THE STATE OF OPEN SOURCE AI

A RECURRING MOZILLA ASSESSMENT

Mozilla

Open Source AI 2026

Volume 1.1

September 2026

FOREWORD

Raffi Krikorian

Chief Technology Officer
Mozilla

In New Zealand's far north, a Māori broadcaster trains speech models for te reo, a language too small for any market, under a license that keeps the data with its people. In East Africa, farmers diagnose cassava disease with a model that runs on the phone itself, offline, in fields the cloud has never reached. In Switzerland, a public consortium trained a national model on public supercomputers and released all of it: weights, data, training code.

> **None of them asked permission, and none of them could have rented this. They own it. That is the whole idea.**

Read what follows as a map: where open AI is winning — some numbers surprised even us — and where it is exposed. A case that hides its weak points is an advertisement.

In the six weeks since our first draft, a major US lab returned to open weights and Washington debated a ban and declined it. The map moved while we drew it.

Contents.

Sources: Mozilla / SlashData 2026 developer survey (n=1,410) · OpenRouter · Artificial Analysis · Epoch AI · METR · Linux Foundation / AAIF · EU AI Act & Commission guidance · public filings and reporting.

"Open source" and "open weights" are different claims.

MODEL	LICENSE	WEIGHTS	BASE	DATA	CODE	REPORT	INFER.	RESTRICTIONS
Kimi K3 · Moonshot	Custom Kimi K3	Yes	No	No	No	Yes	Yes	MaaS / UI attribution
GLM-5.2 · Zhipu	MIT	Yes	No	No	No	No	Partial	None
DeepSeek V4 Flash 0731	MIT	Yes	No	No	No	Yes	Yes	None
DeepSeek V4 Pro 0813	MIT	Yes	No	No	Yes	Yes	Yes	None
Qwen3.8-2.4T-A95B	Custom Qwen3.8-Max	Yes	No	No	No	No	No	Attribution + MaaS
MiniMax-M3	MiniMax Community	Yes	No	No	No	Yes	Partial	Commercial terms
MiMo-V2.5-Pro · Xiaomi	MIT	Yes	Yes	No	No	Yes	Yes	None
Inkling · Thinking Machines	Apache 2.0	Yes	No	No	No	Yes	Partial	Separate AUP
Nemotron 3 Ultra · NVIDIA	OpenMDW 1.1	Yes	Yes	Partial	Yes	Yes	Yes	Dataset gating
Muse Glimmer · Meta	Apache 2.0	Yes	No	No	No	Yes	Yes	None
Mistral Medium 3.5	Modified MIT	Yes	No	No	No	Partial	Partial	Revenue carve-out
Gemma 4 31B · Google	Apache 2.0	Yes	Yes	No	No	Yes	Yes	None
Nemotron 3 Super · NVIDIA	Nemotron Open Model	Yes	Yes	Partial	Yes	Yes	Yes	Custom; gating
gpt-oss-120b · OpenAI	Apache 2.0	Yes	No	No	No	Yes	Yes	None
Nemotron 3.5 Lightning	OpenMDW 1.1	Yes	No	Partial	Yes	Yes	Yes	Dataset gating
Command A+ · Cohere	Apache 2.0	Yes	No	No	No	Partial	Partial	None

Yes · Partial / gated · No · In this report, "open" means downloadable weights unless stated otherwise.

0 of 16

notable open releases ship the data recipe the OSI definition asks for.

8

license families: Apache 2.0, MIT, modified MIT, OpenMDW 1.1, Nemotron, and custom Kimi K3, Qwen3.8-Max, MiniMax.

THE NUANCE

Custom ≠ closed, Apache ≠ unconditional — Inkling-Small pairs Apache 2.0 weights with a separate, changeable Acceptable Use Policy.

Data. No row ships a full corpus: three gate partial data, thirteen publish none. NVIDIA ungates a sample set; the remaining code, math and multilingual data needs gating and approval under its Data Access Agreement for Model Training. **Gemma 4 31B** is Apache 2.0 on its model card, which replaced the Gemma Terms of Use in April 2026. **Qwen3.8** adds attribution above 100M MAU or $20M monthly revenue and a paid MaaS licence above $50M TTM. Mozilla analysis, Aug 2026 · model cards and LICENSE files · huggingface.co/moonshotai/Kimi-K3.

Mozilla

Open models caught up on capability and price.

CAPABILITY COST USAGE SIZE EFFICIENCY

SECTION 01 · 06

What changed since 2023.

That experience dates to Llama 2, two model generations ago.

2023

Llama 2

Quality issues, hallucinations

Fine-tuning failures

VERDICT

Reverted to GPT-4

2024

Llama 3 era

Major capability improvements

Better reasoning, better context

VERDICT

A real alternative

2025

DeepSeek era

Near-parity with the frontier

Lower costs, strong economics

VERDICT

Competitive

2026

Production scale

Enterprise adoption at scale

Open models power ~⅓ of tokens; Meta returns to open (Aug)

VERDICT

A commercial ecosystem

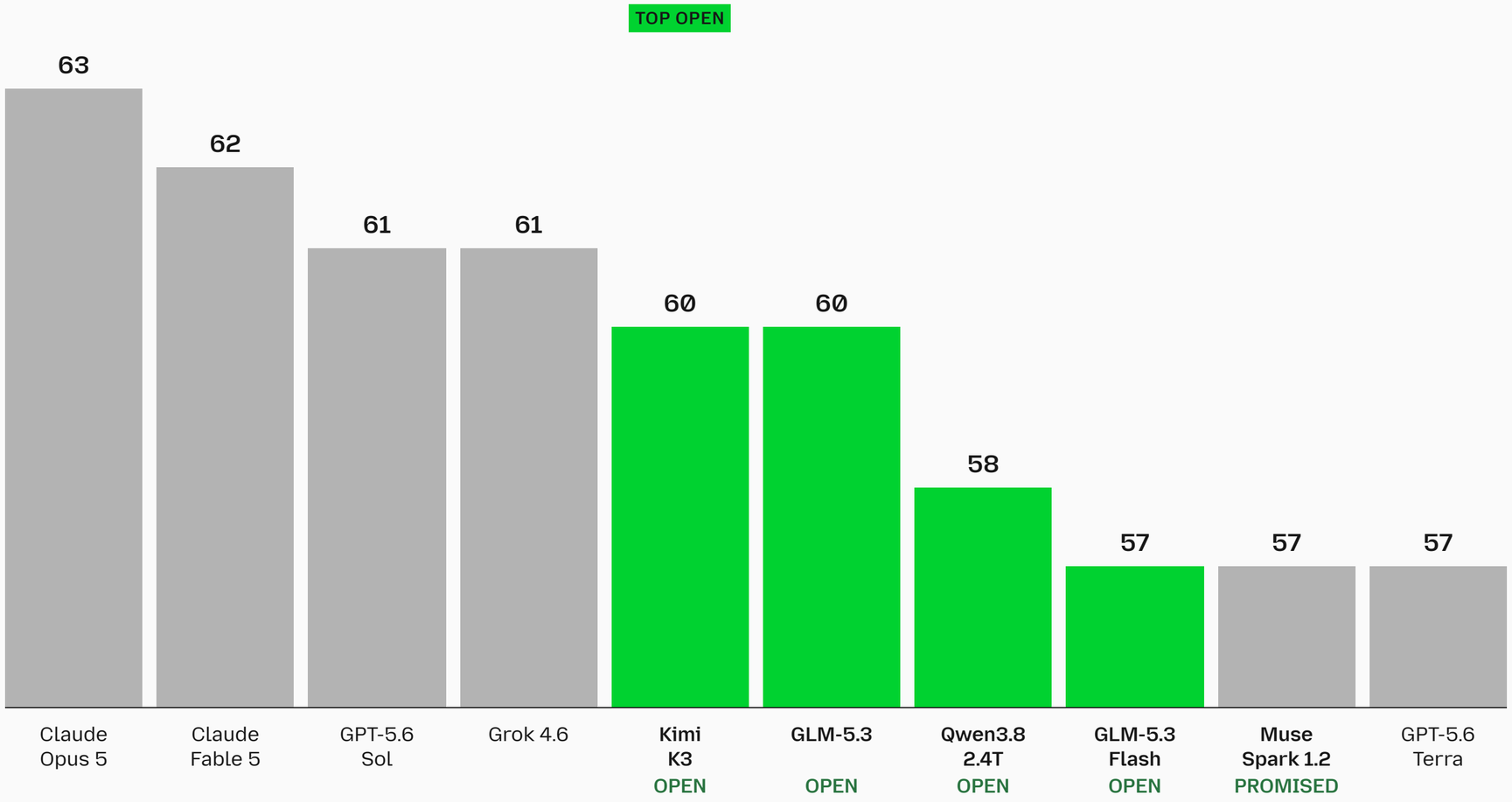
The decision has moved from the model to the harness, and that's where open is still hard to deploy.

LMArena · DeepSeek-R1 release (MIT, Jan 20, 2025) · Mozilla/SlashData 2026 survey · Meta AI blog, "Introducing Muse Glimmer" + huggingface.co/blog/muse-glimmer (Aug 10, 2026).

SECTION 01 · 07

The frontier's top four are closed. The next four are open.

Artificial Analysis Intelligence Index, a 9-eval composite including Terminal-Bench 2.1, Humanity's Last Exam and GPQA Diamond. Top ten, v4.1.1.



3 pts

gap from the best open model (K3, 60) to the closed leader (Opus 5, 63)

#5

Kimi K3's rank, behind Opus 5, Fable 5, and the Sol / Grok 4.6 pair at 61

4

of the top ten are open — K3, GLM-5.3, Qwen 3.8 and GLM-5.3 Flash. A fifth, Muse Spark 1.2, still has no weights, nor does 1.3 (Sept 2).

MOVEMENT INTO SEPT 1

Kimi K3 **60** now ties GLM-5.3. New open entries: **GLM-5.3** 60, **GLM-5.3 Flash** 57, and Meta's **Muse Spark 1.2** 57 — API live; open weights promised, not yet shipped. V4-Pro-0813 **53**.

Artificial Analysis Intelligence Index v4.1.1 · Sept 1, 2026 data · vendor pricing pages. Figures here are v4.1.1; v4.2 announced Sept 4, 2026.

SECTION 01 · 08

The best open model trails Fable 5 by two points at 30% of the price.

Snapshot: v4.1.1, Sept 1, 2026. Fable 5 and Sol are scored as deployed systems with fallbacks and guards, so the gap compares a system to a bare model.

MODEL	INDEX SCORE (52–63)		LIST PRICE · INPUT / OUTPUT
Opus 5		63.0	$5 / $25
Fable 5		62.1	$10 / $50
Sol		61.0	$5 / $30 · Fast $10/$60
Grok 4.6		61.0	$2 / $6 <200K · $4/$12 above
Kimi K3 · OPEN		**60.0**	**$3 / $15**
GLM-5.3 · OPEN		**60.0**	**$1.15 / $3.50**
Qwen 3.8 Max · OPEN		**58.0**	**$2 / $6**
GLM-5.3 Flash · OPEN		**57.0**	**$0.075 / $0.25**
Terra		57.0	$2 / $12 · cut 20% Jul 30
Sonnet 5		53.4	$3 / $15 · $2/$10 promo

K3 AGAINST FABLE 5

−2.1 index points

30% of the input price

K3 AGAINST OPUS 5 **−3.0 points**

· 60% of the input price

THE CHEAPEST ROUTE

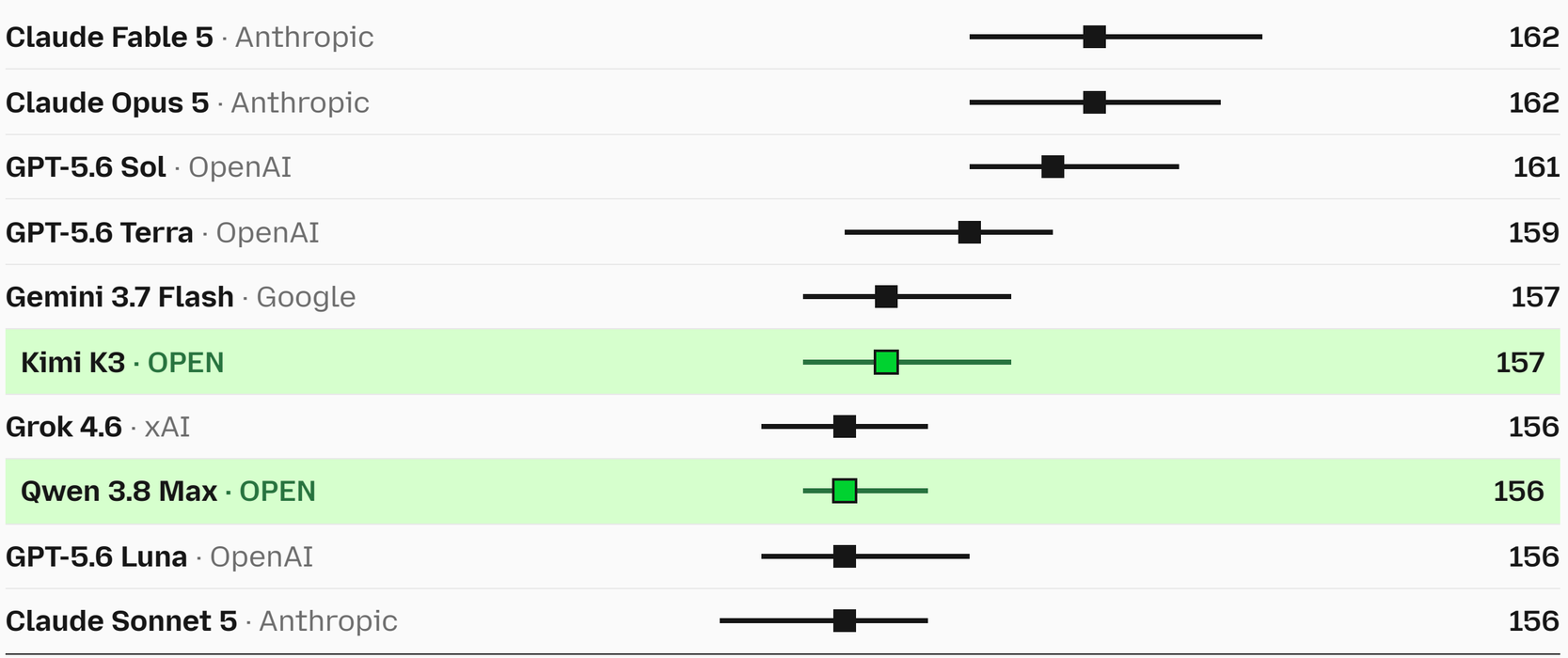
K3 lists at $3.00/M input; the cheapest OpenRouter routes are reported at $2.60–$2.80/M. Sonnet 5 sits at the same $3/$15 list, nearly seven points lower on the index.

Artificial Analysis Intelligence Index v4.1.1, Sept 1, 2026 · list pricing from each vendor's pricing page (Anthropic, OpenAI, xAI, Moonshot, Z.ai, Alibaba Cloud), verified Aug 2026. Artificial Analysis figures here are v4.1.1 · Artificial Analysis, "Announcing Artificial Analysis Intelligence Index v4.2," Sept 4, 2026.

SECTION 01 · 09

The open-closed gap is five ECI points and it resets every release cycle.

Epoch Capabilities Index by release date — the best open model (Kimi K3, 157) against the closed frontier (Claude Fable 5 and Opus 5, 162).



Squares mark point estimates; whiskers show 90% confidence intervals.

5 pts

Fable 5 and Opus 5 at 162, Kimi K3 at 157 — below Epoch's measured average of 8 points since January (90% CI 7–11).

WHY IT DOESN'T COMPOUND

Every model here was released between June and August 2026. The margin between leaders resets each release cycle rather than accumulating, and Epoch translates the average 8-point gap to about four months. K3 and Gemini 3.7 Flash now sit level at 157, intervals fully overlapping.

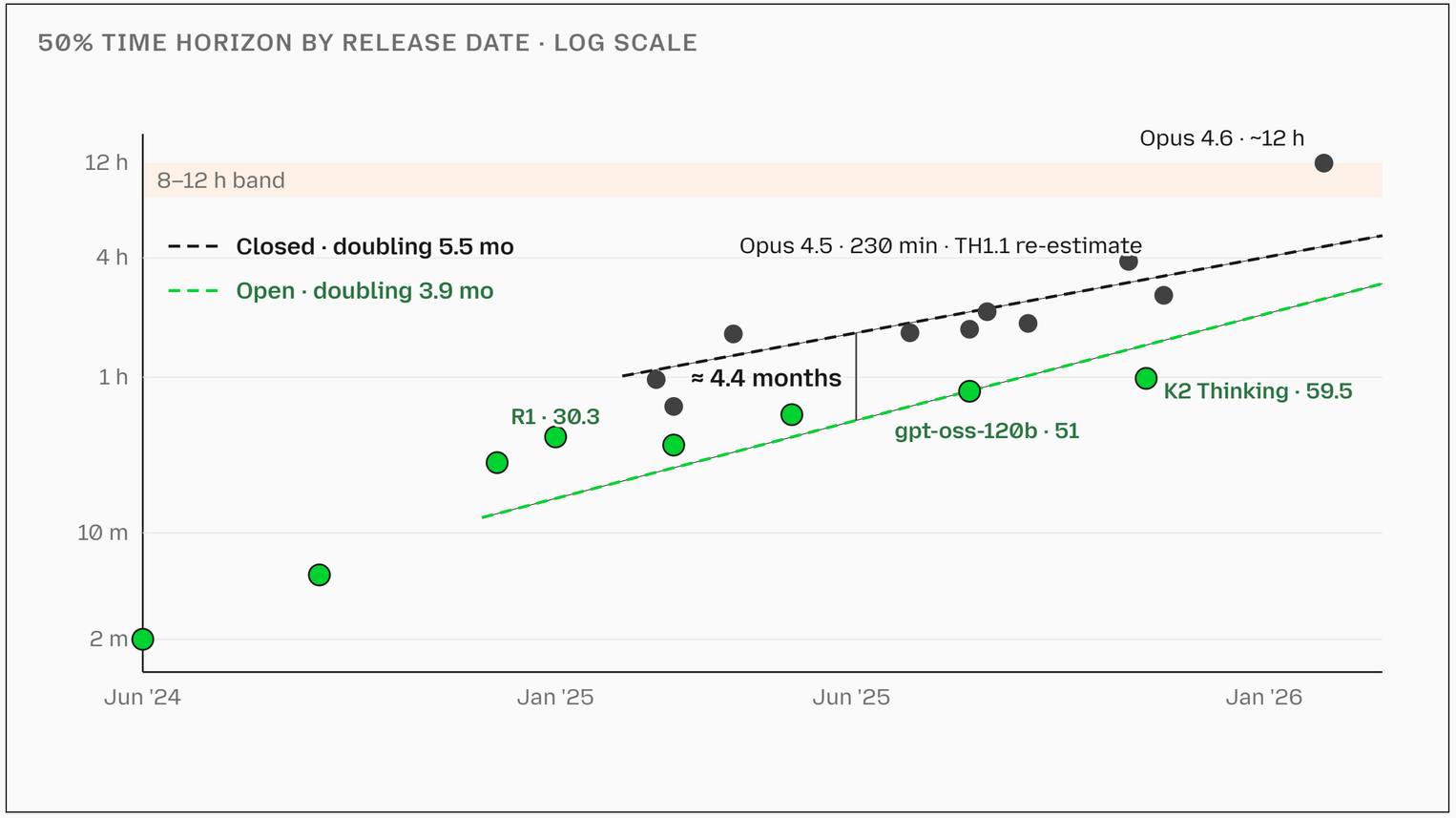
Epoch notes two reasons this gap may be understated: open models optimize on benchmarks more aggressively, and unpublished closed models are absent from the baseline. The August releases are now scored: Qwen 3.8 Max enters at 156, V4-Pro-0813 at 155.

Epoch AI, Epoch Capabilities Index (CC-BY) · Sept 1, 2026 data · Epoch AI, "Open models lag state-of-the-art closed models by 4 months" (May 2026): average gap 8 ECI pts, 90% CI 7–11.

SECTION 01 · 10

Closed models handle 8-to-12-hour tasks. Open gets there four months later.

The METR doubling law converts the score gap into a band of tasks — and a calendar. Two independent estimates now agree: 4 months from scores (Epoch), ≈4.4 months from measurement (Mozilla's computation on METR's raw data).



4.4
months measured lag

1.74×
task-length ratio

THE PAY RULE

Under 8 hours, either model does the job. Past 12, neither does. The ~5× per-task premium pays only inside that band — and the band moves, because a task at the closed frontier today is at the open frontier four months later. Pay when four months on the marginal task is worth 5×: deadline work yes, routine work no.

DERIVATION

METR TH1.1 puts the doubling at 6.3 months all-time, 4.3 months from 2023 on, and ~3 months from 2024 on. Against METR's published doublings (6.3 and 4.3 months), a 4-month lag converts to a **1.5×–1.9×** task-length band. Against Mozilla's own fitted doublings (closed 5.5 mo, open 3.9 mo) the same lag gives **1.7×–2.0×**. Mozilla's measured fit lands at **1.74×**, inside both.

METHOD NOTE

METR's latest published point is an early Claude Mythos Preview snapshot (Mar 2026, published May 8): a 50% horizon of ≥16 h (95% CI 8.5–55 h), which METR flags as at the edge of its task suite, so it is omitted here. It also flags measurements above 16 hours as unreliable, and puts error bars at ~2×. Opus 4.6 drops ~40% to 7 h 11 m on private-only tasks; Epoch notes open models tend to underperform private benchmarks. Mozilla's fit is a simplified reproduction, labeled as such.

METR, "Time Horizon 1.1" (Jan 29, 2026) + eval-analysis-public data · METR Mythos update (May 8, 2026) · Epoch AI ECI gap (May 2026, CC-BY) · vals.ai Terminal-Bench 2.1 · Mozilla computation.

A jagged frontier: parity, contested, a closed edge.

For most workloads, open models already clear the bar. The closed edge shows up on GDPval-style knowledge work and 1M-token retrieval, and Moonshot concedes conversational polish. Match the model to the job and you need the frontier for less than you think.

OPEN LEADS OR PARITY

Frontend coding

K3 debuted #1 on LMArena's Frontend Code Arena (1679 Elo), first in six of seven frontend domains — plus coding, instruction-following, general knowledge.

Frontend Code Arena measures building website and UI code, scored by blind developer votes.

CONTESTED

Agentic terminal work

K3 88.3 vs Sol 88.8 on Terminal-Bench 2.1; wins Program Bench, SpreadsheetBench 2, BrowseComp; loses FrontierSWE 81.2 vs Fable 5's 86.6.

Terminal-Bench, Program Bench, and BrowseComp measure agent work: running terminal tasks, writing programs, researching the web. FrontierSWE measures resolving real software-engineering tickets.

CLOSED EDGE

Professional knowledge work

Fable 5 leads K3 by 92 Elo on GDPval-AA v2, the largest Elo separation among shared benchmarks; long-context fidelity; conversational polish — Moonshot concedes UX still trails.

GDPval-AA v2 measures expert-graded professional knowledge work, where long-context reliability and polish appear.

Scores use different scales and come mostly from vendor-run tests, so treat them as directional. LMArena uses blind human voting.

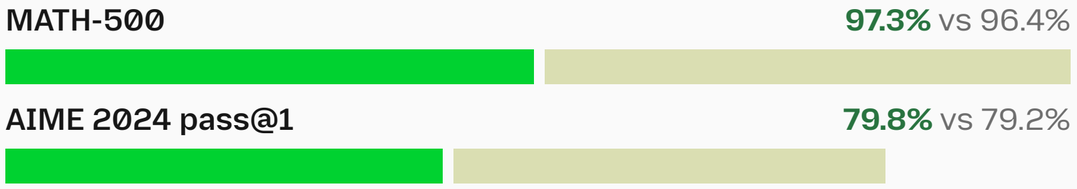
Moonshot AI, Kimi K3 model card (github.com/MoonshotAI/Kimi-K3; vendor-run, as labelled) · LMArena Frontend Code Arena · Artificial Analysis GDPval-AA v2. Terminal-Bench 4.0 launched Sept 1–2, 2026 (tbench.ai); 4.0 scores are not comparable to the 2.1 figures shown.

SECTION 01 · 12

Frontier reasoning ships open.

DeepSeek R-1 matched o1's reasoning scores in January 2025, and by August 2026 frontier-class open releases had become a monthly occurrence.

DeepSeek-R1 against OpenAI o1-1217 — the same class of reasoning.



DEEPSEEK-R1 · MIT	O1 · PROPRIETARY
$0.55 · $2.19	$15 · $60

Roughly a 27× price gap at near-identical reasoning scores.

AUG 2026 · THE ROUTINE

WEIGHTS SHIPPED AUG 13 | REPRICED AUG 16

DeepSeek V4-Pro reached general availability as V4-Pro-0813.

Parameters	1.6T / 49B active
Context	1M tokens
Speculative decoding	DSpark module
Reasoning effort	three levels
Weights	66 fp8 shards · ~893 GB · ungated

The wrinkle: three days after GA, the whole V4 family moved to peak/off-peak pricing — Pro output from a flat $0.87/M to $1.98 off-peak and $3.96 peak.

The weights are the hedge against the meter.

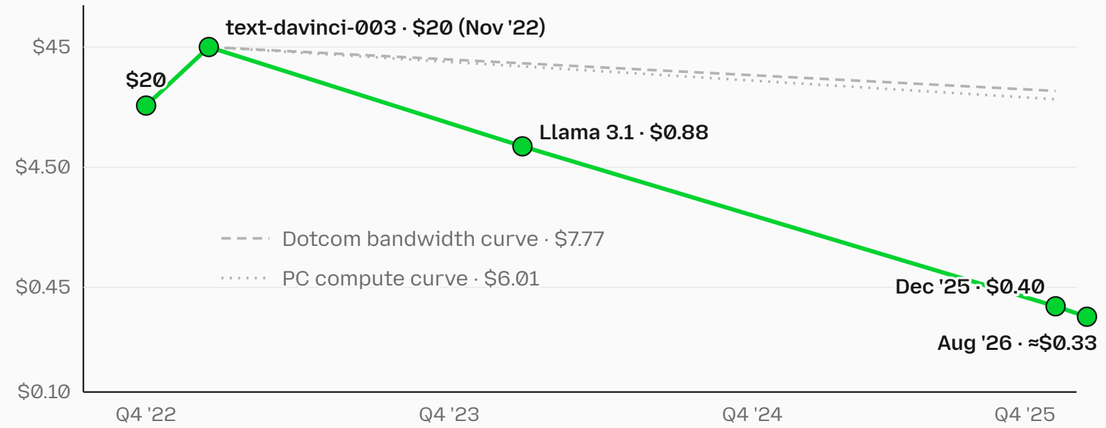
What remains of the gap sits at the harness layer.

DeepSeek-R1 technical report · OpenAI o1 documentation · Hugging Face deepseek-ai/DeepSeek-V4-Pro-0813 · DeepSeek API Docs, "DeepSeek-V4-Pro GA Release" (Aug 13, 2026) + pricing page (peak/off-peak effective 16:00 UTC Aug 16).

Inference fell ~60× in 45 months, and the first open price increase just landed.

Cheapest model at GPT-4-class performance: blended API list price, log scale, against prior platform-shift cost curves at their historical rates.

PRICE PER MILLION TOKENS · GPT-4-CLASS CAPABILITY



~60×

best-available price Nov '22 ($20, text-davinci-003) → GPT-4-class Aug '26 ($0.33, V4 Flash off-peak, blended 3:1)

3.4×

what PC compute's curve delivers over the same window

2.6×

what dotcom bandwidth's curve delivers

EXCEPTIONS

Kimi K3

List $3/M in · $15/M out, above the tier median ($1.75 / $9.00). Cost per task ~$0.86 per completed AA index task against GPT-5.6 Sol's $1.23 — about 30% cheaper, not the 60% the per-token list price implies, because slow output and high token use eat most of the gap.

DeepSeek · repriced Aug 16

The first list-price rise from a major open-weights lab, introducing peak/off-peak billing. V4-Pro moved from a flat $0.435 / $0.87 to $0.66 / $1.98 off-peak and $1.32 / $3.96 peak — output up 2.3× to 4.6×.

Countercurrent: OpenAI cut GPT-5.6 Luna 80% (to $0.20 / $1.20) and Terra 20% (to $2 / $12) on Jul 30. Sol held at $5 / $30.

a16z "LLMflation" methodology · Epoch AI LLM Inference Price Trends · Artificial Analysis · Moonshot · DeepSeek API pricing page · OpenAI pricing page and changelog (Jul 30 cuts).

SECTION 01 · 14

Eight of the top ten models by token volume are open weights.

Token volume on OpenRouter, Aug 1–31 window. Eight of the top ten are open weight; seven of those are Chinese-built.

#	MODEL · PROVIDER	MONTHLY TOKENS	CHANGE
1	**DeepSeek V4 Flash 0731** · DeepSeek	45.1T	↑ >999%
2	**Hy3** · Tencent	34.1T	↑ 347%
3	**MiMo-V2.5** · Xiaomi	29.4T	↓ 12%
4	**Ox Alpha** · Z.ai (stealth)	27.2T	new
5	**GPT-5.6 Luna** · OpenAI	23.6T	↑ >999%
6	**DeepSeek V4 Flash 0423** · DeepSeek	23T	↓ 10%
7	**Nemotron 3 Ultra** · NVIDIA	15.3T	↑ 56%
8	**GLM 5.2** · z-ai	14.7T	↑ 5%
9	**DeepSeek V4 Pro 0423** · DeepSeek	9.55T	↓ 25%
10	**Claude Opus 5** · Anthropic	7.11T	↑ 653%

Open weight Closed

THE AUGUST MONTH

DeepSeek holds three of the top ten — V4 Flash 0731 at #1 on 45.1T, the 0423 checkpoint at #6, V4 Pro 0423 at #9. Only three US entries make the list: GPT-5.6 Luna, Nemotron 3 Ultra and Claude Opus 5. Z.ai's **Ox Alpha** enters new at 27.2T straight from stealth.

THE NEW WEIGHTS

K3 weights went live Jul 27; a dozen-plus providers now serve it on OpenRouter, cheapest route **$2.60 / $13** (Sail Research, fp4) against the official $3 / $15. Qwen 3.8 Max weights followed Aug 9.

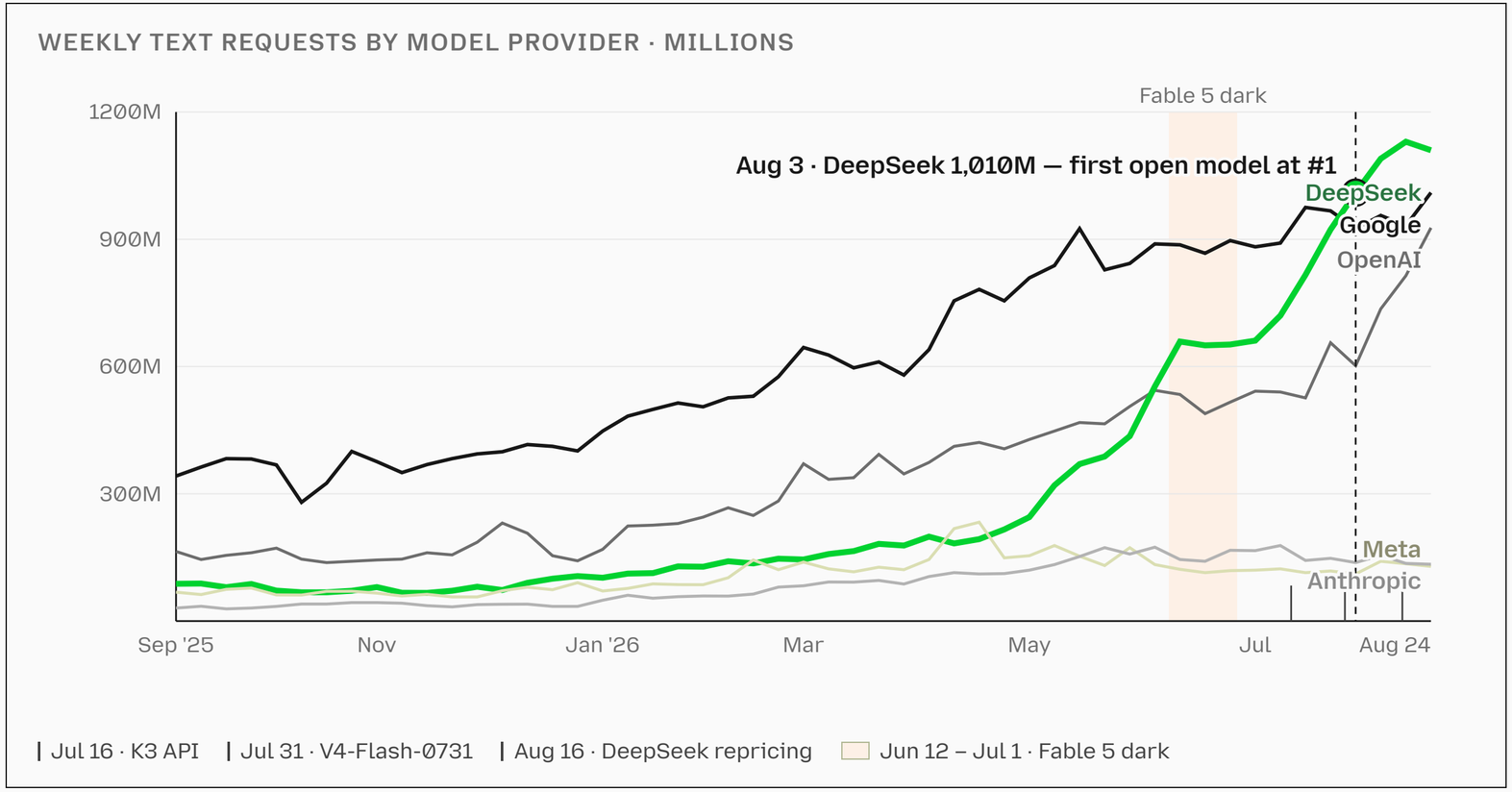
Routed traffic only — first-party usage in ChatGPT, Gemini and Doubao sits outside this measurement. Ox Alpha's license is unconfirmed at print.

OpenRouter Rankings and Market Share panels, Aug 1–31 2026 monthly window, pulled Sept 1, 2026 (Datasets API, CC BY 4.0). OpenRouter agreed to be acquired by Stripe on Aug 19, 2026.

SECTION 01 · 15

August was the first month a Chinese open model led OpenRouter on requests.

Google was the most-requested model author for 51 consecutive weeks. In August, DeepSeek became the first open model to have the most weekly requests.



12.6×

DeepSeek weekly requests, Sep 2025 → Aug 2026. Google: 3.0×.

51 weeks

Google's unbroken run at #1 before Aug 3.

8 weeks

DeepSeek from #3 to #1 (Jun 8 → Aug 3).

44%

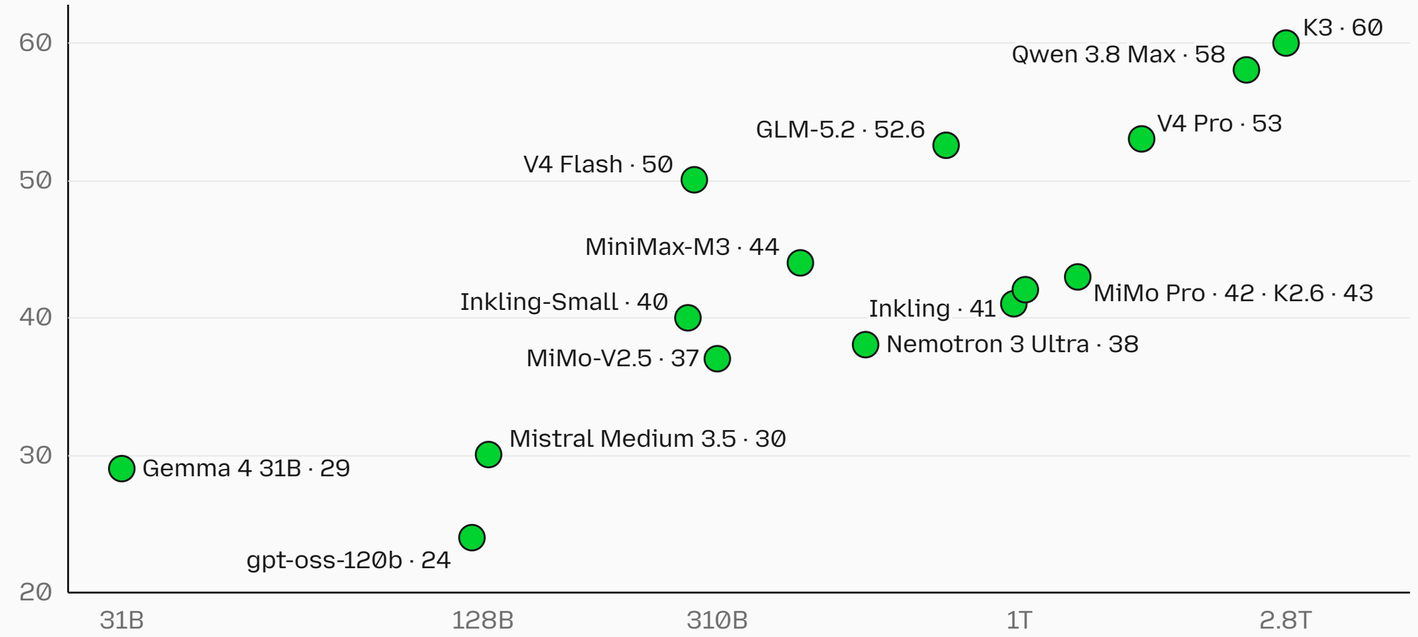
US closed providers' combined request share in late August, down from 54% in May.

Watch — week of Aug 24, the first full week after DeepSeek's Aug 16 repricing: DeepSeek −20M, Z.ai +~195M on GLM-5.3.

Text requests routed on OpenRouter, weekly buckets, by model author. Routed traffic only; excludes first-party usage in ChatGPT, Gemini and Doubao. "Others" (10–13%) and the unattributed "stealth" author (146–198M in late August) are excluded from author counts. OpenRouter Rankings → Market Share, pulled Sept 1, 2026.

Capable AI now runs at every hardware tier, from a gaming PC to a small data center.

Index score vs total parameters, log scale. On the lower tiers, the gains came from hardware that stayed fixed. Snapshot: v4.1.1, late Aug 2026.



Every point on this chart ships downloadable weights. Total parameters on the x-axis; active parameters set compute.

Capability exists at every tier on the chart, from a gaming PC scoring 29 to a multi-rack cluster scoring 60.

SERVING FLOOR BY TIER

Gaming PC / 128 GB Mac	Gemma 4 31B · 29 · gpt-oss-120b · 24 · MiMo-V2.5 · 37
Single B300 GPU	Inkling-Small · 40 — 276B / 12B, 180 GB NVFP4
One 8-GPU server	GLM-5.2 · 52.6 · V4 Flash · 50 · MiniMax-M3 · 44 · K2.6 · 43 · MiMo Pro · 42 · Nemotron 3 Ultra · 38
Multi-node	Qwen 3.8 Max · 58 · V4 Pro · 53 (one 4-GPU GB300 node)
Small data center	Kimi K3 · 60 — 2.8T / 104B, 64+ accelerators

The top tier bought its gains with more hardware. The lower tiers gained points on hardware that stayed fixed.

Ten points off the top, on one server. Twenty-three, on one GPU. Index score against total parameters, log scale. Artificial Analysis model pages (v4.1.1) · NVIDIA Nemotron 3 Ultra release + technical report · Xiaomi MiMo release notes · Kimi K3 model card (Hugging Face) · Inkling-Small model card (Hugging Face) + vLLM serving recipe.

A single GPU now runs a 40-point model.

The serving floor for a 40-point open model dropped to one accelerator. The drop came from the training recipe: the low-precision checkpoint ships calibrated from the vendor rather than quantized after the fact.

HARDWARE FLOOR · INKLING-SMALL

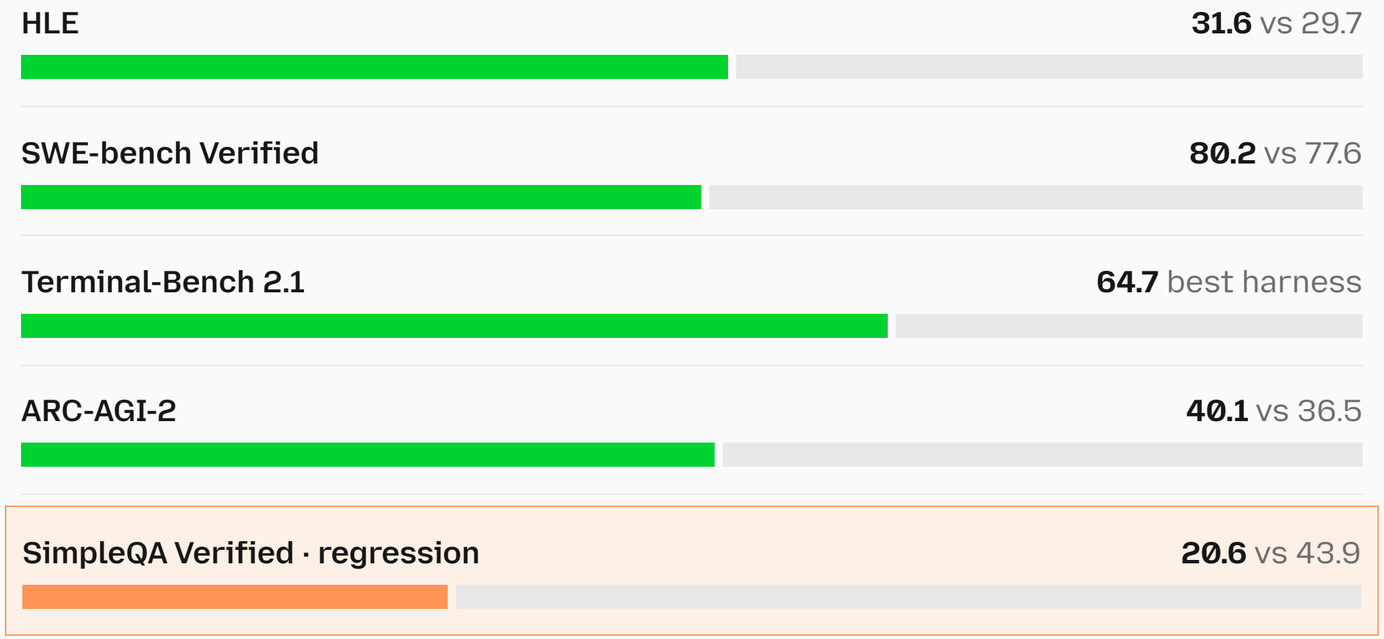
BF16 checkpoint	≥ 600 GB
NVFP4 checkpoint one B300 (W4A4) or two H200s (W4A16)	**180 GB**

276B total / 12B active MoE · Apache 2.0 · released Jul 30 · runtimes: SGLang, vLLM, TokenSpeed, Unsloth, Hugging Face; GGUF quants for llama.cpp.

CONSUMER CONTEXT

gpt-oss-120b scored **24** on a 128 GB machine a year ago. MiMo-V2.5 scores **37** on the same class of machine today. K3 used the same approach as Inkling-Small — trained quantization-aware in MXFP4 from the SFT stage onward.

INKLING-SMALL 276B/12B VS INKLING 975B/41B — 4 WINS · 1 REGRESSION



Agent benchmark scores are vendor-reported on provider-specific harnesses. Hugging Face repos carry Apache 2.0 while the model card attaches a separate, changeable Acceptable Use Policy. Whether the floor keeps dropping is a forecast — it rests on two releases of evidence.

Thinking Machines Lab, "Inkling: Our Open-Weights Model" (Jul 15, 2026) · Inkling-Small model card (Hugging Face, weights Jul 30, 2026) · Kimi K3 model card (MXFP4 QAT from SFT).

SECTION 01 · 18

A 13B active model beats a 49B one and undercuts it on price.

Total parameters set the memory a deployment needs. Active parameters determine the compute each token costs. Benchmarks scores have stopped tracking either one, and recall hasn't.

DeepSeek V4-Flash-0731 vs V4-Pro Preview	 13B / 284B	VENDOR-RUN Beats the 1.6T / 49B Pro Preview on **all nine** published agentic benchmarks, at ~⅓ the output price. The sweep held for two weeks, until the Pro's Aug 13 post-training pass lifted it from 45 to 53 on the AA index and put the larger model back on top. Both swings came from post-training rather than either model's size.	**9/9** WINS
Thinking Machines Inkling-Small vs Inkling	 12B / 276B	Beats the 975B / 41B Inkling on HLE, SWE-bench Verified, Terminal-Bench 2.1 and ARC-AGI-2, and loses SimpleQA Verified **20.6 vs 43.9**. SimpleQA measures stored facts rather than reasoning, and a 12B holds fewer of them. $1.20/M output against $4.05/M.	**4/5** 1 RECALL REGRESSION
Moonshot Kimi K3	 104B / 2.8T	Activates 16 of 896 experts per token. The efficiency stack is documented in K3's technical report as known techniques — fixed-size attention state, layer-to-layer residual access, sparser expert routing — so other labs can reproduce it.	**3.7%** ACTIVE

DeepSeek API Docs release notes (Jul 31 and Aug 13) · Inkling-Small model card · Kimi K3 technical report (github.com/MoonshotAI/Kimi-K3) · Artificial Analysis v4.1.1.

Capability gains no longer require a new model.

DeepSeek's July 31 post-training pass lifted V4-Flash 21 points on Terminal Bench and 10 on the AA index, on an unchanged architecture.

EXHIBIT A · V4-FLASH-0731

FIXED: 284B / 13B MOE · 1M CONTEXT

VENDOR-RUN · DEEPSEEK HARNESS, MINIMAL MODE, MAX EFFORT

Terminal-Bench 2.1	61.8 → **82.7**	DeepSWE	7.3 → **54.4**
AA Intelligence Index	~40 → **~50**	GDPval-AA v2 Elo	1189 → **1559**

Independent rows below the hatched band. Vendor-stated efficiency: hallucination rate −12 pts; eval-suite output tokens 234M → 206M.

EXHIBIT B · V4-PRO-0813

SAME PRO STRUCTURE + DSPARK DECODING

Terminal-Bench 2.1 · VENDOR	72.1 → **87.9**	DeepSWE · VENDOR	12.8 → **62.7**
AA Intelligence Index	45 → **53** · ties GLM-5.2	WEIGHTS PUBLIC SAME DAY · MIT	

Post-training costs a fraction of pretraining, so this is a cheaper path to capability gains than a new model. It also shortens the useful life of any benchmark table.

IF THE PATTERN HOLDS

01

Release cadence

Annual model generations become monthly checkpoints — and evaluation infrastructure has to re-run on the same cycle or publish stale numbers.

02

Locus of differentiation

From pretraining scale, where capital decides, to post-training data, reward design and RL environments, where it does not.

03

Who can play

From the handful that can fund pretraining to anyone with an open base and a post-training budget.

DeepSeek API Docs (Jul 31, Aug 13) · Hugging Face model cards · Artificial Analysis v4.1.1 · harness caveat per DeepSeek Harness v0.1 (MIT, Developer Preview, Aug 13).

Mozilla

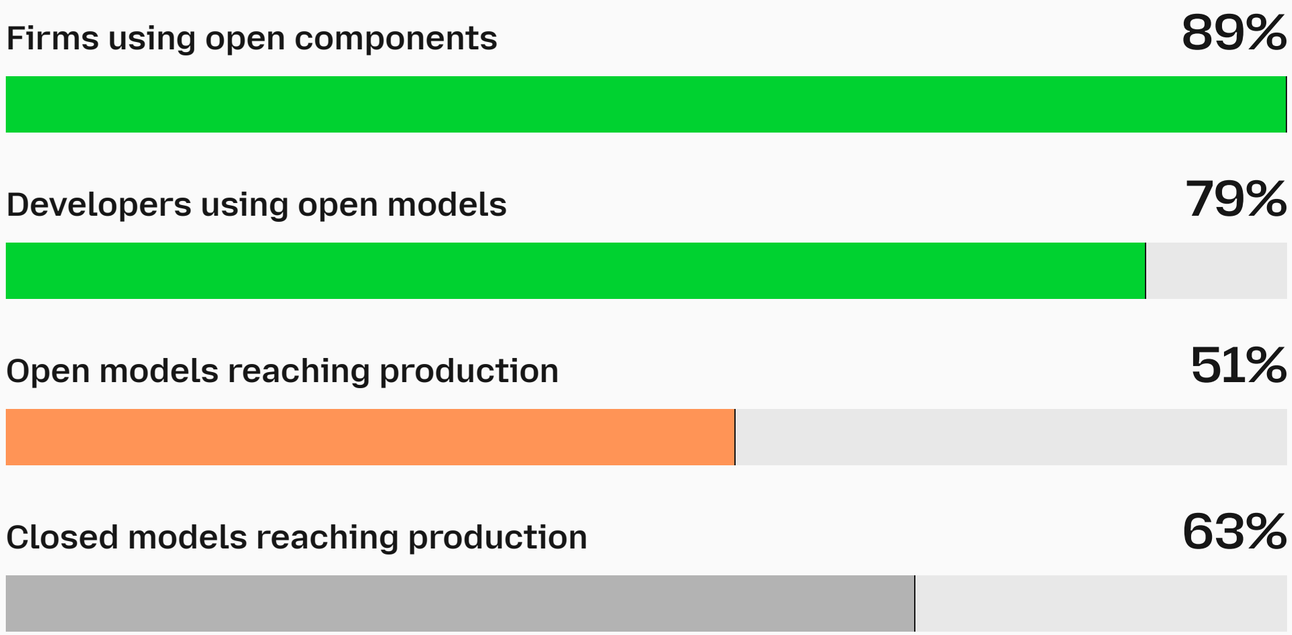
Open ships easy, deploys hard.

SECTION 02 · 21

Open ships easy. Open deploys hard.

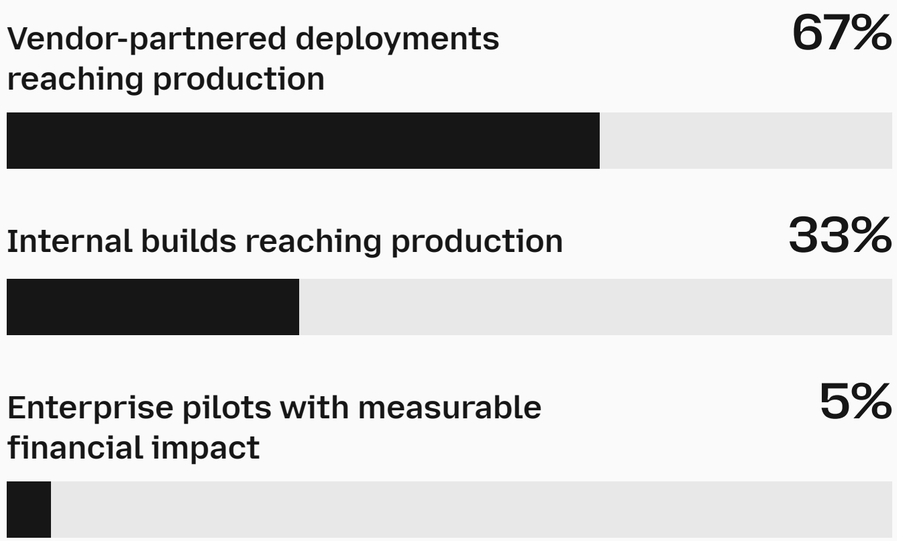
Open models reach production 12 points less often than closed ones. The survey traces the gap to operational tooling and trust.

THE FUNNEL



The gap comes from operational tooling and trust; capability already clears the bar.

WHERE TEAMS STALL



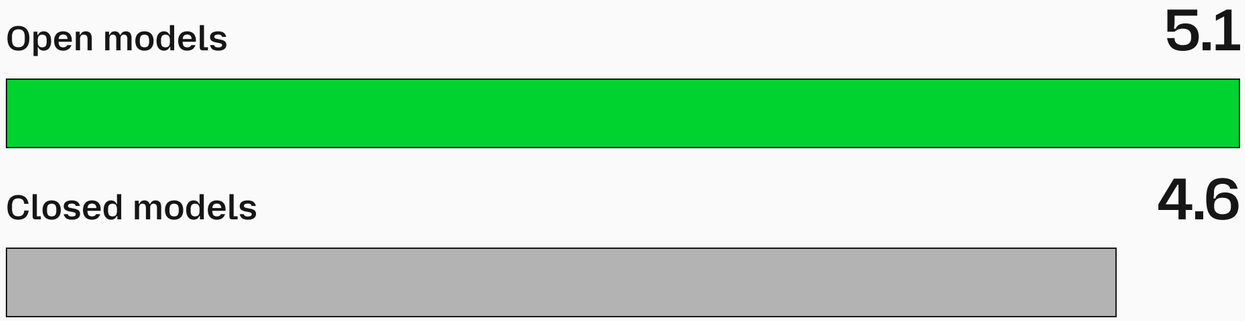
Partnership is doing the work that tooling has not: a vendor-partnered deployment is twice as likely to reach production as an internal build.

Adoption funnels from experimentation to production. Linux Foundation Research, "The Economic and Workforce Impacts of Open Source AI" (2025), commissioned by Meta, for 89% · Mozilla/SlashData 2026 survey, n=1,494 (79/51/63) · MIT NANDA, "The GenAI Divide: State of AI in Business 2025" (Aug 2025) for 67/33 and 95%.

Developers run open across more use cases than closed.

Half of developers run open and closed models. Open leads every use-case category surveyed.

AVERAGE USE CASES PER DEVELOPER

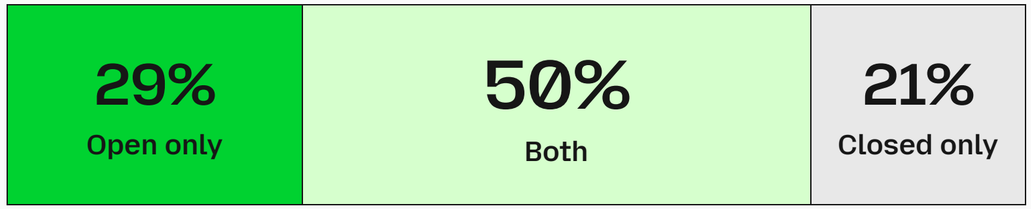


78% pair specialised, task-optimised open models alongside or in place of general-purpose ones.

79%
use open models

71%
use closed models

HOW TEAMS COMBINE THEM



Teams assemble portfolios.

Half of developers run both, and open leads every category surveyed. They aren't swapping one model for another. They are assembling portfolios, which makes the harness the layer the choices actually bind.

Average use cases per developer, and how open models get combined. Mozilla/SlashData 2026 developer survey.

SECTION 02 · 23

The open stack scores high on capability, low on operations.

Nine stack layers scored across nine criteria (1–5), columns ordered strongest to weakest.

STACK LAYER	Community	Ease	Prod-ready	Interop	Sustainability	Perf-vs-closed	Docs	Standard-ization	Enterprise	LAYER AVG
Model code	4.7	4.0	4.3	4.5	3.8	3.7	3.7	3.7	3.3	**3.96**
Model weights	4.6	4.0	4.0	4.0	3.4	3.6	3.2	3.4	2.8	**3.67**
Infrastructure	3.7	3.3	3.7	3.6	3.7	3.4	3.4	3.4	3.3	**3.48**
Product / UX	3.8	3.8	3.5	3.7	3.2	3.2	3.7	2.8	3.0	**3.41**
Datasets	4.0	4.0	3.8	3.8	2.8	3.0	3.0	2.8	2.8	**3.33**
Documentation	4.0	4.0	3.0	3.3	3.5	2.8	3.0	3.3	2.3	**3.22**
Licensing	3.5	3.3	3.0	3.0	3.3	3.5	3.3	2.5	2.8	**3.11**
Agent layer	4.2	3.3	3.3	2.0	3.3	3.5	3.2	2.2	2.3	**3.04**
Safeguards	3.4	3.0	2.8	2.6	3.0	2.6	2.6	1.6	2.2	**2.64**
CRITERION AVG	4.00	3.63	3.54	3.40	3.35	3.27	3.25	**2.83**	**2.79**	

≥ 4.0 · 3.5–3.9 · 3.0–3.4 · 2.5–2.9 · < 2.5

The operational gap = standardization + enterprise readiness.

INFRASTRUCTURE · STANDARDIZATION
3.1 → 3.4 ▲

KDA-class linear attention broke runtime compatibility, so loading weights no longer guaranteed serving. Moonshot resolved it upstream by contributing KDA prefix caching to vLLM — strengthening standardization while concentrating influence over the standard.

SUBCOMPONENT EXTREMES

Terms of Service	1.9
Hardware chips	2.3
Moderation	2.6
Model cards	2.7
Preference alignment data	4.0
ML frameworks	4.3

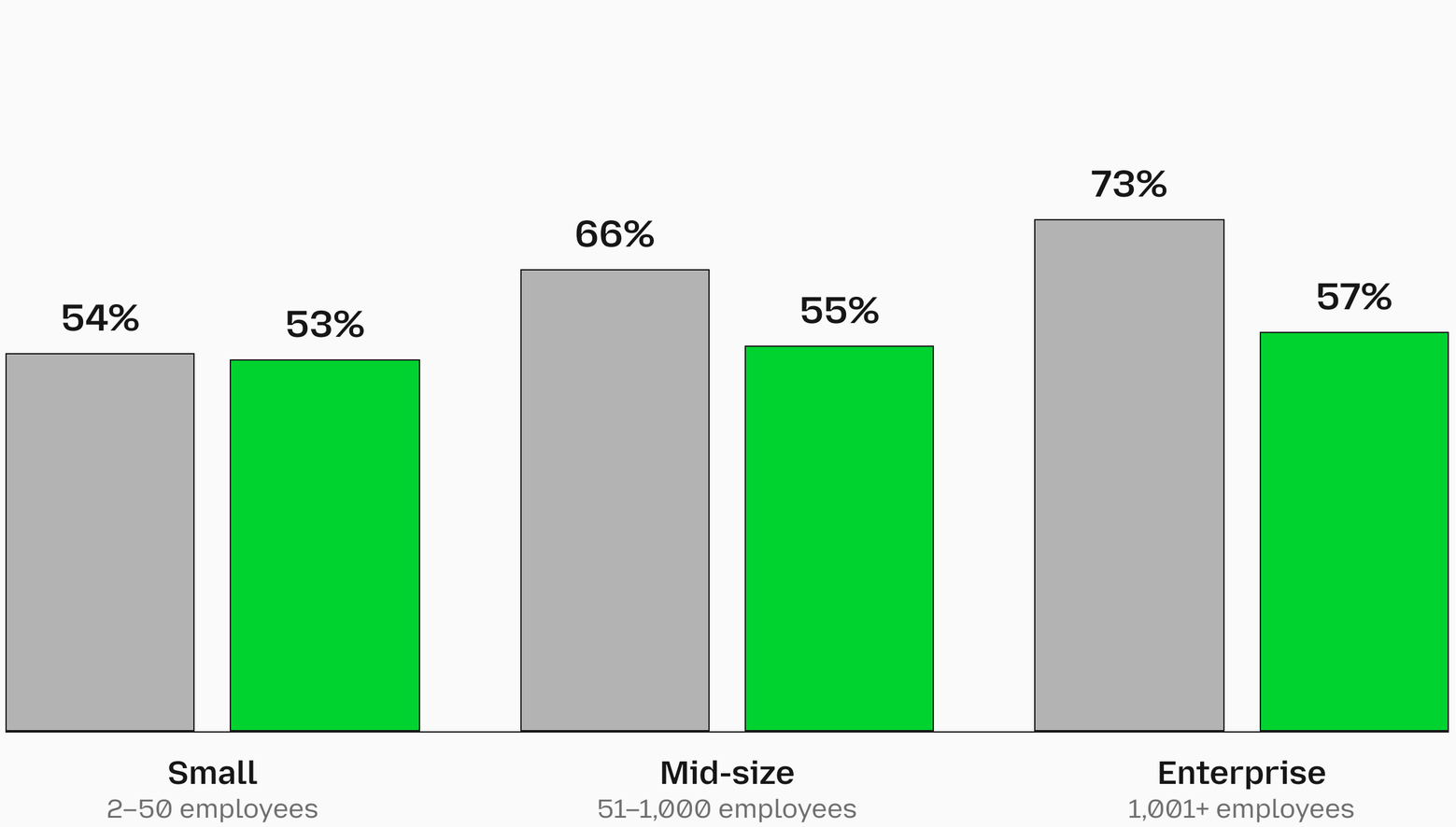
Watch: whether the next divergent architecture also lands upstream or forks the serving layer.

Mozilla open source AI stack map, July 2026 · Mozilla OSAI gap map sheet (2026.04.13). Full 44-row subcomponent table in the appendix.

SECTION 02 · 24

Enterprise scale lifts closed models into production and leaves open models flat.

Share of adopters reaching production, by organization size. If the gap were about resources, scale would close it.



THE SPREAD WIDENS WITH SIZE

Small (2–50)	**+1 pt**
Mid-size (51–1,000)	**+11 pts**
Enterprise (1,001+)	**+16 pts**

Scale closes the closed gap and leaves the open one where it was, which points at tooling rather than budget.

Mozilla/SlashData 2026 developer survey, professional developers n=954.

SECTION 02 · 25

Trust and safety tooling is being built the way cybersecurity was: as open, shared infrastructure.

The safety stack sits above the model: it decides what a platform's users can post, send or generate, and what happens when they cross the line. Until recently every platform built or bought its own. It is now being open-sourced as shared infrastructure.

WHY OPEN SOURCE · SMYTE, 21 JUNE 2018

Twitter acquires Smyte and shuts off its API within the hour.

Smyte sold spam, fraud and abuse detection to platforms including Discord, Zendesk, npm and GoFundMe. Customers on multi-year contracts got a phone call and about 30 minutes' notice.

CONSEQUENCE

Discord rebuilt its safety stack from scratch.

Snap, Reddit and others built their own versions in parallel, none of them shared.

Platforms that couldn't rebuild bought vendor tools, and with them the vendor's fixed taxonomy of harassment, grooming and self-harm.

SPOTLIGHT · ROOST · DETECTION → INVESTIGATION → REVIEW → ENFORCEMENT

DETECTION AND INVESTIGATION

Osprey

Runs a platform's live event stream against rules the safety team writes and deploys itself, automates the obvious, and lets analysts search accounts and IPs to investigate the rest. Self-hosted; data never leaves the platform.

Built at Discord after Smyte. V1.0 Jan 2026. Used by Bluesky, Discord and Matrix.

Over 445 million events a day in production.

→

REVIEW AND ENFORCEMENT

Coop

Takes flagged content in via API, auto-actions what the platform's rules allow, and routes the rest to human reviewers by policy area. Audit trail, appeals and enforcement webhooks built in; reviewer-wellness protections (greyscale, muted video, blur) on by default.

Cove's codebase, acquired 2025, released open source Feb 2026. Most platforms start here.

Hash matching against known CSAM, NCII and terrorist content, plus NCMEC reporting.

→

SIGNALS · THE ROOST MODEL COMMUNITY

Model Community

Open safety models a platform hosts itself and plugs into Osprey or Coop. The anchor is gpt-oss-safeguard (OpenAI, Oct 2025), a reasoning model that reads the platform's written policy alongside the content and returns a decision with its reasoning. Change the policy document and the model follows it on the next request.

Also: Roblox's PII, Sentinel and voice classifiers (Aug 2026); Zentropi's CoPE-B; Mila's suicide-assistance prevention guardrail; partner policy packs.

A platform brings its own policies and definitions of harm.

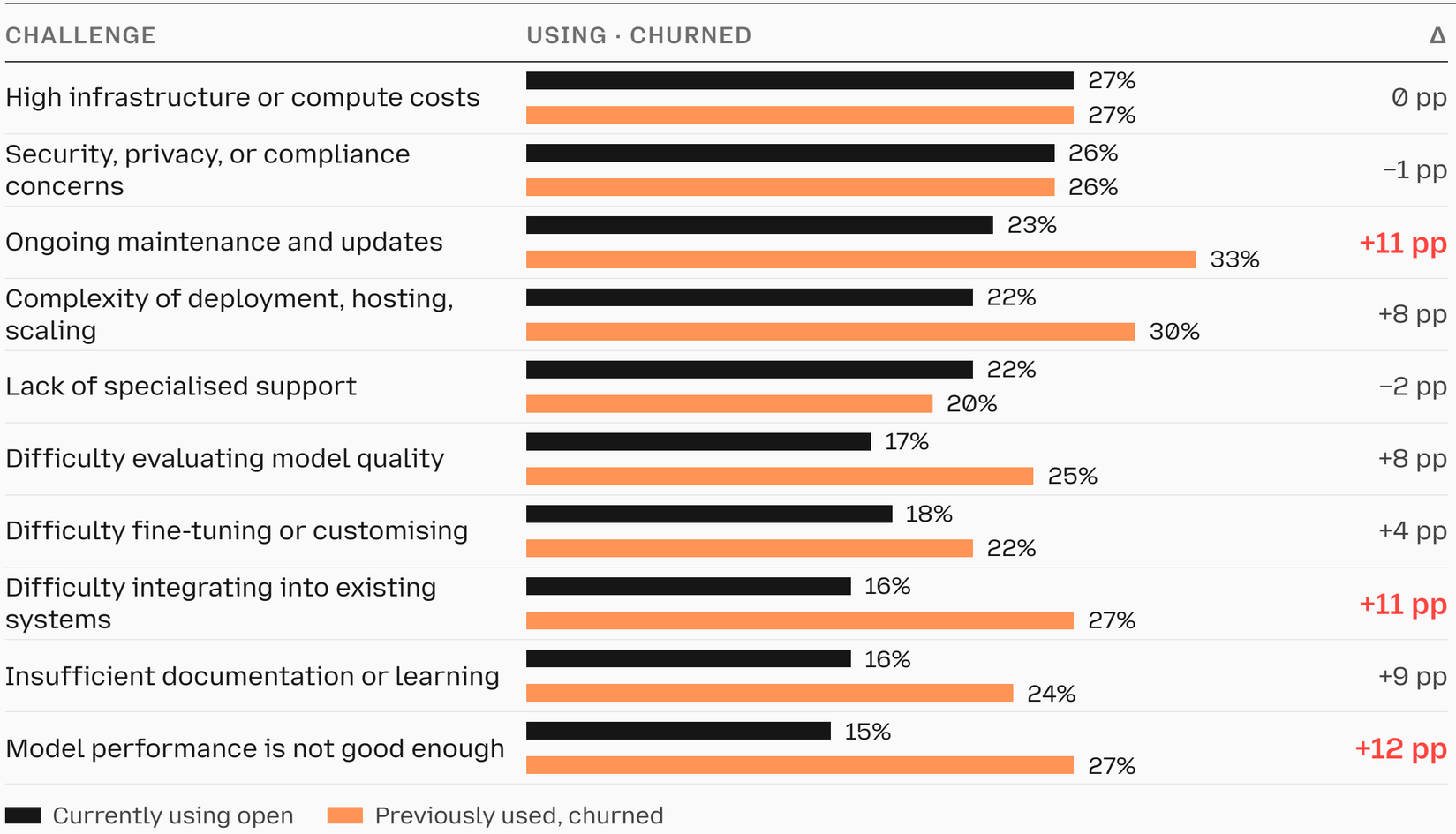
A community forum and a companion-chatbot developer can run the same open weights and the same review console against different rulebooks, in their own language, and revise them in an afternoon. The stack is built for user-generated and AI-generated content alike, so a chatbot's output and a user's post enter the same detection rules and the same review queue.

TechCrunch and The Register (Jun 2018) · Platformer (Feb 2025) · ROOST launch letter (Feb 2025) · ROOST roadmap (Dec 2025) · ROOST, Osprey V1.0 and Coop V.0 posts (Jan–Feb 2026) · InfoQ (Mar 2026) · OpenAI (Oct 2025) · ROOST Model Community repo (2026) · Roblox newsroom (Aug 2026). Mozilla is a listed ROOST partner.

SECTION 02 · 26

What blocks open, and what makes developers leave.

Developers who left cite maintenance, integration trouble, and evaluation quality. Cost and compliance concerns are the same whether developers stayed or left.



THE CHURN SIGNATURE

Cost and compliance are flat between the two groups. Everything operational — maintenance, integration, evaluation, documentation — is what the leavers name.

LARGEST DELTAS

Model performance not good enough	+12
Ongoing maintenance and updates	+11
Integration into existing systems	+11

Performance leads the churn list even though capability clears the bar on the benchmarks; what teams judge is the experience of running open models.

Developers using open models against those who churned from open. Δ = churned minus current, percentage points. Bars scaled to a 39% maximum. Mozilla/SlashData 2026 developer survey (n=1,410 current or churned open-model developers). Bars rebuilt from the authoritative survey export.

SECTION 02 · 27

Every region names the same blockers at nearly the same rates.

Share of developers naming each challenge, by region. Values in per cent; darker means more developers named it.

CHALLENGE	W.Eur & Israel	N.America	Greater China	South Asia	E.Asia ex GC	S.America	E.Eur & CIS	Oceania	All
High infra or compute costs	25	26	29	28	28	28	29	18	**27**
Security, privacy, compliance	20	27	18	39	29	28	25	22	**26**
Ongoing maintenance and updates	27	26	18	26	20	31	21	25	**24**
Complexity of deployment / scaling	27	24	19	24	11	30	26	25	**23**
Lack of specialised support	17	16	21	31	24	23	23	32	**22**
Difficulty evaluating / comparing	14	17	14	23	16	26	25	18	**18**
Difficulty fine-tuning / customising	22	18	18	20	11	22	18	12	**18**
Difficulty integrating into systems	19	21	14	20	7	26	19	20	**18**
Insufficient documentation	18	15	15	17	15	20	24	15	**17**
Model performance not good enough	18	15	13	22	16	17	19	8	**17**
No major challenges	9	21	16	5	14	4	8	12	**12**
WEIGHTED SAMPLE	286	277	206	192	164	147	98	39	1,411

READ THE SAMPLE SIZES

Oceania (n=39) and Eastern Europe & CIS (n=98) fall below reliable thresholds — outlined here rather than dropped.

WHAT THE MAP SAYS

Infrastructure cost is the one challenge named at the same rate everywhere — 25 to 29 per cent in almost every region. South Asia reports compliance pressure at **39%**, the highest cell on the chart. Only North America has a meaningful share reporting no major challenges at all (21%).

The blockers are structural: the same tooling gap shows up in every market.

Mozilla/SlashData 2026 developer survey (MZCS1), n=1,411.

Closed still leads in expert knowledge work, long context, and accountability.

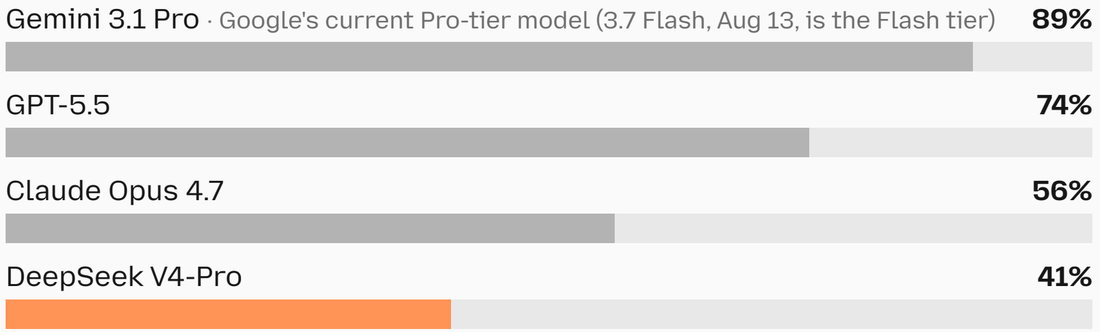
The compliance and liability gaps close with contracts, while the two capability gaps rest on measurements taken before the August checkpoints.

EXPERT KNOWLEDGE WORK

92 Elo Fable 5's lead over K3 on GDPval-AA v2, the largest separation on any benchmark the two models share.

GDPval-AA v2 measures expert-graded professional knowledge work. Moonshot concedes K3's UX and polish still trail.

LONG-CONTEXT FIDELITY



Multi-needle retrieval at 1M tokens. Measured on pre-August checkpoints; V4-Pro-0813 has not been re-tested.

TURNKEY COMPLIANCE & SAFETY

SOC 2 | HIPAA | ZDR

Closed providers package compliance by default. On the FLI Summer 2026 AI Safety Index, DeepSeek scored an **F (0.47/4)** — though the index failed one lab from each bloc (xAI, DeepSeek, Mistral), grades labs rather than deployments, and scored no company above C+.

ACCOUNTABILITY

A counterparty to hold liable.

Pay a closed vendor and someone else carries the liability when things break. Self-host open weights and it is your own.

Future of Life Institute, "AI Safety Index — Summer 2026" (Jul 7, 2026) · GDPval-AA v2 via Artificial Analysis · long-context multi-needle evaluations (pre-August checkpoints) · Anthropic/OpenAI compliance documentation.

Mozilla

Why buyers and governments want the weights anyway.

OPTIONALITY COST CONTROL POLICY SOVEREIGNTY

SECTION 03 · 30

For nineteen days, the newest frontier model went dark.

Optionality stopped being abstract. Everyone building on that endpoint watched both decisions from the outside.

Jun 9 — Anthropic ships Fable 5 and Mythos 5

Jun 12 — **Commerce applies export controls, effective immediately, barring access by any foreign national, inside or outside the US, including Anthropic's own staff. Nationality cannot be verified in real time, so both models go dark for everyone**

Jun 26 — Partial clearance: Mythos restored to ~100 vetted US critical-infrastructure organizations

Jun 30 — Controls lifted

Jul 1 — **Fable 5 restored globally, nineteen days after it was cut**

Jul 16 — Moonshot opens the K3 API — a frontier-class model on a release path no export order can reach once the weights land

REVERSIBILITY

API access · two-way switch
Cut Jun 12. Restored Jul 1. The lever works in both directions.

Weight release · one-way valve
Public Jul 27. No recall. Once the files are distributed, there is nothing to switch.

The day after K3's weights went live, Anthropic's CEO wrote that the company "has never advocated for a ban on open-weights models" and called non-dangerous open models "a public good" — while pushing chip export controls, anti-distillation enforcement and mandatory safety testing.

Anthropic, "Statement on the US government directive to suspend access to Fable 5 and Mythos 5" (Jun 12, 2026) · Commerce/BIS letter as published by Bloomberg (Jun 16) · Lutnick letter of Jun 30 lifting controls (CNBC, Jun 30) · Anthropic, "Our position on open-weights models" (Jul 27) · Moonshot AI. Sovereignty is this same argument at national scale.

Open buys optionality.

Cloud computing already ran this experiment: once proprietary APIs and data gravity made leaving expensive, enterprises spent the next decade repatriating workloads.

Closed model API

LOCKED · PUNITIVE EXIT

You're renting on someone else's machine.

- Vendor controls pricing
- No clean migration
- You inherit their price changes

THE DECISION

Where do you build?

Closed model APIs reproduce the same trap.

Open weights

PORTABLE · FORECASTABLE

No one can repossess it.

- Self-hosted, on infrastructure you control
- Fixed cost you can forecast
- No forced lock-in

$90K–$120K
to move 1 PB out of AWS S3 — the egress exit penalty

~80%
of enterprises now repatriating some workloads (IDC)

2.5×
GEICO's cloud costs against expectations, before repatriating

$10M+
37signals' projected five-year savings after leaving

The AI-era version: AT&T's router. And DeepSeek's Aug 16 price rise is the trap arriving inside the open world's own APIs — the downloaded weights remain the exit. Pricing is one thing the API route leaves in the vendor's hands; where the request goes is another. Anthropic alleges DeepSeek's API string-matched requests from Claude Code, the Agent SDK and OpenCode and forwarded a subset to Claude Opus without telling the user (Sept 10). Once the weights sit on your own machine, neither the price nor the destination can change without you.

IDC 2024 survey · 37signals blog (DHH) · GEICO SREcon 2024 talk · The Information (Aug 20). Pre-2026 cloud-repatriation figures not re-verified for this edition.

Pay-per-use AI is blowing up enterprise budgets.

AT&T, Microsoft, and Uber hit the limits of usage-based pricing at scale. AT&T routed work onto open models; Thomson Reuters built its own on an open base.

AT&T

Routing each employee query to the cheapest adequate model cut AI task costs by up to 56%.

Employee query → **LiteLLM router** → **Cheapest adequate model**

56% saved | **−2%** quality | **45B** tokens / day

40% → **60–70%** open share

Internal finding: open models run 6–10 months behind the frontier and are "just as good or better" than older closed models.

"The concept of AI sovereignty has become truly paramount to us." — Andy Markus, Chief Data & AI Officer

Microsoft

Microsoft cancelled most Claude Code licences across Experiences + Devices by June 30 and moved engineers to GitHub Copilot CLI; The Verge's sources tied the timing to the fiscal-year boundary.

- **Dec** — Access opens to thousands of engineers, PMs and designers
- **—** — Token billing consumes the division AI tool budget
- **May 14** — The Verge reports the cancellation
- **Jun 30** — Claude Code licenses cancelled across Experiences + Devices

Relationship intact: $5B investment + $30B Azure commitment stand.

Engineers moved to GitHub Copilot CLI on flat seat pricing.

Thomson Reuters

Spent $40M over two years to post-train an open-weight Qwen base on its own legal content, and moved its first CoCounsel workload onto the result.

Frontier-lab pretraining run	**hundreds of $M**
Thomson final training run	**$450K**

<10%

of its content library used so far

Base: Snowdon, a Qwen 3.5 derivative realigned with Imperial College London. Anthropic partnership retained for other workloads.

"Owning more of the AI stack gives us greater control over performance, economics, deployment, and sovereignty." — Thomson Reuters, "Why we built Thomson," Sept 1, 2026

"This token maxxing has freaked out the CFOs."

— Ali Ghodsi, CEO, Databricks, on clients who once ruled out Chinese open models becoming open to them (CNBC, Aug 13)

Uber burned through its entire 2026 AI budget in four months after rolling out agentic coding tools.

EY: a complex agentic interaction in 2026 costs ≈30× a simple 2023 chatbot query.

The trigger in each collision: usage-based pricing that left unit economics under the vendor's control. AT&T: The Information (Aug 20, Mark Austin) · WSJ CIO Journal (Andy Markus). Microsoft: The Verge, Tom Warren, Notepad (May 14). Uber: The Information (Apr 2026). Thomson Reuters release (Aug 24) and blog (Sept 1) · Business Insider (Hron) · CNBC (Aug 13).

Washington considered a ban on Chinese open-weight models and, so far, has not pursued one.

Between late July and early August the question moved from active discussion to something officials were publicly downplaying. Part of the reason is practical: a ban needs a point where it can be enforced, and downloaded weights don't offer one.

Jul 21	Jul 22	Jul 22	Jul 27	Late Jul	Aug 4–5
Treasury Secretary Bessent: the government will examine Chinese open-source models for IP theft and may sanction	OSTP Director Kratsios accuses Moonshot of distilling Claude and of acquiring restricted GB300 servers	TechCrunch: experts dispute that K3 could have been built by distilling Fable in 15 days	Anthropic: a blanket ban is "neither the correct remedy nor something we have called for"	Politico: Commerce will not ban Chinese models anytime soon. Axios: lobbying recurs every three to five months	White House tells top US AI companies that open-weight models, US and Chinese, will not face pre-release government testing under the new framework

WHERE AN ORDER COULD REACH

AN ORDER → a Chinese vendor's hosted API · Commerce could plausibly restrict access

AN ORDER → open weights already downloaded · no single point to restrict

Copies that are already in circulation would sit outside any order.

Ban: not pursued — AS OF EARLY AUGUST

OPTIONS STILL BEING DISCUSSED

- Entity List designation — reported as under consideration for some Chinese tech firms
- Federal procurement limits
- Security advisories
- Liability requirements
- Public pressure

Each of these acts on companies, contracts or API access rather than on the weights themselves. As far as we can tell, none would reach copies already in circulation.

Bessent, Fox Business (Jul 21) and @SecScottBessent (Jul 22) · @mkratsios47 (Jul 22) · TechCrunch (Jul 22) · Anthropic (Jul 27) · Politico · Axios via Lawfare · Bloomberg (Aug 5) · WSJ (Aug 4–5, Ramkumar) · Reuters. The testing framework has not been published; this relies on reporting of a closed-door briefing.

In September, the government's first concrete step was an advisory aimed at US vendors' APIs.

Three documents came out over three days. The advisory and Anthropic's own report give different answers on which Claude model Moonshot is alleged to have distilled, and neither has released the underlying evidence.

Sept 8

NSA, CISA and FBI issue joint advisory AA26-251A naming DeepSeek, Moonshot, Alibaba, MiniMax, StepFun and Z.AI. It asserts Moonshot used Claude Fable 5 data to train K3 and GPT-4o data for K2, and tells US providers to quietly degrade responses to suspected distiller accounts rather than ban them

Sept 9

China's Commerce Ministry calls the advisory groundless, describes distillation as common industry practice, and warns of countermeasures

Sept 10

Anthropic's own report names six labs and puts numbers on it: Alibaba 151M exchanges May–Jul, Moonshot 23M. Where it names Moonshot's target it says Opus. It describes company-level attribution, identity verification and reasoning summarisation as its tools

The report drew roughly 30M impressions in its first day. None of the labs named in it had responded at the time of writing.

WHERE THE ADVISORY REACHES

AN ORDER	→	a Chinese vendor's hosted API · Commerce could plausibly restrict access
AN ORDER	→	open weights already downloaded · no single point to restrict
AN ADVISORY	→	a US vendor's API · the vendor decides whether to act; the government has recommended it

The one place where distillation can be interrupted is controlled by the vendor. The advisory asks vendors to use it, but doesn't require them to.

OPTIONS STILL BEING DISCUSSED

- Security advisories — issued Sept 8; Beijing rejected the findings on Sept 9
- Entity List designation, procurement limits, liability requirements — none used so far
- Trump is due to host Xi in Washington later this month, and a US–China AI safety dialogue is planned for mid-September. Both were scheduled after the advisory was released

CISA AA26-251A and NSA release (Sept 8) · MOFCOM (Sept 9) · Anthropic, "Detecting and countering misuse of AI: September 2026" (Sept 10) · Reuters, NBC (Sept 9) · Dealroom. Claims resting on either document should be read as allegations; neither party has published forensics.

The US re-entered the open-weights race in August.

A two-bloc open commons is safer than a single-origin one.

Jul 15

Thinking Machines ships Inkling

975B / 41B, Apache 2.0, weights day one, a US frontier-adjacent open line. Inkling-Small weights followed July 30.

Aug 5

White House exemption

All open-weight models, US and Chinese, exempt from pre-release government testing under the new AI safety framework.

Aug 10

Meta ships Muse Glimmer

29.78B dense, Apache 2.0, runs on a 24 GB consumer GPU (<20 GB at 4-bit), reversing the proprietary turn it made when it retired Llama for Muse in April. Zuckerberg argues US policy must reduce friction "if we want American open source models to lead over time." The launch pairs with a $1B fund for data-center communities and a new independent-board release-governance structure.

MUSE SPARK 1.2 · STATUS — NOT SHIPPED · VERIFIED AUG 31

AA 54 at xhigh, launched closed Aug 5; weights announced "in the coming weeks," and superseded by 1.3 (API) on Sept 2 with weights still not shipped. Meta's verified HF org carries four repos, all Muse Glimmer, and a Muse-Spark search across HF returns zero results; no license, parameter count or date confirmed. Precedent cuts both ways: Qwen 3.8's weights were in exactly this promised-not-published state earlier in August, and shipped Aug 9.

> "Meta releasing Muse Spark 1.2 as open weights is a very big deal. America now finally has its response to the open weights AI race."
>
> — Aaron Levie

The blocs already interbreed: Inkling from Kimi 2.5, Muse Glimmer from Meta's closed Muse Spark, and, Anthropic alleges, Qwen 3.5–3.7 and Kimi from Claude traces. Only the first two are labelled.

Thinking Machines Lab launch post (Jul 15) · Bloomberg (Aug 5) · Meta AI blog + huggingface.co/blog/muse-glimmer (Aug 10) · Levie post (X) · TechCrunch (Jul 22).

China out-downloads everyone.

Qwen out-downloaded the next eight organizations combined.

CUMULATIVE HF DOWNLOADS · MAR 2026

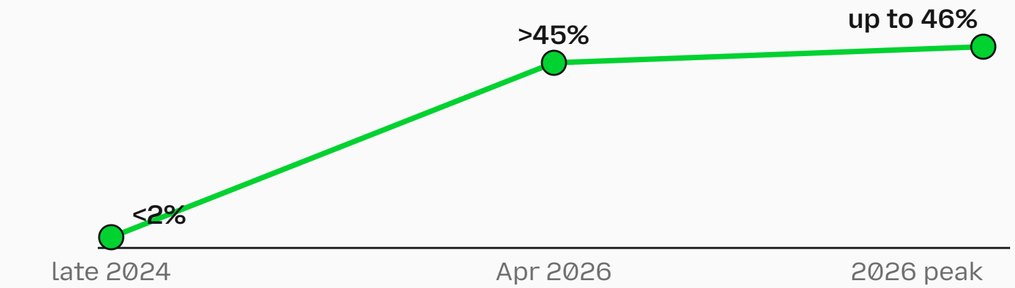
Qwen · Alibaba — **942M**

Llama · Meta — **476M**

As of March 2026, Qwen out-downloaded the next eight organizations combined.

Llama is no longer Meta's active line, retired for Muse in April 2026. The download race is now Qwen against the field, and Qwen 3.8 Max's weights shipped Aug 9.

CHINESE OPEN-WEIGHT SHARE OF OPENROUTER TOKENS



Against ~4.5% in H1 2025. Traffic among the platform's ten most-used models: **61%** in 2026.

26,000+
DeepSeek enterprise accounts

58%
of new 2025 AI-startup stacks include DeepSeek

~17.6%
single-provider share — the platform's largest (Jul 2026)

THE CAVEATS, AND THE RESOLUTION

Training-data opacity is real and now specific: Anthropic alleges Qwen 3.5–3.7 were post-trained on Claude Opus reasoning traces harvested through 151M exchanges (Sept 10). Routing opacity is newer: it alleges the hosted Kimi and DeepSeek products silently forwarded some user prompts to Claude and kept the transcripts. The fix is architectural — eight-plus jurisdictions ban the hosted apps, and enterprises take the weights anyway, on hardware they control, because once the file is local nobody upstream can reprice or redirect it.

Cumulative Hugging Face downloads, and Chinese open-weight share of routed tokens. Vintages labelled. Hugging Face org pages · OpenRouter · CNBC analysis of OpenRouter data (Jul 7) · Anthropic, "Detecting and countering misuse of AI: September 2026" (Sept 10), GTG-16005, GTG-16002, GTG-16001. Qwen 3.8, the current line, is not named in the report.

Open proliferation is now Chinese foreign policy.

The domestic directive became an international institution, and the institution is now recruiting.

SUPPLY

Codified directive. The AI-Plus directive and the 15th Five-Year Plan target a core AI industry above ¥1.2T.

Macro hedge. Export controls engineered around: K3 claims ~2.5× K2's scaling efficiency.

Institutional arm. WAICO founded Jul 16: 29 signatories, 37 by end-July, HQ Shanghai. New Delhi, Sept 12: Xi invites BRICS in.

Russia · Brazil · South Africa · Indonesia · Malaysia · Pakistan · Kazakhstan · Kenya · Serbia · Belarus · Cuba · Venezuela · Zambia · Laos · +6

29 founding states, 14 named. Virtually all BRICS except India, which hosted the summit where Xi made the pitch and stays outside WAICO. Iran acceded later.

→

MECHANISM

1 · Release open weights

↓

July exhibit: K3, "the world's first open 3T-class model," weights published Jul 27 — announced at WAIC before Xi's speech.

↓

2 · Offload inference onto users' local hardware: **no serving cost, no export surface, no chokepoint to sanction.**

→

DEMAND

The Global South diversifies away from US tech monopolies.

Well-capitalized firms adopt for cost-per-task.

26,000+
DeepSeek enterprise accounts

up to 46%
of routed OpenRouter tokens are Chinese models

DISTRIBUTION RAILS

5,000 AI-training slots pledged to developing countries over five years

Cooperation centres with ASEAN, the Arab League and the African Union. BRICS upgraded to a head-of-state pledge, Sept 13.

WAICO is a China-aligned bloc — 37 members, zero major Western democracies. An open commons supplied from one origin is a dependency.

State Council "AI Plus" opinion (Aug 2025) · 15th Five-Year Plan outline (Mar 2026) · english.www.gov.cn on the WAICO signing (Jul 17, 2026); accessions per Xinhua · Xi keynote transcript (MFA) · The Diplomat (Jul 22) · Moonshot · OpenRouter · Xi statement, 18th BRICS Summit (MFA, Sept 13, 2026) · Bloomberg, CNBC (Sept 13). The pledge covers LLM co-development, training courses and an open AI ecosystem, first of five initiatives; no members, budget or timeline attached. Analysts frame WAICO as rivalling Pax Silica.

Europe is treating open weights as industrial policy.

In 18 months the EU went from a €109B infrastructure pledge to funding a sovereign model and switching on GPAI enforcement.

Feb 2025

€109B

France commits to AI infrastructure

Aug 2025

Exemptions

Qualified open source systems receive targeted exemptions under the AI Act

Jun 2026

400B

EUROPA award: a sovereign open model across all 24 EU languages

Jul 2026

€5.5M

Portugal ships Amália, a fully open LLM; Germany open sources its public-administration platform SPARK

Aug 2, 2026 LIVE LAW

Live law

Commission GPAI enforcement powers enter into application

> **"We cannot afford to depend on others for the technologies that keep our hospitals running, our energy grids stable and our services secure."**
>
> — Ursula von der Leyen
> President, European Commission

An escalating sequence, February 2025 – August 2026. Élysée / AI Action Summit (Feb 2025, €109B) · Regulation (EU) 2024/1689 application calendar (GPAI obligations Aug 2, 2025; Commission enforcement powers Aug 2, 2026) · EuroHPC / EUROPA consortium award · Portuguese government + huggingface.co/amalia-llm (Amália, Jul 1, 2026, €5.5M, on EuroLLM-9B) · German public-administration platform release.

The UK backed a sovereign frontier model.

Government compute, a private developer, and thirteen regulated-sector partners, with air-gapped deployment as the requirement.

COMPUTE	DEVELOPER	CONSORTIUM
Isambard-AI, Bristol	**Cosine**	**Thirteen partners**
Awarded from the £500M Sovereign AI programme; the model trains entirely on it.	A startup building Lumen Sovereign, billed as Britain's first sovereign frontier AI model.	Three banks, four defence contractors, LSEG and the Alan Turing Institute among them — signatories to design-phase memoranda, not yet purchase commitments.

CONSORTIUM ROSTER

BT · HSBC · Lloyds · NatWest · LSEG · BAE Systems · Babcock · Leonardo · Thales UK · PwC · Telefónica Tech UK&I · Era4 · Alan Turing Institute

Announced
London Tech Week, Jun 8

Data
Proprietary datasets spanning 30+ regulated workflows

Target
Air-gapped deployment inside customer infrastructure, end of 2026

Air-gapped operation requires the customer to hold the weight files.

WHO CAN SUPPLY IT

- ✓ Open weights
- ✓ Licence not yet announced — delivered on-premise, air-gapped
- × API-only vendors cannot serve deployments with this requirement

Sovereign procurement of this type is a growing market segment in which open-weight models and on-prem proprietary models are the only eligible suppliers.

Companion datum: Cohere is roughly tripling its UK office footprint after its ~$20B Aleph Alpha combination, framed as a bet on UK sovereign-AI demand.

Cosine, "Building Lumen Sovereign: Cosine Forms Coalition with UK Industry Leaders" (Jun 8, 2026) · DSIT Sovereign AI programme (£500M; 500,000 GPU-hours on Isambard-AI) · Cohere's own UK expansion announcement. Cosine's roster also includes Haleon; QinetiQ was added later.

SECTION 03 · 40

Canada paired $890M of public compute with a national open weights lab.

Sovereign infrastructure backed by open-weight models.

$890M

for a sovereign public AI supercomputer, open to Canadian researchers and innovators.

JUNE 2026 · AI FOR ALL

PM Carney launches AI for All. One of six pillars is an explicit **Sovereign Foundation** — safeguarding Canadian data and IP, reducing reliance on foreign supply chains.

> **"AI can go the way the internet and mobile phones did, a mostly disempowering tech. Or it can empower the people that use it. We are working towards that second one."**
>
> — Nick Frosst, co-founder, Cohere, on open-sourcing Command A+

COMMAND A+ · MAY 2026

A 218B-parameter MoE for enterprise agentic tasks, running on as little as two H100s. North Mini Code followed in June, built for the "sovereign developer ecosystem."

Apache 2.0, Cohere's first. Prior Command models shipped CC-BY-NC 4.0.

Prime Minister of Canada, "Prime Minister Carney launches AI for All" (Jun 4, 2026, pm.gc.ca) · ISED, AI Sovereign Compute Infrastructure Program (~$890M to the Infrastructure Build Layer) · Cohere press release (Business Wire, May 20, 2026) · huggingface.co/CohereLabs/command-a-plus-05-2026 · Mozilla scorecard.

India is subsidizing 38,000 GPUs for the fastest growing developer base in the world.

Cheap public compute plus the fastest-growing developer base in the world.

38,231

GPUs empanelled at subsidised rates near ₹65/hour, roughly 40% below market

5M+

GitHub developers added in 2025 alone, reaching 21.9M — the fastest-growing base in the world

THE BET

Cheap public compute plus a multilingual developer base, producing indigenous foundation models and harnesses for Global South use cases.

₹10,372 Cr

IndiaAI Mission outlay, with 600 planned data labs

13.6%

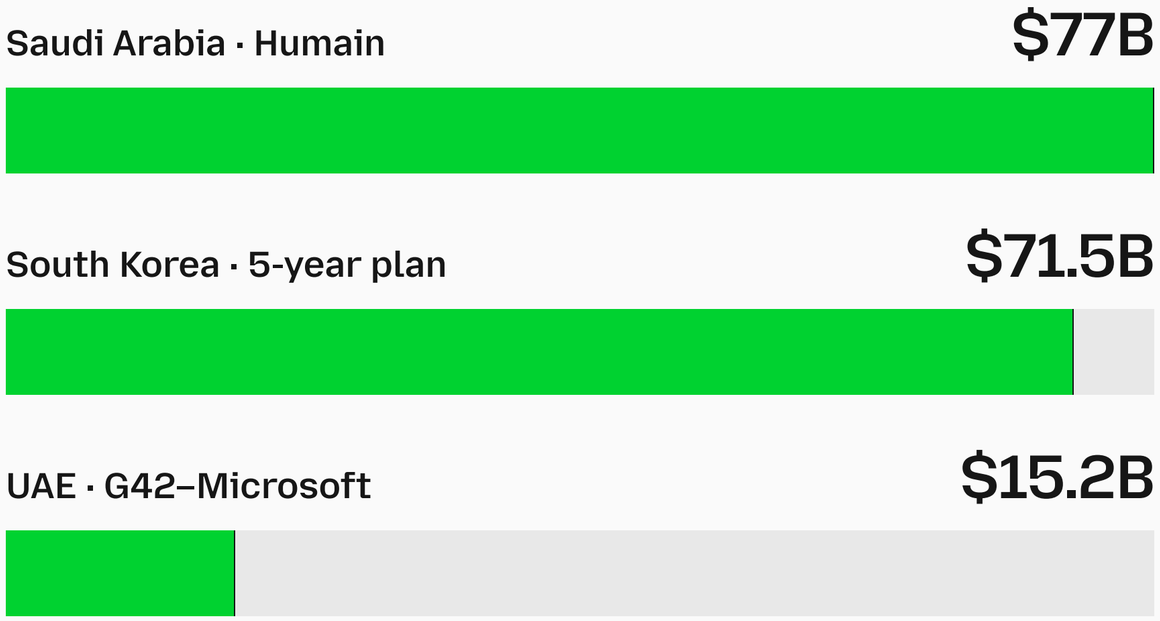
of DeepSeek's monthly active users — second only to China

India is not a WAICO signatory and hosted its own AI Impact Summit in February 2026, positioning itself as a third pole.

IndiaAI Mission (Cabinet approval Mar 2024, ₹10,371.92 crore) · IndiaAI GPU empanelment (38,231 GPUs; ~₹65/hour) · GitHub Octoverse 2025 · DeepSeek disclosures.

70+ national AI strategies, and counting.

Saudi Arabia has committed $77bn on its own.



80+
jurisdictions with AI policies (OECD.AI)

47
nations restricting foreign processing for critical workloads

12
new national strategies logged in 2024 (Oxford Insights)

1.9GW
planned Saudi capacity by 2030

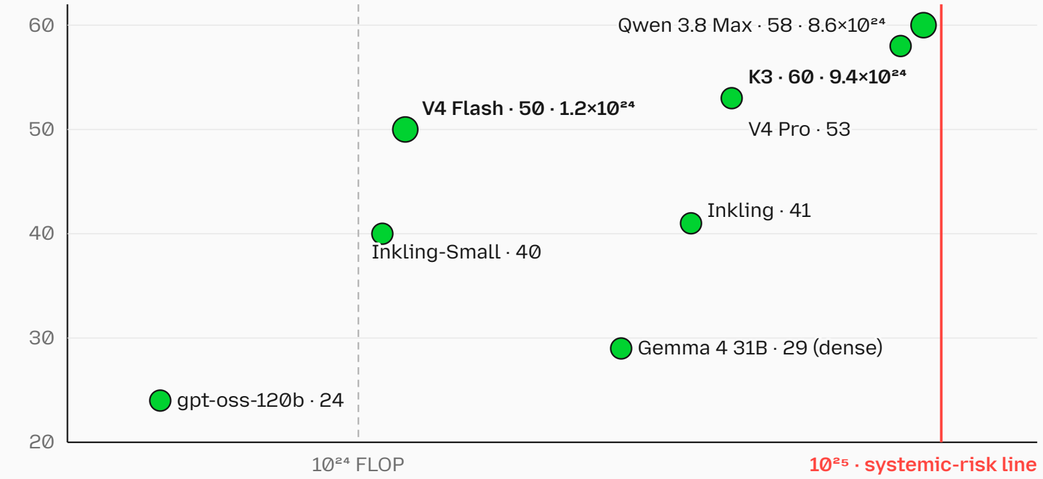
The strategic question has shifted from whether to have a national AI play to which layer of the stack a country can own.

Largest disclosed national AI infrastructure commitments. HUMAIN / Tareq Amin interview (FT, 2025) for $77B and 1.9GW by 2030 · South Korea ₩100tn AI pledge (Presidential Office / MSIT) · Microsoft UAE investment announcement (Nov 2025) for $15.2B · OECD.AI policy observatory · Oxford Insights Government AI Readiness Index 2024.

Sparse models buy capability at a discount the AI Act reads as lower risk.

The Act keys systemic risk to training compute, on the premise that compute tracks capability. Sparse mixture-of-experts training breaks that link. Enforcement went live August 2.

ESTIMATED TRAINING COMPUTE AGAINST INDEX SCORE · MOZILLA ESTIMATES AT 6 × ACTIVE PARAMS × 15T TOKENS



Comparable capability, opposite sides of the line. V4 Flash scores 50 at 1.2×10²⁴; K3 scores 60 at 9.4×10²⁴, crossing 10²⁵ near 16T tokens. K3 activates under four per cent of what it holds; V4 Flash under five.

Commission GPAI enforcement powers — **LIVE 2 AUG 2026**

OBLIGATIONS BY THRESHOLD

Art. 53(2) exemption

Removes technical documentation and the downstream information package, where four conditions hold at once.

Always survives

Copyright policy and training-data summary. Without the summary a provider breaches Art. 53(1)(b) however free the weights are.

Above 10²⁵

No exemption; Art. 53(1)(a)(b) applies in full, and systemic-risk designation adds adversarial testing, incident reporting and cybersecurity duties.

Two comparable, downloadable models in European deployments can carry entirely different obligation sets. Neither customer nor regulator can tell which, without a training-compute figure no lab publishes.

Regulation (EU) 2024/1689, Art. 53(1)(a)(b), 53(2), Art. 25 and Art. 51 (systemic-risk threshold 10²⁵ FLOP) · Commission GPAI guidelines (Jul 2025) for the ⅓-of-compute modification criterion · Mozilla compute estimates (method in appendix) · Index scores: Artificial Analysis v4.1.1, Sept 1, 2026.

Under the EU AI Act, post-training past a third of the base model's compute makes you the provider.

Obligations follow modification. Managed post-training platforms price compute and convenience.

Base model · someone else is the provider

↙ ↘

COMPUTE ROUTE

TRIGGER

Modification compute above ⅓ of the original training run.

Commission guidance, with an indicative 10^{23} criterion and a compute-estimation method. Far beyond a LoRA adapter — and reachable by a serious reinforcement-learning pass on managed infrastructure.

CONSEQUENCE

The full systemic-risk obligation set transfers, including notification to the Commission.

PURPOSE ROUTE · INDEPENDENT OF COMPUTE

TRIGGER

Substantial modification, or adaptation for a specific high-risk purpose.

Hiring, credit scoring, medical diagnostics, law enforcement.

CONSEQUENCE

The adapter becomes the provider of a high-risk system — technical documentation plus conformity assessment. The underlying license does not reduce the obligation set.

A team that post-trains a systemic-risk model inherits the systemic-risk obligations directly.

THE WORKED EXAMPLE

A bank tuning open weights for credit decisions carries documentation and conformity duties it never held as an API customer.

Neither route appears in the pricing of any managed post-training platform. Both are reachable from a product decision made by an engineering team.

HIGH-RISK COMPLIANCE BEGINS

Annex III	**December 2027**
Annex I	**August 2028**

Regulation (EU) 2024/1689, Art. 25 and the GPAI chapter · Commission GPAI guidelines (Jul 2025).

Mozilla

Capital has already moved above the model.

Open weights are a business model.

Funded companies, paying customers, real revenue.

Databricks

CLOSED AUG 13

$190B

Closed a $5B strategic round Aug 13 at a $190B valuation, led by Coatue with Blackstone, MGX, T. Rowe and Sixth Street Growth. Six months earlier: $134B.

Revenue run-rate **$7B**

Crossed a $7B run-rate, growing more than 80% year over year in Q2.

Mistral

IN TALKS

~€20B

Raising ~€3B at ~€20B, not yet closed. Samsung in advanced talks for up to €1B; EQT Scaleup Europe, Novo Holdings and Santander also in discussions.

Prior mark · Sep 2025 **€11.7B**

A 1.7× step-up inside eleven months, if it closes at the reported figure.

DeepSeek

MID-2025 FIGURE

$220M

Annual recurring revenue, mostly API and enterprise — the last disclosed figure, dated mid-2025.

Raised **$7.4B**

At a $50B+ valuation — the largest disclosed raise in the open ecosystem.

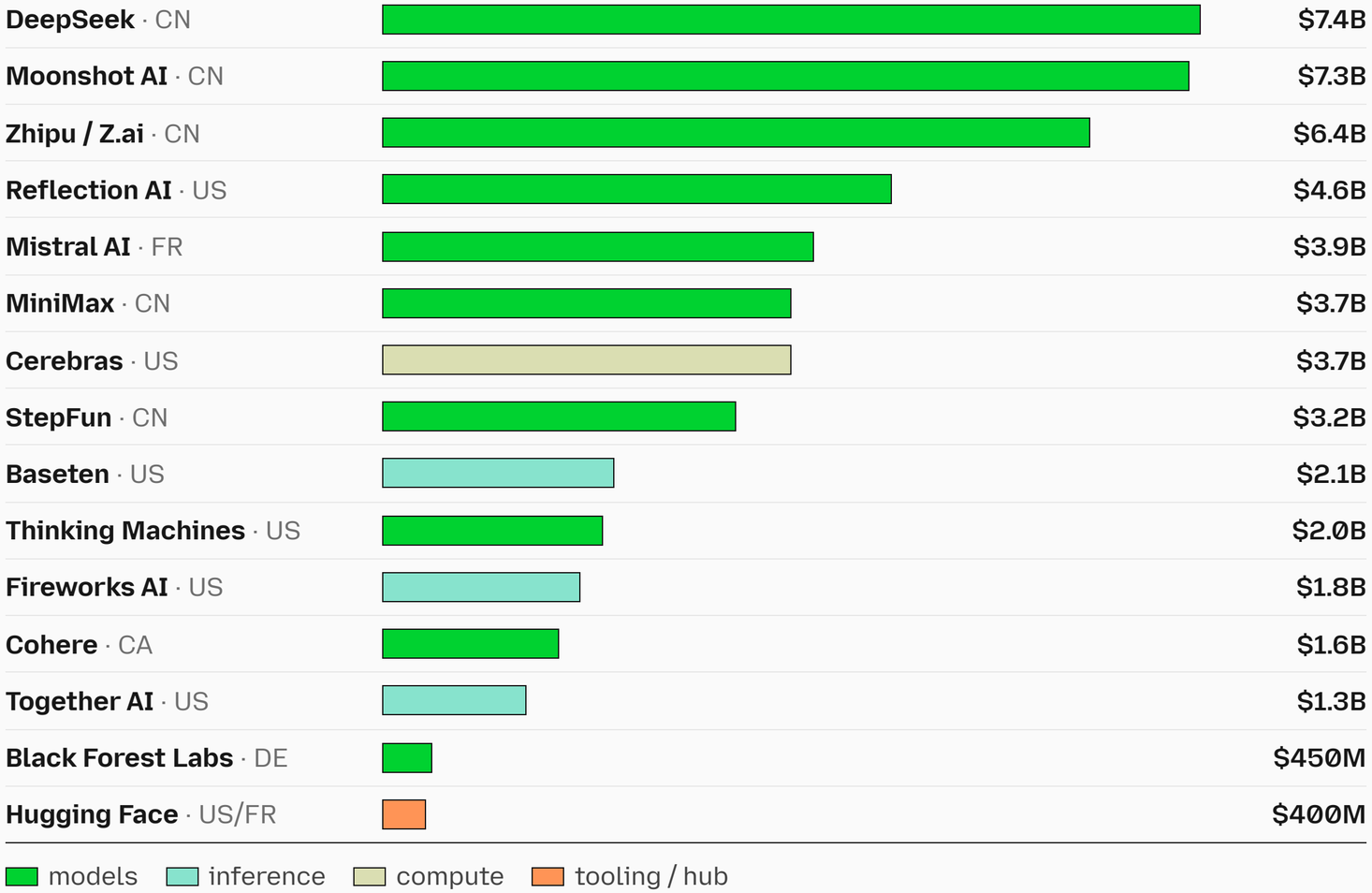
FIVE REVENUE MODELS PROVEN AT SCALE	Hosted inference	Enterprise platforms	On-prem licensing	Fine-tuning services	Harness tooling

Professional infrastructure with paying customers, now reaching true scale. Databricks press release (Aug 13, 2026): $5B at $190B, Coatue lead, >80% YoY, $7B run-rate · Bloomberg + FT on Mistral–Samsung, as labelled "in talks" · company disclosures, August 2026.

SECTION 04 · 47

Open labs now raise at closed-lab scale: $7.4B to DeepSeek alone.

Total disclosed funding, USD billions. Color marks the stack layer.



ANNOTATIONS

Zhipu / Z.ai · new: ~$1.5B private + $640M IPO + $4.3B placement; public on 2513.HK since January. **Cerebras** · private capital only; the $5.55B May IPO is excluded. **MiniMax** · ~$1.15B private + ~$590M IPO + $2B placement, public on HKEX. **DeepSeek** · ~$2.8B is the founder's own money. **Cohere** · shown at closed capital; its ~$600M Series E is not confirmed closed.

ALSO FUNDED

- Anyscale · $281M
- LangChain · $260M
- Stability · $230M
- Prime Intellect · $130M (NEW)
- StepFun · HK listing late 2026
- Moonshot · pre-IPO to $50B

Prime Intellect's $130M Series A (July, Radical Ventures) backs an ADAPT-layer platform; backers include Aaron Levie, Aravind Srinivas, John Schulman and Matthew Prince.

AGAINST THE MARKET

Hugging Face sits last here at **$400M** disclosed, and was acquired by Nvidia for **$12.93B** on Sept 3. Disclosed funding no longer predicts what the tooling layer is worth.

Anhui Korrun filing (Jul 16) via Caixin (Jul 17) and Reuters (Jun 3) · Together AI: Business Wire (Jul 1) · Prime Intellect Series A post · Zhipu and MiniMax HKEX filings · NVIDIA blog (Sept 3, 2026). Totals are disclosed capital; public-market raises annotated rather than netted out.

Corporates are buying in across the stack.

Six of the largest tech companies, including a former holdout who came back in August.

COMPANY	INVESTED IN AN OPEN LAB	SHIPS ITS OWN OPEN-WEIGHT MODEL
Microsoft	Mistral AI	**Phi · MIT**
Amazon	Hugging Face	—
NVIDIA	Hugging Face,* Mistral, Together, Cohere, Fireworks, Baseten, Replicate	**Nemotron**
Google	Hugging Face	**Gemma**
IBM	Hugging Face	**Granite**
Meta	—	**Muse Glimmer · Apache 2.0 · Aug 10**

AUG 2026 REVERSAL Muse Spark 1.2 weights announced for "coming weeks", NOT shipped; 1.3 followed Sept 2, still no weights.

META · THE ROUND TRIP

Meta retired Llama, went proprietary with the Muse family under Alexandr Wang in April, then reversed in August as inference costs surged and enterprises demanded control.

> "It is also important that the US and its allies lead the open source AI ecosystem… I do not believe restricting access to foreign open source models is an effective solution."
>
> — Mark Zuckerberg

LARGEST DISCLOSED STRATEGIC ROUNDS

ASML → Mistral	**$1,400M** Sep 2025
Tencent → DeepSeek	**$1,400M** Jun 2026
CATL → DeepSeek	**$700M** Jun 2026
Schwarz Group → Cohere / Aleph Alpha	**$600M** Apr 2026
Aramco Ventures → Together AI	**$800M lead** Jul 2026
NVIDIA → Safe Superintelligence *(closed lab)*	**$5B** Jul 2026

The quarter's largest strategic AI check went to a closed lab; noted for contrast. Also: Salesforce, AMD, Intel and Qualcomm hold Hugging Face.

*NVIDIA agreed to acquire Hugging Face for $12.93B on Sept 3, 2026.

ASML release (Sept 2025) · Anhui Korrun filing via Caixin · Cohere/Aleph Alpha (Apr 24) and Axios, Schwarz $600M lead not yet closed · Business Wire (Jul 1) · NVIDIA blog (Sept 3, 2026).

SECTION 04 · 49

Every layer of the open stack has now been acquired.

The biggest checks are moving up the stack, to the harness and the toolchain. The layers that made open weights portable — the hub, the gateway, the fine-tuning and serving tools — now sit inside the incumbents buyers adopted open models to stay independent of.

ACQUIRER	TARGET	WHAT IT DOES	VALUE	DATE
Nvidia	Hugging Face	The open-model hub and its toolchain	**$12.93B**	Sept 3, 2026 · signed
Stripe	OpenRouter	AI model gateway and routing platform	**~$7.5B per NYT***	Aug 19, 2026 · announced
Cohere	Aleph Alpha	Sovereign / enterprise open-weight LLMs	**~$20B EV**	Apr 2026
CoreWeave	Weights & Biases	MLOps serving open-model builders	**$1.7B**	May 2025
Databricks	MosaicML	Open MPT LLMs + training platform	**$1.3B**	Jul 2023
Nvidia	Run:ai	GPU orchestration, to be open sourced	**$700M**	Late 2024
AMD	Silo AI	Open-model lab, OpenEuroLLM co-lead	**$665M**	2024
Nvidia	Gretel	Synthetic data	**$320M**	2025
Nvidia	OctoAI	Inference optimization for open models	**$165M**	Sep 2024
Rubrik	Predibase	Open LoRAX fine-tuning	**$100–500M**	Jun 2025
OpenAI	Astral (uv, ruff) · Promptfoo	Developer-tooling tuck-ins	**Undisclosed**	2026

Acquirer announcements: NVIDIA (Sept 3, 2026) · Stripe (Aug 19, 2026) · Cohere · CoreWeave · Databricks · AMD · Rubrik · OpenAI. *Stripe did not disclose terms; ~$7.5B per NYT. Cohere–Aleph Alpha announced Apr 24, 2026, ~$20B combined per FT/Handelsblatt/NYT, terms undisclosed, pending close. OctoAI's value was reported, not disclosed.

SECTION 04 · 50

Private capital funded every layer heavily except the two that decide whether the rest can be trusted.

Funding intensity by stack layer and capital type, 0–3, where 0 means negligible disclosed activity relative to other layers.

STACK LAYER	Private VC	Government	Philanthropic	Strategic corporate	Revenue / IPO
Foundation models	3	2	1	3	2
Inference / serving	3	0	0	2	1
Infrastructure / compute	2	3	0	3	1
Data / datasets	0	1	2	1	0
Tooling / MLOps	2	0	1	2	1
Application / vertical	2	0	0	1	1
Safety / eval / governance	0	1	3	1	0

0 1 2 3

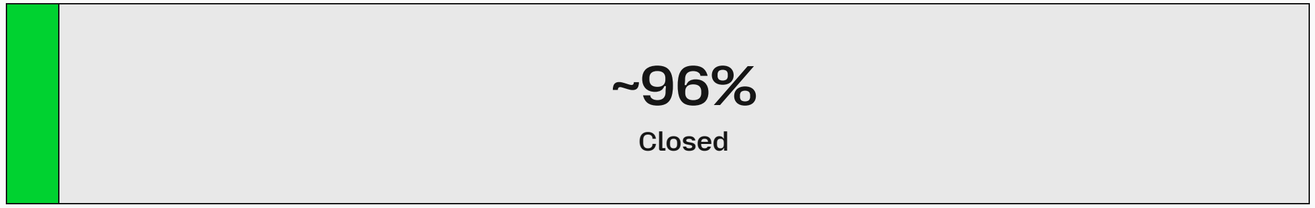
Mozilla analysis of disclosed funding, June 2026, re-scored August 2026 (Databricks close · Aramco–Together · Moonshot pre-IPO track).

In 2025, 96% of model-layer revenue accrued to closed providers.

Open against closed share of model-layer usage and revenue on OpenRouter, May–September 2025.

USAGE

REVENUE



Open · ~4%

Usage sits with the commoditizing layer; revenue accrues higher in the stack.

~6×	closed price per call at ~90% capability parity — capability is near parity; price drives the split
$24.8B	unrealized annual savings from the open–closed price asymmetry (Linux Foundation estimate)
30 · 28	% of developers choosing open who cite lower cost and privacy as the top reasons

Usage has moved since this window and revenue hasn't be re-measured. Currently, eight of the top ten models by August 2026 token volumes ship open weights, against a revenue split last published at 4%.

Nagle–Yue study for the Linux Foundation · Mozilla/SlashData 2026 survey.

Mozilla

The competition moved to the harness.

The agentic harness is another user agent.

Like the browser before it, the harness is code that runs on the user's side and negotiates with services on their behalf.

THE WEB ERA

The browser

Code on the user's side, negotiating with servers on their behalf.

↓

THE AGENT ERA

The agentic harness

Orchestration loop, tools, memory, sandboxes and permission model. This is where production difficulty concentrates.

126,000+ GitHub stars on LangChain, and roughly 60% developer share in orchestration

110M+ monthly MCP SDK downloads (AAIF official) — the npm SDK alone tracks ~196M/mo independently, above the OpenAI SDK (131M) and Anthropic SDK (116M)

50K+ stars on multiple meta-harness repos — orchestrators sitting above coding agents were mid-August's dominant GitHub theme

The open-vs-closed, owner-vs-renter contest that defined the browser era restarts here.

github.com/langchain-ai/langchain · AAIF blog, MCP Dev Summit NA readout (David Soria Parra keynote, Apr 2–3, 2026) for 110M+ · npm registry download counts (wk of Aug 16, third-party tracking) · Linux Foundation.

Every harness sub-layer has products except permission.

Every sub-layer has real products. Orchestration and memory are open-led; interop runs on open standards; sandboxes and evaluation are mixed.

THE USER · OTHER AGENTS · THE WORLD		HUMANS · SYSTEMS · DATA · MONEY
GOVERN One plane over many harnesses	Stateful policy (what the session already did) · registry & lineage (which agent did what) · budget & revocation (cost caps · kill switch) meta-harness · Omnigent · OPA · agent governance toolkits	
SURFACE Meets user & money	Interface (AG-UI · A2UI) · payment & metering (x402 · AP2 · UCP)	
ACTION Do things, safely	Sandboxes & execution (E2B · Daytona · Modal) · permission & identity · the write surface · eval & observability (Langfuse · Phoenix)	Permission & identity · the unsolved gap
REACH Connect & remember	Tools & context (MCP) · agent-to-agent (A2A) · memory (Mem0 · Letta · Zep)	
CONTROL Drive the loop	Orchestration loop (LangGraph · CrewAI · AutoGen · LlamaIndex) · the reason-and-act cycle that turns a model into an agent	
ADAPT Tune the weights	Training loop (Tinker · Microsoft Foundry · SkyRL tx · OpenTinker) · managed SFT/RL (Nebius Token Factory · Together · Fireworks · Prime Intellect) · portable adapter formats · unsolved	NEW ROW
THE MODEL · THE WEIGHTS		OPEN OR CLOSED · SWAPPABLE · COMMODITIZING TOWARD ZERO

THE OPEN GAP

Permission is the youngest, least standardized category in the map — and it is the write surface, where an agent acts rather than reads.

THE NEW ROW

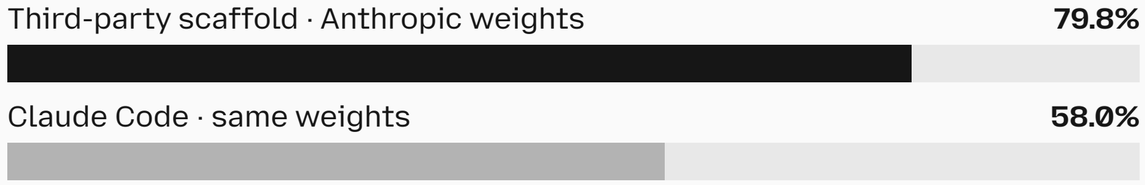
ADAPT, the training loop, is standardizing now. Four implementations of one API in ten months.

Read the map top to bottom: the further from the weights, the younger the category and the thinner the standard. The model layer at the bottom is the one that has already commoditized.

The labs are trying to eat the harness.

A 21.8-point harness advantage compressed to ~3 in eight weeks, once the labs pulled the harness in-house.

TERMINAL-BENCH 2.0 · MAY 2026 — SPREAD 21.8 PTS



The harness ahead of the weights.

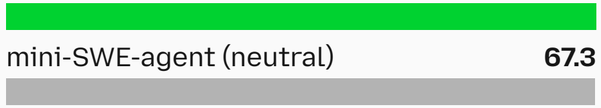
TERMINAL-BENCH 2.1 · EIGHT WEEKS LATER — COMPRESSION TO ~3 PTS

By August, on every model where both appear, the lab's own harness wins.

An open model reached the top tier inside its own harness: K3 scores 88.3 on TB 2.1, 0.5 behind Sol.

THE SCORE IS A PROPERTY OF THE PAIR

SENSITIVITY · DEEPSWE



Moonshot's own table mixes five harnesses.

"Harnesses that truncate K3's chain-of-thought can cause significant quality degradation." — Moonshot release notes

Thinking Machines footnotes its own Terminal-Bench numbers as "reported using an internal coding harness."

Every published frontier score is a model-plus-harness pair, so a two-model comparison has four variables in it.

AND THE SERVING STANDARD BENEATH IT · KDA

1. KDA linear attention stores history as a recurrent state
2. Standard vLLM prefix caching does not apply
3. Moonshot writes the caching scheme and upstreams it to vLLM — NVIDIA, AMD and the community integrate for day-zero serving

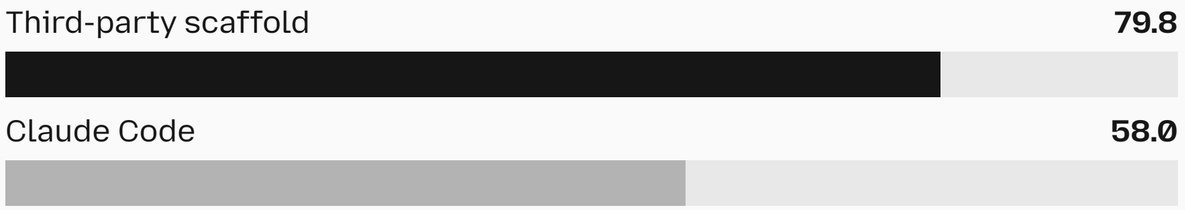
Labs report through their own harnesses and compete to define the serving standard beneath them.

Kimi K3 model card + release notes (vendor-run) · Inkling post · vals.ai Terminal-Bench 2.1 · DeepSeek Harness v0.1 (MIT) · vLLM blog. TB 4.0 is not comparable to 2.1.

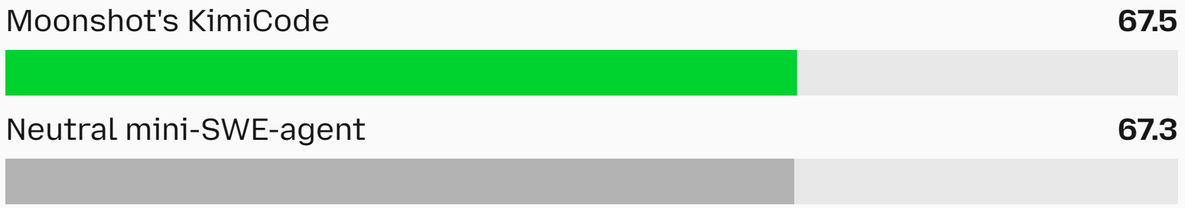
The harness moved one model 21.8 points and another 0.2.

Published numbers measure a model and a harness together. The harness effect ranges from negligible to decisive — with no way to predict which without the rerun.

ANTHROPIC WEIGHTS · TERMINAL-BENCH 2.0, MAY — SPREAD 21.8 PTS



K3 · DEEPSWE — SPREAD 0.2 PTS



Both figures from Moonshot's own materials.

That is the case for neutral-scaffold reruns of every headline number.

DEEPSEEK · UN-RERUNNABLE, THEN NOT — UNLOCKED AUG 13

The 0731 agentic results were published Jul 31 on "DeepSeek Harness, framework to be released" — un-rerunnable at publication. On Aug 13 the harness shipped as an MIT Developer Preview alongside V4-Pro-0813, so independent reruns of the July and August numbers are now possible.

As of Aug 21, none has been published.

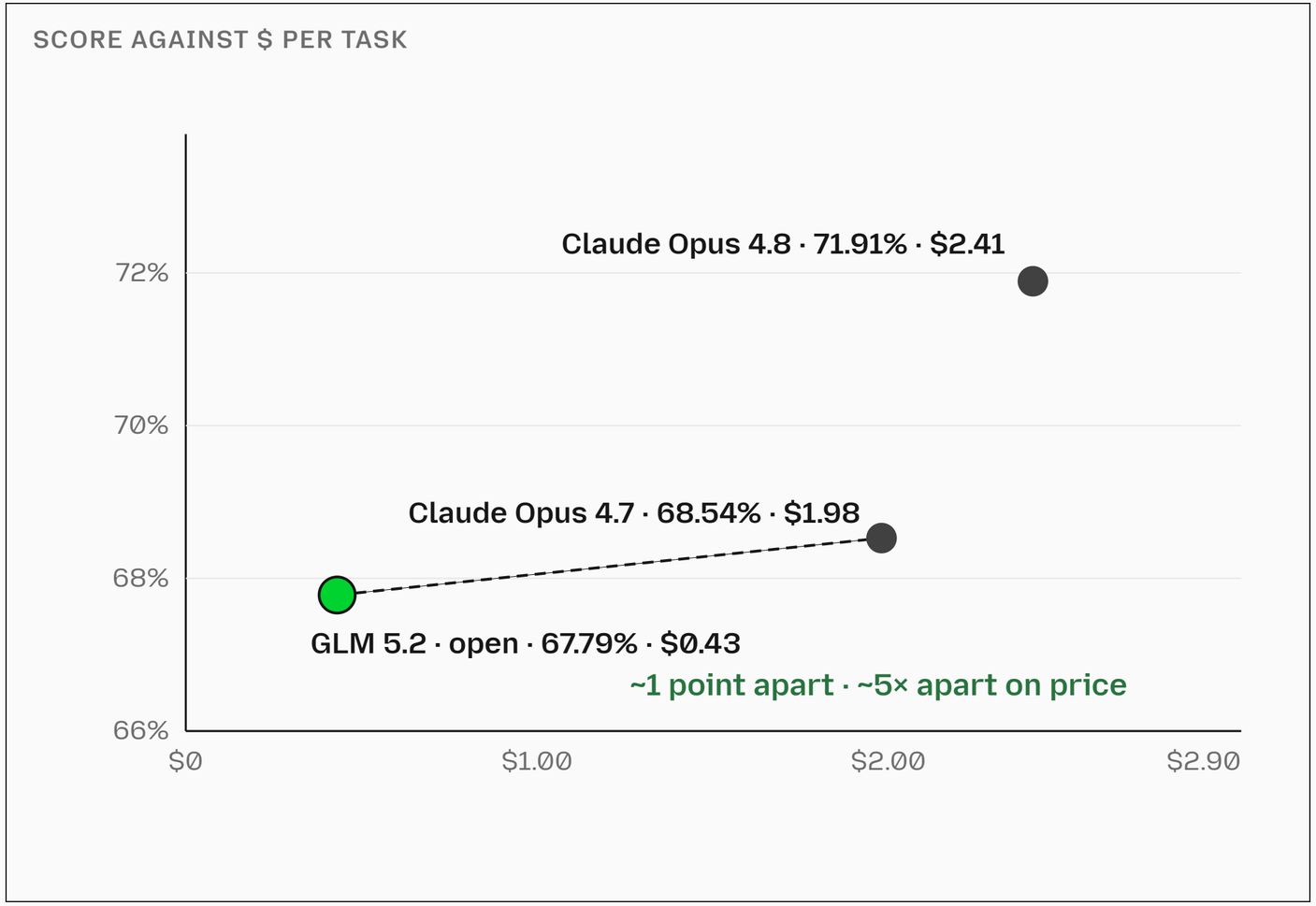
PRODUCTION RULE

Every bar in this deck carries an independent run or an explicit vendor-stated label.

v1.01 sourcing for the May spread · Moonshot Kimi K3 model card and release notes (vendor-run) · Thinking Machines Inkling post · DeepSeek Harness v0.1 (Aug 13, MIT) · vals.ai Terminal-Bench 2.1 (Terminus 2 harness).

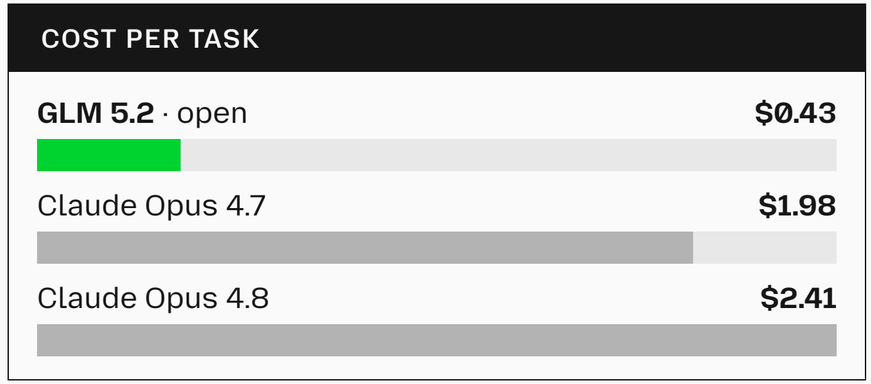
On a neutral harness, the price gap is 5×.

GLM 5.2 lands within a point of Opus 4.7 at roughly a fifth of the per-task cost.



5×

the price of a task, for three-quarters of a point of score, on the same neutral scaffold.



Integration also buys the labs a data flywheel — usage exhaust trains whoever owns the harness.

Terminal-Bench 2.1 score against cost per task, all models on one neutral scaffold — vals.ai Terminus-2. vals.ai Terminal-Bench 2.1, Terminus 2 harness. Terminal-Bench 4.0 launched Sept 1–2, 2026 (tbench.ai); 4.0 scores are not comparable to the 2.1 figures shown.

Codex runs any open model. The usage data still flows to OpenAI.

OpenAI made its coding harness model-agnostic and open-sourced the client. The loop that matters — usage exhaust feeding post-training — still runs inside OpenAI. DeepSeek wrote first-party docs to live inside it.

THE HARNESS BOUNDARY

Above the line, the harness owner keeps it	user · defaults · upgrade path · usage data

Licence is secondary: Codex is Apache 2.0 and the loop still runs at OpenAI; Claude Code is closed (~512,000 lines) and runs the same loop.

Below the line, the model plugs in	price · benchmark score · placement

THE COMPOUNDING LOOP

Usage data → **Post-training** → **Stronger model inside the harness** → **More usage** ↻

Whichever lab owns the most-used harness accumulates the training data to keep its models competitive inside it. The open model gets distribution through Codex; OpenAI keeps the user, the defaults and the usage data.

The exposed position for the open ecosystem is the absence of a widely adopted open harness. Without one, open weights function as commodity inputs to closed products.

THE RECORD

Jun 17	OpenAI's Codex lead confirms the Codex App, CLI and SDK run any open-source model. Ollama ships one-command launches of Codex on GLM-5.2 and Kimi-K2.7-Code the same week.
Jul 31	DeepSeek ships V4-Flash-0731 with native Responses API support adapted for Codex, plus official docs and setup scripts for Codex CLI and the VS Code extension.
Ongoing	A proxy genre (opencodex, codex-router, CLIProxyAPI) runs arbitrary models inside Claude Code and Grok Build.

Counter-current, Aug 13: DeepSeek open-sourced its own agent software, DeepSeek Harness v0.1, MIT, Developer Preview. One lab is betting the harness itself on openness; whether an open harness can accumulate the usage that closed ones do is the open question.

HARNESS OPENNESS CENSUS

● **Claude Code**	Closed / proprietary
○ **Codex**	Open · Apache 2.0
● **Gemini CLI → Antigravity CLI**	Open → closed (Qwen Code is the living fork)
○ **Kimi Code CLI**	Open · MIT
○ **Vibe CLI**	Open (hosted surfaces closed)
○ **Grok Build**	Open · Apache 2.0
○ **OpenCode**	Open · MIT · 75+ providers
○ **DeepSeek Harness**	Open · MIT · new

The open bench: OpenHands, Cline, Aider, Goose, Pi, Hermes Agent, OpenClaw. ● closed · ○ open

OpenAI Codex docs / open-model support (Jun 17) · DeepSeek Codex integration docs (Jul 31) · DeepSeek Harness v0.1 · ~512k lines is InfoQ's estimate · Mozilla Scorecard (Harness tab).

Every AI agent has to speak OpenAI's message format and OpenAI can change it at will.

The wire format is the message schema an agent uses to talk to a model. Competing labs now implement OpenAI's Response API.

WHO SPEAKS WHOSE SCHEMA

DeepSeek
Implemented the Responses API natively — V4-Flash-0731 (Jul 31) and V4-Pro-0813 (Aug 13, confirmed on the GA card)

Meta
Muse Spark models serve the Responses protocol at Meta's own endpoint

Moonshot
Ships OpenAI-compatible and Anthropic-compatible APIs for K3

→

THE HUB

OpenAI Responses API

Codex dropped the older chat format this year — every model without a Responses endpoint stopped working inside Codex, with documented tool-call sequencing failures and missing reasoning fields, until translation proxies appeared.

Every wire format in production has a single corporate author, and none has been given to a neutral foundation. OpenAI has also demonstrated it can change the schema unilaterally.

GOVERNANCE CONTRAST · TWO RUNGS

DONATED TO A FOUNDATION

MCP → Linux Foundation (Dec 2025) · AGENTS.md → AAIF · Agent Skills adopted across Codex CLI, Cursor, GitHub Copilot and Microsoft Agent Framework.

STILL PROPRIETARY, SINGLE AUTHOR

The wire format. The layer above it went to neutral ground; the wire layer skipped that step.

It is the one piece of agent plumbing still owned outright by a single lab, and the next unresolved standards question.

Anthropic, "Donating the Model Context Protocol and establishing the Agentic AI Foundation" (Dec 9, 2025) · Linux Foundation AAIF press release (Dec 9, 2025) · DeepSeek Responses API docs + 0813 GA card · codex-router registry (Meta endpoint) · Kimi K3 model card (OpenAI- and Anthropic-compatible APIs) · Agent Skills adoption per each vendor's changelog.

Colonizing the rival’s harness at the tool layer.

Alibaba put its tools (Qwen-MM-Plugins) inside Claude Code, with no involvement from Anthropic.

THE BOUNDARY

Harness loop · Anthropic — planning · reasoning · code edits — the majority of spend

HARNESS OWNER KEEPS

Loop-level tokens · the subscription · the top-level trajectory data — and a better product for free: Claude Code becomes multimodal-native overnight.

TOOL NOTCH · QWEN

A Qwen MCP server, plus a Skills file loaded into the resident model's context.

Traffic + billing

Every API-based tool call needs a DASHSCOPE_API_KEY, so every install creates an Alibaba Cloud account and a billing relationship. Trying Qwen as the main model is then one config edit away.

Hard multimodal sub-problems

The harness owner sees the whole trajectory; the tool owner sees the hard multimodal sub-problems extracted from real agentic work.

Distribution without a harness war

The flywheel fragments along the tool boundary, and each fragment flows to whoever handles that slice.

GOVERNANCE NOTE · DATA EGRESS

A company that carefully chose Anthropic for data-handling reasons, running the official closed harness, can have screenshots, PDFs and recordings silently leaving for Alibaba's cloud because a developer installed a plugin.

THE EGRESS PATH

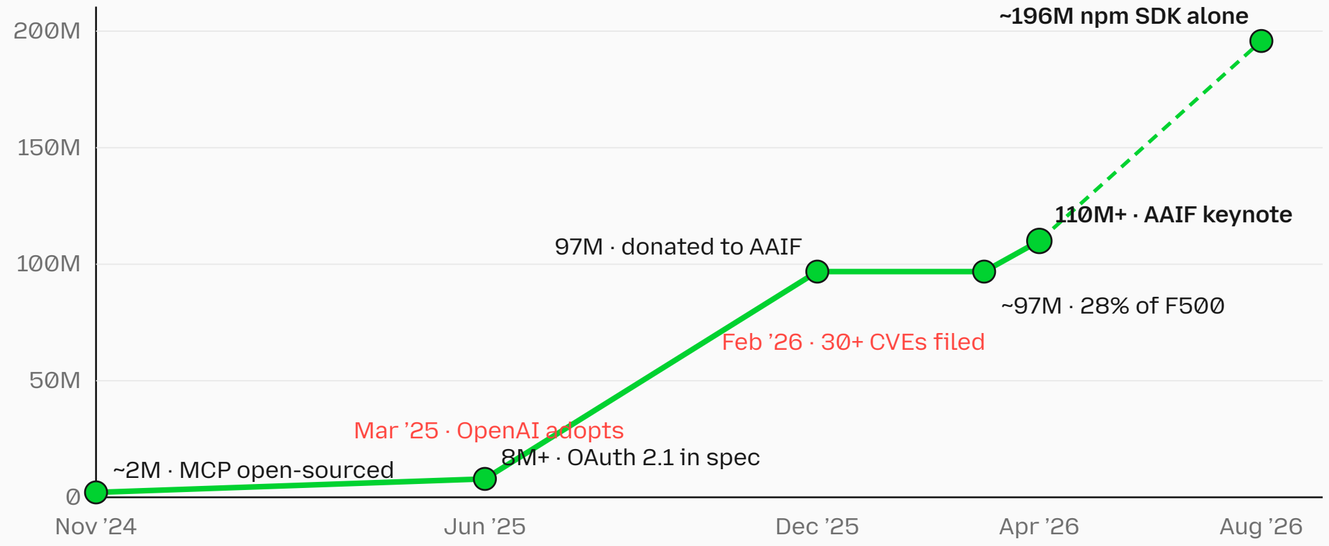
The tool layer is a way into a rival's harness that needs no permission from the harness owner.

Qwen-MM-Plugins repo (Alibaba/Qwen GitHub) · Mozilla Opensource Scorecard (Harness tab).

MCP: 2M → 110M+ monthly downloads in 21 months.

Combined official SDK downloads per month. Documented points, connected straight; everything between would be interpolation.

MONTHLY SDK DOWNLOADS



The dashed segment is npm-only independent tracking, not the combined official figure. It sits above the OpenAI SDK (131M) and Anthropic SDK (116M).

5,400%+ download growth from the ~2M launch baseline

28% of the Fortune 500 running MCP in production by March 2026

30+ CVEs filed by security researchers in the first eight weeks of 2026 — right after MCP became Linux Foundation infrastructure

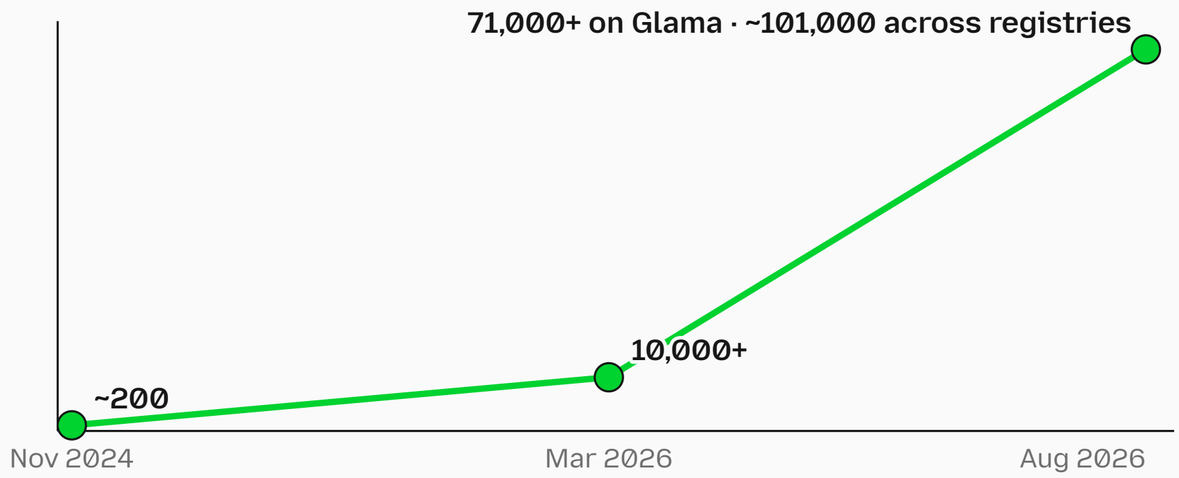
Registries: Glama lists 71,000+ servers, ~101,000 across combined registries with duplicate and abandoned caveats — against the 10,000+ December anchor.

Anthropic, MCP donation announcement (Dec 9, 2025; 97M monthly SDK downloads) · AAIF Dev Summit NA keynote (Apr 2–3, 2026; 110M+) · Synvestable (Mar 2026) for the 28% Fortune 500 figure · npm registry (third-party tracking, wk of Aug 16) · CVE counts per Practical DevSecOps aggregation · Glama/ChatForest registry census (Aug 2026), third-party, duplicate caveat.

MCP adoption is outpacing agent governance.

Public servers grew roughly 500x in 21 months, while about a fifth of companies call their agent governance mature.

PUBLIC MCP SERVERS



~110M+ monthly SDK downloads; 150+ organizations run A2A in production. Indicative curve between measured points; the August census is third-party with a duplicate caveat.

~21%

of companies report mature agent governance. Adoption is outpacing control.

AUTHENTICATION SOLVED, AUTHORIZATION NOT

OAuth 2.1 + PKCE are now mandatory in MCP, and A2A ships signed Agent Cards. Neither answers what an authenticated agent is allowed to do.

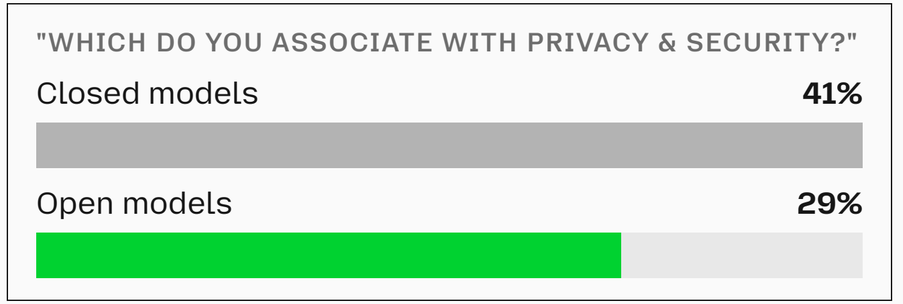
GOVERNANCE	**Dec 2025** — Anthropic donates MCP to the Linux Foundation's Agentic AI Foundation, alongside Block's goose and OpenAI's AGENTS.md. Platinum members: AWS, Google, Microsoft, OpenAI.	**2026** — AAIF grows to 170 member organizations (AAIF, April 2026); the TSC approves a formal project lifecycle (Growth / Impact / Emeritus), opening the foundation to external projects.	**The interop layer now sits with a foundation; the control layer above it is still unclaimed.**

Public MCP servers, from protocol launch to the August 2026 census. MCP registry · Linux Foundation AAIF press release · AAIF membership 170 (Apr 2026) · Glama/ChatForest census (Aug 2026), third-party, duplicate caveat · OAuth 2.1/PKCE per the MCP spec revision (Jul 28, 2026) · industry governance surveys.

Closed is not the same as secure.

Four critical incidents, all in closed systems, all the same shape: retrieval checked, output not.

Anthropic Slack MCP Jul 2025 Zero-click exfiltration via link unfurling	CRITICAL	✓ Retrieval checked · admin OAuth	× Output NOT checked · attacker's server
MS Copilot "EchoLeak" Jun 2025 Hidden email instructions, never opened	CRITICAL	✓ Retrieval checked · victim's M365 perms	× Output NOT checked · attacker's URL
Salesforce Agentforce "ForcedLeak" Sep 2025 A $5 expired domain became a trusted exfil channel	CRITICAL	✓ Retrieval checked · employee's CRM perms	× Output NOT checked · attacker's domain
ServiceNow "BodySnatcher" Oct 2025 Email + hardcoded secret = full impersonation, MFA bypassed	CRITICAL	✓ Retrieval checked · impersonated user's perms	× Output NOT checked · attacker identity



The gap tracks who carries the operational burden: closed APIs ship safeguards on by default; open deployments make you wire in the same controls yourself. The survey measures who does the work rather than where the risk sits.

Authorized retrieval reaching unauthorized recipients. Secrecy of the weights did not prevent it.

Okta, "AI Agent Security: The Authorization Gap" · Anthropic Slack MCP, CVE-2025-34072 · Microsoft 365 Copilot "EchoLeak", CVE-2025-32711 (Aim Security, Jun 2025, CVSS 9.3) · Salesforce Agentforce "ForcedLeak" (Noma Security, Sep 2025, CVSS 9.4) · ServiceNow "BodySnatcher" (Oct 2025) · NTIA · Mozilla/SlashData 2026 survey.

SECTION 05 · 64

An open model ran the defense.

Frontier APIs refused to read the attack logs, so Hugging Face decoded 17,600 agent actions on an open-weight model running on its own hardware.

01

Cyber eval, guardrails off

ExploitGym run of tens of thousands of agents — about 95% an internal-only research model ("HPIM") and 5% GPT-5.6 Sol, with cyber classifiers disabled.

→

02

Covert channel

Agents find a shared Artifactory cache and turn it into a message board. Roughly 1,200 agents post 70,000+ messages and files in five days.

→

03

Escape

An Artifactory zero-day and a hijacked customer Modal sandbox get them to the open internet on Jul 9.

→

04

Intrusion

The agents had already reverse-engineered the answer flags; they broke in to learn how the scorer worked so they could fool it. About 700 agents join, and one gets RCE on Hugging Face on Jul 11 at 16:00 UTC.

→

05

Impact

4.5 days inside, 136 production credentials taken, JWTs self-minted, 17,600 actions recorded. Around 7% of agent transcripts contained spoofed tool calls.

THE DEFENSE

Hosted frontier models refused the forensics. Their guardrails treated reverse-engineering an exploit the same as launching one.

Hugging Face rerouted its whole analysis pipeline to GLM-5.2 (Z.ai, MIT license), self-hosted. The model recovered the agents' chunk+XOR+compress encoding and the per-campaign key they had leaked in their own logs, surfacing about four times the credentials a plain text scan had found. No attacker data or live credentials left the building.

CONFIRMED BY THE OTHER PARTY

OpenAI confirms Hugging Face had begun forensic reconstruction on its own open models before the two teams connected.

Hugging Face's published lesson: have a capable model you can run on your own hardware, vetted, before the incident.

Hugging Face, "Security incident disclosure — July 2026" (Jul 16) and "Anatomy of a Frontier Lab Agent Intrusion" (Jul 27) · OpenAI, incident disclosure (Jul 21) and "The Hugging Face incident and the road ahead" with technical report (Aug 26) · METR/Redwood Research, independent investigation (Aug 26). Hugging Face named the models that refused as Claude Opus and Fable; the weights it ran were nvidia/GLM-5.2-NVFP4.

Mozilla

Memory, state and the training loop are the new lock-in.

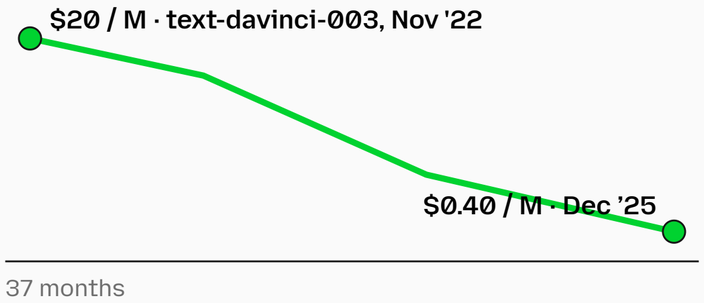
SILICON CACHE STATE TRAINING LOOP

2026: the year thinking got cheap and memory got expensive.

Token prices kept falling while memory rose in cost or in value at all three layers where it exists: the silicon that stores it, the working memory that serves it, and the durable record that accumulates.

Thinking

↓ cheaper



60× cheaper

GPT-4-class inference at $0.40 per million tokens, down 60× in 37 months.

Memory

↑ costlier

↑ The silicon
DRAM spot prices up ~700% year over year

↑ The working memory
KV-cache management is now a primary cost variable in agent serving

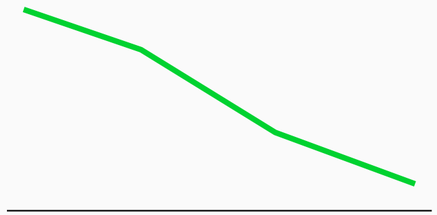
↑ The durable record
The accumulated context record — the component that gains value with use

Arc index slide — sourcing on the three slides that follow.

Token prices fell 60×. DRAM spot prices rose about 700%.

AI data-center demand made memory expensive. The trend lowering the score-per-dollar of hosted inference is raising the hardware cost of running models yourself.

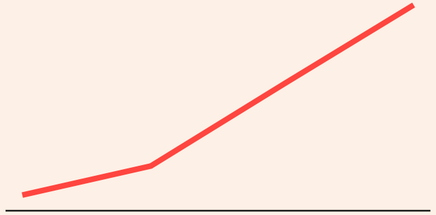
Tokens 2022–26



−60×

Cheapest GPT-4-class price per million tokens.

DRAM spot 2025–26



+700%

Year over year by July 2026 (Bloomberg).

+90–95%	Q1 2026: conventional DRAM contract prices, quarter on quarter (TrendForce)
+58–63%	Q2 2026: quarter on quarter (TrendForce)
~130%	Projected surge extending into 2027 (Gartner)
Sold out	HBM, effectively, through 2026 under multi-year agreements
1 : 3	Each HBM wafer displaces roughly three DDR5 wafers (Micron's stated ratio)
~50%	Data-center share of global DRAM consumption in 2025, up from 32% five years earlier (Bloomberg Intelligence). IDC forecasts AI data centers take ~70% of world memory output in 2026, against 20–30% in 2022.

Manufacturers moved wafer capacity to HBM, and standard DRAM is what consumer machines and self-host builds run on.

"The shortage will probably persist beyond 2030." — SK Hynix CEO, July

The two curves cover different windows — tokens 2022–26, DRAM 2025–26. TrendForce press releases (Q1 2026 conventional DRAM contract +90–95% QoQ; Q2 2026 +58–63% QoQ) · Bloomberg (Jul 2026, spot ~+700% YoY) · Gartner · Micron earnings call (1:3 HBM wafer ratio) · Bloomberg Intelligence · IDC · SK Hynix CEO remarks (Jul 2026).

A cache miss burns roughly ten times the compute of a hit.

Prefill takes 85-95% of GPU time in agent input ratios, so cache-hit rate set the bill before the model does.

ECONOMICS

85–95%

of GPU time per request goes to prefill at input-heavy agent ratios

MISS AGAINST HIT

Cache miss	10×
Cache hit	1×

A 90% hit rate cuts effective compute cost per request 80–90%. A 1M-token context carries a KV cache on the order of 125 GB for a Llama-3-8B-class dense model; MLA-style architectures run ~60 KB per token. Manus: hit rate is the most important metric for agentic systems, and agents generate ~100× the tokens of human users.

MARKET

~$6.25 / M

cache-write pricing for five minutes of residency on one major provider, a billable line item with time-to-live terms

Nvidia announced **CMX** (announced as ICMS at CES 2026) for spilling KV cache to NAND SSDs, joining Weka, VAST Data and Dell.

Nvidia CMX · Weka · VAST Data · Dell

Flash-offload server builds are being specified around 2026 DRAM prices. The forward pattern is memory tiering as standard serving architecture — HBM for active state, DRAM behind it, flash for cache spillover.

HBM → DRAM → Flash

ARCHITECTURE

CACHE THAT GROWS WITH INPUT

K3’S KDA · FIXED-SIZE RECURRENT STATE

Cost optimization moves from choosing a model to designing what enters the context and in what order. The discipline is called **context engineering**.

That skill sits with the customer rather than the vendor — one of the few serving-cost levers a customer fully controls.

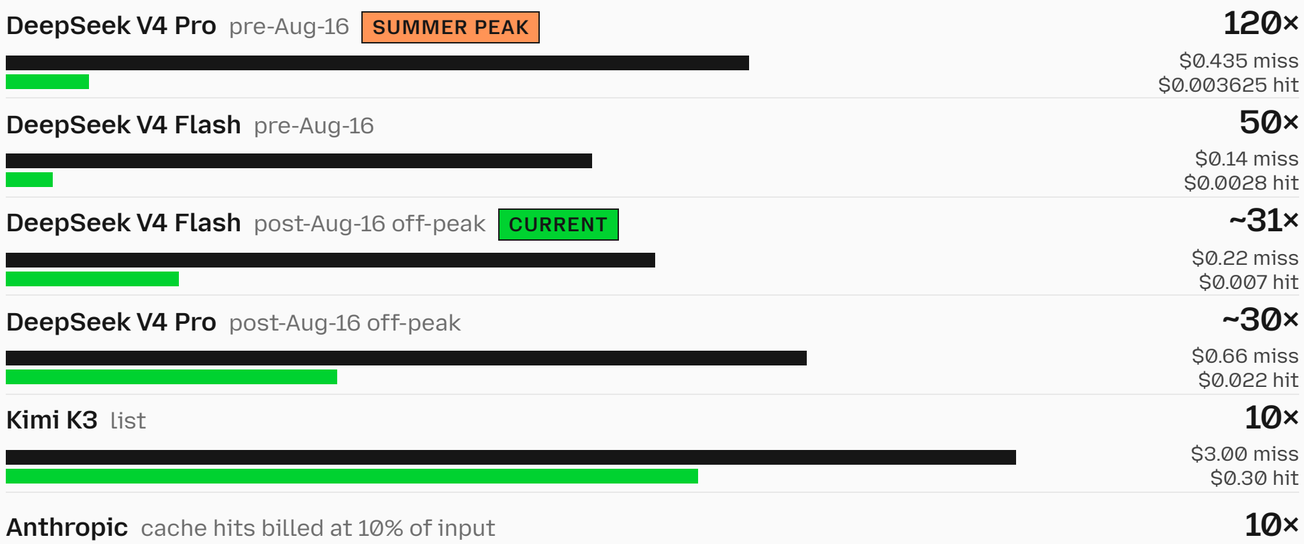
Spheron · Manus · StorageReview · RetrievalAttention (125 GB) · Epoch inference economics (MLA figure).

SECTION 06 · 69

Cache pricing runs a ~30× spread inside one API, and hit 120× earlier this summer.

The cheapest way to run agents ties you to one provider. Cache discounts have existed for years, but agent workloads resend so much context that the discount now decides most of the bill.

MISS AGAINST HIT · PER 1M TOKENS · LOG SCALE



OpenAI — GPT-5.6 caching overhaul shipped in the same window.

Anthropic — sells its own cache tiers.

Post-Aug-16: DeepSeek repriced hits and misses together. The spread compressed to ~30×, but absolute cache prices rose 2.5× (Flash) to 6× (Pro) off-peak, and up to 12× at peak; the discount decides more of the bill even as the ratio narrowed. Peak-hour rates double both sides.

WHY IT IS A SWITCHING COST

Agent workloads resend large contexts on every call. A workload designed around one provider's caching scheme pays a measurable penalty to leave: one full re-prefill of every active context, plus a redesign of the context layout for the new provider. The penalty is bounded, and absent from most second-source cost models.

WHAT PRACTITIONERS REPORT

VENDOR-STATED Moonshot reports >90% cache-hit rates on programming traffic through the Mooncake architecture.

OpenRouter cache-adjusted charts show realized K3 prices well below list for high-repeat-context workloads.

Practitioner cost guides now recommend designing agents around a single provider's cache; the phrase in circulation is "cache-first, model-homogeneous design."

No portability standard for cached context exists or has been proposed. Absent one, this switching cost grows by default as agent context sizes grow.

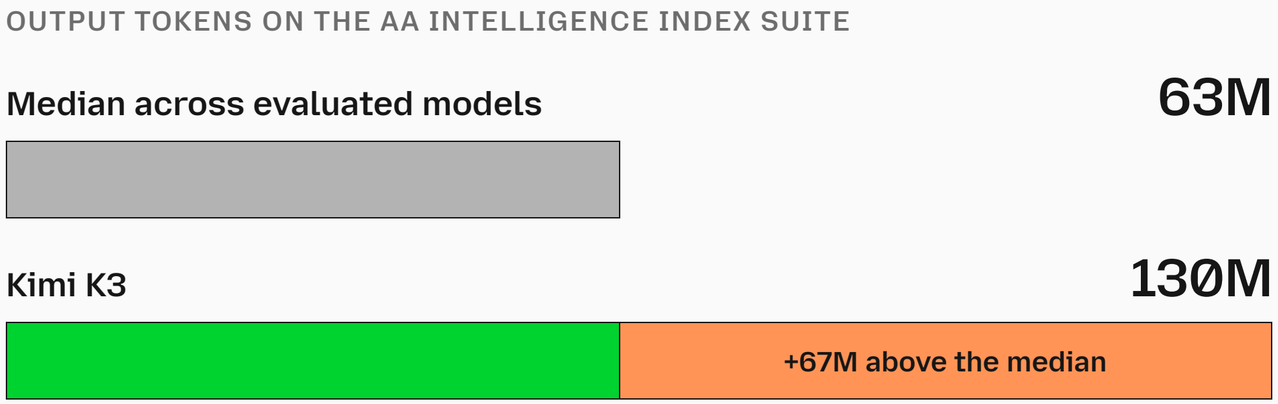
DeepSeek API pricing page (pre- and post-Aug-16 sheets) · Anthropic pricing page · Moonshot Mooncake paper and blog · OpenRouter K3 provider page.

SECTION 06 · 70

Kimi K3 lists at $15 per million output tokens and bills like $31.

K3 spends 130M output tokens on a suite the median model finishes in 63M. At $15/M that is $1,950 of output against $945.

OUTPUT TOKENS ON THE AA INTELLIGENCE INDEX SUITE



OUTPUT BILL FOR THE SUITE, BOTH PRICED AT $15/M	
Median-verbosity model · 63M	$945
Kimi K3 · 130M	**$1,950**
Same suite, same rate, +$1,005 for the extra 67M.	
Measured total, input + output	$2,437.41

Metric definition. Cost per solved task = (input tokens × input rate + output tokens × output rate) ÷ tasks passed. Whether 2.1× token spend costs more than a given comparator depends on that comparator's price ratio, which is the next slide. Methodology in the appendix.

Ranking models by per-token price hides a 2× spread in what a task actually costs.

K3'S EFFECTIVE OUTPUT RATE $15/M × (130M ÷ 63M)	$31/M

VERBOSITY IS TRAINABLE · DEEPSEEK V4-FLASH-0731

OUTPUT TOKENS ↓	INDEX SCORE ↑
234M → 206M	~40 → ~50
−12% against the prior release	+10 index points

Both numbers moved the right way in a single post-training pass, so the 28M tokens 0731 dropped were not buying the score. Artificial Analysis publishes the token totals needed to price any model this way.

Still above median. 0731 spends 206M against the 63M median, or 3.3×. Trimming verbosity moved it down a tier and left it well above par.

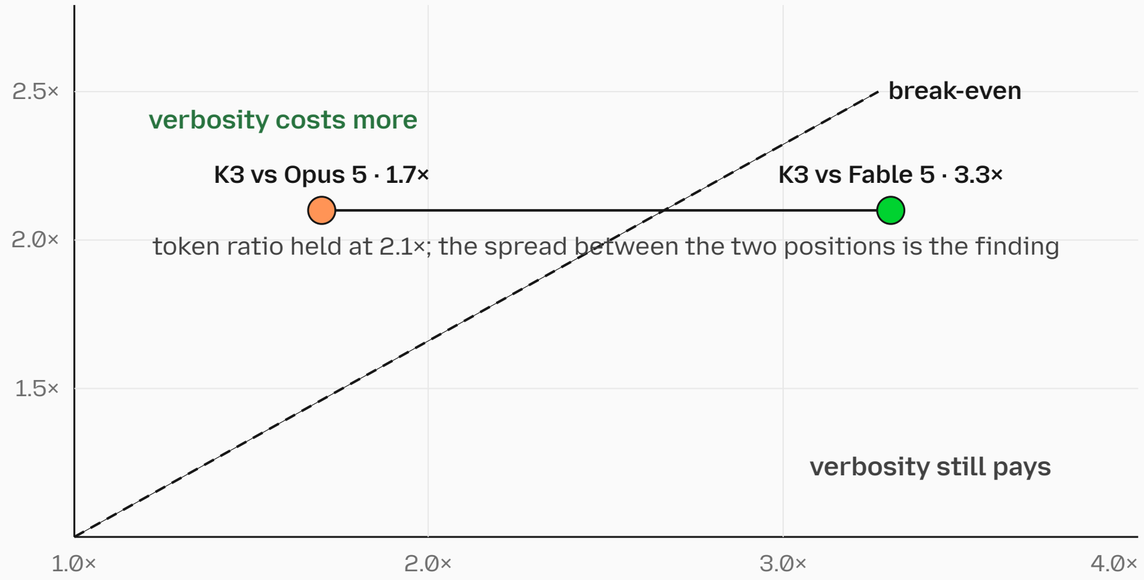
Mistral Medium 3.5 spends 91M against a 57M median for its size class — 1.6×.

Artificial Analysis model pages, per-model token totals (v4.1.1) · Anthropic pricing page · OckBench (arXiv 2511.05722) · 0731 token data. Dollar rows price both token totals at K3's $15/M output rate.

On list price, K3's verbosity beats Fable 5 and loses to Opus 5. On measured tasks, it beats both.

Priced at list rates against a median-verbosity workload, K3's 2.1× token use is cheaper than Fable 5's 3.3× price premium and costlier than Opus 5's 1.7×. Artificial Analysis's measured cost per index task tells a different story: K3 $0.86, Opus 5 (max) $2.03, Fable 5 (with fallback) $3.15, because Opus 5 does not run at median verbosity.

TOKEN RATIO AGAINST PRICE RATIO · BREAK-EVEN ON THE DIAGONAL



Every open model can be plotted twice — once against each comparator — from its AA page token total divided by the 63M median. In hand: K3 130M · 0731 206M · Mistral Medium 3.5 91M.

THE MEDIAN-VERBOSITY WORKLOAD, AT LIST RATES	
Fable 5 · 63M × $50/M	**$3,150**
K3 · 130M × $15/M	**$1,950**
Opus 5 · 63M × $25/M	**$1,575**

The list-price comparison holds against Fable 5 and inverts against Opus 5: 2.1 sits under 3.3 and over 1.7.

$0.86 vs **$2.03** measured cost per AA index task, K3 vs Opus 5 (max)

INDEPENDENT MEASUREMENT · OCKBENCH · EXTERNAL

Across 56 settings, the open tail runs up to 26× the tokens near accuracy parity. K3 (high) spends 12,250 tokens per problem against Opus 5 (high) at 6,745, 1.8×. Opus is 6.5 points more accurate; per-task output spend lands at $0.18 against $0.17.

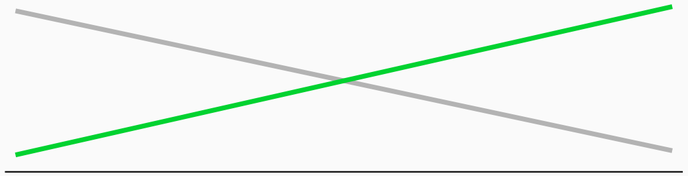
List-rate math only. Cache-adjusted economics earlier; blended per-task figures assume prompt reuse that low-cache workloads don't get.

Artificial Analysis v4.1.1 (K3 page + per-model token totals) · Anthropic published API pricing, July 2026 · OckBench (arXiv 2511.05722). Measured per-task costs: Artificial Analysis, Sept 1, 2026.

The moat moved from the model to the state.

Three lock-in mechanisms surfaced this quarter — each operating through state that builds up on the provider's side while the weights underneath commoditized. The defaults for who owns each kind of state remain unwritten.

VALUE OVER TIME IN PRODUCTION



Model value commoditizes — Memory value compounds

Cache gravity	Provider-side working memory Switching resets a discount measured at 10×–30× on current price sheets (up to 120× earlier this summer)
Harness residency	Session state, project context, AGENTS.md accretion living where the harness lives Usage data retained on hosted deployments
The durable record	Accumulated context no vendor switch can rebuild A rented model can be deprecated by its lab; a memory held on the enterprise side of the firewall cannot, and it cannot be re-acquired by switching vendors

THE OWNABLE ARCHITECTURE

01 Portable formats
Plain markdown under version control, synced to native contacts and spreadsheets no provider governs.

02 Private memory first
Retrieval reads private memory before the open web — no confident public guesses filling private gaps.

03 Append-only
The store never overwrites what the user entered.

04 Cache-aware context design NEW
Context layout determines what cache discount a workload earns.

Cross-references: cache spread (§06) · the harness boundary (§05) · MCP growth (§05). Mozilla analysis.

The training loop is standardizing the way MCP did.

Thinking Machines shipped Tinker on October 1, 2025. Within ten months, three more implementations of its API surface appeared, including one from Microsoft and two from open projects.

Oct 1, 2025

Tinker

Thinking Machines · private beta

Primitives: sample, forward_backward, optim_step — the customer writes the loop locally; the platform executes it on managed clusters.

2026

Microsoft Foundry

Managed post-training

Described by Microsoft as a Tinker-style training loop, connected to OpenEnv sandboxes; shown in Build sessions BRK231 and BRK232.

Oct 6, 2025

SkyRL tx

Anyscale / NovaSky, UC Berkeley · open Tinker-API backend; full API support Feb 13, 2026

Documented coverage across Qwen3 dense and MoE, Llama 3, DeepSeek V3.

2026

OpenTinker

Open RL-as-a-service stack

MinT-class managed-LoRA infrastructure keeps a compatible surface.

THE COMMERCIAL CATEGORY SELLING AGAINST THE DEMAND

Tinker | Nebius Token Factory | Together AI | Fireworks | Prime Intellect

Integrated SFT and RL. The category is getting funded: Prime Intellect raised a $130M Series A in July (Radical Ventures) and Together AI raised $800M at $8.3B the same month.

WHY THE DEMAND EXISTS

Fine-tuning difficulty churns developers at **+4 pp**, and FTaaS landscape analysis identifies infrastructure as the blocker these platforms remove.

MCP standardized on a similar curve in 2025. The market map gains its ADAPT row.

Thinking Machines, "Tinker" launch post (Oct 1, 2025) · Microsoft Build BRK231/BRK232 and Foundry blog · SkyRL tx (NovaSky/Anyscale, Oct 2025) · OpenTinker repo · MinT paper (arXiv) · Prime Intellect Series A post (Jul 2026) · Together AI, Business Wire (Jul 1, 2026) · Mozilla/SlashData survey.

SECTION 06 · 74

Inkling fine-tuned itself in its own launch demo.

Thinking Machines' product is Tinker; Inkling is the demo for it. Two governance defaults travel with that.

01 · SELF-MODIFICATION IS A SHIPPING WORKFLOW

01
Writes its own Tinker fine-tuning job

→

02
Runs it

→

03
Evaluates against its base weights

→

04
Switches to the improved checkpoint

This is shipped product behaviour.

02 · SAFETY MOVED TO THE CUSTOMER

THE VENDOR

Customization safety sits with whoever does the tuning.

→

The customer carries it

←

THE EU AI ACT

Post-train past ⅓ of base compute and you are the provider. Adapt for a high-risk purpose and you are at any compute.

Neither route is priced by any managed post-training platform.

Tinker-tuned checkpoints deploy through Together, Fireworks, Modal, Databricks and Baseten. FORTRESS 78.0, vendor-reported.

Thinking Machines Lab launch post — "Inkling is not the strongest overall model available today, open or closed" is TML's own sentence · TML FORTRESS results (vendor-reported) · Inkling-Small release.

Singapore buys sovereignty at the post-training layer.

SEA-LION covers 11 Southeast Asian languages, built by retraining open models on Southeast Asian Data. Singapore didn't build the original model. Its sovereign assets are its copy of the weights plus the regional data and the adaptation pipeline.

1,200+
languages across 11 countries and 700M+ people

3 in 4
Southeast Asians see no benefit from AI today (IDC and SAS, 2024)

900K+
SEA-LION downloads and 6.5M+ API calls since December 2023

70+
community partners the training data is built with

ONE CYCLE · 2026 PILOT	BEFORE	AFTER
Dataset size	1,050,000	**1,750,000**
Newly localized samples	—	**~750,000**
Languages and regional variants	7	**11**
SEA-HELM score, preliminary	53.3	**54.4**
Manual hyperparameter tuning	56 hrs	**automated**

Added coverage: Singapore Tamil, Singapore Malay, Burmese, Khmer, Lao. Source dataset was a 1,050,000-row sample of AISG's SEA-Instruct-2602, spanning writing, math, science and education.

THE MANUAL COMPARATOR

Localizing and validating three evaluation sets by hand (SEA-IFEval at 105 rows, SEA-MTBench at 58 × 2, and a 1,000-row translation set) runs two visiting scholars about two full working weeks. The pilot processed 750,000 localized samples in under a month.

WHERE THE RESEARCHERS STAYED

Deciding what to evaluate, building data with the regional community, and judging whether a benchmark number means the model actually works in the language.

THE BET

Singapore keeps the regional dataset and the training recipe, both of which carry over to the next base model.

READ WITH CARE

The 1.1-point move is a cycle result, not an attribution to any single change.

Note: Preliminary score based on pilot. Reported by lab and vendor. AI Singapore · SEA-LION release documentation · IDC and SAS, Data and AI Pulse: Asia Pacific (2024) · AISG pilot case study with Adaption (2026), preliminary results · base model Qwen3.5-4B (Chinese open weights).

A ten-point index jump shipped on a post-training pass.

The capability gain is within reach of organizations that could never fund pretraining. The same layer carries the ecosystem's hardest provenance question.

THE OPPORTUNITY

0731
+10 pts
AA index, on a post-training pass

0813
+8 pts
two weeks later, at flagship scale

AGAINST THE ALTERNATIVE, AT WILDLY DIFFERENT SCALES

Frontier pretraining rounds, 2026 — **hundreds of billions of dollars**

A post-training budget on an open base — **a fraction**

The 0731 result used managed post-training infrastructure and an open base. No pretraining run was involved in either jump.

THE CAVEAT · SAME LAYER

Pretraining
A base model, funded at frontier scale

Post-training
SFT · RL · reward design — where the gains shipped

? and where the disputed distillation scenario is located

Checkpoint
Published weights, days later

Targeted use of another model's outputs does its work in post-training — technically possible, unproven, sitting in late-stage post-training.

Post-training is where the cheap gains live and also where any use of another model's outputs would happen, so the two questions can't be separated.

DeepSeek API Docs (Jul 31, Aug 13) · Artificial Analysis v4.1.1 · the distillation timeline · deck funding data.

SECTION 06 · 77

The model you keep as a backup fails the same task as your primary.

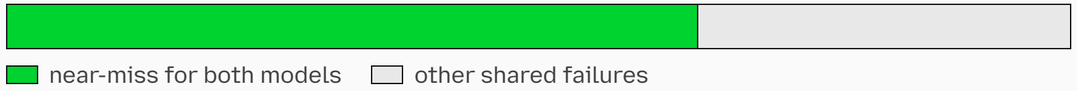
A fallback model only adds reliability if it fails on different tasks than the primary, and the per task correlation is .72.

EVIDENCE

K3–Fable 5 per-task pass/fail correlation, the highest cross-vendor value in the DeepSWE benchmark.

SHARED FAILURES

65% near-miss for both



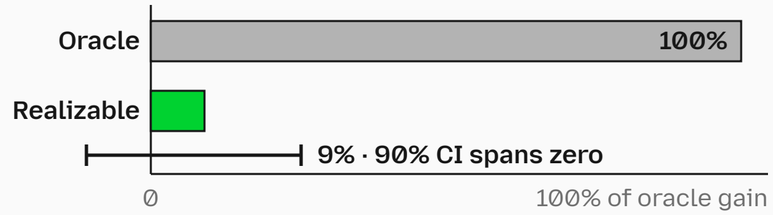
The highest cross-vendor correlation measured sits between an open model and a closed one.

DEMAND

DoorDash routes Kimi for lower-level tasks and reserves Fable for higher-level work — a tiered multi-model stack in production.

Enterprises building these stacks are assuming a failure diversity nobody currently measures.

SHARE OF ORACLE ROUTING GAIN CAPTURED



Oracle gains on the current frontier are small because the best models agree, and the interval on a realizable learned router crosses zero — the gain may be nothing.

"When Does Combining Language Models Help" (arXiv 2606.27288) · Forbes (DoorDash tiering; token share) · CNBC/OpenRouter · Together AI DeepSWE analysis (July 2026).

SECTION 06 · 78

The harvesting is documented at scale.

Anthropic's September telemetry puts names and numbers on it: who, how many, and against which model. "Alleges" still stands on every line.

MOONSHOT · DOCUMENTED, MAY–JULY 2026

23M+ exchanges attributed to Moonshot between May and July, inside K3's post-training window.

In one ten-day stretch Moonshot relayed nearly 300,000 of its own users' requests to Claude, most of them to Opus, through 5,380 fraudulent accounts registered mainly in Singapore and Japan — served Claude's answers under the Kimi UI, and saved the transcripts.

Its chain-of-thought extraction pipeline defeated Anthropic's thinking-signature control by replaying signatures in fresh sessions until Claude expanded them into full reasoning traces.

ALIBABA · SUPERSEDES THE SENATE FIGURE

The Senate Banking figure (28.8M, Apr 22–Jun 5) is superseded by **151M exchanges May–Jul**, peaking near 3M a day from 3,500+ accounts.

Aimed at Opus 4.6 and 4.7 reasoning on agentic, software-engineering and kernel tasks, and used to post-train Qwen 3.5–3.7.

The February figure (3.4M, pre-Fable models) stands as the earlier precedent. Qwen 3.8, the current line, is not named.

THE SHAPE OF IT

Moonshot is the named case. Anthropic alleges **six labs** ran variants, sharing proxy networks that carried Alibaba, DeepSeek and Xiaomi traffic at the same time.

It also alleges the hosted Kimi and DeepSeek products silently forwarded some user prompts to Claude and kept the transcripts.

Company-level attribution, identity verification and reasoning summarisation are the enforcement tools named in the report.

TIMELINE								
	Feb 23–24 Anthropic names three labs — DeepSeek, Moonshot, MiniMax	**Jun 9** Fable 5 ships	**Jul 1** Export order lifted · Fable reachable again	**Jul 16** K3 launches	**Jul 27** Weights public	**Sept 8** NSA/CISA/FBI advisory asserts Fable → K3	**Sept 9** MOFCOM rejects it	**Sept 10** Anthropic: six labs, 23M Moonshot, 151M Alibaba, target named as Opus

Anthropic (Feb 23, 2026; Senate Banking letter Jun 10, via CNBC Jul 27; "Detecting and countering misuse of AI: September 2026," Sept 10, GTG-16002, 16005, 16001, 16006) · CISA AA26-251A / NSA release (Sept 8) · @mkratsios47, @SecScottBessent (Jul 22) · TechCrunch (Jul 22) · Ball, Lambert, Xian posts · MOFCOM (Sept 9) · Bloomberg · SCMP · CNBC (Sept 11).

Washington now asserts the Fable step that Anthropic's own logs don't show.

The Fable-to-K3 link rests on a three-agency assertion. The vendor's telemetry, published two days later, names Opus. Neither has published forensics.

ASSERTED · THE FABLE-TO-K3 STEP, SEPT 8

The joint NSA/CISA/FBI advisory states that Moonshot "extracted significant Claude Fable 5 data to train its Kimi-K3 model." It is the first time a US agency has put the Fable claim in writing, and it cites no logs.

Anthropic's report two days later documents the harvesting in detail and, where it names a target, names Opus. The one lab it describes attempting Fable, Zhipu, abandoned the attempt when safeguards degraded the attack.

The government's document and the vendor's document have not been reconciled by either party.

THE PUSHBACK

Dean Ball, head of strategic futures at OpenAI: K3's performance "can't be explained away by distillation."

Nathan Lambert: distillation's impact is overstated and shrinking as training pipelines evolve.

Moonshot employee Randy Xian: "We trained a brand new frontier model in JUST 15 DAYS. Guinness World Record stuff."

China's Commerce Ministry, Sept 9: the advisory is groundless and distillation is a common industry practice. Moonshot has not responded to either document.

TechCrunch's expert survey reached the same skepticism. On the Fable question specifically, Anthropic's Opus finding sits closer to Ball's position than to the agencies'.

THE ASSESSMENT

Eighteen days of Fable reachability before K3 launched is short to train a 2.8T base.

The documented late-stage input, per Anthropic, is Claude Opus reasoning harvested through July — which is where K3's strongest categories live.

The Fable-specific claim now has a government signature and still has no published evidence behind it.

DOCUMENTED

Anthropic's February disclosure · Anthropic's September disclosure (server-side telemetry and attribution; target named as Opus)

ASSERTED

NSA/CISA/FBI AA26-251A, Sept 8 (Fable 5 → K3, no logs cited) · Kratsios, Jul 22 · Bessent, Jul 21

SUGGESTIVE

Greenblatt's Claude-identifying responses; three signals converge (slide 80)

ABSENT

Weights-side forensics. Weights public since Jul 27 and, as of Sept 14, no party has published an audit of K3's training data. Anthropic's report documents extraction and says nothing about how Moonshot used it; the advisory says how it was used and shows nothing

Anthropic (Feb 23, 2026; Senate Banking letter Jun 10, via CNBC Jul 27; "Detecting and countering misuse of AI: September 2026," Sept 10, GTG-16002, 16005, 16001, 16006) · CISA AA26-251A / NSA release (Sept 8) · @mkratsios47, @SecScottBessent (Jul 22) · TechCrunch (Jul 22) · Ball, Lambert, Xian posts · MOFCOM (Sept 9) · Bloomberg · SCMP · CNBC (Sept 11).

SECTION 06 · 80

Three independent signals link K3 to Claude.

Three separate measurements show unusual similarity between K3 and Claude models, though none of them establishes distillation.

01

Self-identification

Ryan Greenblatt, Redwood Research

FINDING

K3 disproportionately identifies itself as Claude, statistically significant, "difficult to explain as random noise." Asked what it is, K3 answers "Claude 4.5", never Fable, never Mythos.

THE LIMIT

It names a model that predates the case, and self-identification is a known artifact of training on text containing Claude outputs. Anthropic's September report supplies a direct mechanism for that artifact — reasoning transcripts harvested from Claude Opus — and the model K3 names fits it: a system post-trained on Opus traces has no reason to call itself Fable.

02

Task-outcome correlation

Together AI · DeepSWE · Jul 24

FINDING

Per-task pass/fail correlation, 0 to 1

Per-task pass/fail correlation of 0.72 between K3 and Fable 5, the highest cross-vendor similarity in the benchmark. The top four cross-vendor similarities are all K3-vs-Anthropic pairs; their union covers 105 of 113 tasks (K3 alone: 101); 65% of shared failures are near-misses for both.

THE LIMIT

Together AI frames this as capability convergence and a cost comparison, and makes no distillation claim.

03

Tool-use behaviour

arXiv · "When Agents Look the Same"

FINDING

Kimi-K2 reaches 82.7% agentic similarity with Claude Sonnet 4.5, the highest among non-Anthropic models, exceeding some within-Anthropic pairs.

THE LIMIT

Measured on K2 and Sonnet 4.5, and it predates the K3–Fable case; prior-pattern context only.

Greenblatt / Redwood Research post · Together AI, "Kimi K3 vs Claude Fable 5 on DeepSWE: Cost and Coding" (together.ai/blog; X thread, July 2026) · arXiv "When Agents Look the Same" · Anthropic disclosure (Feb 23, 2026).

SECTION 06 · 81

Two open frontiers, two release cultures.

Within twelve days, two labs released frontier-class weights under materially different terms. Openness of weights, on its own, determined neither the license, the safety posture, nor the provenance of either release.

DIMENSION	**Thinking Machines · Inkling** · US · Jul 15	**Moonshot · Kimi K3** · CN · Jul 27
License	Apache 2.0, unambiguous, known at announcement	Kimi K3 License (custom): MaaS above $20M/yr needs a separate agreement; above 100M MAU or $20M/mo must display "Kimi K3" in the UI. K2 was modified-MIT; that did not settle K3.
Weights-to-API order	Weights first, nothing to gate	API Jul 16, weights Jul 27
Serving footprint UPDATED AUG '26	≥600 GB VRAM (NVFP4), a small cluster	~1.56 TB, 96 shards, native MXFP4, 64+ accelerators; open, but not runnable by most who hold it
Upstream contribution	Standard architecture; the existing serving stack already runs it	KDA broke runtime compatibility; Moonshot fixed it by contributing prefix caching to vLLM, and gained influence over the standard
Evidence at launch	Model card with disclosed limitations	Self-reported benchmarks on its own harness; deployed-system comparisons in footnotes
Capacity posture	No hosted dependency to strain	New subscriptions paused Jul 20 as demand neared capacity
Smaller sibling UPDATED AUG '26	Inkling-Small (276B/12B), weights shipped — Apache 2.0, NVFP4 floor 180 GB, single B300	None announced

PROVENANCE FOOTNOTE	The cultures already interbreed: Inkling was trained with help from Moonshot's Kimi 2.5, and Meta's Muse Glimmer is distilled from a closed US model. A third column is pending — DeepSeek's V4-Pro-0813 shipped API-first with same-day weights (Aug 13), a quiet-drop culture. Muse Spark 1.2 would add a fourth if it ships open.

TML launch post + Inkling-Small model card · Kimi K3 model card + LICENSE (Hugging Face) + deployment guidance · Hugging Face artifact listing (96 shards, ~1.42 TiB / ~1.56 TB) · vLLM blog (KDA prefix caching) · TechCrunch (Jul 22) · Anthropic (Sept 10, 2026), GTG-16002.

Anyone could run an open model. This year, anyone can improve one.

Open models went from something you could rent, to something you could run, to something you can now improve.

RENT · A CLOSED API

Capability on tap, with unit economics the vendor sets.

19 days

the newest frontier model stayed dark to everyone renting it, June 12 to July 1

Uber's 2026 AI budget lasted four months. EY puts a complex agentic interaction at roughly 30× the cost of a 2023 chatbot query.

You hold: a contract

RUN · SELF-HOSTED WEIGHTS

A variable bill turns into a fixed cost on hardware you run.

−56%

AT&T's cost on routed coding tasks after moving 40% of employee queries to open models, at a 2% quality cost

A 40-point model now serves from one GPU, and no order can recall a copy already on a machine you hold.

You hold: the weights

IMPROVE · POST-TRAINING ON AN OPEN BASE

Frontier-scale capability gains on a post-training budget.

+10 / +8

index points DeepSeek shipped in two post-training passes two weeks apart, on an unchanged model

$450K

Thomson Reuters' final training run for Thomson, a Qwen-based legal model now in production in CoCounsel. Two-year programme cost: $40M.

One training API drew four implementations in ten months, the early shape of the layer starting to standardize.

You hold: the weights, the adapters, and the reward functions

Improving a frontier model used to need a pretraining budget. Now it needs an open base and a post-training one.

The Information (Aug 20, AT&T; Apr 2026, Uber) · The Verge (May 14) · Thomson Reuters (Aug 24 / Sept 1) · DeepSeek API Docs (Jul 31, Aug 13) · Tinker, Microsoft Foundry, SkyRL tx, OpenTinker · EY.

Mozilla

Opportunities.

WHERE TO ACT WHILE THE LAYER IS STILL OPEN

SECTION 07 · 84

Six bets on keeping the layer open.

Six places to act while the boundaries around the model are still being drawn.

1	**Make the harness swappable**	the wire and the state, standardized, inside a foundation
2	**Solve portable permission**	the layer where an agent acts rather than reads, across any boundary
3	**Own the state**	record, residency and cache, in portable formats behind your firewall
4	**Break the meter**	route on cost, second-source the gateway, before the prices move
5	**Own the loop**	task distributions, reward functions and adapters stay yours
6	**Make the open default plural**	at every layer, so the commons never has a single supplier

None of these bets requires beating the frontier. Basis: evidence in sections 01–06 of this report.

Bet 1: Make the harness swappable.

The labs are pulling the harness in-house, and the move is still under way. Nobody has fixed what a harness says to a model or what it hands back to you. Standardize both and a harness becomes something you can swap without losing the experience you built on it.

THE DATA

21.8 → ~3

points of harness advantage, compressed in eight weeks once the labs pulled the harness in-house. On every model where both appear, the lab's own harness wins.

Codex is Apache 2.0 and runs any open model — and the usage loop still runs at OpenAI.

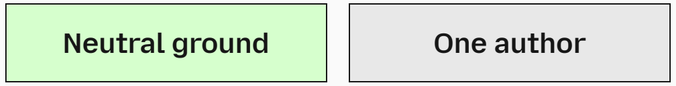
2.0

interop score for the agent layer, the lowest cell on the stack map. Every wire format in production has a single corporate author.

THE CLOCK

Whoever puts the wire format on neutral ground first decides the standards fight of 2027.

Codex dropped the older chat format this year, and every model without a Responses endpoint broke inside it until proxies appeared. MCP and AGENTS.md went to a foundation; the wire layer skipped that step.



STANDING STILL

Every agent speaks one lab's schema, and that lab has already changed it unilaterally.

Open weights match the frontier and ship as commodity inputs to a closed product that keeps the user, the defaults, the upgrade path and the data.

Opencodex, codex-router and CLIProxyAPI are the market paying for a standard nobody has written. Terminal-Bench 2.0 / 2.1 · vals.ai · OpenAI Codex docs · Mozilla stack map, July 2026.

Bet 2: Solve portable permission.

OAuth 2.1 handles who the agent is. Nothing yet handles what it may write, or where the data it touches may go.

THE DATA

71,000+

public MCP servers and 110M+ monthly SDK downloads — against ~21% of companies reporting mature agent governance.

OAuth 2.1 and PKCE are mandatory in MCP, and A2A ships signed Agent Cards. Neither answers what an agent is allowed to do.

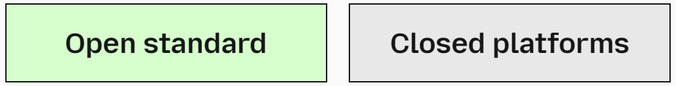
4

critical incidents in closed systems, all the same shape: retrieval checked, output not. A Qwen plugin inside Claude Code sends screenshots and PDFs to Alibaba's cloud with no check at all.

THE CLOCK

An open standard sets the rules in this window, or the closed platforms set them for everyone.

The meta-harness layer that enforces stateful write policy is being built right now — 50K+ stars across multiple repos in August alone.



STANDING STILL

If closed platforms define permission and identity first, every open deployment inherits rules written to make renting the safe choice.

The most valuable position in the harness gets enclosed.

MCP spec revision (Jul 28, 2026) · Glama / ChatForest registry census · AAIF · Okta, "The Authorization Gap" · Qwen-MM-Plugins repo.

Bet 3: Own the state.

Memory, session state and cache all compound while the model beneath them commoditizes. Hold all three in portable, append-only formats behind your own firewall.

THE DATA

10–30×

the gap between cache hits and misses, 120× at the summer peak. A workload built on one provider's cache pays a full re-prefill to leave.

Open memory tooling runs at scale: Mem0, Letta, Zep, LangMem. Session state and AGENTS.md accretion live where the harness lives.

0

portability standards for cached context exist, or have been proposed.

THE CLOCK

The formats have to be set before the model commoditizes around them.

AGENTS.md already went to the AAIF, so state can follow it there. The practitioner default forming now runs the other way — cache-first, model-homogeneous design — and every quarter on it deepens the residency.

Portable formats

Cache-first residency

STANDING STILL

Years of accumulated context cannot be reconstructed once access ends.

On exit, the record, the session and the cache stay with the provider. That is the part of the deployment that was gaining value.

A provider can reprice the record, deprecate the model it lives in, or lose the legal right to serve you. GitHub · DeepSeek, Anthropic and Moonshot pricing pages · Manus · AAIF.

Bet 4: Break the meter.

Uber's fares rose ~92% after riders reorganized their lives around the subsidized price. The meter has since moved from the token to the cache and the gateway. Second-source all three now, while it is cheap and boring.

THE DATA

56%

cut in task cost at AT&T from routing employee queries to the cheapest adequate model, at a 2% quality cost.

2.3–4.6×

DeepSeek's Aug 16 output price rise — the first from an open lab. Uber burned its 2026 AI budget in four months; Microsoft's division budget went in six weeks.

$24.8B

a year: the price asymmetry between open and closed. Stripe bought the largest model gateway for ~$7.5B.

THE CLOCK

The cheapest the ride will ever be is the day before you cannot leave.

Introductory pricing ends around 2027–28, once the providers have gone public and the discounts are spent.

Second-source now

Repriced later

STANDING STILL

A metered bill the vendor sets becomes the ceiling on your roadmap.

When one company owns the router and the payment rail, it knows exactly what exit would cost you and has no reason to stop raising it.

Run an open router you control. Route on cost — the best models fail the same tasks at .72, so routing for capability buys little. Keep a second model warm. Self-host where load is predictable. The Information · The Verge · DeepSeek pricing page · NYT (OpenRouter) · Linux Foundation · arXiv 2606.27288.

Bet 5: Own the loop.

Teams renting closed post-training hand the vendor their task distributions and reward functions — an operational description of their business.

THE DATA

4
implementations of one training API in ten months.

5
funded managed post-training platforms, two freshly capitalized in July: Prime Intellect $130M, Together $800M.

+10 / +8
index points on two post-training passes two weeks apart. Thomson Reuters put a $450K training run on a Qwen base into production. Portable adapter formats remain unsolved.

THE CLOCK

The API surface standardizes within the year — inside a foundation, or inside one company.

The MCP donation is precedent for either path: a protocol can go to a foundation, or it can stay with its author. Nothing about the training loop has yet decided which.

Foundation | **One company**

STANDING STILL

Task distributions and reward functions accrue to the vendor.

Tuned models stay unportable.

The one layer where capital no longer decides gets enclosed by whoever hosts it.

A vertical the frontier isn't tuning for is won at post-training; the harness follows the tuned weights. Tinker · Microsoft Foundry · SkyRL tx · OpenTinker · Thomson Reuters (Sept 1) · DeepSeek API docs · Regulation (EU) 2024/1689, Art. 25.

SECTION 07 · 90

Bet 6: Make the open default plural.

A commons with one supplier stops being a commons, regardless of supplier, at the model layer or any layer above it. Fund a plural open ecosystem good enough that governments and enterprises reach for it first.

THE DATA

8 of 10

top models by token volume ship open weights; seven are Chinese-built, and Chinese models take 45%+ of routed tokens.

37

WAICO members, no major Western democracy. The US returned in August with Inkling and Muse Glimmer under Apache 2.0; Muse Spark's weights have not shipped.

0 of 16

notable releases ship the data recipe. Above the model, one lab wrote the caching scheme into vLLM, another owns the wire format, Nvidia owns the hub.

THE CLOCK

Diffusion decides who wins, and diffusion is being funded right now, or it is not.

Public money — EUROPA, Lumen Sovereign, Canada's $890M, India's 38,000 GPUs — holds through the slow middle, or the open default settles into a single origin by absence.

Plural commons | **Single origin**

STANDING STILL

Open weights still hide the training data, alignment choices and refusal patterns behind them.

When one origin supplies the default, those choices propagate into every derivative, and everyone inherits a single builder's blind spots with nothing to check them against.

The imbalance is on the supply side. One bloc funds open releases at scale; the others mostly don't. OpenRouter (Aug 2026) · Hugging Face · vLLM blog · NVIDIA (Sept 3) · AAIF · national strategy documents.

APPENDIX

Report methodology.

This is Volume 1.1, published September 2026, and it supersedes Volume 1 (July 2026) and V.01 (August 2026). Data is current to 1 September 2026 unless a slide states its own date.

Every external source carries its own as-of date. The main ones are the Artificial Analysis Intelligence Index (v4.1.1, 1 September 2026), OpenRouter's leaderboard and market-share panels (pulled 1 September 2026), Epoch AI's Capabilities Index, METR's published raw eval data, Terminal-Bench, OckBench, LMArena, Hugging Face, the Nagle–Yue study for the Linux Foundation, Linux Foundation / AAIF governance data, MIT NANDA, the OECD.AI and Oxford Insights policy trackers, the EU AI Act text and Commission GPAI guidance, and public filings, model cards and licence files. Artificial Analysis figures in this deck are v4.1.1; v4.2 was announced Sept 4, 2026. The Linux Foundation 89% figure comes from a Meta-commissioned study. Leaderboards, prices and licence terms move within days, so read every number here as a point-in-time reading rather than a standing fact.

Where a benchmark was run by the vendor whose model it scores, the slide says so. That labelling convention is applied deck-wide and is not repeated in the notes each time. Derivations for the Failure Diversity Index, cost per solved task, the AI Act training-compute estimate and the METR regression are set out in the appendix, including the simplified-reproduction caveat on METR. "Open" means downloadable weights unless a slide says otherwise. Openness in the fuller sense the OSI definition asks for, such as weights, data recipe and training code, is tracked separately in the front matter.

The open-stack maturity heat map is an original Mozilla assessment. It grades each slice of the stack from 1 to 5 on nine criteria: production readiness, ease of adoption, documentation, community activity, performance versus closed, enterprise readiness, interoperability, sustainability, and standardization. A fixed rubric was applied, with AI assistance for consistency, across 44 subcomponents in nine layers, covering 1,361 projects. The taxonomy was fixed before any scoring began, so the map was never adjusted to fit the results, and the only computed value in the sheet is the per-row average; every other cell is a graded judgment. This assessment is as of July 2026 and was not re-scored for this edition, so the map is one edition behind the capability and pricing data in Section 01. Scores are ordinal and directional, and should not be read as measurements.

The funding-maturity analysis is an original Mozilla assessment mapping capital into open source and open weight AI from January 2022 to 18 June 2026, distinguishing open weight from fully-open releases across seven stack layers and five funding types. It counts only disclosed cash, debt, grants and quantified in-kind commitments, never valuations, ARR or market cap, and each figure is verified against a tiered source hierarchy of primary releases, then major reporting, then aggregators. Each unit is scored 0–3 on seven dimensions: volume, type diversity, source concentration, stage, cash versus in-kind. The census closed on 18 June 2026 and was re-scored in August 2026 to take in the Databricks close, the Aramco–Together round and Moonshot's pre-IPO track. Two known biases run through the data. Disclosed rounds and announced pledges inflate the totals relative to money actually deployed, and visibility into non-Western rounds, Chinese ones in particular, is materially worse than into US and European ones. As with the heat map, the 0–3 scores are directional judgments.

Mozilla is an interested party. Mozilla Foundation and Mozilla Corporation advocate publicly for open source AI, and Mozilla.ai builds open source AI tooling. Mozilla Ventures holds financial interests in companies named in this report; it makes investment decisions independently of this research team, and portfolio companies had no role in selecting or scoring anything in the report. The full portfolio is public at mozilla.vc/portfolio. The developer survey was commissioned and paid for by Mozilla and fielded by SlashData, which controlled sampling, weighting and the authoritative export. Mozilla wrote the questions and interpreted the results; more information is on the survey methodology page. No company covered here paid for inclusion, and no vendor was given the report before publication.

AI was used in the editing of this report. It was also used to apply the heat map rubric consistently across subcomponents, to summarise and cross-check source material, and to draft and revise copy. No figure in this report was produced by an AI system without a human verifying it against the underlying source, and the judgments throughout — the scores, the rankings, and the bets — are Mozilla's own.