Semiconductors

DeepSeek V4 pricing offer extended, signaling China's LLM inference cost advantage

OUTPERFORM (Maintain)

China Semiconductors Sector

Saiyi HE, CFA
(852) 3916 1739
hesaiyi@cmbi.com.hk

Aaron GUO
(852) 3916 3715
aaronguo@cmbi.com.hk

Kevin ZHANG
(852) 3761 8727
kevinzhang@cmbi.com.hk

DeepSeek (unlisted) released V4 on 24 Apr 2026, comprising V4-Pro (1.6T total / 49B activated parameters) and V4-Flash (284B / 13B activated), both natively supporting a 1M-token context window. On 25 Apr 2026, the company further announced a 75% promotional discount on V4-Pro API rates, from US$1.74 / 3.48 to US$0.44 / 0.87 per 1M input/output token, originally valid through 5 May and subsequently extended through May 31 2026. V4-Flash is priced at US$0.14 / 0.28 per 1M input/output tokens.

V4-Pro's standard rate now sits in the middle of the domestic flagship range, **while Chinese flagship LLM API pricing is now at 25–35% of overseas flagship rates on a blended-3:1 1 ((3 x Input + Output) / 4) basis**. DeepSeek V4's pricing indicates two structural shifts: 1) Supply-chain depth across most publicly listed Chinese GPU/GPGPU vendors lowering the inference cost basis, and 2) V4's compressed output-to-input pricing ratio (2.0x vs c.4.5x overseas leaders) pointing to a model-side architectural compression**.** Within the domestic field, there is a 5–10x price spread across domestic model tiers: 1) Commodity-tier flagships at US$0.50–1.30 blended-3:1, and 2) premium-tier at US$3.0–5.0, indicating tier-1 names retain the pricing power vs the commodity tier. We remain constructive on Chinese cloud platforms with own-model, cloud distribution and domestic chip exposure, and on selected local GPGPU vendors leveraged to inference localization.

- **We attribute the cross-border gap to two drivers: A chip-side cost edge and a model-side architecture compression.** V4-Pro benchmarks are competitive vs. Claude Opus 4.6/4.7 and are ahead of GPT-5.4 and Gemini 3.1 Pro on selected coding benchmarks (which do not fully proxy enterprise adoption, latency, stability or throughput). **On the chip side, Day-0 inference adaptation is now visible across most major Chinese GPU/GPGPU stacks for V4 and other domestic flagships**, allowing inference at a structurally lower cost basis. **On the model side, V4's output-to-input ratio at 2.0x is well below the c.4.5x overseas players' average**, attributed by DeepSeek to a hybrid attention design. We treat the model-side argument as a working hypothesis pending DeepSeek's technical report.
- **Promo extended, Ascend ramp, and WAIC as near-term catalysts, with cost basis remaining the durable element.**
  1. **V4-Pro 75% Promo extended from 5 May to 31 May 2026:** A soft signal that the promotional rate is sustainable on the current cost basis, while DeepSeek mgmt. has mentioned ability to further lower the pricing with future Ascend deployment.
  2. **Volume ramp of Huawei Ascend 950 SuperPods through 2H26, which DeepSeek mgmt. has indicated could further reduce V4-Pro's underlying inference cost basis.** Huawei has guided AI chip revenue to reach ~US$12bn in 2026, up from US$7.5bn in 2025 (+60% YoY) per *Financial Times* (FT) (1 May 2026, link); the majority of orders are for the 950PR (mass production from March 2026), with a 950DT upgrade slated for 4Q26. FT also reports DeepSeek used the 950PR for V4 with improved inference efficiency at reduced costs, providing third-party confirmation of the chip-side cost-edge argument.
  3. **WAIC Shanghai in July 2026, where next-generation domestic GPU launches are expected.** We see the underlying domestic cost basis, rather than the headline promo, as the more durable element of the China LLM pricing story, the within-domestic price spread shows tier-1 names retain pricing power.

- **Tencent and Alibaba as primary beneficiaries; Iluvatar CoreX, Biren on the supply side.** Tencent (700 HK, BUY, covered by our Internet team), with Hunyuan own-model exposure, cloud distribution, and Ascend-aligned infrastructure, and Alibaba (9988 HK, BUY, covered by our Internet team), with Qwen 3.6 own-model exposure across Max / Plus / A3B, cloud distribution, and Ascend-aligned infrastructure work. On the supply side, the V4 launch and Day-0 readiness across most Chinese chip stacks is consistent with the thesis behind our 22 Apr 2026 initiation on Iluvatar CoreX (9903 HK, BUY).
- **Key risks:** 1) NVIDIA H-series shipment resumption to China at scale would compress the Day-0 cost-of-inference gap, 2) overseas frontier providers cutting pricing further (blended-3:1 down 30%+ on Opus / GPT-5.5 class) to defend share, narrowing the cross-border spread from the top; 3) earlier-than-expected price competition across multiple tier-1 domestic LLMs (more than two vendors stacking 70%+ promotional discounts concurrently for over 30 days, or DeepSeek standard rate cut by more than 50% post-promo).

Industry Overview

DeepSeek V4 sits within the first tier of global open-source models across model capability, cost-effectiveness, and adaptation to domestic compute

We believe DeepSeek V4 delivers a relatively comprehensive capability upgrade and remains within the first tier of global open-source models. The release includes two MoE models, V4-Pro (1.6T total / 49B activated parameters) and V4-Flash (284B total / 13B activated), both natively supporting a 1M-token context window. By dimension, V4 performs notably well on coding and competitive programming benchmarks (Codeforces rating of 3,206 vs GPT-5.4's 3,168; SWE-bench Verified at 80.6%, 0.2pts below Claude Opus 4.6; LiveCodeBench at 93.5% vs Gemini's 91.7% and Claude's 88.8%).

Its agentic capabilities have moved into the same competitive range as closed-source peers; world-knowledge results trail only the leading closed-source models; and the 1M long-context window represents a meaningful point of differentiation.

Pricing matrix reveals a two-level split: 25–35% cross-border, 5–10x within-domestic

The cross-border price gap is stark and visible across every layer of the API capability stack. As detailed below, V4-Pro operates at a fraction of the cost of overseas peer's top tier LLM model across input, output, and blended-3:1 metrics.

Figure 1: Deepseek V4 Pro pricing compared to Overseas LLM flagship

Model	Input	Output	Blended-3:1 Pricing
Deepseek V4-Pro (Standard Rate)	$1.74	$3.48	$2.18
Claude Opus 4.7	$5.00	$25.00	$10.00
GPT-5.5 Pro (Short Context)	$30.00	$180.00	$67.50
Gemini 3.1 Pro (Short Context)	$2.00	$12.00	$4.50

Source: Company data, CMBIGM

This structural divergence extends across the broader market, with the domestic flagship average sitting significantly below the overseas average. As outlined in table below, the cross-border spread is most extreme at the top of the capability stack, yet it remains highly visible down to the commodity tier, where domestic models maintain a roughly 11x price advantage.

Figure 2: Pricing Range for major China and overseas LLM & DS V4 promo rate

Model	Input Range	Output Range	Note/Input Length
Domestic	Minimax M2.7 $0.3 ~ Qwen 3.6 Max $2.2	Minimax M2.7 $1.2 ~ Qwen 3.6 Max $13.2	Qwen 3.6 Max: Input(128K,256K]
Overseas	Grok 4.2 $2.0 ~ GPT 5.5 Pro $50	Grok 4.2 $6.0 ~ GPT 5.5 Pro $270	GPT 5.5 Pro: Long Context
Deepseek V4-Pro (Promo Rate)	$0.44	$0.87	Valid through May 31

Source: Company data, CMBIGM

While API pricing alone could theoretically reflect temporary promotional strategies for customer acquisition or ecosystem reasons , we view this as an observable signal of a deeper, structural cost advantage. We base this conclusion on three additional confirmations: 1) Day-0 domestic chip adaptation, 2) V4's unusually low output-to-input pricing ratio, 3) and the domestic super-node roadmap.

Figure 3: Major LLM providers' pricing matrix

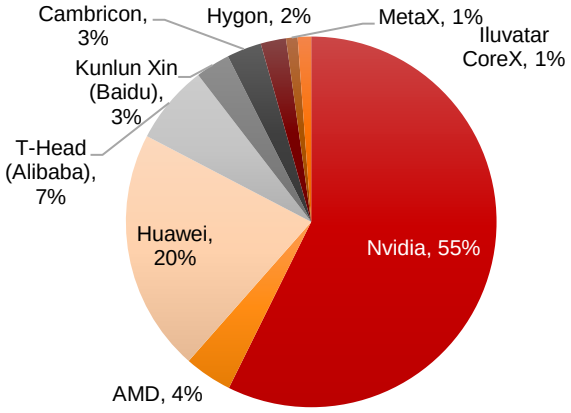
Region	Provider	Model	Flag ship	Latest	USD / 1M tokens					Index			
					Input	Cache R	Cache W	Storage (1M/h)	Output	Blended 3:1 ($)	I/O Ratio	Context Length	Output Length
CHINA	DeepSeek	**DeepSeek V4 Pro**	★	Y	$1.74	$0.01	$1.74		$3.48	$2.1750	2.00x	1M	384K
		DeepSeek V4 Flash		Y	$0.14	$0.00	$0.14		$0.28	$0.1750	2.00x	1M	384K
		DeepSeek V3.2			$0.28	$0.03	$0.28		$0.42	$0.3150	1.50x		
	MiniMax	**MiniMax-M2.7**	★	Y	$0.30	$0.06	$0.38		$1.20	$0.5250	4.00x	~200K	
		MiniMax-M2.7-highspeed		Y	$0.60	$0.06	$0.38		$2.40	$1.0500	4.00x	~200K	
		MiniMax-M2.5			$0.30	$0.06	$0.38		$1.20	$0.5250	4.00x	~200K	
		MiniMax-M2.5-highspeed			$0.60	$0.06	$0.38		$2.40	$1.0500	4.00x	~200K	
		MiniMax-M2.1			$0.30	$0.03	$0.38		$1.20	$0.5250	4.00x	~200K	
		MiniMax-M2			$0.30	$0.03	$0.38		$1.20	$0.5250	4.00x	~200K	
	GLM Knowledge Atlas	**GLM-5.1 Input [0, 32)**	★	Y	$0.88	$0.19		-	$3.51	$1.5372	4.00x	200K	128K
		GLM-5.1 Input [32+)	★	Y	$1.17	$0.29		-	$4.10	$1.9032	3.50x	200K	128K
		GLM-5-Turbo Input [0, 32)			$0.73	$0.18		-	$3.22	$1.3542	4.40x	200K	128K
		GLM-5-Turbo Input [32+)			$1.02	$0.26		-	$3.81	$1.7202	3.71x	200K	128K
		GLM-5 Input [0, 32)			$0.59	$0.15		-	$2.64	$1.0980	4.50x	200K	128K
		GLM-5 Input [32+)			$0.88	$0.22		-	$3.22	$1.4640	3.67x	200K	128K
		GLM-4.7 Input [0, 32)/输出长度 [0,			$0.29	$0.06		-	$1.17	$0.5124	4.00x	200K	128K
		GLM-4.7 Input [0, 32)/输出长度			$0.44	$0.09		-	$2.05	$0.8418	4.67x	200K	128K
		GLM-4.7 Input [32, 200)			$0.59	$0.12		-	$2.34	$1.0248	4.00x	200K	128K
	Moonshot	**Kimi K2.6**	★	Y	$0.95	$0.16	$0.95		$4.00	$1.7125	4.21x	262K	
		Kimi K2.5			$0.60	$0.10	$0.60		$3.00	$1.2000	5.00x	262K	
		Kimi K2			$0.60	$0.15	$0.60		$2.50	$1.0750	4.17x	262K	
	Qwen Alibaba	**Qwen3.6 Max Preview Input[0,128K]**	★	Y	$1.32	$0.13	$1.65		$7.91	$2.9646	6.00x	256K	64K
		Qwen3.6 Max Preview Input[128K,256K]	★	Y	$2.20	$0.22	$2.75		$13.18	$4.9410	6.00x	256K	64K
		Qwen3.6 Plus Input[0,256K]		Y	$0.29	$0.03	$0.37		$1.76	$0.6588	6.00x	1M	64K
		Qwen3.6 Plus Input[256K,1M]		Y	$1.17	$0.12	$1.46		$7.03	$2.6352	6.00x	1M	64K
	MiMo Xiaomi	**MiMo-V2.5-Pro Input[0,256K]**	★	Y	$1.02	$0.20	$1.02		$3.07	$1.5372	3.00x	1M	128K
		MiMo-V2.5 Pro Input[256K,1M]	★	Y	$2.05	$0.41	$2.05		$6.15	$3.0744	3.00x	1M	128K
		MiMo-V2-Pro Input[0,256K]			$1.02	$0.20	$1.02		$6.15	$2.3058	6.00x	1M	128K
		MiMo-V2-Pro Input[256K,1M]			$2.05	$0.41	$2.05		$6.15	$3.0744	3.00x	1M	128K
		MiMo-V2.5 Input[0,256K]		Y	$0.41	$0.08	$0.41		$2.05	$0.8198	5.00x	1M	128K
		MiMo-V2.5 Input[256K,1M]		Y	$0.82	$0.16	$0.82		$4.10	$1.6397	5.00x	1M	128K
	Doubao ByteDance	**doubao-seed-2.0-pro Input [0, 32]**	★	Y	$0.47	$0.09		$0.00	$2.34	$0.9370	5.00x	256K	128K
		doubao-seed-2.0-pro Input (32, 128]	★	Y	$0.70	$0.14		$0.00	¥3.51	¥1.4054	5.00x	256K	128K
		doubao-seed-2.0-pro Input (128, 256]	★	Y	$1.41	$0.28		$0.00	¥7.03	¥2.8109	5.00x	256K	128K
		doubao-seed-2.0-lite Input [0, 32]			$0.09	$0.02		$0.00	¥0.53	¥0.1976	6.00x	256K	128K
		doubao-seed-2.0-lite Input (32, 128]			$0.13	$0.03		$0.00	¥0.79	¥0.2965	6.00x	256K	128K
		doubao-seed-2.0-lite Input (128, 256]			$0.26	$0.05		$0.00	¥1.58	¥0.5929	6.00x	256K	128K
		doubao-seed-2.0-mini Input [0, 32]			$0.03	$0.01		$0.00	¥0.29	¥0.0952	10.00x	256K	128K
		doubao-seed-2.0-mini Input (32, 128]			$0.06	$0.01		$0.00	¥0.59	¥0.1903	10.00x	256K	128K
		doubao-seed-2.0-mini Input (128, 256]			$0.12	$0.02		$0.00	¥1.17	¥0.3806	10.00x	256K	128K
		doubao-seed-2.0-code Input [0, 32]			$0.47	$0.09		$0.00	¥2.34	¥0.9370	5.00x	256K	128K
		doubao-seed-2.0-code Input (32, 128]			$0.70	$0.14		$0.00	¥3.51	¥1.4054	5.00x	256K	128K
		doubao-seed-2.0-code Input (128, 256]			$1.41	$0.28		$0.00	¥7.03	¥2.8109	5.00x	256K	128K
		doubao-seed-1.8 Input [0, 32] & Output [0, 0.2]			$0.12	$0.02		$0.00	¥0.29	¥0.1610	2.50x	256K	32K
		doubao-seed-1.8 Input [0, 32] & Output (0.2,			$0.12	$0.02		$0.00	¥1.17	¥0.3806	10.00x	256K	32K
		doubao-seed-1.8 Input (32, 128]			$0.18	$0.02		$0.00	¥2.34	¥0.7174	13.33x	256K	32K
		doubao-seed-1.8 Input (128, 256]			$0.35	$0.02		$0.00	¥3.51	¥1.1419	10.00x	256K	32K
	Hunyuan Tencent	**Hunyuan 2.0 Think Input (32k, 128k]**	★	Y	$0.78				$3.10	$1.3579	4.00x	256K	64K
		Hunyuan 2.0 Think Input [0, 32k]		Y	$0.58				$2.33	$1.0184	4.00x	256K	64K
		Hunyuan 2.0 Instruct Input (32k, 128k]		Y	$0.66				$1.63	$0.9020	2.47x	256K	16K
		Hunyuan 2.0 Instruct Input [0, 32k]		Y	$0.47				$1.16	$0.6401	2.50x	256K	16K
		Hunyuan-TurboS			$0.12				$0.29	$0.1610	2.50x	256K	16K
		Hunyuan-T1			$0.15				$0.59	$0.2562	4.00x	256K	64K
OVERSEAS	OpenAI	**GPT-5.5 (Short Context)**		Y	$5.00	$0.50			$30.00	$11.2500	6.00x	1.05M	128K
		GPT-5.5 (Long Context)		Y	$10.00	$1.00			$45.00	$18.7500	4.50x	1.05M	128K
		GPT-5.5 Pro (Short Context)	★	Y	$30.00	-			$180.00	$67.5000	6.00x	1.05M	128K
		GPT-5.5 Pro (Long Context)	★	Y	$60.00	-			$270.00	$112.5000	4.50x	1.05M	128K
		GPT-5.4 (Short Context)			$2.50	$0.25			$15.00	$5.6250	6.00x	1.05M	128K
		GPT-5.4 (Long Context)			$5.00	$0.50			$22.50	$9.3750	4.50x	1.05M	128K
	Claude Anthropic	**Claude Opus 4.7**	★	Y	$5.00	$0.50	$6.25		$25.00	$10.0000	5.00x	1M	128K
		Claude Opus 4.6		Y	$5.00	$0.50	$6.25		$25.00	$10.0000	5.00x	1M	128K
		Claude Sonnet 4.6		Y	$3.00	$0.30	$3.75		$15.00	$6.0000	5.00x	1M	64K
		Claude Haiku 4.5		Y	$1.00	$0.10	$1.25		$5.00	$2.0000	5.00x	200K	64K
	Gemini Google	**Gemini 3.1 Pro Preview Input (0, 200k]**	★	Y	$2.00	$0.20		$4.50	$12.00	$4.5000	6.00x	1.05M	65.5K
		Gemini 3.1 Pro Preview Input (200k, ++)	★	Y	$4.00	$0.40		$4.50	$18.00	$7.5000	4.50x	1.05M	65.5K
		Gemini 3 Pro			$1.25				$10.00	$3.4375	8.00x	1.05M	65.5K
		Gemini 3 Flash			$0.50	$0.05		$1.00	$3.00	$1.1250	6.00x	1.05M	65.5K
		Gemini 2.5 Pro Input (0, 200k]			$1.25	$0.13		$4.50	$10.00	$3.4375	8.00x	1.05M	65.5K
		Gemini 2.5 Pro Input (200k, ++)			$2.50	$0.25		$4.50	$15.00	$5.6250	6.00x	1.05M	65.5K
	Grok	**Grok 4.2**	★	Y	$2.00	$0.20			$6.00	$3.0000	3.00x	2M	
		Grok 4.1 Fast			$0.20	$0.05			$0.50	$0.2750	2.50x	2M	
	Llama	**Llama 4 Maverick**	★	Y	$0.15				$0.60	$0.2625	4.00x	1M	
		Llama 4 Scout			$0.15				$0.60	$0.2625	4.00x	10M	
	Command Cohere	**Command A**	★	Y	$2.50				$10.00	$4.3750	4.00x	256K	8K
		Command R		Y	$0.15				$0.60	$0.2625	4.00x	128K	4K
		Command R7B		Y	$0.04				$0.15	$0.0656	4.00x	128K	4K

Source: Company data, CMBIGM

Localisation runs from chip-side adaptation to forward-looking super-node ramp, and model-side compression upgrade delivers an architectural cost edge

Reuters, citing IDC, reported that Chinese GPU and AI chipmakers captured about 41% of China's AI accelerator server market in 2025, while total AI accelerator shipments in China reached roughly 4.0mn units. Domestic vendors collectively shipped about 1.65mn units, versus 2.2mn for Nvidia, implying that local adoption has already moved well beyond a niche substitution story.

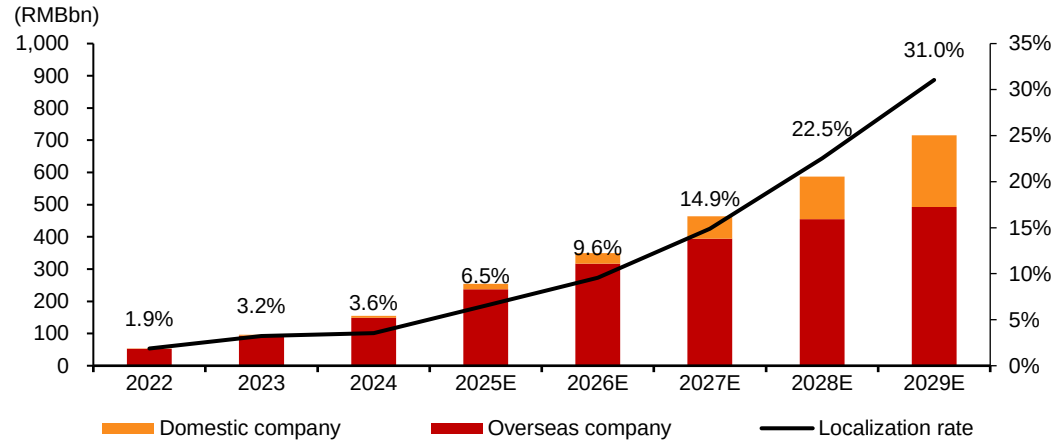
Figure 4: China's AI GPU market share by shipment in 2025



Source: IDC, Reuters, CMBIGM

Per F&S, China's GPGPU market is projected to grow at a 29.5% CAGR over 2025-29E, with domestic vendors growing at 91.4% CAGR vs 20.0% for overseas suppliers. The localization rate is projected to rise from 6.5% in 2025E to 31.0% by 2029E.

Figure 5: China's GPGPU market size and localization rate, 2022–2029E



Source: F&S, CMBIGM

Inference-side localization moves faster than training-side: 3.0% in 2024 to 37.9% by 2029E for inference, vs 4.0% to 16.8% for training, based on F&S estimates.

■ Day-0 readiness across most listed Chinese GPU/GPGPU stacks

Per the updated Day-0 adaptation matrix as of 1 May 2026, seven of nine listed Chinese GPU/GPGPU stacks now show at least one Day-0 V4-series serving path. The median Chinese vendor in the matrix delivers Day-0 across four to six model families.

Figure 6: Day-0 adaptation

Company	DeepSeek V4 Pro	DeepSeek V4 Flash	MiniMax-M2.7	MiniMax-M2.5	GLM-5.1	GLM-5	Kimi K2.6	Qwen3.6-35B-A3B	Qwen3 Series	Mimo-V2.5-Pro	Doubao-Seed	Hunyuan 3 Preview
Huawei Ascend	Day-0	Day-0	Day-0			Day-0		Day-0	Day-0			
Cambricon	Day-0	Day-0				Day-0			Day-0			
Hygon	Day-0	Day-0			Day-0	Day-0		Day-0		Day-0		Day-0
MetaX	Day-0	Day-0	Day-0	Day-0	Day-0	Day-0		Day-0				
Moore Threads	Day-0	Day-0	Day-0	Day-0	Day-0	Day-0			Day-0			
Iluvatar CoreX	Day-0	Day-0	Day-0		Day-0	Day-0				Day-0		
Biren	Day-0*	Day-0*		Day-0	Day-0		Day-0	Day-0				Day-0
T-Head												
Kunlunxin			Day-0		Day-0	Day-0				Day-0		

Source: Company data, CMBIGM; data as of 1 May 2026

We believe domestic hardware ecosystems have reached a critical inflection point, transitioning from isolated pilots to generalized industry readiness for frontier LLM inference. Most GPU vendors published Day-0 support for V4-Pro and V4-Flash within the first 72 hours, demonstrating a rapid maturation of the underlying software stacks. This swift, cross-vendor adaptation suggests that domestic chip readiness is no longer a single-vendor anomaly, but a sector-wide capability.

While the localization progress is currently concentrated on the inference side, where addresses immediate commercial pressure. In the near term, model inference costs represent a major operating expense for LLM providers. Looking longer-term, we believe domestic substitution on the training side is also progressing.

■ V4's I/O ratio at 2.0x suggests an architectural cost saving

Output-to-input pricing ratio reflects the prefill / decode cost asymmetry, and output tokens are memory-bandwidth-bound and structurally more expensive per token than input. Higher ratios pass more decode cost through to the customer; lower ratios imply either a lower underlying decode cost or a decision to absorb it.

Across the field, overseas flagship I/O ratios cluster in the 4.0x to 6.0x range, and most Chinese flagships sit in the 3.0x to 6.0x range, with selected commodity-tier ByteDance Doubao SKUs reaching 10x or higher. DeepSeek V4-Pro and V4-Flash both are priced at 2.0x, with V3.2 lower at 1.5x, well outside the field on the low side.

Figure 7: Input/Output Ratio across Major LLM

Domestic		Overseas	
Model	**I/O Ratio**	**Model**	**I/O Ratio**
DeepSeek V4 Pro	2.0x	GPT-5.5 Pro (Short Context)	6.0x
MiniMax-M2.7	4.0x	GPT-5.5 Pro (Long Context)	4.5x
GLM-5.1 (Input ⩾32K)	3.5x	Claude Opus 4.7	5.0x
Kimi K2.6	4.21x	Gemini 3.1 Pro Preview (⩽200K)	6.0x
Qwen3.6 Max Preview	6.0x	Gemini 3.1 Pro Preview (>200K)	4.5x
MiMo-V2.5-Pro	3.0x	Grok 4.2	3.0x
doubao-seed-2.0-pro	5.0x	Llama 4 Maverick	4.0x
Hunyuan 2.0 Think	4.0x	Command A	4.0x

DeepSeek attributes the V4 compression to a hybrid attention design improving long-context efficiency. The architectural specifics have not been disclosed in detail. We treat this argument as a working hypothesis pending DeepSeek's technical report, expected within roughly two weeks per prior release cadence. If the technical report does not substantiate the architectural claim, this argument weakens and the structural reading rests on the chip-side argument alone.

- **Domestic super-node buildout reinforces the localization trajectory and lowers China LLM cost basis**

The super-node concept, originally framed by NVIDIA as a tightly-coupled multi-GPU compute unit, has been picked up across the domestic compute stack through 2H25 and 1H26.

Huawei anchors the high end. Atlas 950 SuperPoD (Up to 8,192 cards, announced in September 2025, scheduled for 4Q26) and Atlas 960 SuperPoD (Up to 15,488 cards, announced September 2025, scheduled for 4Q27) are built on the proprietary UnifiedBus (灵衢) interconnect, with the earlier CloudMatrix 384 reaching around 300 deployed units across approximately 20 customers. Atlas 950 SuperCluster aggregates 64 SuperPoDs into a 500k-card cluster (4Q26), with a 1mn-card Atlas 960 SuperCluster targeted for 4Q27. Underlying silicon is the Ascend 950 series, splitting workloads across two die variants: 950 PR (1Q26, prefill-oriented, HiBL 1.0 memory) and 950 DT (4Q26, decode and training-oriented, 144GB HiZQ 2.0 memory at 4 TB/s). Commercial momentum is corroborated by Huawei's guidance that AI chip revenue reaches ~US$12bn in 2026, up from US$7.5bn in 2025 (+60% YoY) per FT (1 May 2026, link), with the majority of orders directed to the 950 PR; FT also reports DeepSeek used the 950 PR for V4 with improved inference efficiency at reduced costs, providing third-party confirmation of our chip-side cost-edge argument.

Figure 8: Atlas 950 SuperPoD

Source: Company data, CMBIGM

Alibaba is the commercial alternative on the cloud-platform side. T-Head's Panjiu 128 (磐久 128) SuperPoD launched at Yunqi 2025 supports 128 AI chips per rack, with claimed inference performance roughly 50% above traditional architectures at equivalent compute. PPU has reached deployment scale on Alibaba Cloud, with the China Unicom Sanjiangyuan green-compute project disclosing 16,384 PPU cards for total compute of 1,945 PFLOPS, the largest disclosed domestic-chip deployment to date.

Among listed GPGPU vendors, Biren Tech (6082 HK, NR) brought LightSphere X SuperPod to market in 2025 on the BR10X series, using optical interconnect and distributed optical circuit switching (dOCS), with a 2,048-card deployment at a national-level computing platform. An earlier 1,024-card Nanjing cluster delivered 95% linear scalability for large-model training. 2nd-gen BR20X targets 2026 commercial launch with native FP8/FP4 support and a Blink 2.0 SuperPod architecture scaling to thousand-card clusters.

Cluster-level efficiency continues to drive down the underlying cost basis for Chinese LLM APIs. As inference workloads are routed to these clusters and Ascend 950 SuperPods scale up in 2H26, we expect per-token cost basis to decline further. Notably, DeepSeek's management has explicitly stated that broader Ascend deployment will enable them to continue lowering API prices in the future. We believe this trajectory will also benefit other localized LLM providers like Alibaba's Qwen Max running on its Panjiu 128 infrastructure, alongside cloud service providers like Tencent Cloud.

Disclosures & Disclaimers

Analyst Certification

The research analyst who is primary responsible for the content of this research report, in whole or in part, certifies that with respect to the securities or issuer that the analyst covered in this report: (1) all of the views expressed accurately reflect his or her personal views about the subject securities or issuer; and (2) no part of his or her compensation was, is, or will be, directly or indirectly, related to the specific views expressed by that analyst in this report.
Besides, the analyst confirms that neither the analyst nor his/her associates (as defined in the code of conduct issued by The Hong Kong Securities and Futures Commission) (1) have dealt in or traded in the stock(s) covered in this research report within 30 calendar days prior to the date of issue of this report; (2) will deal in or trade in the stock(s) covered in this research report 3 business days after the date of issue of this report; (3) serve as an officer of any of the Hong Kong listed companies covered in this report; and (4) have any financial interests in the Hong Kong listed companies covered in this report.

CMBIGM Ratings

BUY	: Stock with potential return of over 15% over next 12 months
HOLD	: Stock with potential return of +15% to -10% over next 12 months
SELL	: Stock with potential loss of over 10% over next 12 months
NOT RATED	: Stock is not rated by CMBIGM
OUTPERFORM	: Industry expected to outperform the relevant broad market benchmark over next 12 months
MARKET-PERFORM	: Industry expected to perform in-line with the relevant broad market benchmark over next 12 months
UNDERPERFORM	: Industry expected to underperform the relevant broad market benchmark over next 12 months

CMB International Global Markets Limited

Address: 45/F, Champion Tower, 3 Garden Road, Hong Kong, Tel: (852) 3900 0888 Fax: (852) 3900 0800

CMB International Global Markets Limited ("CMBIGM") is a wholly owned subsidiary of CMB International Capital Corporation Limited (a wholly owned subsidiary of China Merchants Bank)

Important Disclosures

There are risks involved in transacting in any securities. The information contained in this report may not be suitable for the purposes of all investors. CMBIGM does not provide individually tailored investment advice. This report has been prepared without regard to the individual investment objectives, financial position or special requirements. Past performance has no indication of future performance, and actual events may differ materially from that which is contained in the report. The value of, and returns from, any investments are uncertain and are not guaranteed and may fluctuate as a result of their dependence on the performance of underlying assets or other variable market factors. CMBIGM recommends that investors should independently evaluate particular investments and strategies, and encourages investors to consult with a professional financial advisor in order to make their own investment decisions.
This report or any information contained herein, have been prepared by the CMBIGM, solely for the purpose of supplying information to the clients of CMBIGM or its affiliate(s) to whom it is distributed. This report is not and should not be construed as an offer or solicitation to buy or sell any security or any interest in securities or enter into any transaction. Neither CMBIGM nor any of its affiliates, shareholders, agents, consultants, directors, officers or employees shall be liable for any loss, damage or expense whatsoever, whether direct or consequential, incurred in relying on the information contained in this report. Anyone making use of the information contained in this report does so entirely at their own risk.
The information and contents contained in this report are based on the analyses and interpretations of information believed to be publicly available and reliable. CMBIGM has exerted every effort in its capacity to ensure, but not to guarantee, their accuracy, completeness, timeliness or correctness. CMBIGM provides the information, advices and forecasts on an "AS IS" basis. The information and contents are subject to change without notice. CMBIGM may issue other publications having information and/ or conclusions different from this report. These publications reflect different assumption, point-of-view and analytical methods when compiling. CMBIGM may make investment decisions or take proprietary positions that are inconsistent with the recommendations or views in this report.
CMBIGM may have a position, make markets or act as principal or engage in transactions in securities of companies referred to in this report for itself and/or on behalf of its clients from time to time. Investors should assume that CMBIGM does or seeks to have investment banking or other business relationships with the companies in this report. As a result, recipients should be aware that CMBIGM may have a conflict of interest that could affect the objectivity of this report and CMBIGM will not assume any responsibility in respect thereof. This report is for the use of intended recipients only and this publication, may not be reproduced, reprinted, sold, redistributed or published in whole or in part for any purpose without prior written consent of CMBIGM.
Additional information on recommended securities is available upon request.

For recipients of this document in the United Kingdom
This report has been provided only to persons (I)falling within Article 19(5) of the Financial Services and Markets Act 2000 (Financial Promotion) Order 2005 (as amended from time to time)("The Order") or (II) are persons falling within Article 49(2) (a) to (d) ("High Net Worth Companies, Unincorporated Associations, etc.,) of the Order, and may not be provided to any other person without the prior written consent of CMBIGM.

For recipients of this document in the United States
CMBIGM is not a registered broker-dealer in the United States. As a result, CMBIGM is not subject to U.S. rules regarding the preparation of research reports and the independence of research analysts. The research analyst who is primary responsible for the content of this research report is not registered or qualified as a research analyst with the Financial Industry Regulatory Authority ("FINRA"). The analyst is not subject to applicable restrictions under FINRA Rules intended to ensure that the analyst is not affected by potential conflicts of interest that could bear upon the reliability of the research report. This report is intended for distribution in the United States solely to "major US institutional investors", as defined in Rule 15a-6 under the US, Securities Exchange Act of 1934, as amended, and may not be furnished to any other person in the United States. Each major US institutional investor that receives a copy of this report by its acceptance hereof represents and agrees that it shall not distribute or provide this report to any other person. Any U.S. recipient of this report wishing to effect any transaction to buy or sell securities based on the information provided in this report should do so only through a U.S.-registered broker-dealer.

For recipients of this document in Singapore
This report is distributed in Singapore by CMBI (Singapore) Pte. Limited (CMBISG) (Company Regn. No. 201731928D), an Exempt Financial Adviser as defined in the Financial Advisers Act (Cap. 110) of Singapore and regulated by the Monetary Authority of Singapore. CMBISG may distribute reports produced by its respective foreign entities, affiliates or other foreign research houses pursuant to an arrangement under Regulation 32C of the Financial Advisers Regulations. Where the report is distributed in Singapore to a person who is not an Accredited Investor, Expert Investor or an Institutional Investor, as defined in the Securities and Futures Act (Cap. 289) of Singapore, CMBISG accepts legal responsibility for the contents of the report to such persons only to the extent required by law. Singapore recipients should contact CMBISG at +65 6350 4400 for matters arising from, or in connection with the report.