

THE STATE OF OPEN SOURCE AI

AN INAUGURAL MOZILLA ASSESSMENT

Mozilla | 

Open Source AI 2026

Volume 1.0

July 2026



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.

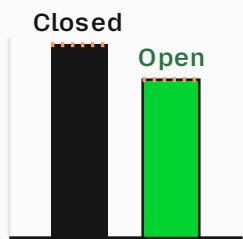
Read what follows as a map: where open AI is winning — some numbers surprised even us — and where it is exposed. A case that hides its weak points is an advertisement.

Contents.

01	The current state	Capability, cost, and adoption — the evidence
02	Who's betting on it	The capital, the companies, the economics
03	Why it's happening everywhere	Optionality, sovereignty, national strategies
04	The harness is the new frontier	The layer above the model, and its open gap
05	Five bets	Where to act while the layer is still open

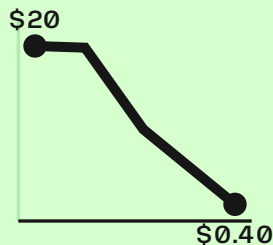
The current state of open source AI.

Open source AI in 2026, in four numbers.



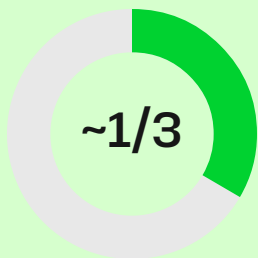
3.3 pts

capability gap to the top closed models — open near parity, an average over a jagged frontier



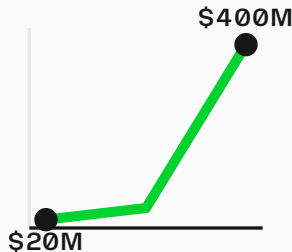
50x

fall in inference cost in 36 months — \$20 to \$0.40 per million tokens



~1/3

of all tokens on OpenRouter served by open weights by late 2025



What changed since you tried it in 2023.

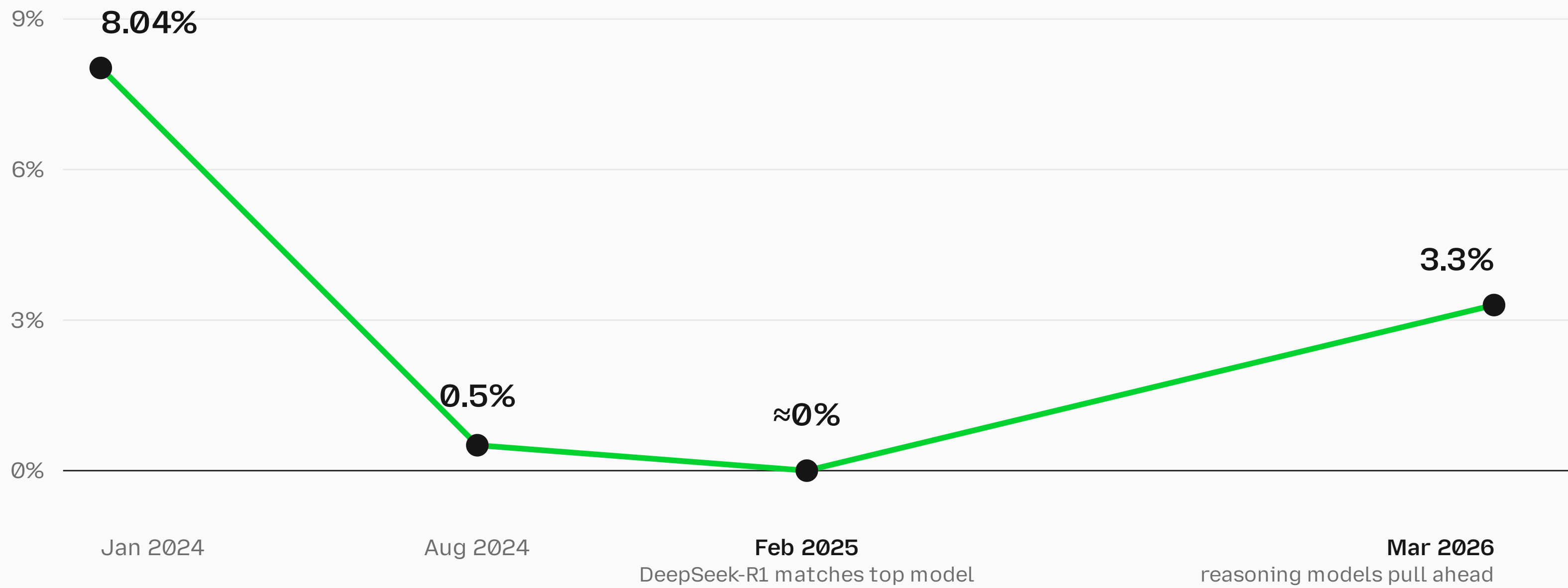
That experience was real. It is also two model generations old.

2023	2024	2025	2026
● Llama 2 ● Quality issues, hallucinations ● Fine-tuning failures ● Reverted to GPT-4	● Llama 3 era ● Major capability improvements ● Better reasoning, better context ● A real alternative	● DeepSeek era ● Near-parity with the frontier ● Lower costs, strong economics ● Competitive	● Production scale ● Enterprise adoption at scale ● Open models power ~33% of tokens ● A commercial ecosystem

The real decision has moved off the model and onto the harness that still makes open hard to deploy. Sources: LMSYS Chatbot Arena · DeepSeek-R1 release (MIT, Jan 2025) · Mozilla / SlashData 2026 survey.

The capability gap: 8.04% → 0.5% → 3.3%.

Closed lead over open on Chatbot Arena, in three phases over 24 months.



Source: LMSYS Chatbot Arena. Six of the top ten Arena slots are now closed models.

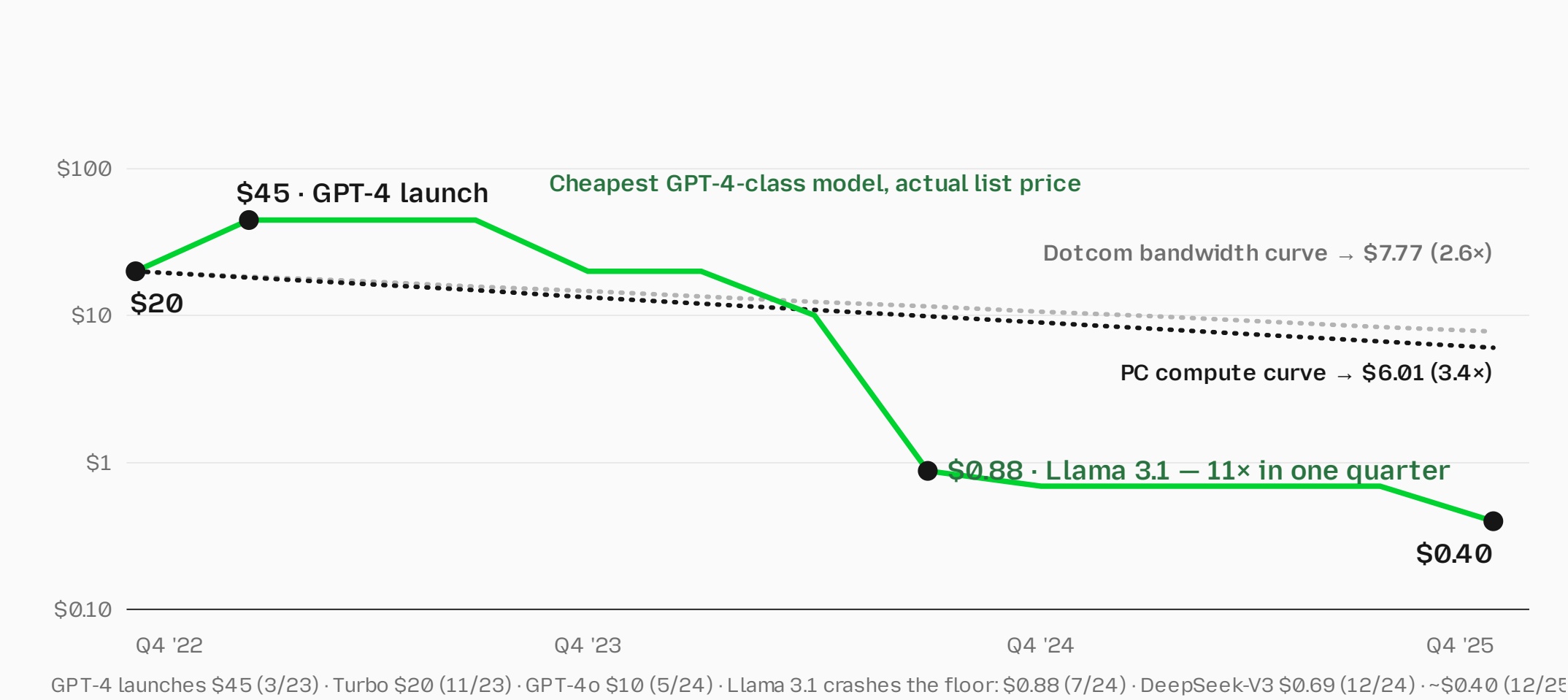
A 3.3% average hides a jagged frontier.

The question is no longer whether open models are good enough. It's what this workload needs.

AT OR NEAR PARITY	CLOSED RETAINS AN EDGE
Coding	Reasoning
Instruction-following	Long-context retrieval
General knowledge	Agentic tasks

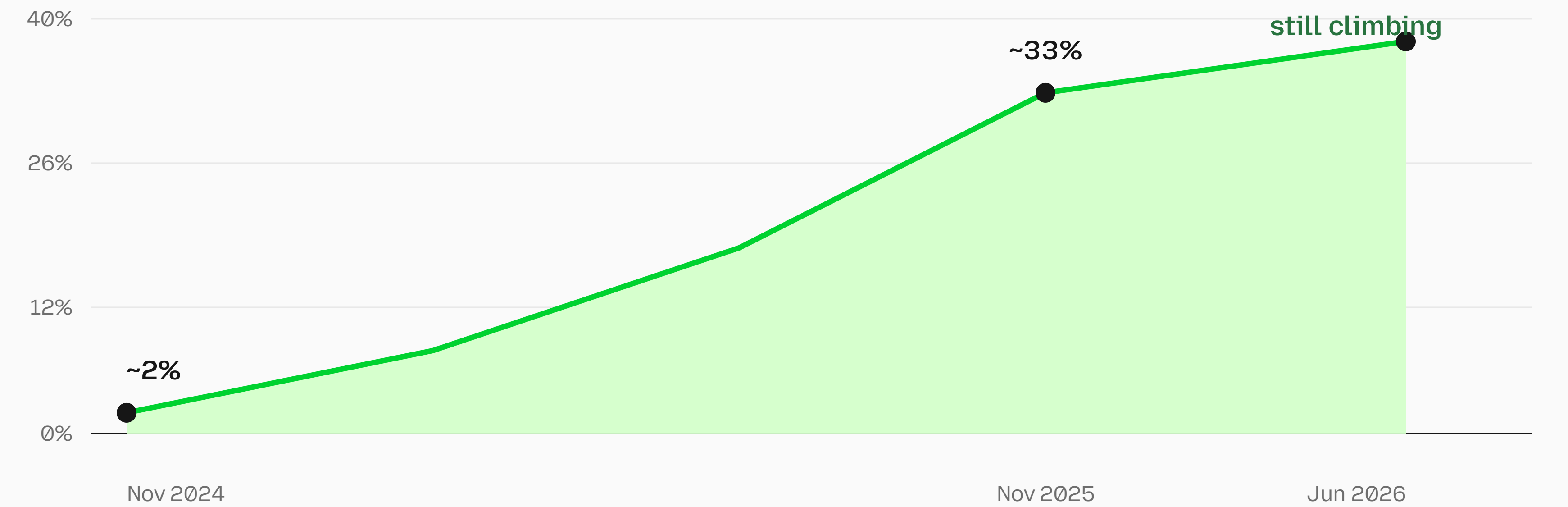
Inference fell 50× in 36 months — \$20 → \$0.40 per 1M tokens.

Cheapest model at GPT-4-class performance, blended API list price (avg in/out), log scale — vs. prior platform-shift cost curves at their historical rates.



Open weights now route a third of all tokens.

Open weight share of all tokens routed on OpenRouter, Nov 2024 – Jun 2026.



Source: OpenRouter 100T-token study (Nov 2024 – Nov 2025); OpenRouter live leaderboard, H1 2026. Indicative curve between measured points.

Open weights dominate token usage.

Monthly token volume on OpenRouter · June 2026. The top five models by volume are all open weight.

1	DeepSeek V4 Flash · DeepSeek		18.4T	↑ 95%
2	MiMo-V2.5 · Xiaomi		14.9T	↑ >999%
3	Hy3 preview · Tencent		14.8T	↑ 67%
4	MiniMax M3 · MiniMax		14.3T	new
5	Owl Alpha · OpenRouter stealth		11T	↑ 200%
6	Claude Opus 4.7 · Anthropic		9.02T	↑ 26%
7	DeepSeek V4 Pro · DeepSeek		8.55T	↑ 112%
8	Claude Sonnet 4.6 · Anthropic		7.33T	↑ 3%
9	Claude Opus 4.8 · Anthropic		6.18T	↑ >999%
10	DeepSeek V3.2 · DeepSeek		4.31T	↑ 5%

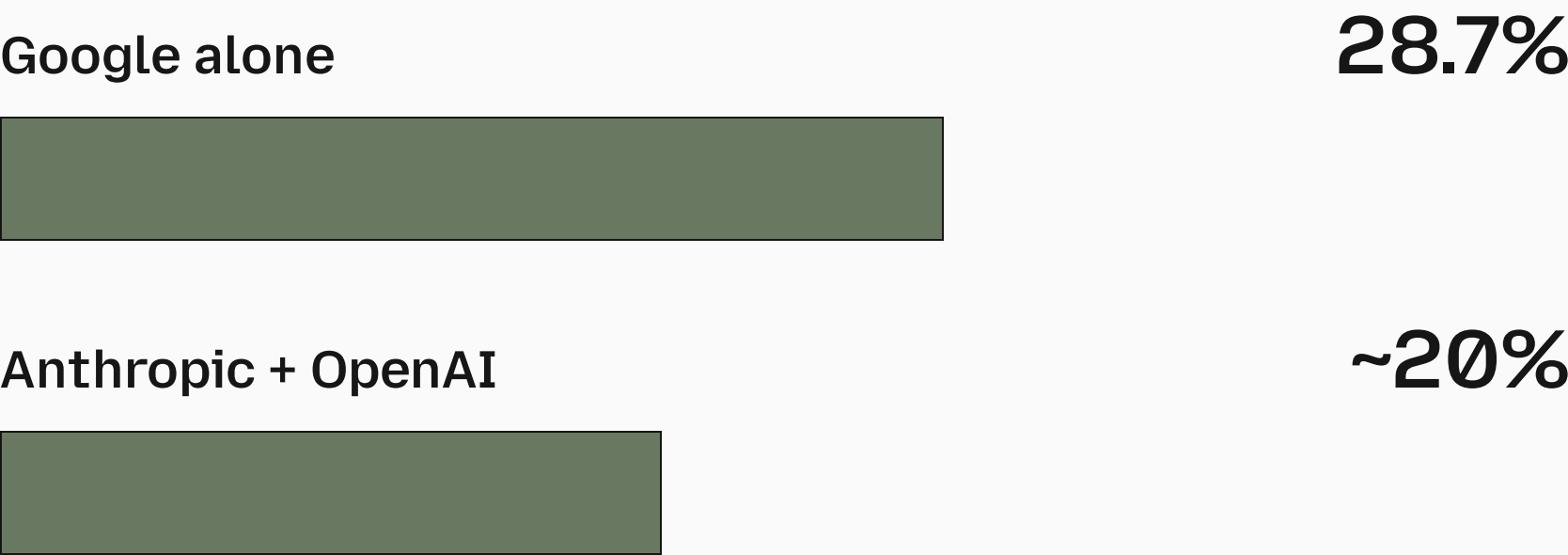
Open wins the tokens. Closed still wins the requests.

MONTHLY TOKENS · TOP NINE MODELS



More than 3.5:1. The five highest-volume models are all open weight, led by DeepSeek V4 Flash (18.4T), Owl Alpha / LongCat-2.0 (11T) and MiMo-V2.5 (14.9T).

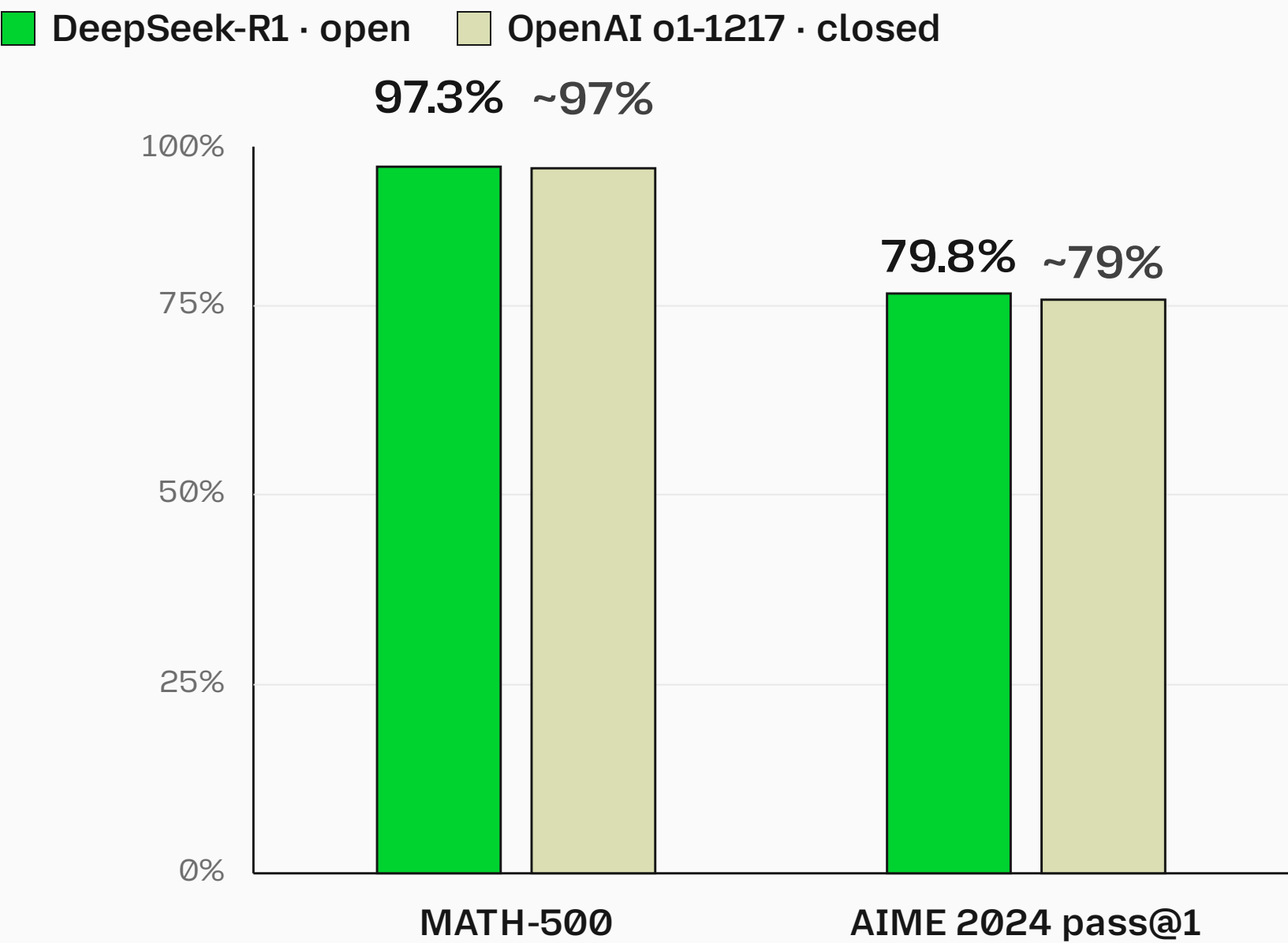
REQUESTS · BY PROVIDER SHARE



By request count, closed US providers still lead. The open lead concentrates in token-heavy coding and agentic workloads — where the spend

Frontier reasoning now ships under MIT license.

DeepSeek-R1 (Jan 2025) against OpenAI o1-1217 — same class of reasoning, a fraction of the price.



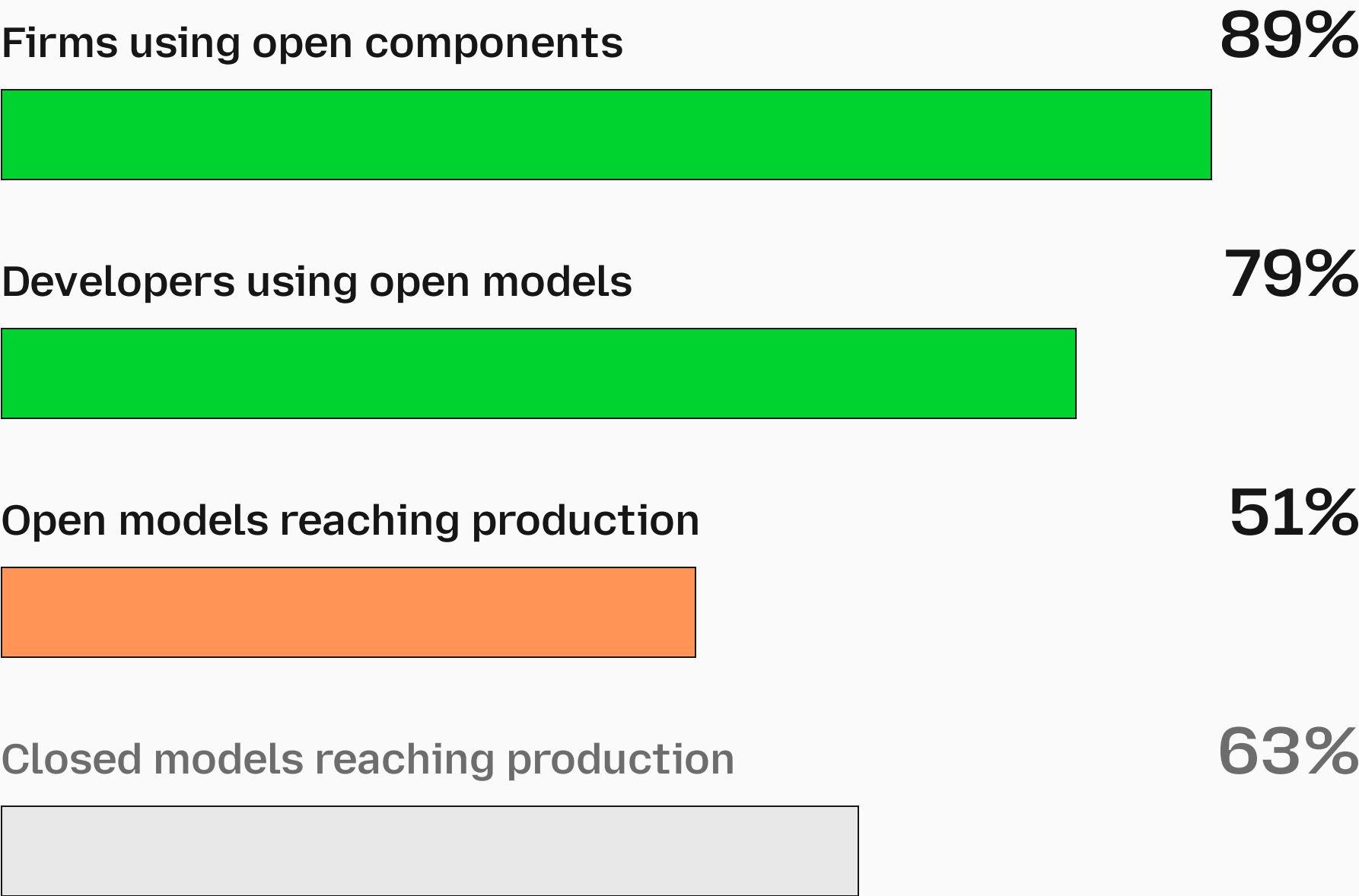
PRICE & LICENSE

\$0.55 · \$2.19 DeepSeek-R1 price per 1M tokens, in · out — MIT license

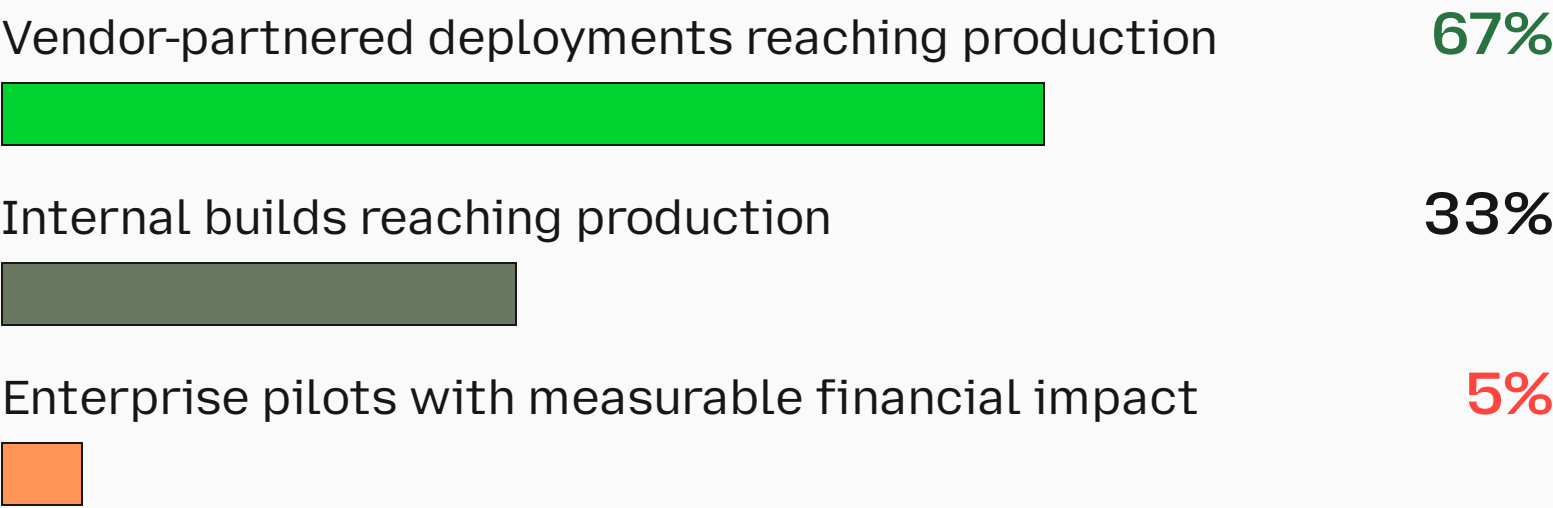
\$15 · \$60 OpenAI o1-1217 price per 1M tokens, in ·

Open ships easy. Open deploys hard.

Adoption funnels from experimentation to production.



WHERE TEAMS STALL



The gap is operational tooling and trust, not model capability.

Sources: Linux Foundation (89%); Mozilla / SlashData 2026 survey, n=1,494 (79% · 51% · 63%); MIT NANDA 2025 (67% · 33%); Stanford 2026 Enterprise AI Playbook (5%).

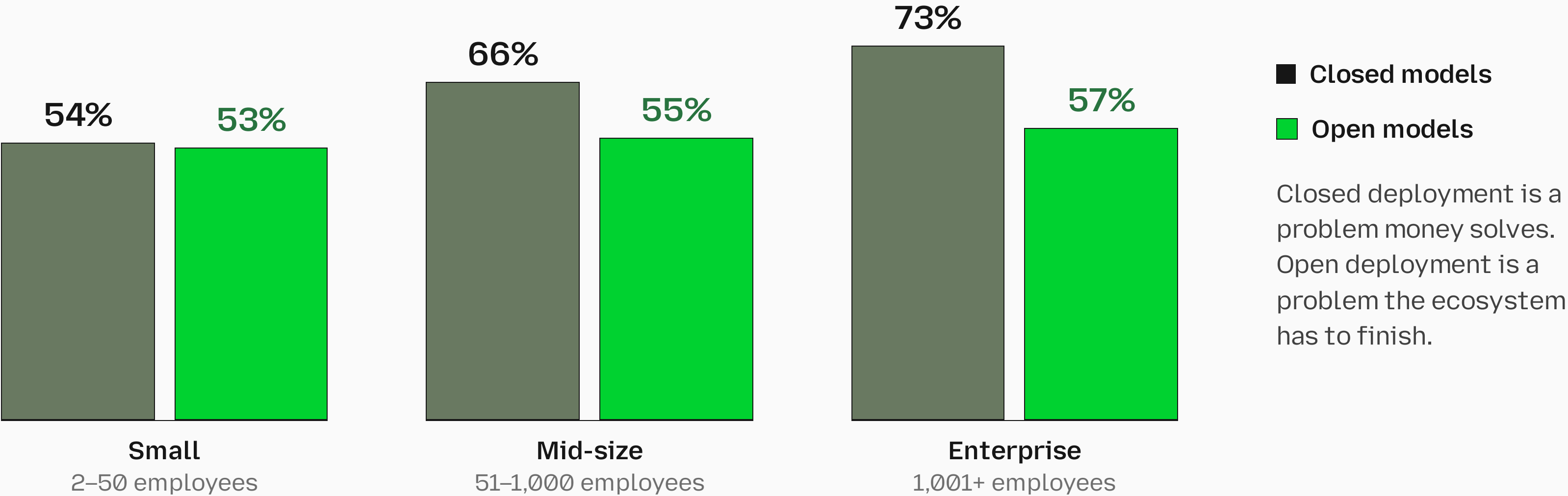
The open stack scores high on capability, low on operations.

Nine stack layers scored across nine criteria (1–5), columns ordered strongest to weakest. The two coldest columns – standardization and enterprise readiness – repeat down every layer.

Layer · avg	Community	Ease of adopt	Prod. ready	Interop	Sustain-ability	Perf. vs closed	Docs	Standard-ization	Enterprise ready	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	3.1	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			

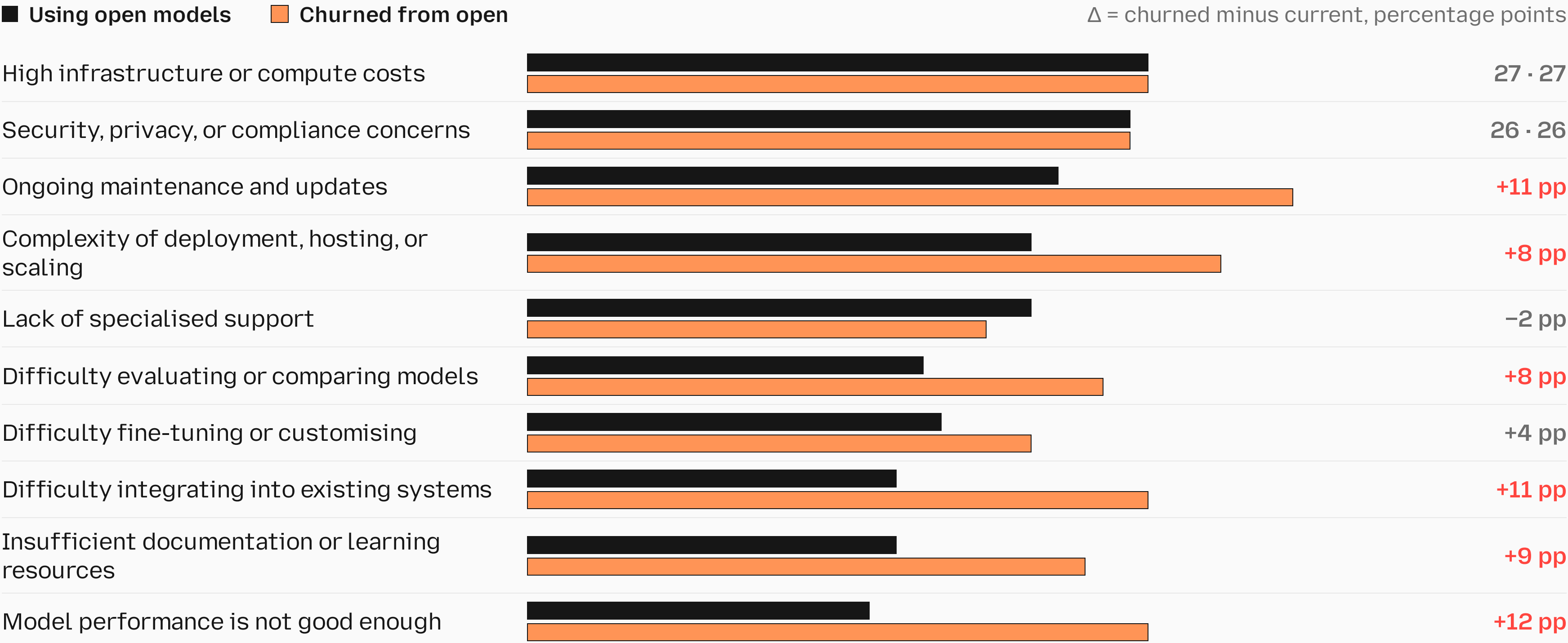
Scale unblocks closed deployment, not open.

Share of adopters reaching production, by organization size. If the gap were about resources, scale would close it — it doesn't.



Source: Mozilla / SlashData 2026 developer survey, professional developers n=954. Production rates among adopters of each model type.

What blocks open – and what makes developers leave.



Who's betting on it.

Open weights is a business model.

Funded companies, paying customers, real revenue.

Databricks \$5.4B revenue run-rate, >65% YoY; weighing a raise at a \$165B+ valuation	Mistral \$400M ARR, up 20× in twelve months; in talks to raise €3B at a €20B valuation	DeepSeek \$220M ARR by mid-2025, mostly API and enterprise; raised \$7.4B at a \$50B+ valuation
---	---	--

FIVE REVENUE MODELS PROVEN AT SCALE

Hosted inference

Enterprise platforms

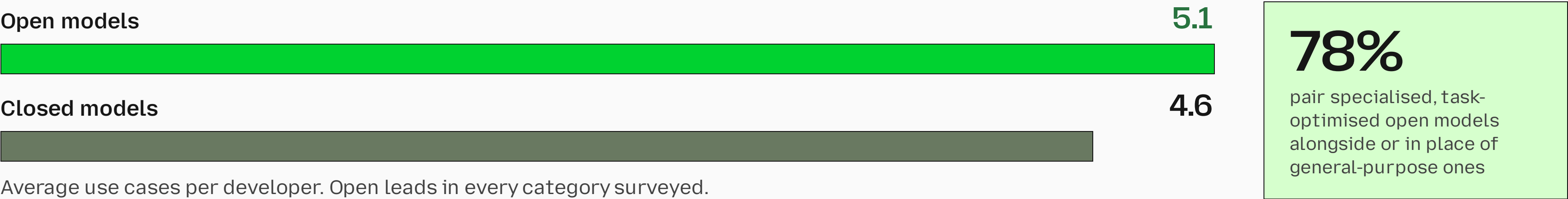
On-prem licensing

Fine-tuning services

Harness tooling

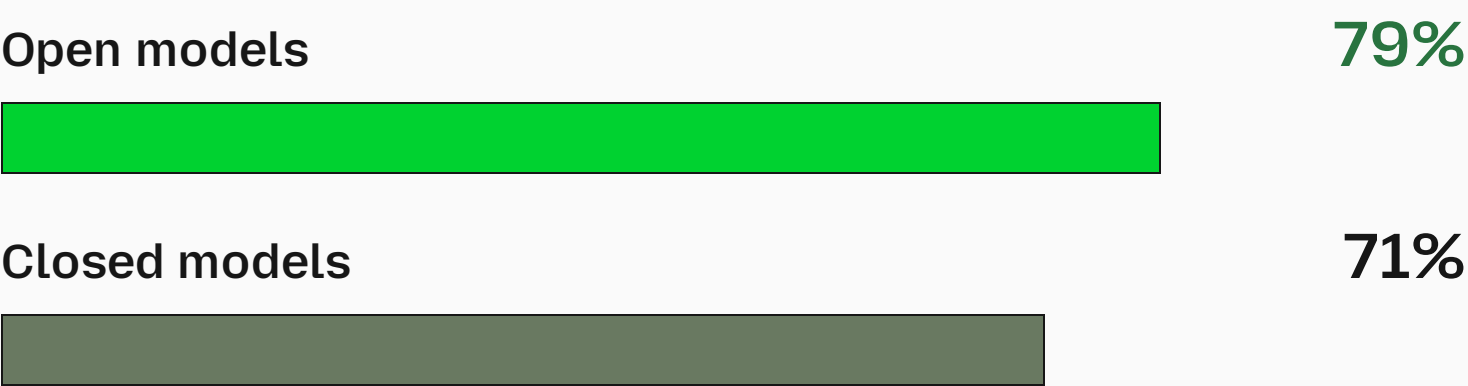
Developers run open across more use cases than closed.

Average use cases per developer, and how open models get combined.



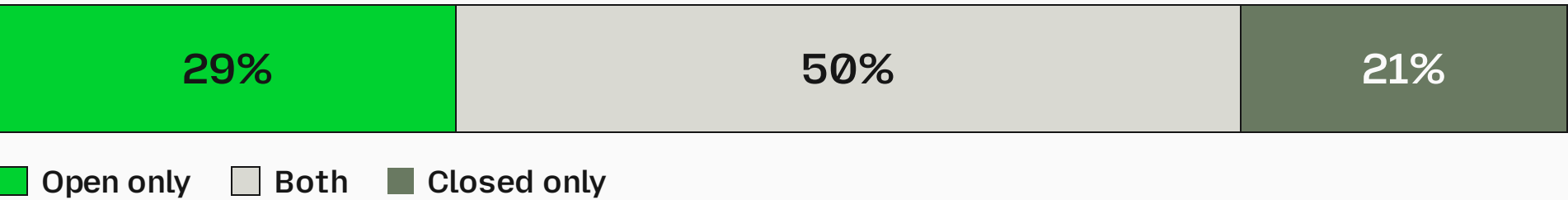
Average use cases per developer. Open leads in every category surveyed.

WHO USES EACH MODEL TYPE



Share of developers adding AI functionality who currently use each model type.

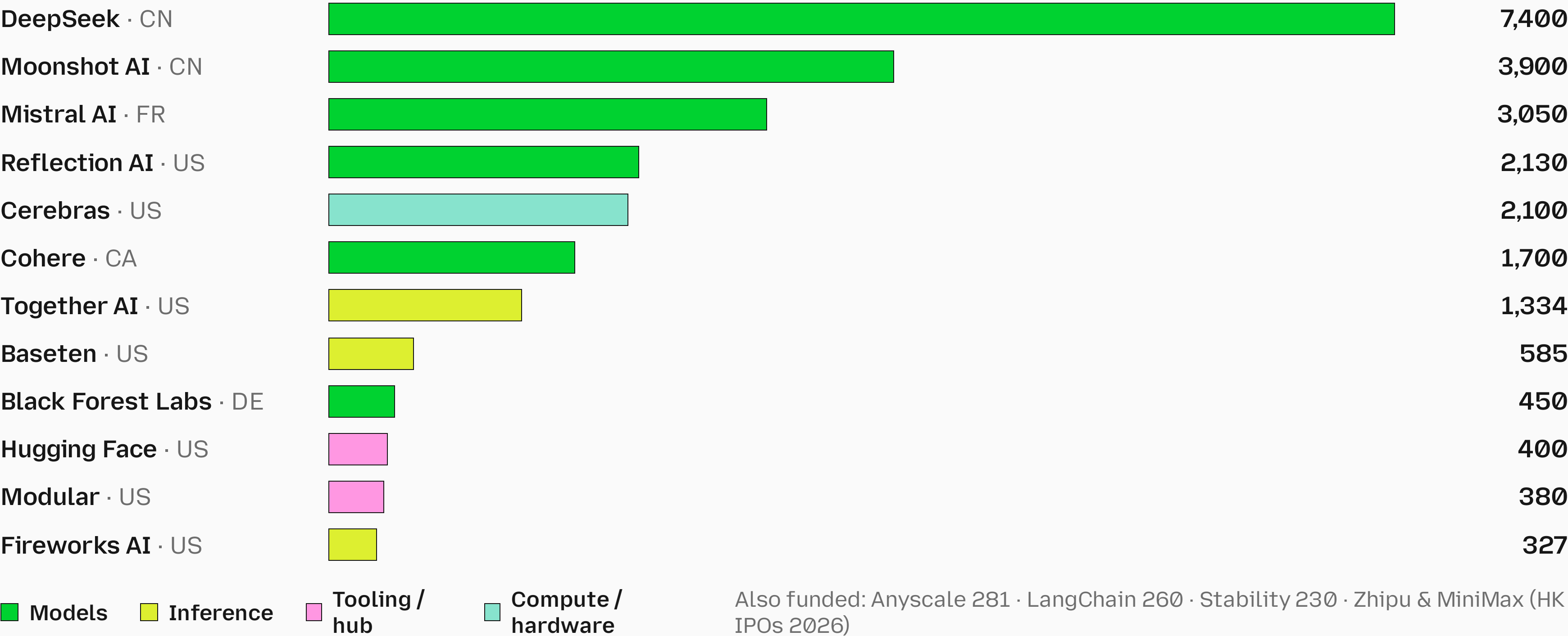
HOW THEY COMBINE



Open leads in adoption – but open and closed aren't substitutes. Most teams run both.

The venture-funded open source ecosystem.

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



Corporates are buying in across the stack.

Six of the largest tech companies, all in.

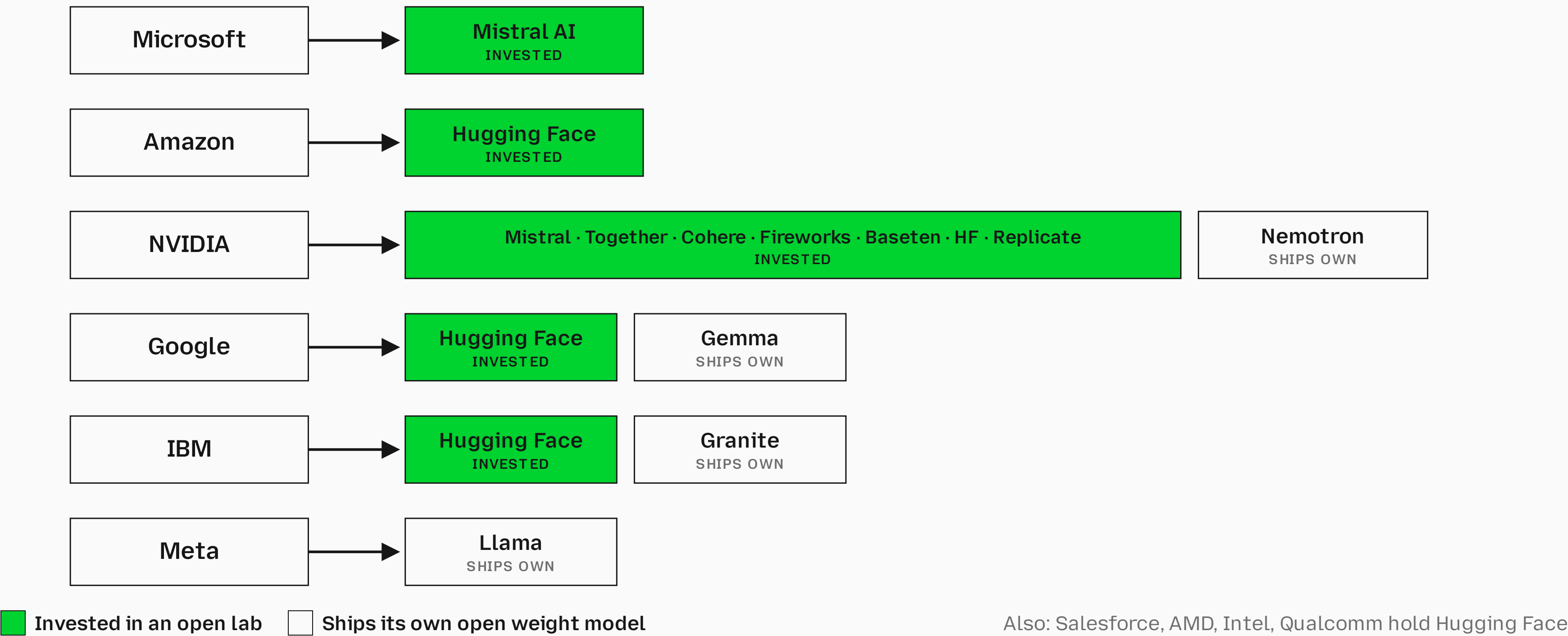


Diagram shows participation, not dollar magnitude. Largest disclosed rounds: ASML → Mistral AI \$1,400M (Sep 2025) · Tencent → DeepSeek \$1,400M (Jun 2026) · CATL → DeepSeek \$700M (Jun 2026) · Schwarz Group → Cohere/Aleph Alpha \$600M (Apr 2026). Source: public filings and reporting, 2023–2026.

Consolidation has started.

Notable acquisitions in and around the open source AI space, 2023–2026.

ACQUIRER	TARGET	WHAT THE TARGET DOES	USD M	DATE
CoreWeave	Weights & Biases	MLOps serving open-model builders	1,700	May 2025
Databricks	MosaicML	Open MPT LLMs + training platform	1,300	Jul 2023
Nvidia	Run:ai	GPU orchestration, to be open sourced	700	Late 2024
AMD	Silo AI	Open-model lab, OpenEuroLLM co-lead	665	2024
Nvidia	Gretel	Synthetic data	320	2025
Nvidia	OctoAI	Inference optimization for open models	165	Sep 2024
Rubrik	Predibase	Open LoRAX fine-tuning	100–500	Jun 2025
Cohere	Aleph Alpha	Sovereign / enterprise open weight LLMs	~\$20B ent.	Apr 2026

Also: ClickHouse–Langfuse, Mintlify–Helicone, Mistral–Koyeb & Emmi AI, Baseten–Parsed, three Hugging Face tuck-ins, three more by Nvidia — mostly undisclosed.

The funding stack has holes.

Funding intensity by stack layer and capital type, 0–3. Darker green means more capital.

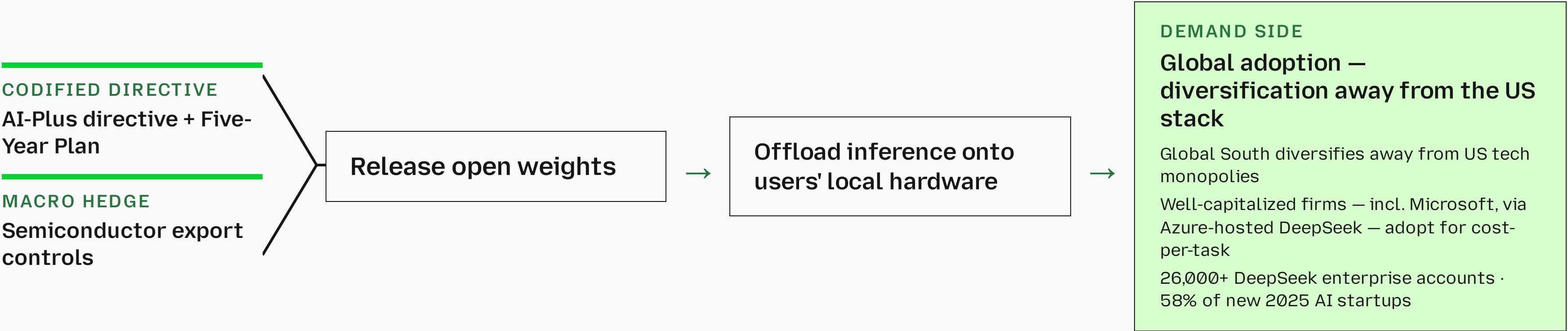
Stack layer	Private VC	Government	Philanthropic	Strategic corporate	Revenue / IPO
Foundation models	3	2	1	3	2
Inference / serving	3	0	0	2	1
Infrastructure / compute	2	3	0	3	1
Data / datasets	0	1	2	1	0
Tooling / MLOps	2	0	1	2	1
Application / vertical	2	0	0	1	1
Safety / eval / governance	0	1	3	1	0

Models, inference and compute are well capitalized. Data, safety and evaluation — the trust layer — run on philanthropy and government alone.
Source: Mozilla analysis of disclosed funding, June 2026.

Why it's happening everywhere.

Open proliferation is Chinese industrial policy.

The migration to Chinese open weights is an intentional policy outcome.



Global demand

Europe is treating open weights as industrial policy.

An escalating sequence, February 2025 – July 2026.

Feb 2025 €109B France commits to AI infrastructure	Aug 2025 AI Act exemptions Qualified open source systems receive targeted exemptions under the AI Act	Jun 2026 EUROPA award Sovereign 400B-parameter open model across all 24 EU languages	Jul 2026 Amália & SPARK Portugal ships a fully open LLM on €5.5M; Germany open sources its public-administration AI platform
--	---	--	--

"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

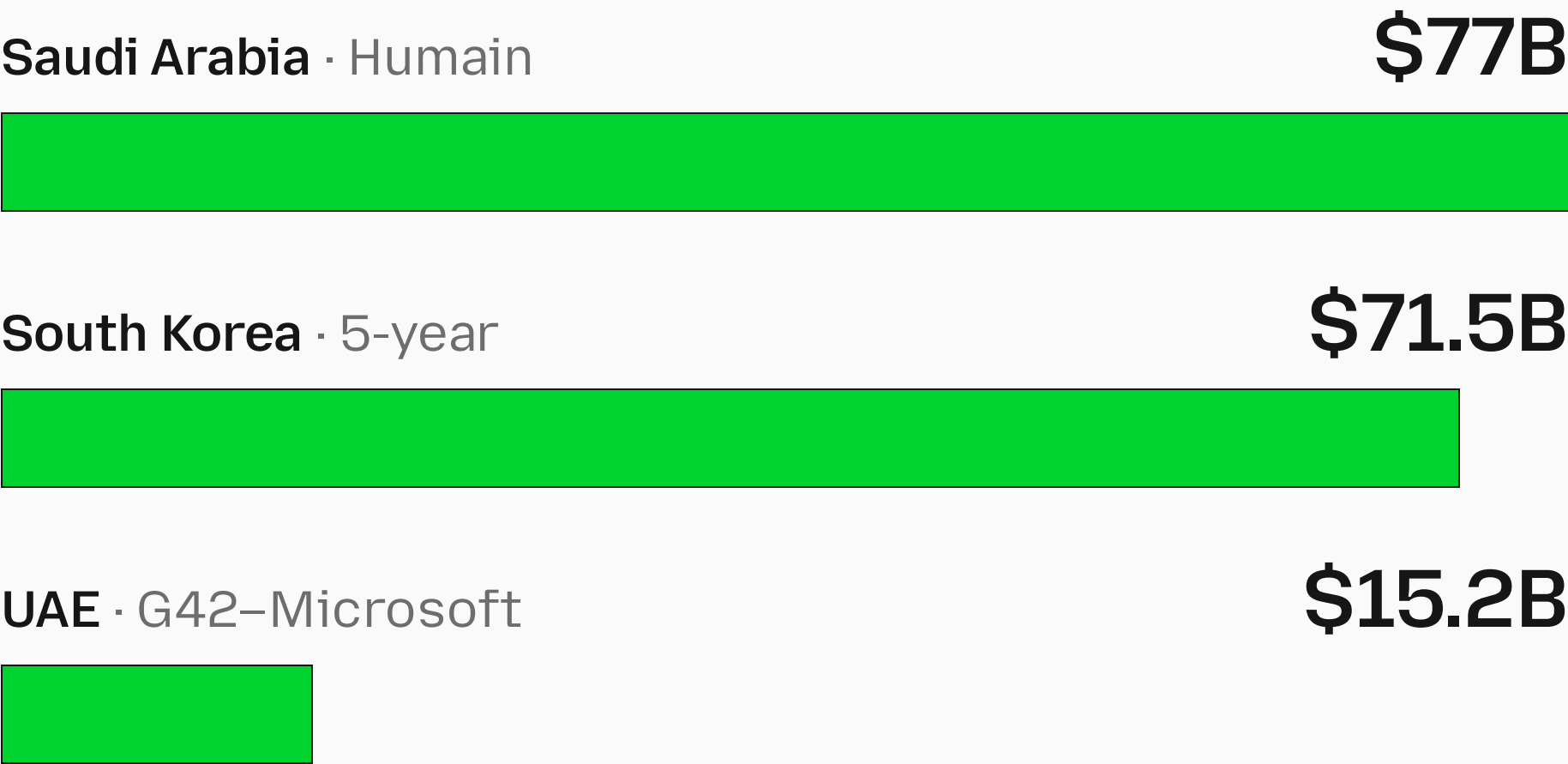
India: 38,000 GPUs and five million new developers.

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
₹10,372 Cr IndiaAI Mission outlay, with 600 planned data labs	13.6% of DeepSeek's monthly active users — second only to China

The bet: cheap public compute plus a multilingual developer base producing indigenous foundation models and harnesses for Global South use cases. Sources: IndiaAI Mission · GitHub Octoverse 2025 · DeepSeek disclosures.

70+ national AI strategies, and counting.

Largest disclosed national AI infrastructure commitments.



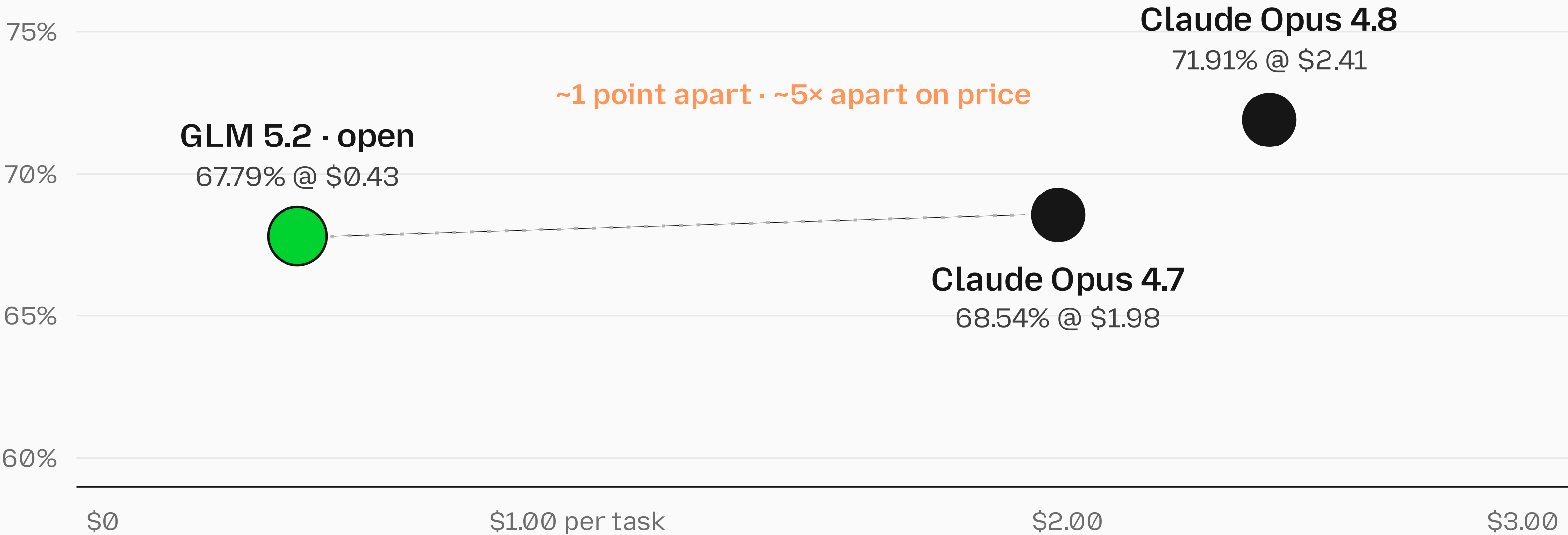
80+ jurisdictions with AI policies — OECD.AI	47 nations restricting foreign processing for critical workloads
12 new national strategies logged in 2024 — Oxford Insights	1.9GW planned Saudi capacity by 2030

The strategic question has shifted from whether to have a national AI play to which layer of the stack a country can own. Sources: OECD.AI · Oxford Insights · public reporting.

The harness is the new frontier.

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

Terminal-Bench 2.1 score vs. cost per task, all models on one neutral scaffold (vals.ai Terminus-2).



Integration also buys the labs a data flywheel — usage exhaust

Opportunities

Five bets on keeping the layer open.

1	Build the open harness	co-designed with open weights, general or vertical
2	Own the memory	portable, append-only, behind your firewall
3	Solve portable permission	the write surface, standardized across boundaries
4	Break the meter	second-source now, while it is cheap and boring
5	Make the open default plural	so the commons never has a single supplier

None of these bets requires beating the frontier. They require owning the layers above it — while those layers are still open. Basis: evidence in sections 01–04 of this report.

Build the open harness.

Co-designed with open weights and tuned to them the way Codex is tuned to GPT-5.5 — general, or purpose-built for a vertical the frontier isn't tuning for.

THE DATA

The harness now outweighs a model generation, and the labs pulled it in-house. Frontier-model rounds absorbed hundreds of billions in 2026; the open-harness category took a rounding error.

THE CLOCK

Vertical integration is forming this year and is not finished. The window closes when the closed stacks weld model and scaffold into a single rented product.

STANDING STILL

Once a lab owns model and harness together, the harness becomes the moat. Open weights can match the frontier and still lose every deployment that matters.

Own the memory.

The one thing in the stack that compounds. Hold it in portable, append-only formats behind your own firewall.

THE DATA

Open memory tooling already runs at scale — Mem0 past 47,000 stars, with Letta, Zep and LangMem. The accreted record gains value with every interaction, and no vendor switch can rebuild it.

THE CLOCK

The architecture has to be set before the model commoditizes around it. Every quarter on a closed endpoint hands the one appreciating asset to a vendor you don't control.

STANDING STILL

Years of accumulated context cannot be reconstructed once access ends. The asset designed to appreciate becomes the one you are least able to protect — discovered the day you try to leave.

