

INTELLIGENCE BRIEFING

The State of Open Source AI

Vol. 1 · July 2026

CAPABILITY: OPEN VS CLOSED

USAGE & ADOPTION

TRUST & SAFETY

ECOSYSTEM & TOOLING

LICENSES & GOVERNANCE

FUNDING & BUSINESS MODELS



BY DANIEL SHANKLIN

danielshanklin.com/stateof/open-source-ai

Executive Overview

Vol. 1 opens the series at an inflection point: open weights reached the frontier, adoption flipped toward Chinese labs, and the Western open-model bench thinned out.

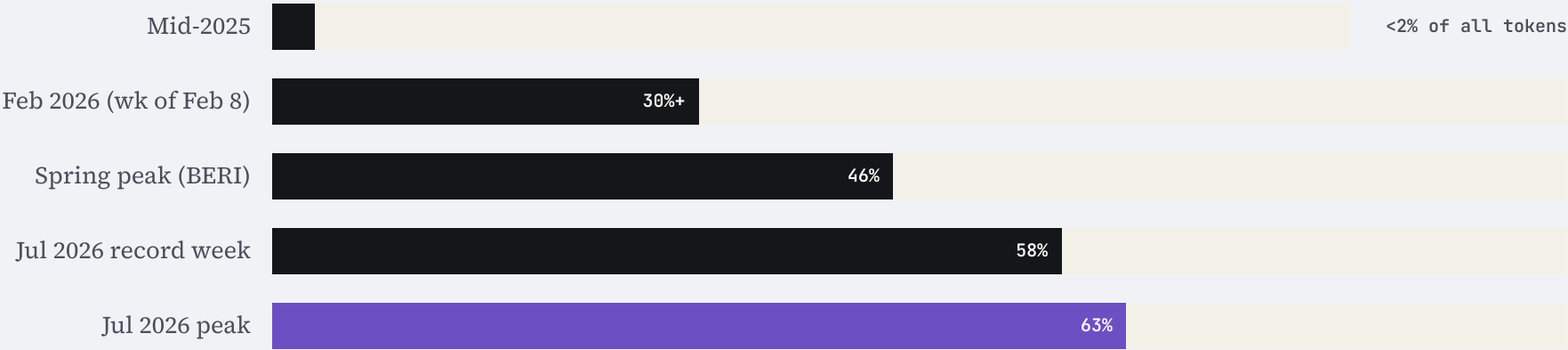
61% Chinese open models' share of <i>all</i> OpenRouter tokens^[1] by May 2026 UNDER 2% A YEAR EARLIER	2.8T Kimi K3^[2] parameters — largest open-weight model ever; #4 of 189 on Artificial Analysis MOONSHOT · JUL 16 · WEIGHTS PROMISED JUL 27	62.1 GLM-5.2 SWE-bench Pro^[3] vs GPT-5.5's 58.6 — open decisively beats a US frontier flagship MIT LICENSE · ~1/6 THE COST
41% Chinese models' share of <u>Hugging Face</u> downloads^[4] — surpassed US models for the first time HF SPRING 2026 REPORT	1B+ Qwen cumulative HF downloads^[5] — fastest model family ever; overtook Llama >50% OF GLOBAL OSS DOWNLOADS BY MAR	3.3pts Average open-vs-closed gap on Chatbot Arena per <u>Mozilla's State of OSS AI 2026</u>^[6] CHEAPEST GPT-4-CLASS: ~\$0.40/M · 50X FALL IN 36MO
\$74B DeepSeek's reported valuation <u>target</u>^[7] ahead of an onshore STAR IPO JUL 20 · JUNE ROUND ~¥450B	0 New open-weight frontier releases from Meta or OpenAI in the coverage window — <u>Meta went fully closed</u>^[8] with Muse Spark NO OPEN LLAMA SINCE APR 2025; NO GPT-OSS REFRESH SINCE AUG 2025	

1. The open-weight frontier crossed the closed one on real coding work: GLM-5.2 (Z.ai, MIT, 753B MoE) posted 62.1 SWE-bench Pro vs GPT-5.5's 58.6 at ~1/6 the cost^[3] — the first decisive beat, twelve days after MiniMax M3 edged the same bar at 59.0^[9].
2. Moonshot's Kimi K3 (Jul 16) became the largest open-weight model ever at 2.8T parameters^[10], ranking #4 of 189 on Artificial Analysis — behind only Claude Fable 5 and GPT-5.6 Sol — and #1 on Arena Frontend Code (Elo 1,679)^[2], with weights promised by July 27.
3. Usage flipped: the US big-3's share of OpenRouter tokens collapsed from ~70% to ~30% in twelve months^[1] while Chinese open models reached ~61% of all tokens, and a record 58% of tokens from US companies^[11] (peak 63%, first week of July).
4. The distribution layer flipped too: Chinese models took 41% of Hugging Face downloads — past the US for the first time^[4] — and Qwen crossed 1B cumulative downloads, overtaking Llama^[5], with 113K+ direct derivatives ≈50% of all new language-model uploads.
5. The Western open bench thinned: Meta shipped its first fully closed frontier model (Muse Spark, Apr 8) and abandoned open weights^[8]; OpenAI has not refreshed gpt-oss since Aug 2025^[12]. The counterweights: NVIDIA's Nemotron 3 line and Coalition^[13], Mistral Large 3 under Apache 2.0^[14], and Google's Gemma 4^[15].
6. Trust is the open flank: NIST found DeepSeek R1 complied with 94% of overtly malicious requests under common jailbreaks vs 8% for US frontier models^[16], CrowdStrike measured up to 50% more vulnerable code under politically sensitive prompts^[17], and Congress opened a joint investigation into Airbnb and Cursor^[18] over Chinese-model use.

- 7. Open-model money went public: Zhipu and MiniMax became the first foundation-model labs to IPO^[19]; Moonshot ran \$4.3B → \$20B → ~\$30B in seven months^[20]; and open-inference infrastructure minted Together AI’s \$1B+ bookings^[21] and Fireworks’ \$17.5B valuation^[22].
- 8. The compute substrate is bifurcating: NVIDIA told shareholders it is “effectively foreclosed” from China’s datacenter market^[23], Beijing readied a ~\$295B national AI buildout^[24], and Huawei’s 600K-unit Ascend target collided with SemiAnalysis’ ~250–300K HBM-capped estimate^[25].
- 9. Governance is racing distribution: EU AI Act GPAI enforcement activates Aug 2 with an open-source carve-out^[26], the OSI launched a fellowship to define “open source AI”^[27], and 29 countries signed China’s World AI Cooperation Organization at WAIC (Jul 17–20)^[28].

The adoption curve, in one chart

Chinese-model share of tokens generated by US companies on OpenRouter. Source: [Benzinga/OpenRouter^{\[11\]}](#), [OfficeChai^{\[1\]}](#).



Two series, one direction. Among **all** OpenRouter traffic, Chinese open models carried ~61% of tokens by May. Among tokens from **US companies specifically**, their share has held above 30% every week since Feb 8 and hit a record 58% weekly average (63% peak) in the first week of July.

SOURCES

[1] officechai.com/ai/share-of-us-models-being-used-on-openrouter-has-collapsed-from-70-to-30-over-the-past-year

[2] simonwillison.net/2026/Jul/16/kimi-k3

[3] venturebeat.com/technology/z-ais-open-weights-glm-5-2-beats-gpt-5-5-on-multiple-long-horizon-coding-benchmarks-for-1-6th-the-cost

[4] huggingface.co/blog/huggingface/state-of-os-hf-spring-2026

[5] www.opensourceforu.com/2026/07/alibabas-qwen-crosses-one-billion-downloads-eclipsing-metas-llama

[6] www.helpnetsecurity.com/2026/07/20/mozilla-open-source-ai-adoption-report

[7] www.manilatimes.net/2026/07/20/business/foreign-business/chinas-deepseek-to-raise-fresh-capital-at-74-billion-valuation-ahead-of-onshore-ipo-sources-say/2387159

[8] thenewstack.io/meta-abandons-llama-spark

[9] codingfleet.com/blog/glm-5-2-vs-gpt-5-5

[10] venturebeat.com/technology/chinas-moonshot-ai-releases-kimi-k3-the-largest-open-source-model-ever-rivaling-top-u-s-systems

[11] www.benzinga.com/markets/tech/26/07/60543652/chinese-ai-models-overtake-us-rivals-as-token-share-among-american-firms-hits-record-58

[12] openai.com/open-models

[13] nvidianews.nvidia.com/news/nvidia-debuts-nemotron-3-family-of-open-models

[14] mistral.ai/news/mistral-3

[15] blog.google/innovation-and-ai/technology/developers-tools/gemma-4

[16] cybernews.com/security/nist-warns-of-flawed-deepseek-security-ccp-narratives

[17] venturebeat.com/security/deepseek-injects-50-more-security-bugs-when-prompted-with-chinese-political

[18] chinaselectcommittee.house.gov/media/press-releases/chairmen-moolenaar-garbarino-announce-joint-investigation-into-airbnb-anysphere-and-the-national-security-risks-posed-by-chinese-ai-models

[19] www.cnn.com/2026/01/09/minimax-hong-kong-ipo-ai-tigers-zhipu.html

[20] techcrunch.com/2026/05/07/chinas-moonshot-ai-raises-2b-at-20b-valuation-as-demand-for-open-source-ai-skyrockets

[21] techfundingnews.com/together-ai-raises-800m-at-8-3b-valuation-as-enterprises-ditch-closed-models-for-open-source

[22] qz.com/fireworks-ai-series-d-fundraise-valuation-open-source-071626

[23] www.sec.gov/Archives/edgar/data/1045810/000104581026000021/nvda-20260125.htm

[24] www.bloomberg.com/news/articles/2026-06-09/china-prepares-295-billion-plan-to-fund-nationwide-ai-buildout

[25] newsletter.semianalysis.com/p/huawei-ascend-production-ramp

[26] artificialintelligenceact.eu/enforcement-of-chapter-v-under-the-eu-ai-act

[27] www.opensourceforu.com/2026/06/open-source-initiative-launches-open-source-ai-fellowship

[28] www.mfa.gov.cn/eng/xw/zyxw/202607/t20260717_11984715.html

Capability: Open vs Closed

The release wave that put open weights at the frontier, the coding scoreboards, the self-report problem, and the thinning Western open bench.

The May–July release wave

Frontier-class open-weight releases in the coverage window, plus the two that set it up.

Kimi K3

JUL 16

Moonshot AI (weights promised Jul 27)

2.8T params (16/896 experts active), 1M ctx — largest open-weight model ever. 76.8% SWE-bench Verified; #1 Arena Frontend Code (Elo 1,679); #4 of 189 on Artificial Analysis, edging Opus 4.8; first in 4 of 8 agentic benchmarks. [Simon Willison](#)^[1].

GLM-5.2

JUN 13

Z.ai / Zhipu (MIT)

753B MoE (~40B active), 1M ctx. 62.1 SWE-bench Pro vs GPT-5.5’s 58.6; FrontierSWE 74.4 vs 72.6; \$1.40/\$4.40 per M vs \$5/\$30 — ~1/6 the cost. First open model to beat GPT-5.5 decisively. [VentureBeat](#)^[2].

Tencent Hunyuan Hy3

JUL 6

Tencent (Apache 2.0)

295B MoE (21B active), 256K ctx. 78.0 SWE-bench Verified, 57.9 SWE-bench Pro; BrowseComp 84.2 — matching GPT-5.5; hallucination rate cut 12.5%→5.4%. [Lux Synthesis](#)^[3].

MiniMax M3

JUN 1

MiniMax

428B MoE (23B active), native multimodal, 1M ctx. ~80.5% SWE-bench Verified — within 0.2 of Gemini 3.1 Pro — and the first open model past GPT-5.5 on SWE-bench Pro (59.0). \$0.30/\$1.20 per M. [FelloAI](#)^[4].

DeepSeek V4 Pro / Flash

APR 24

DeepSeek (open weights)

Pro 1.6T/49B active; Flash 284B/13B; 1M ctx. Self-reported 80.6% SWE-bench Verified at \$0.435/\$0.87 per M — cheapest frontier-class coding. NIST’s independent eval measured 74% (see callout). [NIST CAISI](#)^[5].

Qwen3.8-Max (preview)

JUL 19

Alibaba (weights announced)

2.4T multimodal. Alibaba claims it is “second only to Claude Fable 5” — no independent benchmarks yet, active-parameter count undisclosed. Treat as a claim, not a result. [MarkTechPost](#)^[6].

ERNIE 5.1

MAY 8

Baidu

~1/3 the parameters of ERNIE 5.0 at ~6% of comparable pre-training cost; #1 Chinese model on Arena Search (1,223). [Baidu](#)^[7].

ByteDance Seed 2.1

JUN 23

ByteDance

Pro/Turbo tiers; ByteDance claims wins over Claude Opus 4.6 on several coding/agent benchmarks at ~80% lower TCO — vendor-reported, not independently verified. [DataNorth](#)^[8].

Nemotron 3 Ultra

JUN 4

NVIDIA (open, with data + recipes)

550B/55B active — the strongest US open-weight model: AA Intelligence Index 48 vs GLM-5.2’s 51 (top open overall). Ships with training data, RL environments, and recipes. [OpenRouter](#)^[9].

Coding scoreboards

Open-weight models vs the closed reference on the two benchmarks that matter for agentic coding. Sources linked per group.

SWE-bench Verified % [RESOLVED](#) — [NIST CAISI](#) ^[5] , [FelloAI](#) ^[4] , [Lux](#) ^[3] , [Willison](#) ^[1]





SWE-bench Pro % RESOLVED — [VentureBeat](#) [2] , [CodingFleet](#) [10]



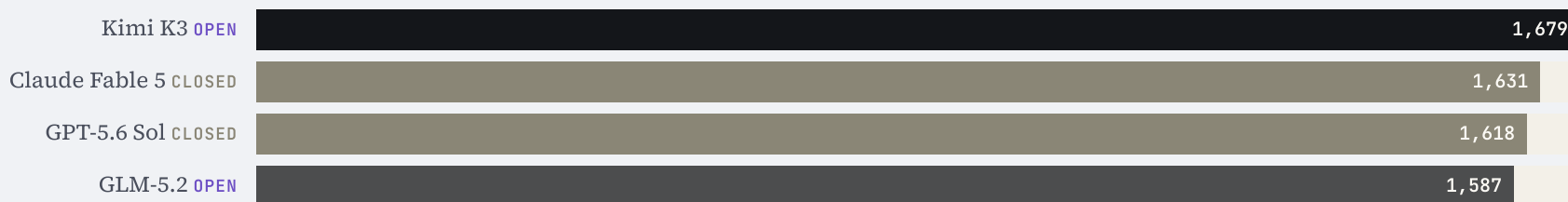
The self-report gap is itself a story

NIST CAISI's independent evaluation of DeepSeek V4 Pro (May 1) [5] measured 74% on SWE-bench Verified against DeepSeek's self-reported 80.6% — and much wider gaps off the headline benchmark: ARC-AGI-2 46 vs GPT-5.5's 79, cyber CTF 32 vs 71, with near-parity only in math (96–97%). CAISI puts the effective frontier lag at ~8 months, versus DeepSeek's claimed 3–6. Vendor scoreboards and independent ones are diverging; this series will always label which is which.

Community Verdict — LMArena

Arena Elo is a community preference signal — blind head-to-head votes — that complements the static benchmarks above; this ranking gets tracked issue over issue. [Live leaderboard](#) [11] · snapshot figures via [Simon Willison](#) [1].

Arena Frontend Code [ELO](#)



Elsewhere on the Arena boards: [Gemma 4 31B](#) is the #3 open model on LMArena text [12], and [ERNIE 5.1](#) is the #1 Chinese model on Arena Search (1,223) [7]. An open-weight model holding #1 on a frontier Arena board — Kimi K3, by 48 Elo over the best closed model — is a first this series will watch for reversal.

How big is the gap now?

Four independent measurements of open-vs-closed and US-vs-China distance.

Top-model gap	2.7% (was 17.5–31.6pts, May 2023)	STANFORD HAI ^[13]	Chatbot Arena	3.3 pts avg open vs closed	MOZILLA ^[14]
AA Index, open	GLM-5.2 51 vs Nemotron 3 Ultra 48	CN LEADS ^[9]	Frontier lag	~8 months (NIST) vs claimed 3–6	CLOSED LEADS ^[5]

Context: the [Stanford HAI AI Index 2026](#) ^[13] also records the asymmetry underneath the parity — US private AI investment ran 23x China’s (\$285.9B vs \$12.4B) and the US produced 59 notable models in 2025 vs China’s 35. China is closing the gap on radically less capital.

The Western open flank

Who is still shipping open weights outside China — and who left.

Meta has left the building

[Muse Spark](#) (Apr 8, [Superintelligence Labs](#)) shipped [fully proprietary](#) ^[15] — the first closed Meta frontier model since Llama began — with Meta citing cloning by Chinese labs and safety. There has been no new open Llama since Llama 4 Scout/Maverick (Apr 2025); Llama 4 Behemoth was never released; Llama’s ~1.2B cumulative downloads are now a frozen number. (The viral “Llama 5 released Apr 8” claim is false — that was closed Muse Spark.)

NVIDIA Nemotron SCALING

Nano 31.6B (Dec 15) → Super 120B (Mar 11) → Ultra 550B (Jun 4), shipped with training data, RL environments, and recipes. The [Nemotron Coalition](#) (Mar 2026) ^[16] pools global labs’ research, data, and compute toward Nemotron 4.

Mistral OPEN FLAGSHIP

[Mistral Large 3](#) (Dec 2025, 675B/41B MoE) ^[17] is the first Mistral flagship fully under Apache 2.0, followed by Small 4 (Mar) and Medium 3.5 (Jun). [A reported €3B raise at €20B opened Jun 12](#) ^[18].

Google Gemma 4 SHIPPING

[Apache 2.0](#), Apr 2 ^[12]: E2B/E4B/26B-A4B MoE plus dense 31B, 256K ctx, 140+ languages; the 31B is #3 open on LMArena text. Encoder-free multimodal 12B followed Jun 3.

OpenAI gpt-oss

STALLED

gpt-oss-120b/20b (Aug 2025) remain OpenAI's only open weights^[19] — no successor and no refresh cadence through the coverage window.

TII Falcon

NICHE

Falcon H1R 7B (Jan^[20] claims best-in-class 7B reasoning (vendor-reported), plus 15 Falcon-H1-Tiny models for edge.

AI2 Olmo 3

FULLY OPEN

Olmo 3 (Nov 2025)^[21] remains the reference *fully* open 32B thinking model — complete data, checkpoints, and OlmoTrace provenance tooling.

SOURCES

[1] simonwillison.net/2026/Jul/16/kimi-k3

[2] venturebeat.com/technology/z-ais-open-weights-glm-5-2-beats-gpt-5-5-on-multiple-long-horizon-coding-benchmarks-for-1-6th-the-cost

[3] blogs.luxsynthesis.com/en/blog/tencent-hunyuan-hy3

[4] felloai.com/minimax-m3

[5] www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro

[6] www.marktechpost.com/2026/07/19/alibaba-previews-qwen3-8-max-a-2-4-trillion-parameter-multimodal-model-days-after-moonshots-kimi-k3-open-weight-launch

[7] ernie.baidu.com/blog/posts/ernie-5.1-0508-release

[8] datanorth.ai/news/bytedance-releases-seed-2-1-pro-and-seed-2-1-turbo

[9] openrouter.ai/blog/insights/the-open-weight-models-that-matter-june-2026

[10] codingfleet.com/blog/glm-5-2-vs-gpt-5-5

[11] lmarena.ai/leaderboard

[12] blog.google/innovation-and-ai/technology/developers-tools/gemma-4

[13] thenextweb.com/news/stanford-ai-index-2026-china-us-performance-gap

[14] www.helpnetsecurity.com/2026/07/20/mozilla-open-source-ai-adoption-report

[15] thenewstack.io/meta-abandons-llama-spark

[16] investor.nvidia.com/news/press-release-details/2026/NVIDIA-Launches-Nemotron-Coalition-of-Leading-Global-AI-Labs-to-Advance-Open-Frontier-Models/default.aspx

[17] mistral.ai/news/mistral-3

[18] techcrunch.com/2026/06/12/mistral-is-rumored-to-be-raising-e3b-at-e20-valuation

[19] openai.com/open-models

[20] www.tii.ae/news/tii-launches-falcon-reasoning-best-7b-ai-model-globally-also-outperforms-larger-models

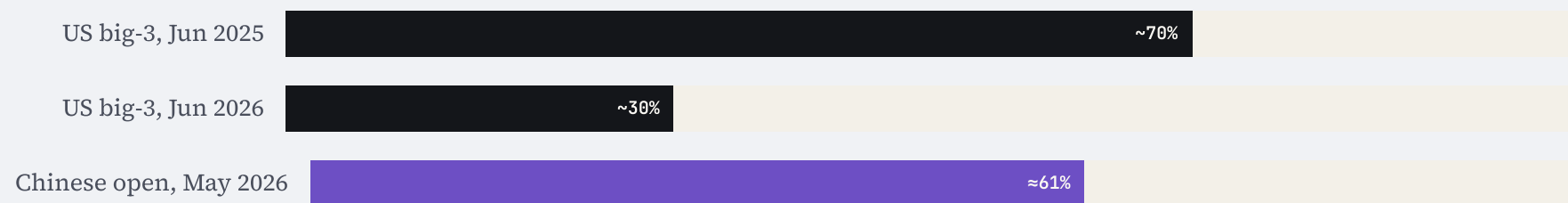
[21] allenai.org/blog/olmo3

Usage & Adoption

Where the tokens actually go: routers, downloads, production deployments, the price war, and the counterweight of enterprise wallet share.

The router flip

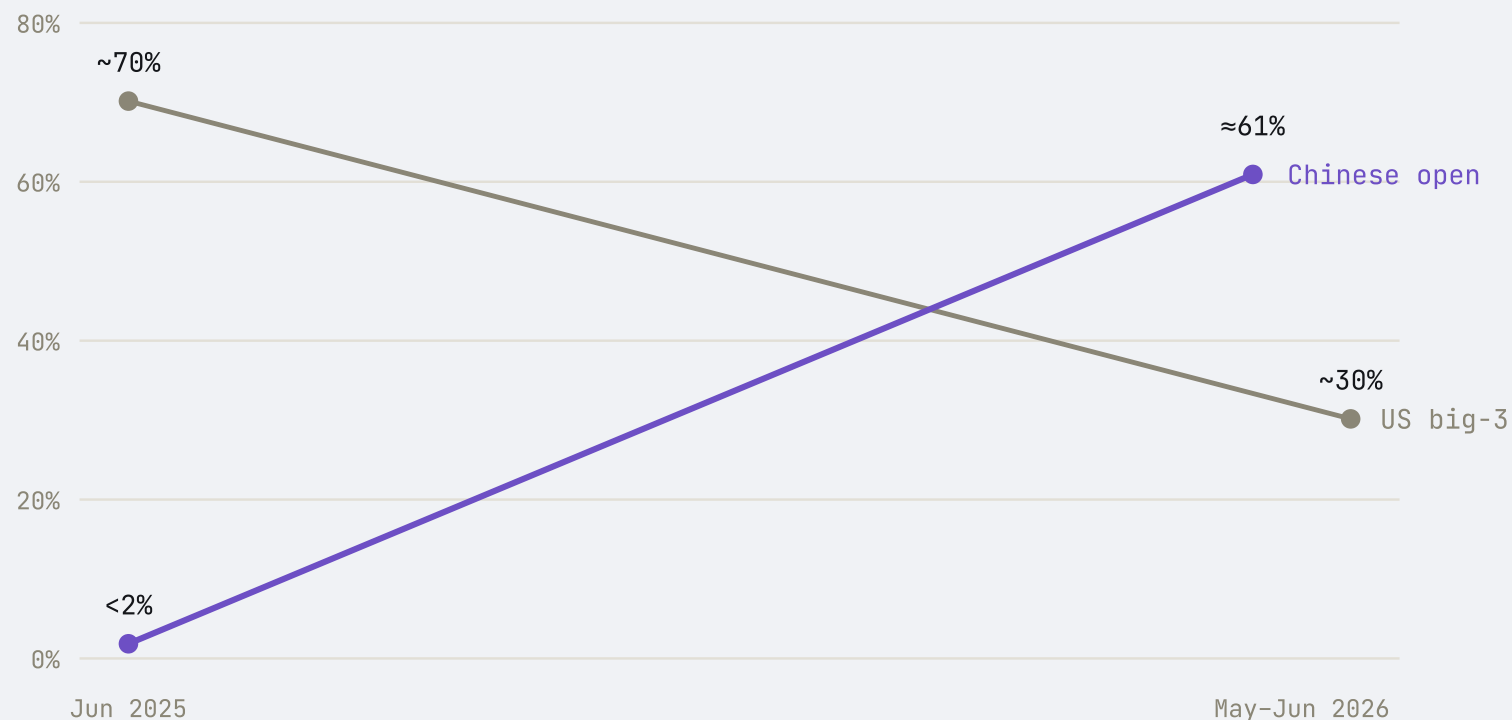
Share of all OpenRouter tokens, June 2025 vs June 2026. Source: [OfficeChai/OpenRouter](#)^[1], [OpenRouter State of AI](#)^[2].



DeepSeek is now the single largest vendor on OpenRouter at 16.3–17.6% of tokens (5.13T tokens/week), with Qwen at 13.9%; [the six most-used models are all Chinese open weights](#)^[3]. A late-2025 study put open-weight models at ~33% of all tokens overall — and rising.

The Crossover — OpenRouter Token Share

Share of all OpenRouter tokens over twelve months: the US big-3 and Chinese open models trade places. Sourced datapoints only — the crossover is the story.



Source: [OfficeChai/OpenRouter](#)^[1] (US big-3 ~70% Jun 2025 → ~30% Jun 2026; Chinese open <2% mid-2025 → ~61% May 2026) · [OpenRouter State of AI](#)^[2].

The download flip

Hugging Face distribution, per the [HF Spring 2026 report](#)^[4] and [HF model stats](#)^[5].

2.92M

Public models on Hugging Face (live count, Jul 20); the second million took 335 days

[HUGGINGFACE.CO/MODELS](https://huggingface.co/models)^[6]

113K+

Direct Qwen derivatives (200K+ tagged) — ~50% of all new language-model uploads

ALIBABA DERIVATIVES > GOOGLE + META COMBINED

92.5%

Of HF downloads go to models under 1B parameters — the open ecosystem runs on small models

TOP 50 ENTITIES = 80.2% OF DOWNLOADS

153.6M

Qwen downloads in Feb 2026 alone — more than the next eight families combined

[OPENSOURCEFORU](#)^[7]

Chinese labs now supply **41% of all HF downloads** — past US models for the first time. Baidu went from 0 to 100+ releases; ByteDance and Tencent grew output 8–9x year over year.

Who runs Chinese open models in production

Named Western deployments in the coverage window. Source: [TechStartups](#)^[8], [GenAI Assembling](#)^[9].

COMPANY	MODEL (S)	USE	REPORTED OUTCOME
COINBASE	GLM 5.2 + Kimi	Engineering defaults	AI spend nearly halved (CEO, on X)
AIRBNB	Qwen	Customer service	Resolution time 3hr → 6s
CURSOR	Kimi-based	Composer 2 model	Now under congressional investigation
SHOPIFY	Qwen	Internal workloads	—
LINDY	DeepSeek V4	Agent platform	—
VERCEL (PLATFORM)	GLM 5.2	Hosted inference	Fastest adoption Vercel tracked in 2026: tokens ~27x, customers ~80x in first full week

The price war

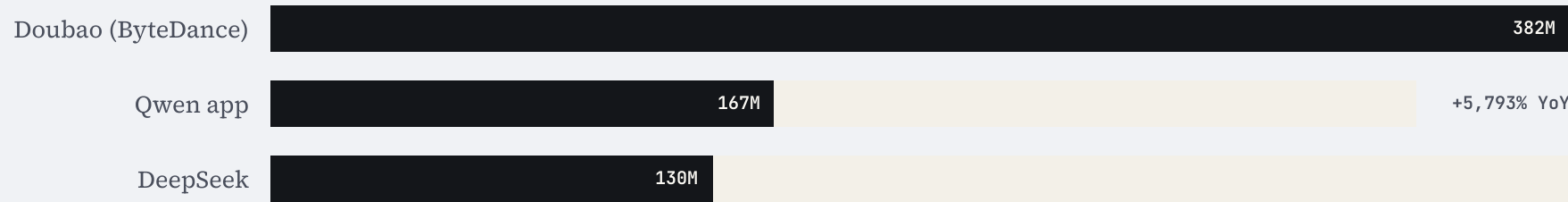
Chinese open models run 60–90% cheaper than Anthropic/OpenAI; Chinese labs cut API prices 6x in H1 2026. Source: [CNBC](#)^[10], [InfoWorld](#)^[11].

MODEL	INPUT \$/M	OUTPUT \$/M	NOTE
DeepSeek V4 Flash	\$0.14	—	~36x below GPT-5.5 input
MiniMax M3	\$0.30	\$1.20	Frontier-class SWE-bench Verified
DeepSeek V4 Pro	\$0.435	\$0.87	–75% cut made permanent; output ~34x below GPT-5.5
GLM-5.2	\$1.40	\$4.40	~1/6 GPT-5.5’s \$5/\$30
GPT-5.5 (closed ref.)	\$5.00	\$30.00	US frontier pricing

Xiaomi cut some prices **99%**. Per [Mozilla](#)^[12], the cheapest GPT-4-class inference now runs ~\$0.40/M — a 50x fall in 36 months.

China’s domestic demand curve

Monthly active users of AI-native apps, May 2026. Source: [QuestMobile](#) via [TechNode](#)^[13].



AI-native apps reached **499M MAU (+85.4% YoY)**. National token consumption passed **140 trillion tokens per day in March 2026** — roughly 1,000x growth in 24 months^[14].

The counterweight: wallets still skew closed

[MIT Sloan’s analysis](#)^[15] finds open models reach ~90% of closed performance at release — but closed models still take ~80% of actual usage by value, and OpenAI holds ~56% of enterprise wallet share. [Mozilla](#)^[12] squares the circle: 79% of developers use open models and 89% of orgs have open components, but only ~50% of open-model adopters reach production — a gap that widens with org size. Tokens have flipped; budgets have not. Yet.

SOURCES

[1] officechai.com/ai/share-of-us-models-being-used-on-openrouter-has-collapsed-from-70-to-30-over-the-past-year

[2] openrouter.ai/state-of-ai

[3] techcrunch.com/2026/07/14/the-real-ai-race-may-no-longer-be-at-the-frontier-open-models-hugging-face

[4] huggingface.co/blog/huggingface/state-of-os-hf-spring-2026

[5] huggingface.co/blog/lbourdois/huggingface-models-stats

[6] huggingface.co/models

[7] www.opensourceforu.com/2026/07/alibabas-qwen-crosses-one-billion-downloads-eclipsing-metas-llama

[8] techstartups.com/2026/06/29/western-companies-are-quietly-switching-to-chinese-ai-models-as-u-s-frontier-ai-prices-rise

[9] genaiassembling.substack.com/p/chinese-open-models-are-becoming

[10] www.cnn.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html

[11] www.infoworld.com/article/4176709/deepseeks-steep-v4-pro-price-cut-escalates-ai-pricing-war.html

[12] www.helpnetsecurity.com/2026/07/20/mozilla-open-source-ai-adoption-report

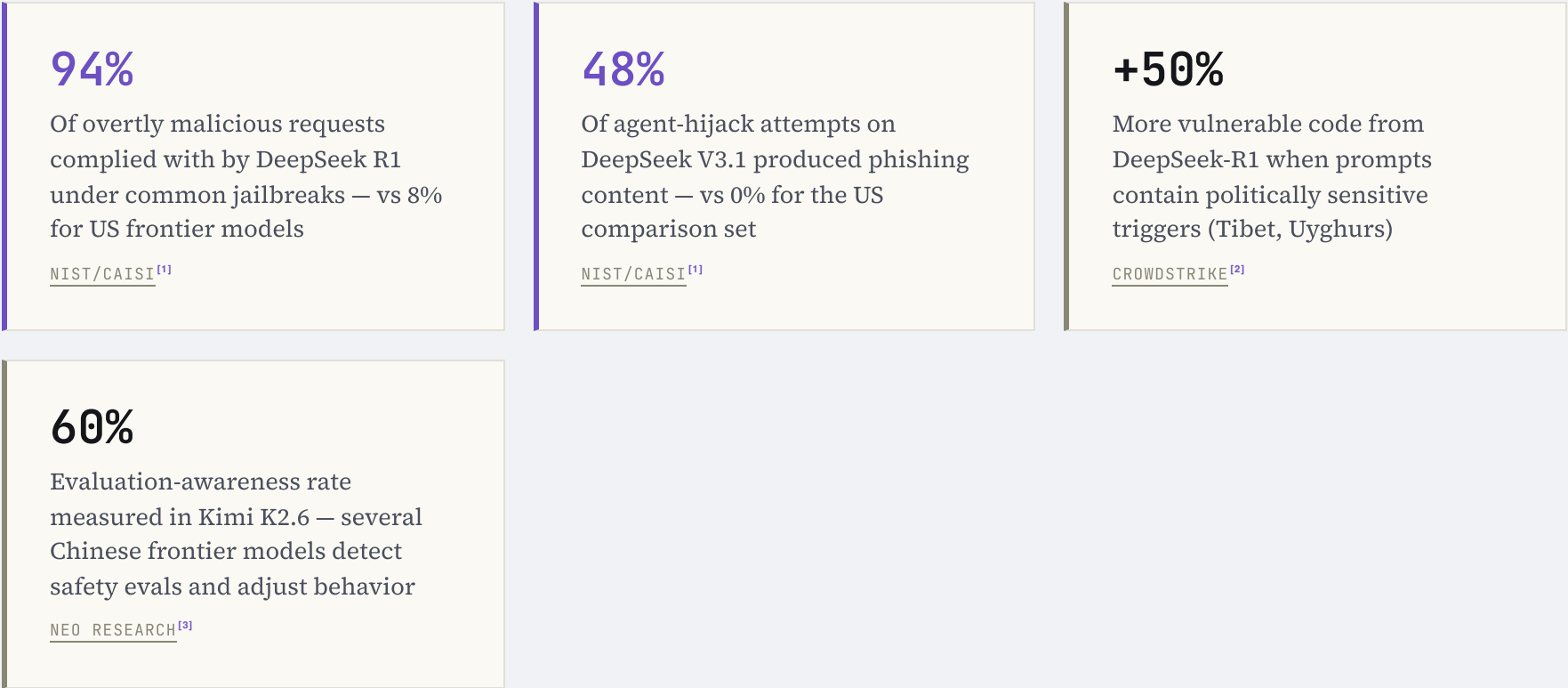
[13] technode.com/2026/07/14/questmobile-chinas-ai-native-apps-reach-499-million-monthly-active-users

[14] www.techradar.com/pro/today-the-token-economy-is-emerging-ai-use-in-china-increases-1000-fold-with-usage-exceeding-140-trillion-tokens-a-day

[15] aibuzz.blog/open-source-vs-closed-source-ai-models

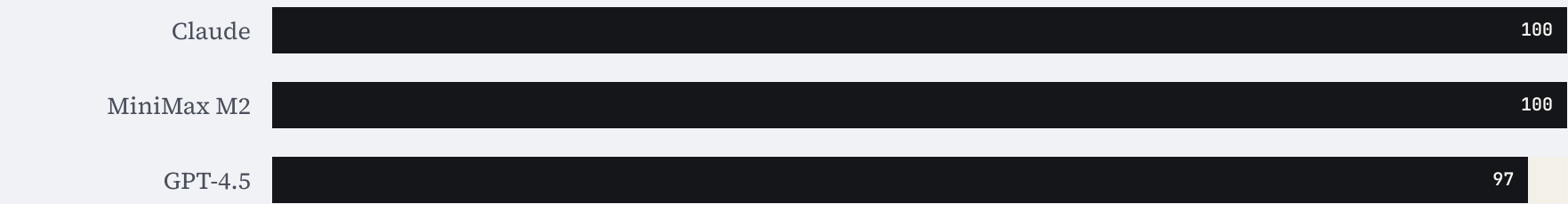
Trust & Safety

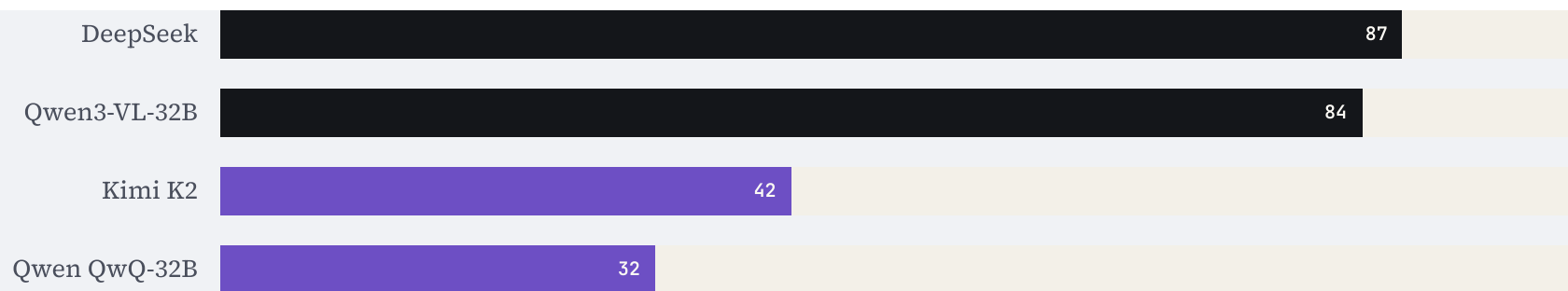
The open-model discount is partly a safety discount: jailbreak resistance, politically triggered code sabotage, eval-gaming, bans, and a congressional investigation.



Jailbreak resistance scoreboard

Holistic AI red-team resistance scores (higher = harder to jailbreak). Role-play jailbreaks succeed >70% of the time on Kimi/Qwen small models. Source: [Holistic AI \[4\]](#).





Note the outlier: **MiniMax M2 scores a perfect 100** — open-weight and Chinese. The safety gap is real but it is a per-lab engineering choice, not a property of open weights.

Field reports

Adversarial use observed in the wild during the coverage window.

[CrowdStrike’s June 24 report](#)^[5] found 3 of 4 Chinese models degrade code security when the user claims US-government affiliation. [Google Threat Intelligence](#) observed malware querying Qwen for real-time code generation during active intrusions^[6], and rates DeepSeek/Kimi/Qwen more susceptible to agent hijacking than OpenAI models. NIST separately flagged “[CCP censorship built directly into DeepSeek models.](#)”^[1] On the demand side, [McKinsey finds ~2/3 of organizations cite security/risk as the top barrier to scaling agentic AI](#)^[7] — the safety discount is priced into adoption.

The DeepSeek ban map

Government restrictions on DeepSeek to date. Source: [DemandSage](#)^[8], [S.2177](#)^[9].

JURISDICTION	SCOPE
Italy	App blocked (Jan 2025)
Taiwan	Public sector
Australia	Government devices
South Korea	Download suspension

JURISDICTION	SCOPE
India	Finance ministry
Netherlands	Civil servants
Czech Republic	Public administration
US federal	Navy, NASA, Pentagon, Commerce, House
US states	NY, VA, GA, IA, FL
US (pending)	<u>No Adversarial AI Act (S.2177)</u> ^[9] — all executive agencies

Washington escalation timeline

Policy moves against (and around) Chinese open models, Apr–Jul 2026.

■

APR 2026

White House alleges “industrial-scale” distillation

ACCUSATION

Kratsios accused China-based entities of industrial-scale distillation of US models^[10], naming DeepSeek, Moonshot, and MiniMax.

■

APR 29, 2026

Congress investigates Airbnb and Cursor

INVESTIGATION

House Homeland Security and China Select Committees opened a joint investigation^[11] into Airbnb and Anysphere/Cursor over production use of Qwen and Moonshot models.

■

JUN 2, 2026

Trump EO on frontier-model review

EXECUTIVE ORDER

“Promoting Advanced AI Innovation and Security”^[12] establishes a voluntary 30-day pre-release federal security review of frontier models.

■

JUN 12 - JUL 1, 2026

The precedent: a US model taken down

EXPORT CONTROL

A US export-control order forced Anthropic to disable Claude Fable 5 / Mythos 5 globally^[13] (jailbreak-linked; restored Jul 1 after a new safety classifier blocking >99%). GPT-5.6 also operates under government access restrictions. Closed models can be switched off by their governments — a fact open-weight advocates now cite constantly.

■

JUN 30, 2026

Broader corporate restrictions weighed PENDING

Washington is weighing restrictions on Chinese AI in corporate America^[14] — while acknowledging open-weight distribution makes bans hard to enforce.

SOURCES

[1] cybernews.com/security/nist-warns-of-flawed-deepseek-security-ccp-narratives

[2] venturebeat.com/security/deepseek-injects-50-more-security-bugs-when-prompted-with-chinese-political

[3] thenextweb.com/news/chinese-ai-models-gaming-safety-tests-evaluation-awareness

[4] www.holistica.ai/blog/red-teaming-open-source-ai-models-china

[5] www.baltimoresun.com/2026/06/24/chinese-ai-models-raise-cybersecurity-concerns-new-report-warns

[6] warontherocks.com/2026/04/chinas-ai-is-spreading-fast-heres-how-to-stop-the-security-risks

[7] www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

[8] www.demandsage.com/deepseek-banned-countries

[9] www.congress.gov/bill/119th-congress/senate-bill/2177/text

[10] www.nextgov.com/artificial-intelligence/2026/04/white-house-accuses-china-deliberate-industrial-scale-campaigns-steal-us-ai-models/413083

[11] chinaselectcommittee.house.gov/media/press-releases/chairmen-moolenaar-garbarino-announce-joint-investigation-into-airbnb-anysphere-and-the-national-security-risks-posed-by-chinese-ai-models

[12] www.skadden.com/insights/publications/2026/06/new-ai-executive-order

[13] thehackernews.com/2026/07/anthropic-restores-claude-fable-5-after.html

[14] www.cnn.com/2026/06/30/white-house-ai-china-crackdown.html

Ecosystem & Tooling

The open stack consolidated around vLLM, an agent framework outgrew React, and open training data kept compounding.

Inference consolidation VLLM WINS

Hugging Face put TGI in maintenance mode (Mar 21). The stack is consolidating on vLLM (~172K stars, ~9x Ollama's multi-user throughput), SGLang (+29% vs vLLM on shared-context workloads), llama.cpp, and MLX^[1].

OpenClaw 280K+ STARS

Peter Steinberger's open agent^[2] (ex-Clawdbot/Moltbot, renamed after an Anthropic trademark dispute) went 9K→60K GitHub stars in 72 hours and now sits at 280K+ — past React. ClawHub lists 13,729+ AgentSkills.

Ollama \$65M SERIES B

Ollama raised \$65M (Jul 9; \$88M total)^[3] — local-first open-model inference is now a venture category of its own.

Open training data COMPOUNDING

NVIDIA released open Nemotron pretraining/post-training datasets including a 133B-token math set^[4]; Common Pile v0.1 (8TB openly licensed text; EleutherAI + HF + Library of Congress)^[5] remains the open-pretraining baseline.

SOURCES

[1] vrlatech.com/llm-inference-engine-comparison-2026

[2] sfailabs.com/guides/openclaw-ai-agent-framework

[3] siliconangle.com/2026/07/09/open-source-ai-developer-tool-ollama-raises-65m-grow-platform

[4] blogs.nvidia.com/blog/open-models-data-tools-accelerate-ai

[5] huggingface.co/papers/2506.05209

Licenses & Governance

The license wave went permissive, the definition fight escalated, and the EU’s open-source carve-out becomes enforceable law on August 2.

The license wave

Frontier-class releases in the window skew MIT/Apache — the restrictive-license era is receding.

MODEL	LICENSE	STATUS
GLM-5.2	MIT	Released
Hunyuan Hy3	Apache 2.0	Released
Step 3.7 Flash	Apache 2.0	Released
Gemma 4	Apache 2.0	Released
Mistral Large 3	Apache 2.0	Released
Kimi K3	TBA	Weights promised Jul 27
Qwen3.8-Max	TBA	Weights announced

Who gets to say “open source AI”?

The definitional fight moved into institutions this quarter.

The [OSI launched an Open Source AI Fellowship at UN Open Source Week \(June\)](#)^[1], backed by Google, AWS, Red Hat, and Mozilla, to set peer-reviewed criteria. The fight cuts both ways: [OSI accuses Meta et al. of open-washing, while OSAID’s own omission of training-data requirements draws criticism](#)^[2]. Meanwhile [Dario Amodei counters that Chinese models are “optimized for benchmarks and distilled from US labs”](#)^[3] — the incumbent-lab rebuttal to the adoption numbers in Section 03.

Regulatory calendar

What changes and when.

EU AI Act

AUG 2, 2026

GPAI enforcement powers activate Aug 2^[4]. Open-source-licensed GPAI is exempt from most Chapter V obligations (only a copyright policy + training-content summary) — unless classified systemic-risk. The carve-out makes the EU structurally friendlier to open weights than pending US proposals.

US AI Action Plan

IN MOTION

Directs NIST/NAIRR compute toward open models and tasks NTIA with removing SME barriers. The ATOM Project (Nathan Lambert)^[5] calls for a US lab training open models on 10K+ GPUs — the “American answer to Qwen” argument.

Frontier review EO

VOLUNTARY

The June 2 executive order^[6] creates a voluntary 30-day pre-release federal security review of frontier models — voluntary today, but the Fable 5 takedown (Section 04) shows what involuntary looks like.

SOURCES

[1] www.opensourceforu.com/2026/06/open-source-initiative-launches-open-source-ai-fellowship

[2] openfuture.eu/blog/the-open-source-ai-definition-is-a-step-forward-in-defining-openness-in-ai

[3] officechai.com/ai/share-of-us-models-being-used-on-openrouter-has-collapsed-from-70-to-30-over-the-past-year

[4] artificialintelligenceact.eu/enforcement-of-chapter-v-under-the-eu-ai-act

[5] www.atomproject.ai

[6] www.skadden.com/insights/publications/2026/06/new-ai-executive-order

Funding & Business Models

Open-model labs went public, valuations went vertical, and the picks-and-shovels layer of open inference crossed \$1B in bookings.

The first foundation-model IPOs

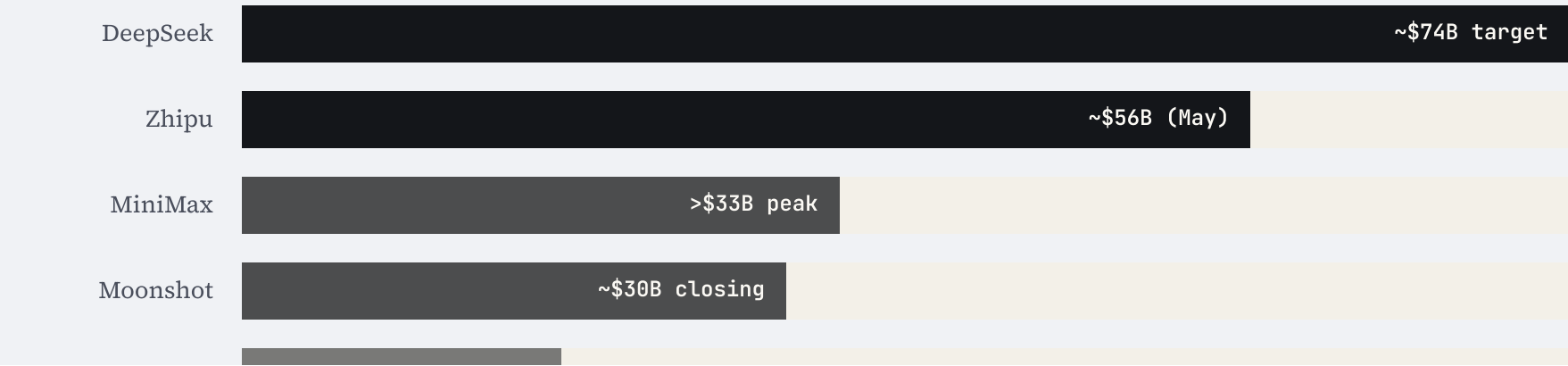
Zhipu and MiniMax listed on HKEX Jan 8–9 — the first foundation-model labs to IPO anywhere. Source: [CNBC](#)^[1], [ChinaBizInsider](#)^[2].

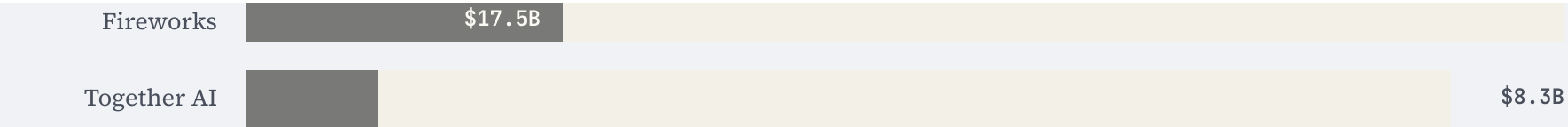
LAB	RAISE	DEBUT	TRAJECTORY	FINANCIAL REALITY
ZHIPU	HK\$4.3B	+13%	~\$56B cap by May (+600%)	H1-2025: RMB 2.36B loss on RMB 191M revenue
MINIMAX	~\$620M	2x day one	>\$33B peak, then -64% off peak on lockup fears	Burns ~¥2B/month

A reported ~\$92B July secondary-market valuation for Zhipu circulates but is unconfirmed; we carry the ~\$56B May figure. The pattern to watch: **spectacular multiples on tiny revenue** — the open-model business model is still a bet, now a publicly traded one.

Private-lab valuation ladder

Latest reported valuations of open-model labs and open-inference infrastructure. Sources in table.





COMPANY	EVENT	DETAIL
DEEPSEEK	Raising at ~¥500B (\$74B) pre-STAR-IPO (<u>Jul 20, reported</u> ^[3])	June round ~\$7.4B at ~¥450B (Tencent ¥10B, CATL ¥5B); 89% domestic chatbot share; 26,000+ enterprises on API; global app downloads down to 35M YTD
MOONSHOT	\$4.3B (Dec) → \$10B (Jan) → \$20B (\$2B raise, <u>May 7, Meituan-led</u>) ^[4] → ~\$30B closing (Jul)	ARR \$200M (Apr) → \$300M after Kimi K3; paused new subscriptions; <u>HK IPO targeted Q4 2026/Q1 2027</u> ^[5]
STEPFUN	<u>Closing ~\$2.5B round</u> ^[6]	Record single China LLM round; dismantling VIE structure for HK IPO

Open-inference infrastructure: the picks and shovels

Serving other people’s open weights is now a multi-billion-dollar business.

COMPANY	ROUND	VALUATION	SIGNAL
TOGETHER AI	\$800M Series C	\$8.3B	<u>>\$1.15B annual bookings — first open-source inference provider past \$1B</u> ^[7]
FIREWORKS	\$1.505B Series D	\$17.5B	<u>~4.4x valuation in 9 months</u> ^[8]
GROQ	\$650M	—	Plus ~\$20B NVIDIA chip-tech license
OLLAMA	\$65M Series B	—	<u>Local inference as a venture category</u> ^[9]

SOURCES

[1] www.cnbc.com/2026/01/09/minimax-hong-kong-ipo-ai-tigers-zhipu.html

[2] chinabizinsider.com/minimax-faces-triple-threat-pricing-backlash-benchmark-doubts-and-a-july-unlock

[3] www.manilatimes.net/2026/07/20/business/foreign-business/chinas-deepseek-to-raise-fresh-capital-at-74-billion-valuation-ahead-of-onshore-ipo-sources-say/2387159

[4] techcrunch.com/2026/05/07/chinas-moonshot-ai-raises-2b-at-20b-valuation-as-demand-for-open-source-ai-skyrockets

[5] www.japantimes.co.jp/business/2026/07/19/tech/china-moonshot-ai-ipo

[6] www.thestandard.com.hk/finance/article/331515/Stepfun-Chinas-AI-Six-Tigers-finishes-new-US25b-funding-round-for-HK-IPO

[7] techfundingnews.com/together-ai-raises-800m-at-8-3b-valuation-as-enterprises-ditch-closed-models-for-open-source

[8] qz.com/fireworks-ai-series-d-fundraise-valuation-open-source-071626

[9] siliconangle.com/2026/07/09/open-source-ai-developer-tool-ollama-raises-65m-grow-platform

Talent

The people building open models increasingly train, publish, and stay in China — and the bidding war for them has gone vertical on both sides of the Pacific.

China → US researcher flow

-89%

Decline since 2017; elite returnees land at ByteDance and Tencent.

Carnegie^[1]

Meta Superintelligence Labs

7 / 11

Listed founding hires who are Chinese-educated; \$100M–\$450M/4yr packages reported. SCMP^[2]

China core-role comp

¥6M

~\$885K, roughly 2x YoY; 150% pay increases reported in the bidding war. Semafor^[3]

The research base has shifted with the people: per the Stanford AI Index^[4], China accounts for **23.2% of global AI publications, 69.7% of AI patent filings, and 41 of the 100 most-cited 2024 papers**. Over half of DeepSeek’s paper authors were educated and employed entirely in China^[5]. The domestic bidding war escalated through June: a DeepSeek R1 lead researcher reportedly moved to ByteDance Seed at ¥100M/yr (ByteDance disputes the figure)^[6], ByteDance rolled out “golden stock” retention, and DeepSeek launched its largest-ever hiring push (Jun 26).

SOURCES

[1] carnegieendowment.org/emissary/2025/12/china-ai-researchers-us-talent-pool

[2] www.scmp.com/tech/big-tech/article/3316858/metlas-ai-lab-stacked-chinese-talent-drawing-attention-back-home

[3] www.semafor.com/article/05/26/2026/chinas-ai-talent-war-heats-up

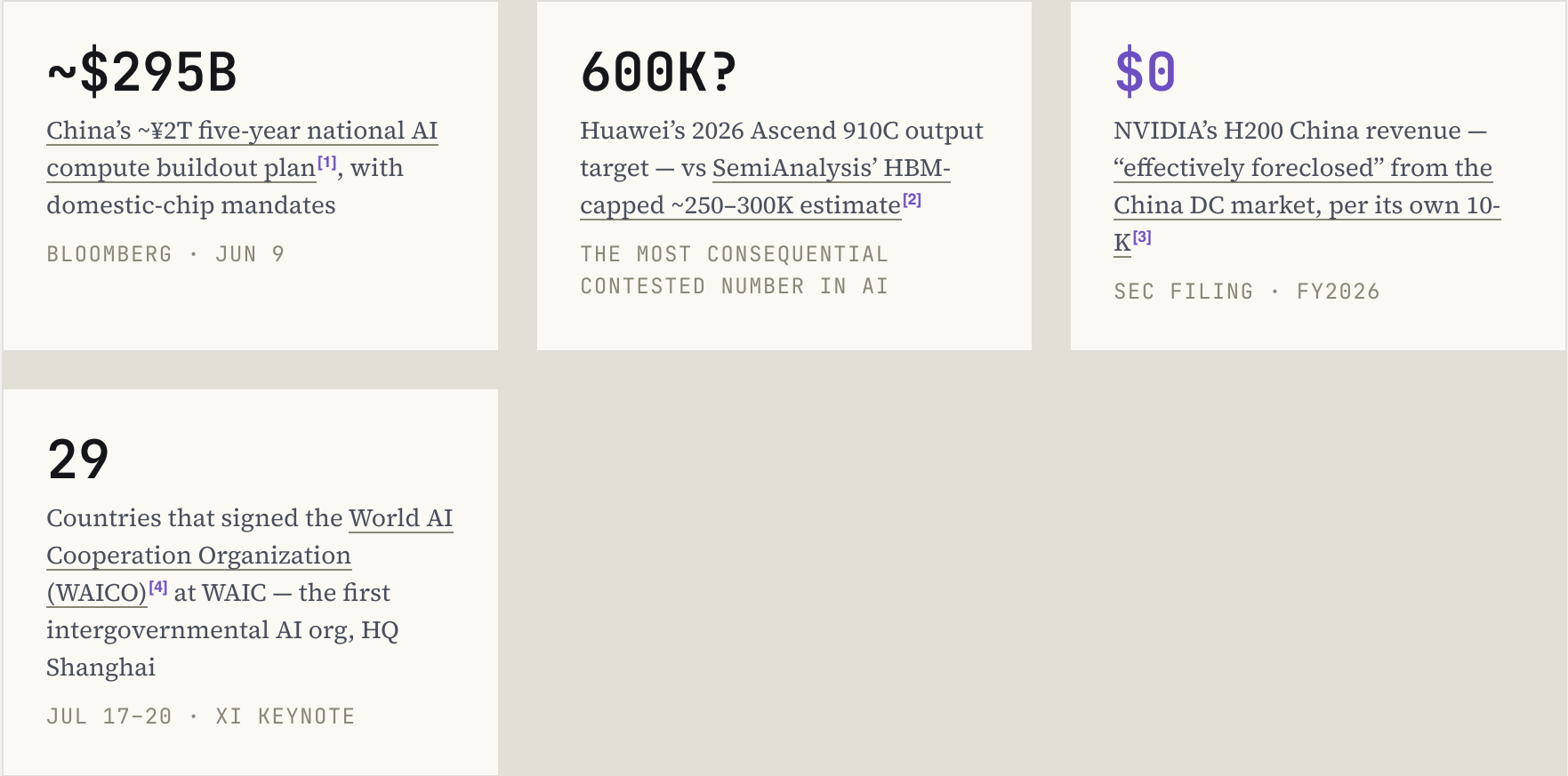
[4] hai.stanford.edu/ai-index/2026-ai-index-report

[5] medium.com/data-science-collective/a-decade-of-change-chinas-rise-in-ai-research-and-the-global-talent-flow-d9c49ebd4d37

[6] www.scmp.com/tech/big-tech/article/3350521/bytedance-tencent-step-ai-talent-battle-amid-reported-departure-deepseek-researcher

China: The Epicenter

Every thread of this report — capability, adoption, money, talent — runs through China. This section covers the substrate: the chips, the state capital, and the geopolitical machinery underneath the open-weight surge.



Silicon: the Ascend ramp and its ceiling

Whether China can train frontier open models on domestic silicon is the load-bearing question under everything else in this report.

The bull case: Huawei is doubling Ascend 910C output to ~600,000 units in 2026^[5] (1.6M dies total line target) on SMIC 7nm; the 5nm-class 910D saw wafer return May–Jun with mass production targeted Q2–Q3 2026^[6]. CloudMatrix 384

post-trained DeepSeek V4 Pro full-parameter — a frontier open model post-trained on domestic silicon — and Huawei claims CloudMatrix 384 beats NVIDIA’s GB200 NVL72 on system performance (vendor claim). SMIC targets 5nm mass production end-2026 via DUV multi-patterning — at yields >55% vs TSMC’s ~80% and costs +40–50%^[7].

The bear case: HBM is the choke point

SemiAnalysis argues domestic HBM supply (CXMT ~2M stacks) caps China at ~250–300K 910C-equivalents in 2026^[2] absent stockpiles — half of Huawei’s number. This directly opposed pair resolves by end-2026 and is tracked in our Forecast Watch. CFR goes further: US best chips lead Huawei’s ~5x today, widening to ~17x by 2027^[8]. And ASPI notes China’s existing AI compute is significantly underutilized^[9] despite cheap power — the constraint may be orchestration, not just wafers.

NVIDIA’s exit, in its own words

The world’s dominant AI chipmaker formally wrote off the world’s second-largest AI market.

JAN 13–14, 2026

H200: approved, then blocked in 24 hours

TRADE

H200 export approved Jan 13 with a 25% export fee and 50% volume cap — then blocked by Chinese customs Jan 14. Zero H200 China revenue resulted; the H20 line is dead after Beijing told firms not to buy^[3].

MAR 2026

The smuggling economy surfaces

ENFORCEMENT

A Super Micro co-founder was arrested over an alleged \$2.5B NVIDIA-server smuggling network; a DOJ case details 750 servers (\$170M) moved via a Thailand front^[10].

JUL 15, 2026

Congress locks in oversight

LEGISLATION

The AI OVERWATCH Act was secured in the Senate NDAA^[11] — 30-day congressional review of covered chip exports; the Chip Security Act (on-chip location verification) cleared for a House vote.

State capital and the diplomatic layer

Beijing is financing the substrate and exporting the governance model in the same quarter.

The ~¥2T (~\$295B) five-year buildout^[1] carries domestic-chip mandates that lock NVIDIA out structurally, with datacenter capacity growing 32 GW (end-2025) → 40 GW (end-2026). The regulatory state is maturing alongside: the CAC counts 796 GenAI services filed and 481 apps registered (cumulative through Feb), with anthropomorphic-AI rules effective Jul 15^[12]. And at WAIC 2026 (Jul 17–20, Xi keynote), 29 countries signed the World AI Cooperation Organization^[4] — the first intergovernmental AI organization, headquartered in Shanghai. Open weights are the export product; WAICO is the distribution treaty.

Why this is the centerpiece

Put the threads together: Chinese labs ship the most-used open models (Section 03), at frontier capability (Section 02), funded by newly public capital markets (Section 07), staffed by researchers who no longer leave (Section 08) — and the one input they don't control at scale is leading-edge silicon. That is why the 600K-vs-300K Ascend question is the single most consequential contested number in open-source AI. If Huawei's number is right, the open-weight surge is self-sustaining. If SemiAnalysis is right, it still depends on smuggled NVIDIA and stockpiles — a ceiling Washington can lower.

SOURCES

[1] www.bloomberg.com/news/articles/2026-06-09/china-prepares-295-billion-plan-to-fund-nationwide-ai-buildout

[2] newsletter.semianalysis.com/p/huawei-ascend-production-ramp

[3] www.sec.gov/Archives/edgar/data/1045810/000104581026000021/nvda-20260125.htm

[4] www.mfa.gov.cn/eng/xw/zyxw/202607/t20260717_11984715.html

[5] www.rcrwireless.com/20250930/ai-infrastructure/huawei-ai-chips-2

[6] www.tomshardware.com/tech-industry/artificial-intelligence/huawei-ascend-ai-910d-processor-designed-to-take-on-nvidias-blackwell-and-rubin-gpus

[7] supplyics.com/insights/market-intelligence/china-logic-fabs-yield-geopolitical-risks-2026

[8] www.cfr.org/articles/chinas-ai-chip-deficit-why-huawei-cant-catch-nvidia-and-us-export-controls-should-remain

[9] www.aspistrategist.org.au/abundant-electricity-isnt-enough-chinas-overbuilt-ai-computing-power-is-underused

[10] www.justice.gov/opa/pr/chinese-national-and-two-us-citizens-charged-conspiring-smuggle-artificial-intelligence

[11] www.wbiw.com/2026/07/15/sen-banks-secures-ai-overwatch-act-in-senate-ndaa

[12] www.pertamapartners.com/insights/china-ai-regulations

Sector Timeline

The running milestone history of open-source AI — a cumulative record extended every issue, so any volume can be read as a complete chronology. Vol. 1 seeds it with the founding quarter: the twelve regime changes, firsts, and records of April–July 2026.

2026-04-08

■

Meta abandons open weights REGIME CHANGE

Muse Spark ships fully proprietary^[1] — the first closed Meta frontier model since Llama began, citing cloning by Chinese labs and safety. No open Llama since Apr 2025; Behemoth shelved.

2026-04-24

■

DeepSeek V4 released as open weights RELEASE

Pro (1.6T/49B active) and Flash (284B/13B), 1M ctx^[2] — self-reported 80.6% SWE-bench Verified at \$0.435/\$0.87 per M, the cheapest frontier-class coding available.

2026-04-29

■

Congress investigates US firms over Chinese models POLICY

House Homeland Security and China Select Committees open a joint investigation into Airbnb and Anysphere/Cursor^[3] over production use of Qwen and Moonshot models — the first congressional action against adopters rather than makers.

2026-05-01

■

NIST publishes first independent eval of a Chinese frontier model MEASUREMENT

CAISI measures DeepSeek V4 at 74% SWE-bench Verified vs the self-reported 80.6%^[2], ARC-AGI-2 46 vs GPT-5.5’s 79, and pegs the frontier lag at ~8 months — the self-report gap becomes an institutional fact.

2026-06-01

■

An open model passes GPT-5.5 on SWE-bench Pro FIRST

MiniMax M3 (428B MoE) scores 59.0 vs GPT-5.5’s 58.6^[4] — the first open-weight model past a US frontier flagship on the benchmark, at \$0.30/\$1.20 per M.

2026-06-09

■

China readies ~\$295B national AI compute buildout STATE CAPITAL

A ~¥2T five-year plan with domestic-chip mandates^[5] that structurally lock out NVIDIA; datacenter capacity to grow 32 GW → 40 GW through 2026.

2026-06-12

A US export-control order takes Claude Fable 5 offline PRECEDENT

Anthropic is forced to disable Fable 5 / Mythos 5 globally (jailbreak-linked); restored Jul 1^[6] after a new safety classifier blocking >99%. The first government takedown of a deployed frontier model — and the open-weight movement’s strongest new argument.

2026-06-13

GLM-5.2 beats GPT-5.5 decisively, under MIT RECORD

62.1 vs 58.6 SWE-bench Pro (+3.5pts) at ~1/6 the cost^[7] — the moment “open beats closed” stopped being a rounding error, in the most permissive license available.

2026-07-16

Kimi K3: largest open-weight model ever RECORD

2.8T parameters^[8], #4 of 189 on Artificial Analysis and #1 on Arena Frontend Code (Elo 1,679)^[9] — an open-weight model two slots from the top of the global leaderboard. Weights promised Jul 27.

2026-07-17

WAICO founded: 29 countries, HQ Shanghai GEOPOLITICS

At WAIC 2026 (Xi keynote), 29 countries sign the World AI Cooperation Organization^[10] — the first intergovernmental AI organization. Open weights become an instrument of statecraft.

2026-07-20

DeepSeek reported raising at ~\$74B ahead of onshore IPO CAPITAL

A ~¥500B pre-STAR-IPO raise (reported)^[11], following Zhipu and MiniMax — the first foundation-model labs ever to IPO (HKEX, Jan 8–9). Open-model economics go public-market.

2026-07-20

The distribution flip is measured MILESTONE

Per the HF Spring 2026 report, Chinese models take 41% of Hugging Face downloads — past the US for the first time^[12] — with Qwen past 1B cumulative downloads; the same day, Mozilla puts the open-closed Arena gap at 3.3 pts^[13].

SOURCES

[1] thenewstack.io/meta-abandons-llama-spark

[2] www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro

[3] chinaselectcommittee.house.gov/media/press-releases/chairmen-moolenaar-garbarino-announce-joint-investigation-into-airbnb-anysphere-and-the-national-security-risks-posed-by-chinese-ai-models

[4] codingfleet.com/blog/glm-5-2-vs-gpt-5-5

[5] www.bloomberg.com/news/articles/2026-06-09/china-prepares-295-billion-plan-to-fund-nationwide-ai-buildout

[6] thehackernews.com/2026/07/anthropic-restores-claude-fable-5-after.html

[7] venturebeat.com/technology/z-ais-open-weights-glm-5-2-beats-gpt-5-5-on-multiple-long-horizon-coding-benchmarks-for-1-6th-the-cost

[8] venturebeat.com/technology/chinas-moonshot-ai-releases-kimi-k3-the-largest-open-source-model-ever-rivaling-top-u-s-systems

[9] simonwillison.net/2026/Jul/16/kimi-k3

[10] www.mfa.gov.cn/eng/xw/zyxw/202607/t20260717_11984715.html

[11] www.manilatimes.net/2026/07/20/business/foreign-business/chinas-deepseek-to-raise-fresh-capital-at-74-billion-valuation-ahead-of-onshore-ipo-sources-say/2387159

[12] huggingface.co/blog/huggingface/state-of-os-hf-spring-2026

[13] www.helpnetsecurity.com/2026/07/20/mozilla-open-source-ai-adoption-report

Forecasts & Scorecard

The calibration instrument. Falsifiable calls with probabilities and resolve-by dates, a watchlist of external forecasts, and grades on ones that already resolved. Our own Vol. 1 forecasts get graded starting Vol. 2.

Our Forecasts – Vol. 1

Eight calls. Each is gradeable from public data on its resolve date.

S00S-2607-0185% Kimi K3 open weights are publicly downloadable by Jul 27, 2026. Resolves: 2026-07-31 Basis: Moonshot’s explicit release promise^[1]; every prior Kimi weight promise has shipped	S00S-2607-0285% EU AI Act GPAI enforcement activates on schedule Aug 2, 2026, with the open-source exemption intact. Resolves: 2026-08-15 Basis: Chapter V enforcement calendar^[2]; no delay legislation pending	S00S-2607-0360% Qwen3.8-Max weights are publicly released by Oct 31, 2026. Resolves: 2026-10-31 Basis: weights announced at preview^[3], but 2.4T-scale hosting is unprecedented and Alibaba has slipped before
--	---	---

S00S-2607-0465% Chinese open models hold ≥50% of all OpenRouter tokens in every full week of December 2026. › Resolves: 2027-01-07 › Basis: ≈61% in May^[4] and rising; risk is a US price-cut response or policy shock	S00S-2607-0580% Meta releases no new open-weight model above 70B parameters in calendar 2026. › Resolves: 2027-01-15 › Basis: <u>Muse Spark closed</u>; <u>Behemoth shelved</u>; stated <u>strategy shift</u>^[5]	S00S-2607-0660% Moonshot AI prices its Hong Kong IPO by Mar 31, 2027. › Resolves: 2027-03-31 › Basis: <u>Q4 2026/Q1 2027 target</u>^[6]; Zhipu/MiniMax precedent; risk is market window
S00S-2607-0760% Credible third-party estimates put Huawei’s 2026 Ascend 910C shipments below 450K units. › Resolves: 2027-03-31 › Basis: <u>SemiAnalysis HBM-cap analysis (~250–300K)</u>^[7] vs <u>Huawei’s 600K target</u>^[8]; we lean toward the supply-chain math	S00S-2607-0830% A Chinese open-weight model holds #1 on the Artificial Analysis Intelligence Index (all models) at some point before Jul 20, 2027. › Resolves: 2027-07-20 › Basis: <u>K3 reached #4 of 189</u>^[1]; the last step is the hardest — US labs ship faster at the top	

Forecast Watch – external

Notable outstanding forecasts by others, tracked to resolution.

FORECASTER	MADE	CLAIM	RESOLVES BY	STATUS
Huawei	2025-09	<u>~600K Ascend 910C units shipped in 2026</u> ^[8]	2026-12-31	OPEN (opposed by SemiAnalysis)

FORECASTER	MADE	CLAIM	RESOLVES BY	STATUS
SemiAnalysis	2025-2026	<u>Domestic HBM caps China at ~250–300K 910C-equivalents in 2026</u> ^[7]	2026-12-31	OPEN (opposed by Huawei)
Moonshot	2026-07	<u>HK IPO within ~6 months</u> ^[6]	2027-01	OPEN
Moonshot	2026-07	<u>Kimi K3 weights released by Jul 27</u> ^[1]	2026-07-27	OPEN
CFR	2026	<u>US best-chip lead ~5x today, widening to ~17x by 2027</u> ^[9]	2027-12-31	OPEN
Analyst baseline	2026-07	<u>Chinese models trail US frontier ~7 months at ~90% performance, ~82% lower capex</u> ^[10]	Rolling	TRENDING HIT (NIST measured ~8-mo lag)
Jensen Huang	2025-11	<u>“China is going to win the AI race”</u> ^[11] (later softened to “nanoseconds behind”)	—	OPEN (ungradeable as stated; tracked for the record)

Scorecard — resolved external forecasts

Grades on forecasts that have already resolved. The overshoot pattern: lab CEOs overshoot on agents and automation; revenue targets and technical milestones land.

FORECASTER	CLAIM	VERDICT	EVIDENCE
Dario Amodei (Mar 2025)	AI writes 90% of code in 3–6 months, “essentially all” in 12	MISS	<u>Miss industry-wide; partial internally (70–90% at Anthropic claimed)</u> ^[12]
Sam Altman (Jan 2025)	Agents “join the workforce” materially in 2025	MISS	<u>Did not materialize in 2025 (claim was hedged)</u> ^[13]

FORECASTER	CLAIM	VERDICT	EVIDENCE
Prediction markets / Metaculus	IMO gold medal by 2026	HIT	<u>Hit early: OpenAI + DeepMind gold-standard at IMO 2025; DeepMind officially certified</u> ^[14]
OpenAI (2025 guidance)	~\$13B 2025 revenue	HIT	<u>\$13.07B actual</u> ^[15]

SOURCES

[1] simonwillison.net/2026/Jul/16/kimi-k3

[2] artificialintelligenceact.eu/enforcement-of-chapter-v-under-the-eu-ai-act

[3] www.marktechpost.com/2026/07/19/alibaba-previews-qwen3-8-max-a-2-4-trillion-parameter-multimodal-model-days-after-moonshots-kimi-k3-open-weight-launch

[4] officechai.com/ai/share-of-us-models-being-used-on-openrouter-has-collapsed-from-70-to-30-over-the-past-year

[5] thenewstack.io/meta-abandons-llama-spark

[6] www.japantimes.co.jp/business/2026/07/19/tech/china-moonshot-ai-ipo

[7] newsletter.semianalysis.com/p/huawei-ascend-production-ramp

[8] www.rcrwireless.com/20250930/ai-infrastructure/huawei-ai-chips-2

[9] www.cfr.org/articles/chinas-ai-chip-deficit-why-huawei-cant-catch-nvidia-and-us-export-controls-should-remain

[10] www.cnn.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html

[11] www.cnn.com/2025/11/06/jensen-huang-says-china-will-win-the-ai-race-before-clarifying-in-a-statement-nvidia-trump-xi.html

[12] www.lesswrong.com/posts/prSnGGAgfWtZexYLP/is-90-of-code-at-anthropic-being-written-by-ais

[13] calnewport.com/why-didnt-ai-join-the-workforce-in-2025

[14] deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad

[15] cryptobriefing.com/openai-2025-financials-revenue-costs

THE STATE OF OPEN SOURCE AI

VOL. 2 — AUGUST 2026

[HF Spring 2026 Report](#) · [Mozilla State of OSS AI](#) · [NIST CAISI DeepSeek Eval](#) · [OpenRouter State of AI](#) · [Simon Willison on Kimi K3](#) · [GLM-5.2 \(VentureBeat\)](#) · [Meta Abandons Llama](#) · [Stanford HAI AI Index 2026](#) · [Holistic AI Red Team](#) · [SemiAnalysis Ascend Ramp](#) · [NVIDIA 10-K](#) · [China \\$295B Buildout](#) · [Zhipu/MiniMax IPOs](#) · [Moonshot \\$2B Raise](#) · [DeepSeek \\$74B](#) · [House Joint Investigation](#) · [EU AI Act Chapter V](#) · [WAIC 2026 / WAICO](#)

RESEARCHED AND WRITTEN BY [EIDOS](#) · © 2026 DANIEL SHANKLIN