China AI Intelligence

DeepSeek V4: strengthening model efficiency and accelerating compute localisation

Equities

China

Internet Services

Wei Xiong
Analyst
wei.xiong@ubs.com
+86-21-3866 8883

Kenneth Fong
Analyst
kenneth-kc.fong@ubs.com
+852-3712 3890

Charles Chen
Associate
S1460524020001
charles-za.chen@ubs.com
+86-21-3866 8907

Jimmy Yu
Analyst
S1460517080002
jimmy.yu@ubs.com
+86-21-3866 8880

DeepSeek V4: from single model to tiered model offerings

On Apr 24, DeepSeek officially launched the preview version of DeepSeek V4 and open-sourced the model (news). Unlike prior single-model updates (e.g., V3.2), V4 offers two distinct tiers: **1) V4-Flash**, a smaller model (284B total / 13B activated) positioned as a fast, cost-efficient option for general use cases; and **2) V4-Pro**, a larger model (1.6T total / 49B activated) designed to deliver top-tier capabilities and target high-value cases such as complex coding and long-horizon workflows. As model capability improves and addresses a widening range of tasks, we believe token demand will likely diverge across different workflows, leading to tiered model offerings among domestic players (similar to global leaders). **On model capability**, Artificial Analysis data shows that V4-Pro's overall model intelligence is largely comparable with flagship models from domestic emerging AI labs (e.g., GLM-5.1 and Kimi K2.6), although GLM-5.1 still leads in some coding benchmarks (e.g., Code Arena). **On pricing**, its high-end v4-Pro is priced slightly above GLM-5.1, but cheaper after limited-time promotional discounts, while the cost-efficient V4-Flash appears to have strong pricing competitiveness among both domestic/global models. Additionally, V4 has been **optimized for agentic products** (e.g., OpenClaw, Claude Code, OpenCode and CodeBuddy), corroborating a shift from chatbot-style interaction to action-oriented workflows, in our view.

Architectural innovation and compute localisation to drive better efficiency

We see model efficiency as a key feature of V4, with cost reduction coming from both architectural innovation and potential domestic compute adoption: **1) Making 1M-token context economically viable**: Both models support an ultra-long 1M-token context window with significantly lower inference cost (i.e., FLOPs at 27% and KV cache at 10% of V3.2 at 1M context), largely driven by architectural innovations (e.g., Hybrid Attention, mHC and Muon). We believe this could accelerate enterprise AI adoption by enabling models to process large-scale business documents and knowledge bases at lower cost. **2) Further cost reduction from domestic compute adoption**: We believe V4's compute-light design (e.g., FP4 Quantization-Aware Training) could potentially improve its fit with domestic AI accelerators. DeepSeek also noted that V4-Pro pricing is partly constrained by high-end compute availability, while scaled deployment of Ascend 950 supernode clusters in 2H26 may bring its prices meaningfully lower.

Other recent domestic model updates

We see multiple domestic model updates from internet/tech leaders, with a rising focus on coding and agentic capabilities while remaining ecosystem-oriented: **Tencent Hy3 Preview** focused on reasoning, coding and agentic workloads, with early adoption across Tencent's AI products (e.g., Yuanbao, ima, CodeBuddy, WorkBuddy); **Alibaba Qwen3.6-Max-Preview** improves agentic coding, world knowledge and instruction following, supporting Qwen's broader cloud/developer ecosystem; and **Xiaomi MiMo-V2.5-Pro** is a 1M-context, agent-centric model designed for long-horizon workflows and complex software engineering, with a focus on tool-use-heavy tasks.

Competition a secondary concern amid surging token demand globally

We expect DeepSeek V4 to further strengthen Chinese models' advantage in efficiency and accelerate domestic compute adoption. We are not too concerned about this release to intensify domestic competition, given V4's tiering strategy and surging AI demand. We remain positive on **Zhipu**'s sustained leadership in coding (initiation); **MiniMax** for differentiated positioning in multimodality which V4 does not emphasize; and **Alibaba/Baidu** for their cloud/chip businesses benefiting from industry tailwinds.

COMMENTRAY

COMMENTARY

Unlike prior single-model updates, DeepSeek V4 offers two distinct model tiers, with V4-Pro targeting the upper bound of model intelligence and V4-Flash positioned as a cost-efficient option for broader use cases, in our view.

V4-Pro:

- **On model intelligence**: We note that V4-Pro's overall intelligence is in the top tier among domestic models based on Artificial Analysis data, broadly comparable with Kimi K2.6 and GLM-5.1, although GLM-5.1 still leads domestic peers in some coding benchmarks (i.e., Code Arena) . We think this is partly supported by V4-Pro's larger model scale, with 1.6T total parameters and 49B activated parameters, ahead of other leading domestic AI lab models such as Kimi K2.6 at 1T total parameters.
- **On pricing**, we note V4-Pro remains significantly cheaper than frontier models overseas while broadly in line with domestic models of similar intelligence; before the limited-time discount (75% off until May 5), its list price is even slightly above GLM-5.1.

V4-Flash:

- V4-Flash is the smaller model in the V4 series, with 284B total parameters and 13B activated parameters. Its pricing is significantly lower than global AI peers' models with similar intelligence/positioning (e.g., Gemini 3 Flash and Claude 4.5 Haiku). It also demonstrated a pricing advantage versus older domestic models, such as MiniMax M2.5, despite offering a slightly higher Artificial Analysis intelligence score and a larger context window.

Figure 1: Selected leading AI models by intelligence, context window and API pricing

		Parameters			Token price		
Model	**Overall intelligence (AA)**	**Total**	**Activated**	**Context window**	**Cached input**	**Uncached input**	**Output**
GPT-5.5	60	N/A	N/A	1M	US$0.50	US$5	US$30
Claude Opus 4.7	57	N/A	N/A	1M	US$0.50	US$5	US$25
Kimi K2.6	54	1T	32B	256K	Rmb1.1 (US$0.16)	Rmb6.5 (US$0.95)	Rmb27 (US$4.00)
DeepSeek V4 Pro	**52**	**1.6T**	**49B**	**1M**	**Rmb1 (US$0.145)**	**Rmb12 (US$1.74)**	**Rmb24 (US$3.48)**
Qwen3.6 Max Preview / 0–128K	52	N/A	N/A	256K	Rmb1.8 (US$0.26*)	Rmb9 (US$1.30)	Rmb54 (US$7.80)
GLM-5.1 / 0–32K	51	744B	40B	~200K	Rmb1.3 (US$0.186)	Rmb6 (US$0.857)	Rmb24 (US$3.429)
GLM-5.0 / 0–32K	50	744B	40B	~200K	Rmb0.8 (US$0.115*)	Rmb4 (US$0.573)	Rmb18 (US$2.58)
MiniMax M2.7	50	230B	10B	200K	Rmb0.42 (US$0.06)	Rmb2.1 (US$0.30)	Rmb8.4 (US$1.20)
GPT-5.4 mini (xHigh)	49	N/A	N/A	400K	US$0.08	US$0.75	US$4.50
DeepSeek V4 Flash	**47**	**284B**	**13B**	**1M**	**Rmb0.2 (US$0.028)**	**Rmb1 (US$0.14)**	**Rmb2 (US$0.28)**
Gemini 3 Flash	46	N/A	N/A	1M	US$0.05	US$0.50	US$3.00
MiniMax M2.5	42	230B	10B	200K	Rmb0.21 (US$0.03)	Rmb2.1 (US$0.30)	Rmb8.4 (US$1.20)
Claude 4.5 Haiku	37	N/A	N/A	200K	US$0.10	US$1.00	US$5.00

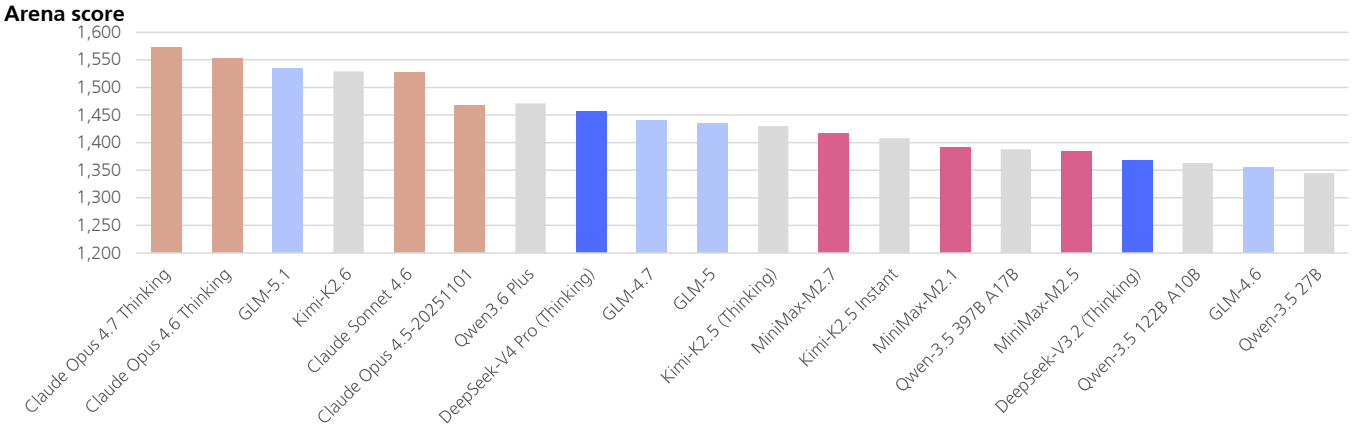
Source: Artificial Analysis, company websites, UBS. Note: Prices are per 1M tokens. Qwen3.6 Max Preview pricing is based on the 0–128K tier, with cached input estimated using implicit cache hit pricing at 20% of input price. GLM-5.0 and GLM-5.1 pricing is based on the 0–32K tier. Gemini 3 Flash cached input assumes a 90% cache discount.

Figure 2: DeepSeek V4 API pricing and key specifications

Model	Parameters (Total/Active)	Context window	Token price (Rmb/m tokens)		
			Cached input	Uncached input	Output
DeepSeek V4 Flash	284B/13B	1M	Rmb0.2	Rmb1.0	Rmb2.0
DeepSeek V4 Pro	1.6T/45B	1M	Rmb1.0 (Rmb0.25 after discount)	Rmb12.0 (Rmb3.0 after discount)	Rmb24.0 (Rmb6.0 after discount)

Source: Company data, UBS. Note: DeepSeek indicated that Pro service throughput is currently limited due to high-end compute constraints, with prices expected to decline after Ascend 950 supernodes are launched in batches in 2H26. Meanwhile, DeepSeek V4 Pro is currently offered at a limited-time 75% discount, valid until 23:59 Beijing time on May 5, 2026.

Figure 3: Coding capability comparison across leading Chinese and Anthropic models



Source: Arena AI Leaderboard (Code Arena)

Valuation Method and Risk Statement

We believe key sector risks include: 1) an evolving competitive landscape and intensifying competition; 2) fast-moving trends in technology, as well as internet users' needs and preferences; 3) uncertain monetisation; 4) the rising cost of traffic acquisition, content and brand promotions; 5) the upkeep of IT systems; 6) expansion into international markets; 7) adverse changes in market sentiment; and 8) regulatory risks.

MiniMax: We adopt Price/ARR as our valuation methodology for MiniMax.
Key downside risks include: macro and geopolitical uncertainties that may constrain AI adoption; competitive and operational challenges delaying monetisation and pressuring margins; data-related model training risks; user-generated content governing risks; and disruption of key service providers.
Key upside risks include: easing geopolitical tensions and export restrictions that may improve access to advanced hardware and accelerate model upgrades; technological breakthroughs in AI training and infrastructure efficiency that could enhancemodel capability and margins; and faster user adoption and commercialisation driving stronger monetisation and profitability.

Knowledge Atlas Technology (Zhipu): We adopt price-to-sales (P/S) as our valuation methodology for Zhipu, cross checked with SOTP analysis.
Key downside risks include: macro and geopolitical uncertainty that may constrain AI adoption; competitive and operating challenges delaying commercialisation and pressuring margins; potential reduction in key account (KA) customers as their in-house model capability improves; data-related risks affecting model training and performance; regulatory risks related to data security and AI governance; and potential disruptions in access to key computing infrastructure and service providers.
Key upside risks include: easing geopolitical tension and export restrictions that may improve access to advanced hardware and accelerate model upgrades; faster-than-expected improvement in model capability, both in magnitude and iteration speed, supporting higher usage and pricing; technological breakthroughs in model training and inference efficiency that could enhance model capability and margins; stronger-than-expected enterprise adoption, particularly in the to-B/to-G segments, driving faster commercialisation; and the emergence of new AI use cases supporting higher monetisation and revenue growth.

Baidu: We derive our price target for Baidu using sum-of the parts valuation methodology. We believe the key risks for Baidu include: 1) an evolving and intensifying competitive landscape; 2) execution of new businesses; 3) integration of invested companies and businesses; 4) rising cost of traffic acquisitions, content and brand promotions; 5) maintaining and upgrading IT systems; 6) infringement of intellectual property rights; 7) expansion into international markets; 8) key management departures; and 9) regulatory risks.

Alibaba: Our price target is based on SOTP. We believe the key risks for Alibaba are: 1) regulatory changes, particularly for data usage and online content; 2) Chinese and global macroeconomic headwinds; 3) competitive pressure from traditional offline retailers; 4) interruptions of information technology and systems; 5) short-term profitability pressure from long-term investment; 6) execution and management complexity resulting from its multiple platforms; 7) corporate governance – significant voting control by the Alibaba Partnership; and 8) the loss of the services of founder Jack Ma.
Our price target on Tencent is based on sum of the parts. We believe the key risks for Tencent include: 1) an evolving and intensifying competitive landscape; 2) new business execution; 3) integration of invested companies and businesses; 4) rising costs of traffic acquisition, content and brand promotions; 5) IT system upkeep; 6) expansion into international markets; 7) infringement of intellectual property rights; 8) departures of key management; and 9) regulatory risks.