions and execution determine what remains difficult for an agent to replace or bypass; they do not by themselves determine who earns the incremental revenue and profit.

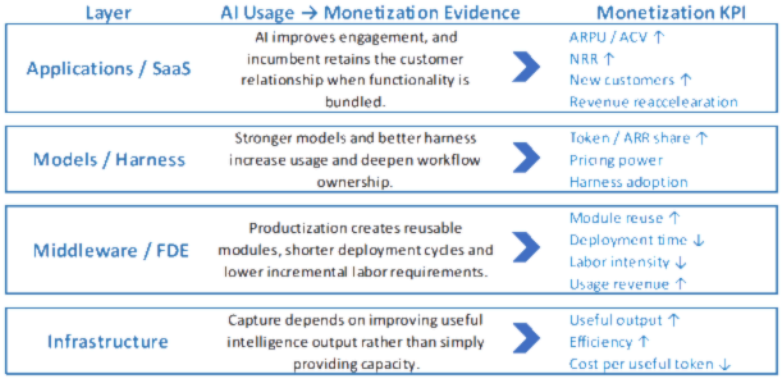
Olivia Xu
(86-21) 6106 6138
olivia.w.xu@jpmorgan.com

Value Capture: Scarcity has to translate into company economics

A scarce capability creates equity value when it changes what customers pay and allows a company to retain attractive economics after serving that demand. **Value Capture therefore spans monetization, customer and workflow retention, gross profit, cash conversion and returns on incremental capital.** Industry position and boundary movement matter because they determine which company controls the charging point and how much of the end-customer spending remains after paying other layers of the stack.

The boundary movements set out under Scarcity are where this is decided in practice. An infrastructure provider deepens capture by improving intelligence output per unit of compute; a model provider expands capture through a successful proprietary harness; middleware increases capture by converting implementation expertise into reusable production modules; an application improves capture by monetizing a larger amount of workflow activity while using models as an input. The distinction from Scarcity is that these moves establish **access** to the workload – Value Capture asks how much of the resulting customer payment the company actually retains after paying the other layers of the stack.

Figure 27: Monetization of AI usage for companies in different layers



Source: J.P. Morgan estimates.

For models and applications specifically, the interaction between **intelligence scarcity and workflow ownership** provides a useful starting point:

Table 8: Interaction between intelligence scarcity and workflow ownership

	Workflow mainly owned by model	Workflow mainly owned by application
Intelligence remains scarce	Model captures the highest value through frontier capability plus a proprietary harness	Model retains strong pricing power; application retains workflow margin
Intelligence commoditizes	Model-owned harness can preserve the customer relationship	Application can capture more value as models become lower-cost, dynamically routable inputs

Source: J.P. Morgan estimates.

Agent adoption adds a dynamic monetization dimension to this framework. Traditional SaaS revenue is predominantly linked to human users and can be simplified as **seats × price per seat**. If agents perform more work per employee, seat demand can come under pressure. At the same time, machine activity can grow much faster than human usage as agents repeatedly retrieve data, update records, trigger workflows and complete transactions. A lower number of human users can therefore coexist with substantially higher underlying system activity.

This creates additional charging points through **AI premium SKUs, usage-based pricing, per-action fees, transaction fees and outcome-based pricing**. The revenue outcome depends on whether vendors can convert rising machine activity into incremental customer spending faster than any decline in seat revenue. **Applications already linked to transaction or workflow volume have a more direct route to this economics**, while predominantly seat-priced vendors need to establish a new monetization mechanism.

The profit impact can differ materially from the revenue impact. AI introduces model and inference costs that were much less relevant to traditional software economics. During the early adoption phase, AI usage can rise faster than monetization, particularly when AI functionality is bundled into existing subscriptions. Revenue can therefore improve while incremental gross margin remains weak. Gross profit economics become more attractive as vendors charge directly for AI usage, improve model routing and inference efficiency, or benefit from lower underlying model prices.

A third pool can accrue to the **agent or harness** itself. If an agent controls where users initiate work, which models and applications are called and where incremental AI spending is directed, it can capture part of the economics even when the underlying SaaS system remains heavily used. An application can therefore retain its role as a system of record while losing part of the incremental wallet around the workflow.

We therefore see value distribution as dynamic across three variables: intelligence scarcity, workflow ownership and control of monetization. In the early phase, incremental spending can skew toward models and agent products as enterprises add a new AI budget on top of existing software. As machine usage scales, applications can capture more if they charge for usage, actions or transactions. If intelligence becomes more interchangeable, bargaining power can shift further toward applications that control proprietary context and execution. Conversely, a model or agent that sustains differentiated intelligence and owns the customer workflow can retain a larger share of the economics.

Evidence of Value Capture: From monetization to financial returns

We use observable operating indicators to determine whether theoretical positioning is translating into actual Value Capture.

For **applications**, the clearest evidence is **incremental ARPU/ACV, AI paid conversion, usage revenue, NRR, workflow penetration, new customers and revenue reacceleration**. As agents scale, we would increasingly add **API/tool calls, agent actions, transaction or workflow volume and machine-generated usage** where disclosed. The key financial test is whether incremental AI revenue exceeds any seat cannibalization and generates gross profit after model and inference costs.

For **foundation models**, we focus on **token and ARR growth trajectory, pricing, relative capability, inference economics and proprietary harness adoption**. Strong harness adoption indicates that a model provider is extending its position from intelligence supply into customer workflow, while sustained capability leadership supports a larger model take rate.

For **middleware**, **module reuse, deployment efficiency, cross-industry replication and usage-based revenue** show whether implementation expertise is becoming scalable software economics. For **infrastructure and cloud**, **utilization, optimization attach, serving efficiency and intelligence output per unit of compute** show whether economics are extending beyond raw capacity.

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

Agent adoption therefore changes both the unit of usage and the unit of capture. **Human seats can decline while machine activity rises; revenue can move from licenses toward usage and transactions; and gross profit depends on the spread between customer monetization and the cost of intelligence required to deliver the work.**

Valuation: Attractive economics still need to offer stock upside

Financial conversion sits inside Value Capture. The path we look for is: **Revenue reacceleration → Gross profit growth → Operating leverage → CFO/FCF → Returns on incremental capital**. AI revenue can carry meaningful inference, compute, deployment and support costs, so top-line monetization provides only the first layer of evidence. Upstream and model companies need to show attractive economics relative to capital intensity; middleware needs productization that allows revenue to scale faster than delivery labor; and applications need AI spending to generate incremental gross profit after model and inference costs.

Valuation then determines whether these economics still offer attractive per-share upside. A strategically attractive AI position can generate limited stock returns when the market already capitalizes a large share of future value – which is the primary reason JST enters coverage at Neutral despite workflow assets we regard as valuable.

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

Table 9: Applying the framework to our coverage

Company	Rating	Opportunity	Company-specific scarcity	Value Capture / evidence	Financial / valuation read-through
Alibaba	Overweight	Compute demand across all three layers	Integrated silicon (T-Head), cloud capacity and model franchise; model/infra co-design potential	c.26.6% after-tax incremental project IRR vs 10% normalized cohort; higher cloud EBITA and lower serving cost per workload	Cash conversion determines shareholder benefit; preferred integrated exposure
Tencent	Overweight	Consumer and commercial activity at scale	Distribution, proprietary context and transaction execution	Incremental commercial profit and paid AI adoption after development and inference costs	AI application assets we believe are largely unpriced within group valuation
Zhipu	Overweight	Strong intelligence-consumption growth	Competitive frontier capability; relative leadership requires continued validation	Paid API growth; serving optimization and proprietary harness can deepen customer capture	Gross profit sufficient to absorb model development; preferred China model exposure
SenseTime	Overweight	Training, inference and enterprise AI demand	Broad exposure across compute, models and applications	SenseCore connects capacity with production AI workloads; external adoption is the key test	Profitability and cash trends improving
Phancy	Overweight	Enterprise AI and heterogeneous-compute demand	Optimization and deployment address persistent China infrastructure complexity	Scheduling, optimization and deployment provide multiple monetization points	Current AI demand supports base case; operating conversion remains key
Xunce	Overweight	Enterprise AI deployment and workflow automation	Context access, integration depth, governance and production complexity	Token/TokenOS and module reuse can shift economics from implementation toward scalable usage	AI already contributing to revenue; productization can improve margin and cash conversion
Kingsoft Office	Overweight	Large knowledge-work pool	Distribution, document state/context and native workflow; medium model attackability	AI paid conversion, WPS 365 NRR and broader workflow penetration	Existing profitability and cash support AI investment; incremental AI gross profit provides upside
MiniMax	Neutral	Strong intelligence-consumption growth	Relative capability and pricing power less established	Meaningful usage and harness optionality; stronger relative capability needed to sustain capture	Capability, pricing and gross-profit conversion remain the re-rating triggers
JST	Neutral	Large commerce workflow and labor pool	Live commerce state, system of record, permissions, integrations and execution	AI can expand monetization from software usage toward work performed while retaining execution	Subdued e-commerce conditions, growth expectations and valuation offset strong workflow assets
Meitu	Neutral	Large visual-creation and marketing opportunity	Distribution and creator workflow; individual features increasingly replicable by general models	Demonstrated AI monetization provides an existing charging point; incremental growth is less certain	Uncertainty over incremental growth and monetization limits upside from current levels
Kingdee	Neutral	Large enterprise AI opportunity	Enterprise state, permissions, business rules and system of record	AI attach, incremental ACV, NRR and workflow usage remain the key tests	Further re-rating requires clearer AI-driven revenue acceleration
Manycore	Neutral	Large design and spatial intelligence opportunity	Professional workflow, proprietary 3D/spatial assets and design-to-production integration	Spatial Intelligence can extend workflow capture; commercialization remains early	Core growth, enterprise NRR and new AI revenue need further evidence
Yonyou	Underweight	Large enterprise AI opportunity	Enterprise state, permissions and complex workflow remain strategically valuable	AI product and contract activity has yet to show strong consolidated capture	Revenue, margin and cash conversion continue to lag the AI narrative

Source: Alibababa, Tencent, Z.AI, MiniMax, SenseTime, Phancy, Xunce, JST, Kingdee, Yonyou, Kingsoft Office, Meitu, Manycore; 2025 annual reports, company websites (as of Sep 11, 2026), J.P. Morgan estimates.

Preferred exposures: Alibaba, Tencent and Zhipu

Our three preferred exposures are deliberately positioned at different points in the value chain, reflecting the sequencing in our industry model. **Alibaba captures the earlier-recognized upstream and cloud economics with company-specific returns above the industry cohort; Tencent holds application assets we believe are more valuable as model costs fall; and Zhipu is the differentiated exposure whose economics depend on capability persisting.**



Alibaba: Integrated chip, cloud and model exposure with above-cohort capacity returns

The investment case rests on integration, rather than participation in any single layer. T-Head broadens Alibaba's hardware participation beyond the buyer side, while cloud and its model franchise connect infrastructure to customer demand. Few operators in China can attempt model/infra co-design – the end state of the cloud differentiation progression – across owned silicon, owned serving capacity and owned models simultaneously.

The evidence we would need is what to disclose financially: We expect Alibaba's business spread across the supply chain to show up through **higher cloud EBITA margin, lower serving cost per workload and growing paid model usage**. Cash conversion then determines the shareholder benefit – a high project IRR on incremental capacity does not create equity value if it is offset by a rising investment base that defers free cash flow. We treat cloud segment margin progression and group FCF as the two disclosures that would confirm or challenge the reconstruction.

Downside risks include competitive pricing compressing cloud revenue yield toward the industry cohort; internal silicon substituting external purchases without improving unit economics; and sustained capacity investment deferring cash returns beyond our forecast horizon.

Tencent: Scarce AI application assets we believe are largely unpriced

Tencent is our preferred expression of the view that application scarcity survives model improvement. Distribution, proprietary context and transaction execution give Tencent direct access to existing consumer and commercial activity. These are precisely the assets our Claudeforce analysis identifies as remaining necessary when an agent performs the work – live state, permissions, business rules and the ability to execute – rather than the interface functionality that general models increasingly absorb.

Falling model costs are a tailwind, rather than a threat to this position. As the cost of intelligence embedded in each workflow declines, more workflows become economical to serve across Tencent's existing user and merchant base, while the cost of serving each one falls. If pricing remains linked to the value of the workflow or transaction, rather than to the intelligence input, the spread widens. This is the "benefit twice" dynamic described in our applications analysis, and Tencent's distribution scale means it applies across an unusually large activity base.

We believe these assets are largely unpriced. We believe the market value estimate that puts models + applications at c.14% of China-AI-attributable listed value reflects limited capitalization of AI-driven incremental profit within large platform businesses where AI revenue is not separately disclosed. **We focus on incremental commercial profit and paid AI adoption after development and inference costs**, tying our valuation view to reported financial results within the group, rather than to announced capability.

Downside risks include AI activity increasing engagement without a commensurate increase in advertising, merchant or transaction revenue; inference and development costs absorbing the incremental gross profit; and disclosure remaining insufficient for the market to recognize the contribution within the forecast period.

Zhipu: Differentiated intelligence driving paid usage

Zhipu is our preferred differentiated-intelligence exposure, and we remain Overweight. Sustained relative capability can support API pricing, paid calls and

customer retention, while serving optimization, proprietary harnesses and enterprise deployment can lower cost per token. The position depends on capability persisting, rather than on capability existing today – this is the layer where our framework explicitly expects scarcity to erode over time.

The specific financial test is paid API growth and gross profit sufficient to absorb model development. Our industry model illustrates why this is the binding constraint: Training and model R&D absorb US$9.0bn in 2026E against downstream gross profit of US$4.7bn, which is what makes the layer loss-making before 2028E. At the company level, the equivalent question is whether gross profit compounds fast enough to fund continued frontier participation without dilutive external funding. **Funding capacity through the earnings ramp is a selection criterion here, not a footnote**, given that approximately three-quarters of modeled downstream five-year profit arrives in 2030E.

Downside risks for Zhipu include faster convergence in model capability; price competition from open-weight distribution; and training costs rising faster than paid usage.

MiniMax remains at Neutral. It participates in the same demand pool and has demonstrated strong usage growth, but we need to see durable capability, pricing and gross profit conversion before that usage translates into a differentiated financial outcome. Usage growth and value capture can move separately, and at this stage we regard the evidence for MiniMax as weighted toward the former.

Upside risks include better-than-expected new model iteration, clearer monetization from MiniMax's multimodal portfolio, and more durable API economics.

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