

Open models: The lens in China, adoption trends, and new usage patterns

21 MAY 2026

Nathan
Lambert

Center for Security
and Emerging
Technology (CSET)

A mix of a trip report
and reading the open
model research
ecosystem.



What's on my mind

I think about:

1. How to make genuinely useful open models, to broaden AI diffusion, access, and safety
2. How usage of open models is shifting, who's leading, who's fading, etc.
3. Learning how the Chinese labs make models, why, and how to lower the temperature on US-China competition within the *industry*



What's on my mind

I think about:

1. How to make genuinely useful open models, to broaden AI diffusion, access, and safety
2. How usage of open models is shifting, who's leading, who's fading, etc.
3. Learning how the Chinese labs make models, why, and how to lower the temperature on US-China competition within the *industry*

This talk:

1. My current thinking about the evolution of open models, what people should build
2. Most recent data on adoption of open models
3. My visit to Chinese open model labs (Kimi, Z.ai, Qwen, Ant Ling, etc.)



Part 1: A phase transition for open models



2025: Everyone built a reasoning model.

2026:



2025: Everyone built a reasoning model.

2026: The year agents change how we work.



2025: Everyone built a reasoning model.

2026: The year agents change how we work.

An LLM today is: weights + tools + harness

The model alone is no longer the product.

- **Closed labs (GPT, Claude, Gemini)** — Full control of tools, experience, some generalization as a “platform”
- **Open labs (DeepSeek, Qwen, GLM, Kimi)** — Designed for broad, diverse use cases, and continued training



2025: Everyone built a reasoning model.

2026: The year agents change how we work.

An LLM today is: weights + tools + harness

The model alone is no longer the product.

- **Closed labs (GPT, Claude, Gemini)** — Full control of tools, experience, some generalization as a “platform”
- **Open labs (DeepSeek, Qwen, GLM, Kimi)** — Designed for broad, diverse use cases, and continued training

GPT-OSS is a great example of a successful, tool-oriented open model. Many more such models have emerged with the rise of OpenClaw et al. An exciting area for agents that run just at the cost of electricity (e.g. on a DGX Spark).



2025: Everyone built a reasoning model.

2026: The year agents change how we work.

An LLM today is: weights + tools + harness

- Research today is defined by horizontal, systems research, building agents across the entire stack OR research ablating changes in one of the three pieces
- Open research is heavily limited by lack of frontier-ready training codebases, cloud sandboxes, environments, etc.



2025: Everyone built a reasoning model.

2026: The year agents change how we work.

“Post-training” as a field to craft models is maturing into two sections:

1. General post-training recipes: Giving base models strong reasoning abilities
2. Agentic specialization: How to craft a harness, tools, and fine-tuning recipe around an existing reasoning model for a new domain

This is new and very exciting to me (but not what this talk is about)!



Part 2: measuring the open language model ecosystem



The ATOM Project

The American Truly Open Models (ATOM) Project was a memo and community movement launched in the summer of 2025 to galvanize support for open models built in the U.S. At the time of launch, Kimi K2, Qwen 3 Coder, GLM-4.5 and StepFun Step 3 had shown that the Chinese ecosystem was producing excellent open models and the U.S. labs had little to show for it.



What have we done for ATOM?

American Truly Open Models (ATOM).

Support for The ATOM Project

- 300+ signatories across policy, research, and industry
- Recognized by USCC's *Two Loops* report and discussed with OSTP
- Cited or reused by Stanford HAI, a16z, and others
- 30+ media quotes across 4 countries, including Washington Post coverage and Global Times response



What have we done for ATOM?

American Truly Open Models (ATOM).

Support for The ATOM Project

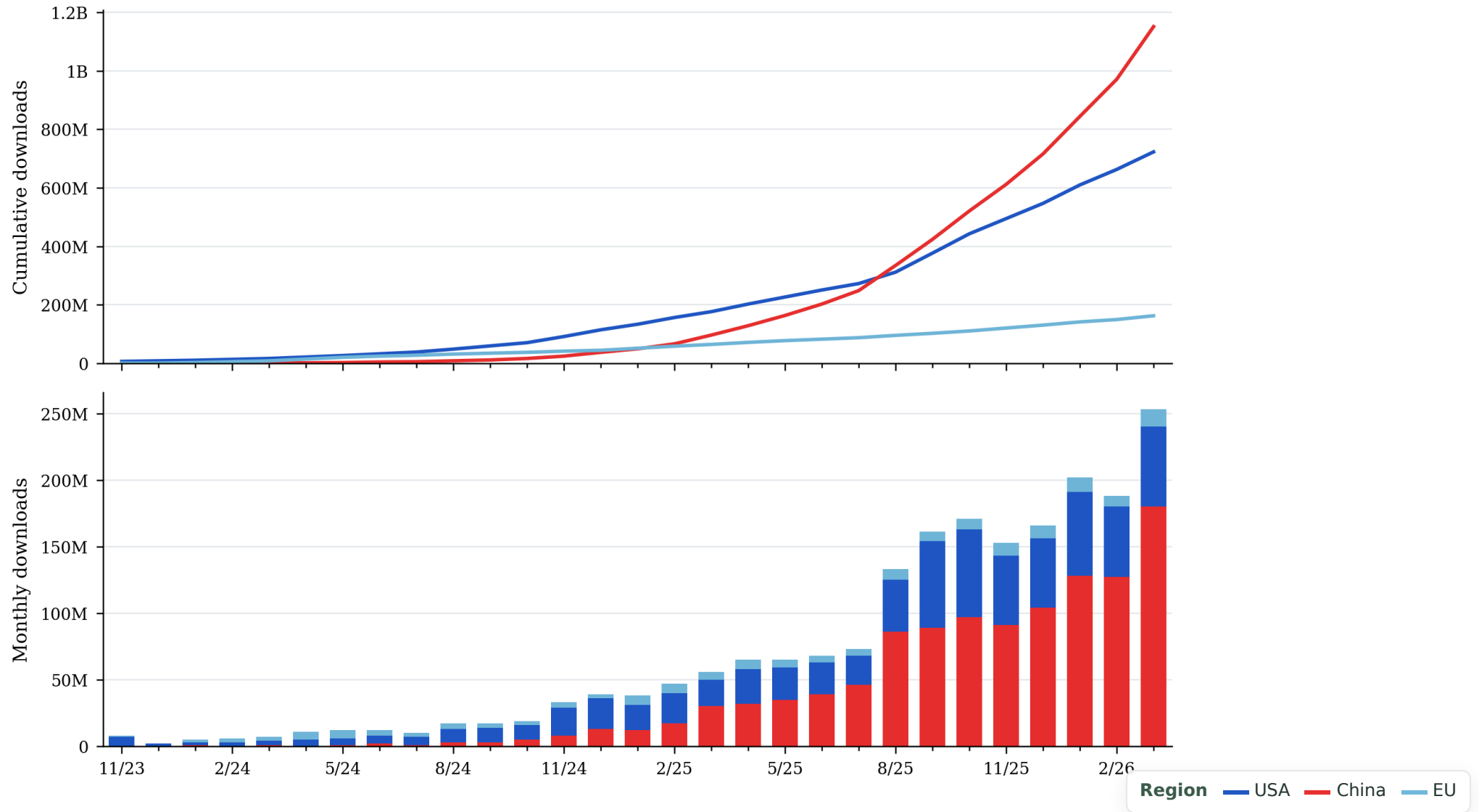
- 300+ signatories across policy, research, and industry
- Recognized by USCC's *Two Loops* report and discussed with OSTP
- Cited or reused by Stanford HAI, a16z, and others
- 30+ media quotes across 4 countries, including Washington Post coverage and Global Times response

Methods research in The ATOM Report

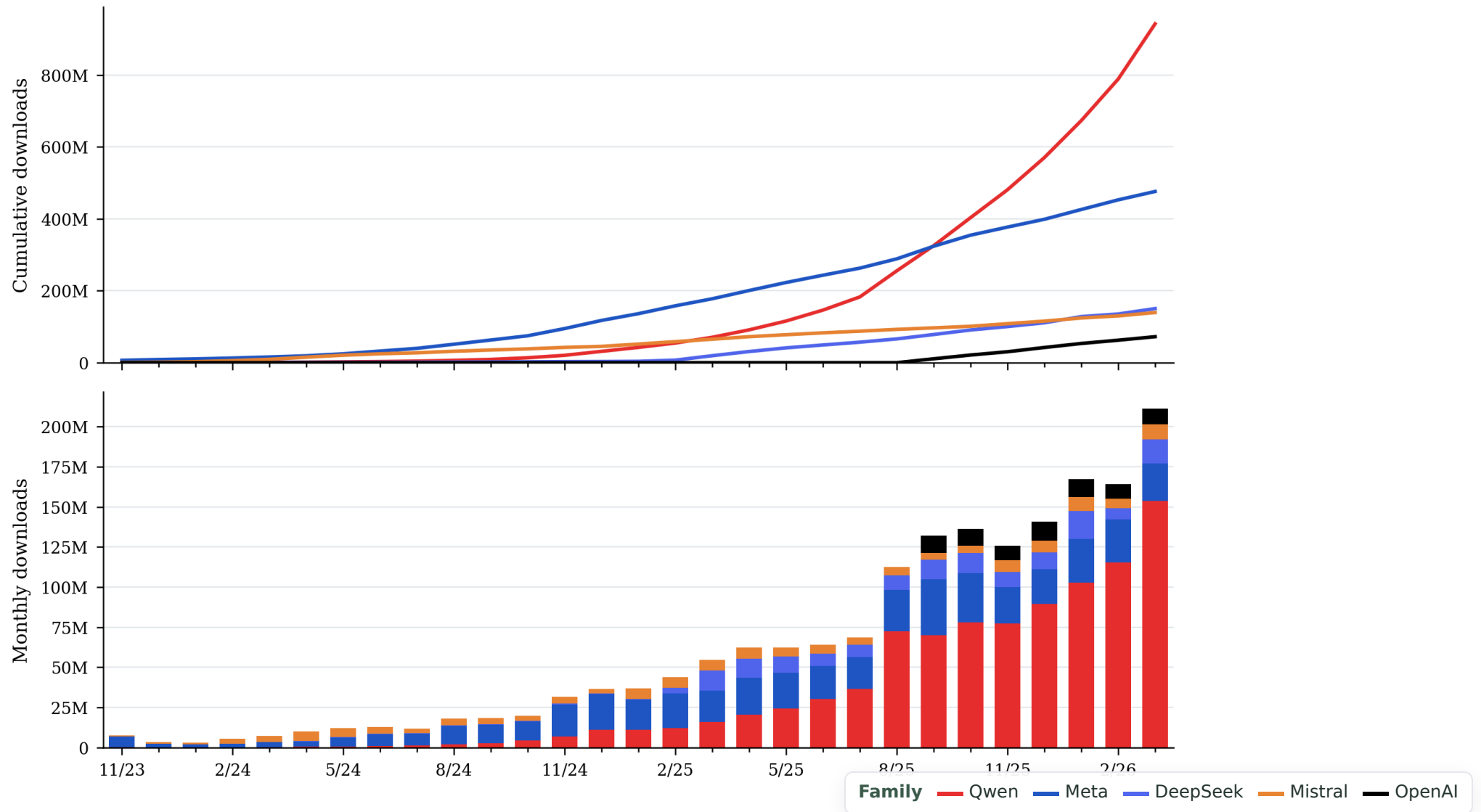
- Maintain a growing list of the 1-2K mainline open-weight models
- Scrape Hugging Face daily to track downloads over time (started July 2025)
- Building an API for researchers to access downloads over time data (**let us know if you want to explore collaborations**)
- Research new methods for measuring the open model ecosystem (e.g. Relative Adoption Metric, RAM)



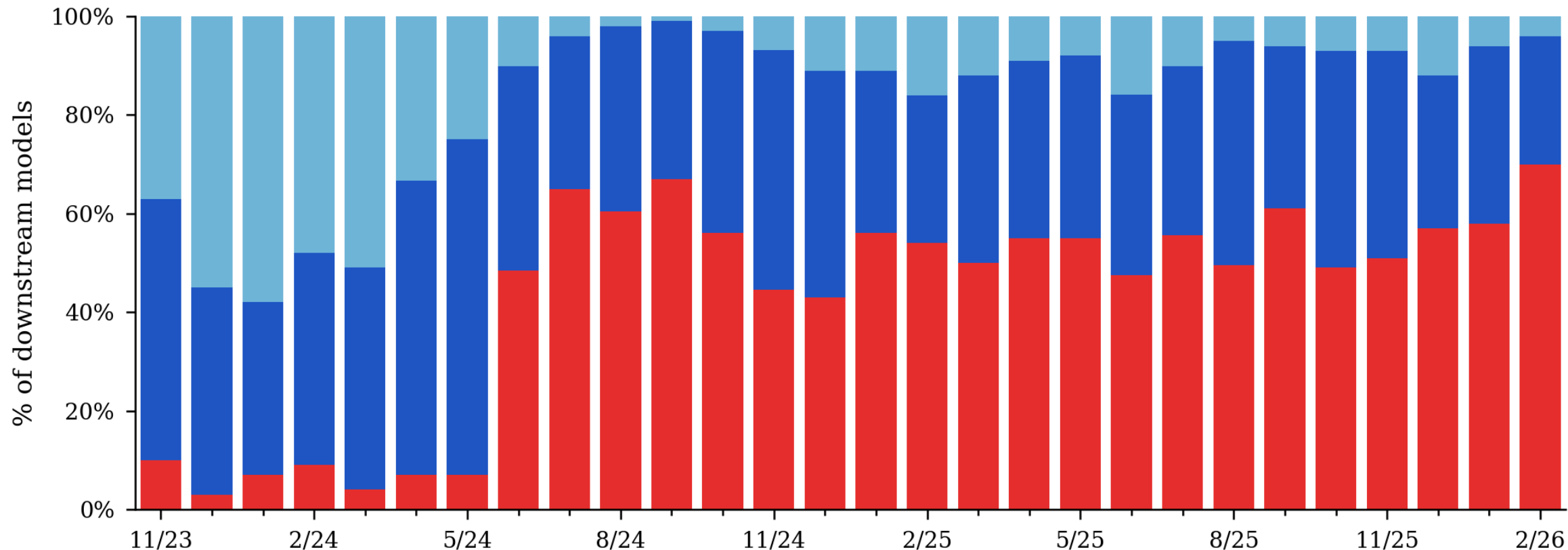
Cumulative downloads by region (China's lead)



Cumulative downloads by organization (Qwen's lead)



China's continued lead in fine-tuning derivative models



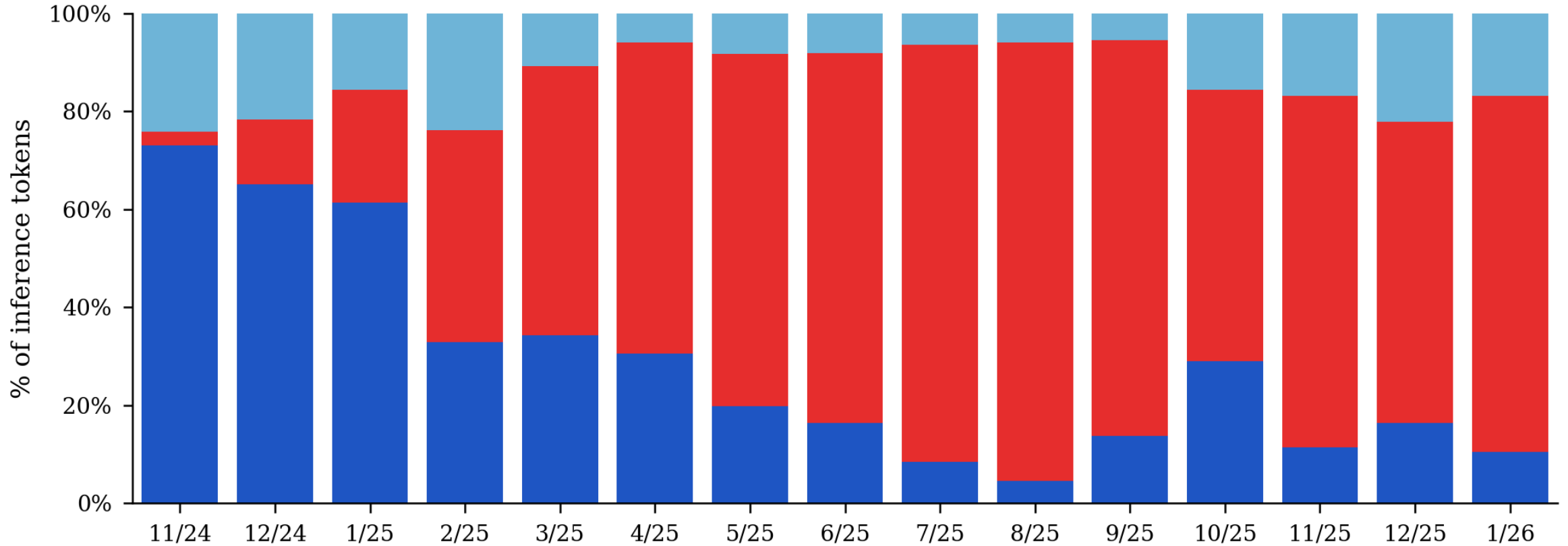
All models with >5 downloads, derived from likes of Qwen, Llama, Mistral, etc. models. China

10% → 70% of derivatives. Europe **58% → 4%**. Derivative share leads downloads as a forward

Region China USA Europe



Inference token share from OpenRouter

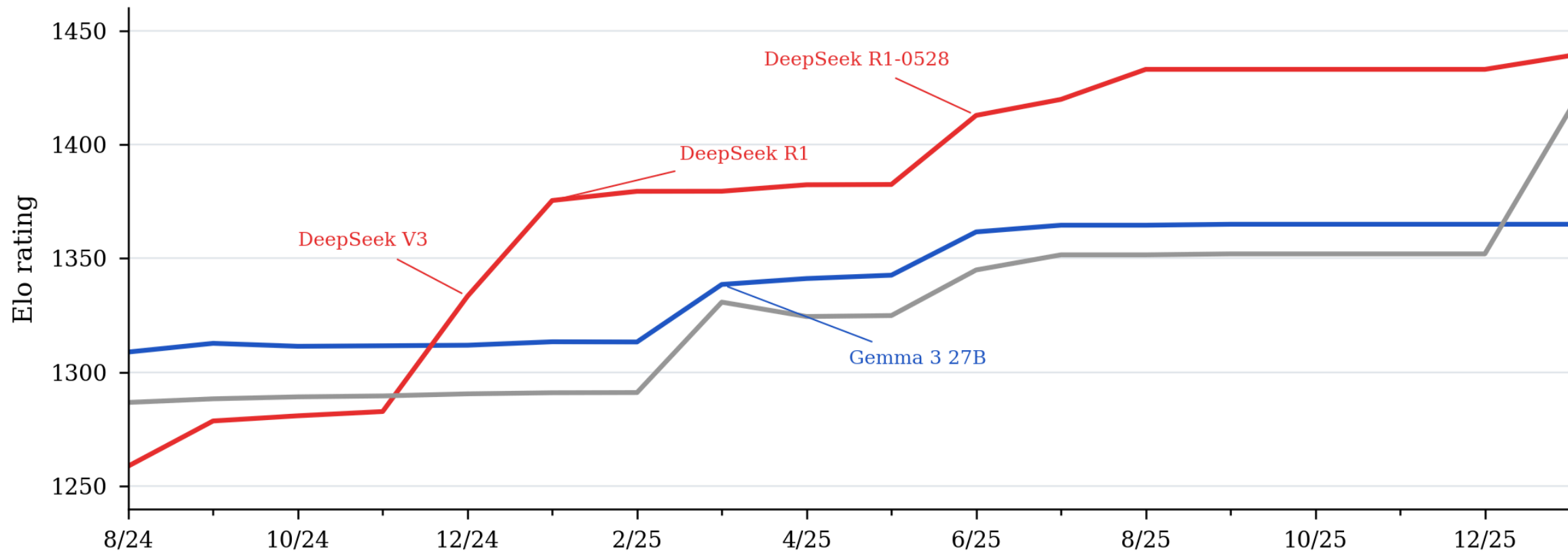


Tracking top 10 open models per month, and their share of inference on OpenRouter. China **2.8%** (Nov '24) → **72.7%** (Jan '26).

Region USA China EU



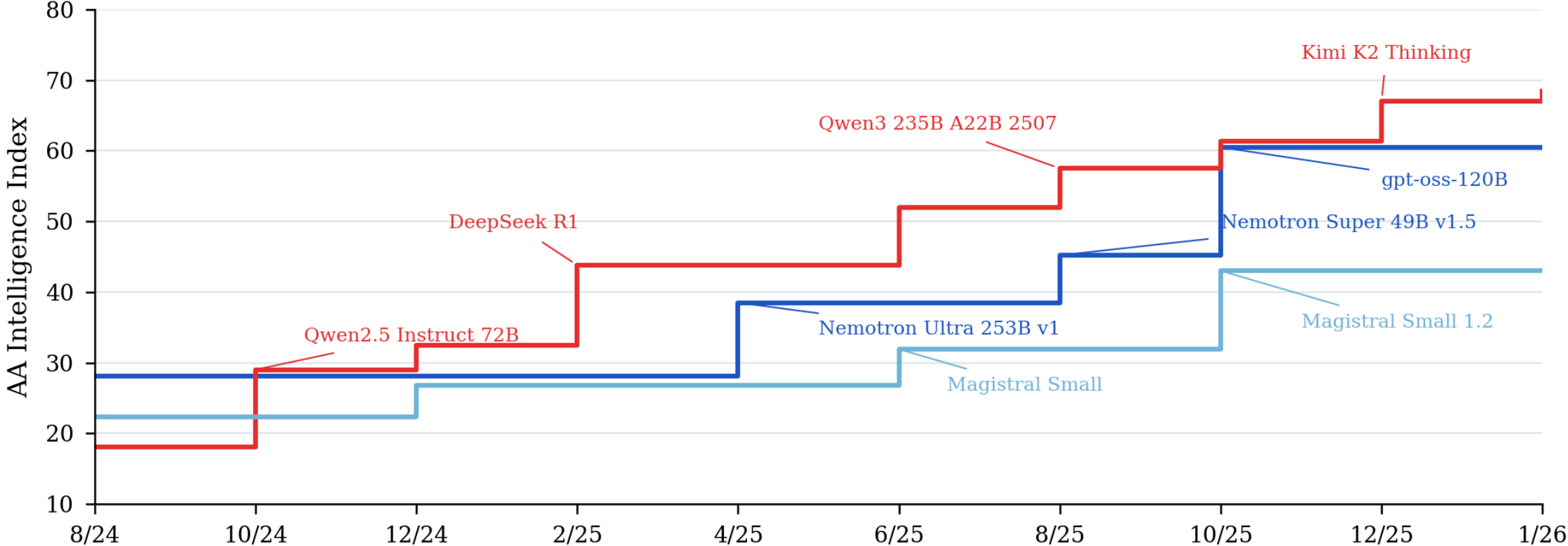
Performance — Arena Elo



Region — USA — China — Other



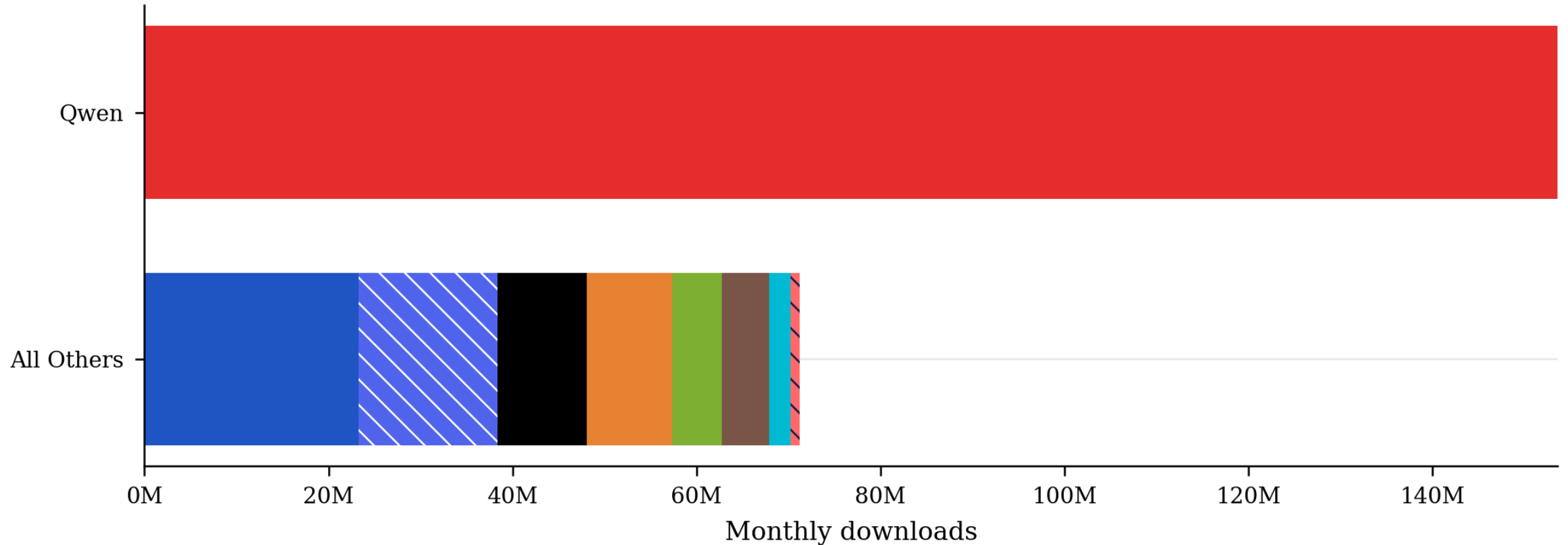
Performance – Artificial Analysis Intelligence Index



Region — USA — China — EU



Qwen is on top of the adoption world



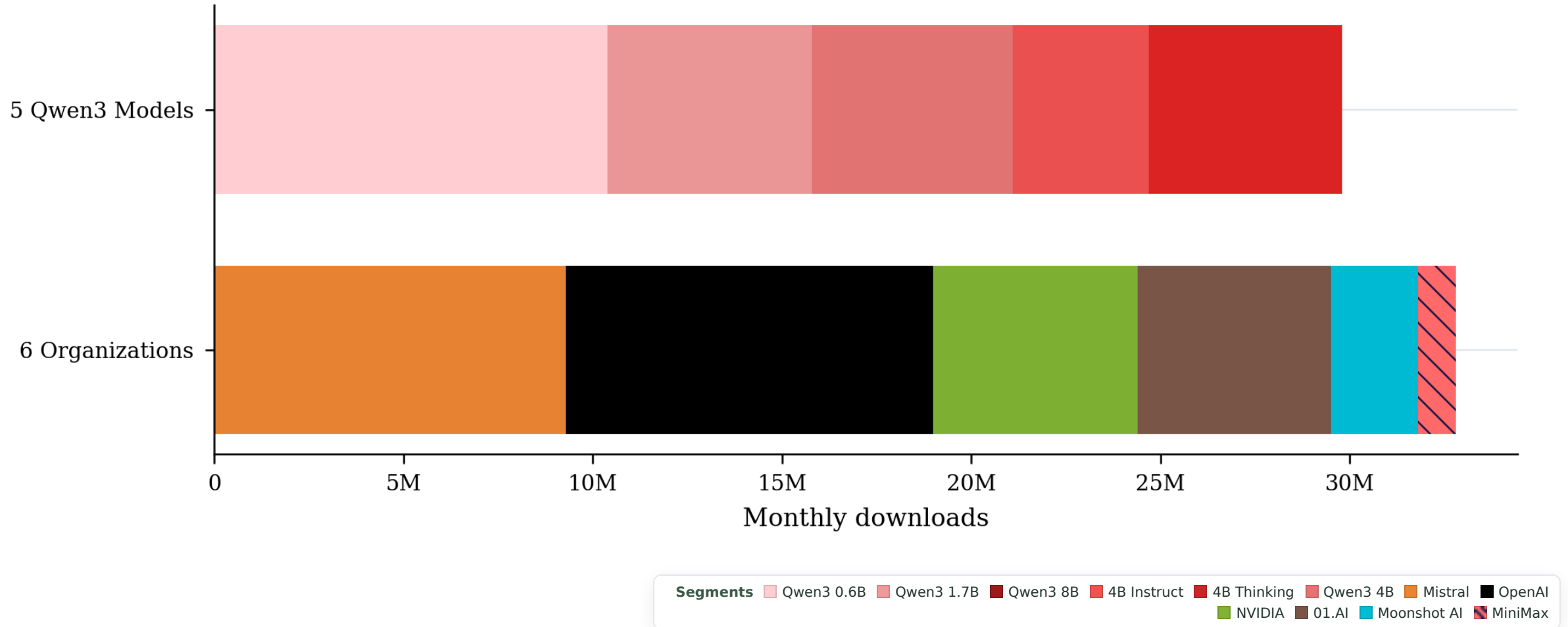
153.6M in February for Qwen vs **71.2M** for the next eight orgs combined.

Org ■ Qwen ■ Meta ■ DeepSeek ■ OpenAI ■ Mistral ■ NVIDIA ■ 01.AI ■ Moonshot AI ■ MiniMax

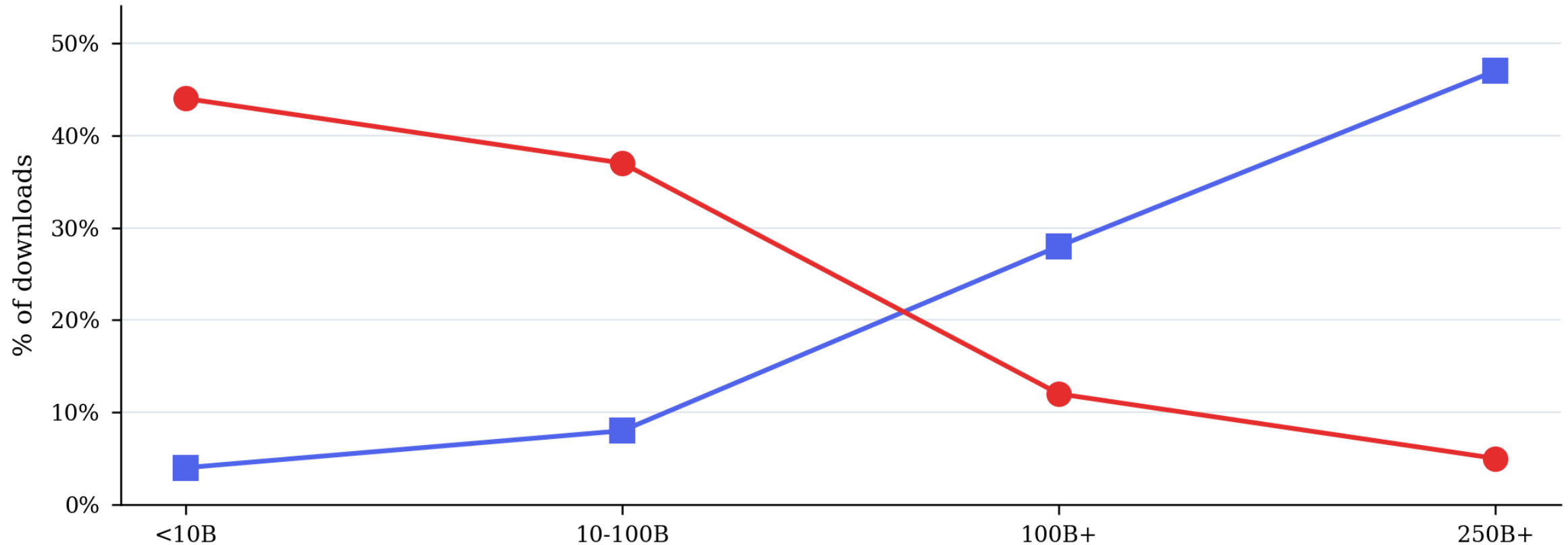


The Qwen3 small-model effect

The top 5 small Qwen3 models dominate monthly download numbers (again, February).



Qwen's one adoption weakness is large models



DeepSeek = **47% of 250B+ downloads**. Two coexisting strategies: Qwen breadth across sizes, DeepSeek flagship at the frontier.

Org ■ DeepSeek ■ Qwen



Part 3: China's AI labs' own perspectives on open models



My China trip

To learn more about the open models I cover:

- About a week across Beijing & Hangzhou (rest of group continued to Shanghai & Shenzhen)
- Sit-down conversations at frontier labs, mid-size players, and the application giants
- Varied from off-the-record conversations with top executives to casual chats with researchers over tea

Trip report:

<https://www.interconnects.ai/p/notes-from-inside-chinas-ai-labs>

Labs visited:

- Alibaba / Qwen
- Z.ai (Zhipu)
- Moonshot (Kimi)
- Tsinghua University
- Meituan
- Xiaomi
- 01.AI
- Ant Group
- ModelScope (China's equivalent to Hugging Face)
- ByteDance (informal/dinner)
- DeepSeek (informal/dinner)



Cultural commonalities across the labs

Our visits were primarily formal, so the labs were not likely to give incendiary takes about competitors or sensitive topics (e.g. distillation). Some commonalities emerged.

1. **Fast-follower culture.** Many Chinese tech industries have started in this way, follow the American leader and make it cheaper. It de-risks a lot of costs in expensive development technologies like LLMs.
2. **Raised to be less vocal on broader impacts.** We tried to nudge researchers to share opinions on the future of work, economic risks, model morality, etc. and got few opinions. Chinese education doesn't reward this type of speculation, where among frontier labs it is a staple.
3. **Hyper-competitive, but less dramatic.** Researchers are very respectful, all praise DeepSeek, all fear ByteDance's market share and closed nature with Doubao chatbot. Less creeps into the discourse as intensely as U.S. dynamics with talent moves and lawsuits.
4. **Similar, practical perspectives on why they openly release models.** Many labs will not release every model, e.g. security concerns on multi-modal generation or lack of



Cultural commonalities across the labs

Our visits were primarily formal, so the labs were not likely to give incendiary takes about competitors or sensitive topics (e.g. distillation). Some commonalities emerged.

5. **Student driven & very young.** The populations are incredibly young, especially around hubs like Tsinghua. Far more so than the U.S. labs who attract a few prodigies.
6. **Desperate for Western attention.** I expect more collaborations to merge across Western AI startups in the open-source space and Chinese model builders. Chinese AI startups are all reading X and following the American labs.



Observations of the AI industry

1. **Early signs of domestic AI demand.** Payment for AI services seems possible, like the cloud, rather than being limited akin to the SaaS market in China.
2. **Most developers are Claude-pilled.** Some use their own models, fewer mentioned Codex.
3. **Chinese companies have a technology ownership mentality.** Examples include Meituan or Ant Group, who build and release strong general models, to later fine-tune to their own products.
4. **Government aid is real, but unclear how big.** E.g. Beijing Academy of Artificial Intelligence (BAAI), helping with office space, etc.
5. **The data industry is far less developed.** Worries of quality, build most in house, including RL environments which is the current focus of global model builders.
6. **Desperation for more Nvidia chips.** Huawei can be used for inference, but a desperation for more training chips is real. E.g. one lab told us their latest big pretraining run took 6 months, which is an incredible amount of risk.



My thoughts on distillation

1. **Jailbreaking uncertainty.** It is fair to ask for help if the Chinese labs are systematically jailbreaking APIs in order to extract reasoning traces that aren't provided in the vanilla API. Unclear why labs cannot prevent it (i.e. is it a property of the model).
2. **Poisoning the word distillation is a major loss.** Distillation broadly is an industry standard technique. Poisoning the entire well with connotation of IP theft will punish small players & researchers.

Read more on Interconnects: The distillation panic · How much does distillation really matter?



My thoughts on the open-closed performance gap

1. **Open models appear to benchmark slightly more.** This is largely a hunch. Closed models outperform slightly more often on out-of-domain benchmarks, but not all the time. Knowledge work frontier domains, e.g. APEX from Mercor, open models lag further.
2. **Closed models are much more robust (less jagged).** I try open models on and off and never stick as much. Unclear if this is just me being finicky or how to measure this.
3. **6-9 months of a gap is a LONG TIME right now!** We are still on the clock to see when an open model has an Opus 4.6 like moment, and it would still be on time! This could kneecap Anthropic's growth.

Read more on Interconnects: Open models in perpetual catch-up · Reading today's open-closed performance gap



Conclusions

The future of the open ecosystem, and strong open models in the next 1-2 years is largely an economics question – how long can companies raise? How long can companies dedicate compute to building open models which can be monetized more effectively elsewhere? Will the open ecosystem coordinate better and unlock further efficiency?

I'm stressed about this! Folks that want open models to be viable need to speak up now and often.

Read more on Interconnects: How open model ecosystems compound

Slides: <https://natolambert.com/slides/china-atom-2026/talk.html>

More: atomproject.ai/report · arxiv.org/abs/2604.07190 · interconnects.ai · [@natolambert](https://twitter.com/natolambert) · nathan@natolambert.com



References

Lambert, N., and Brand, F.. “*The ATOM Report: Measuring the Open Language Model Ecosystem.*” 2026. [\[link\]](#)

{The ATOM Project}. “*The ATOM Project.*” 2026. [\[link\]](#)



Extra slides follow



RAM overview

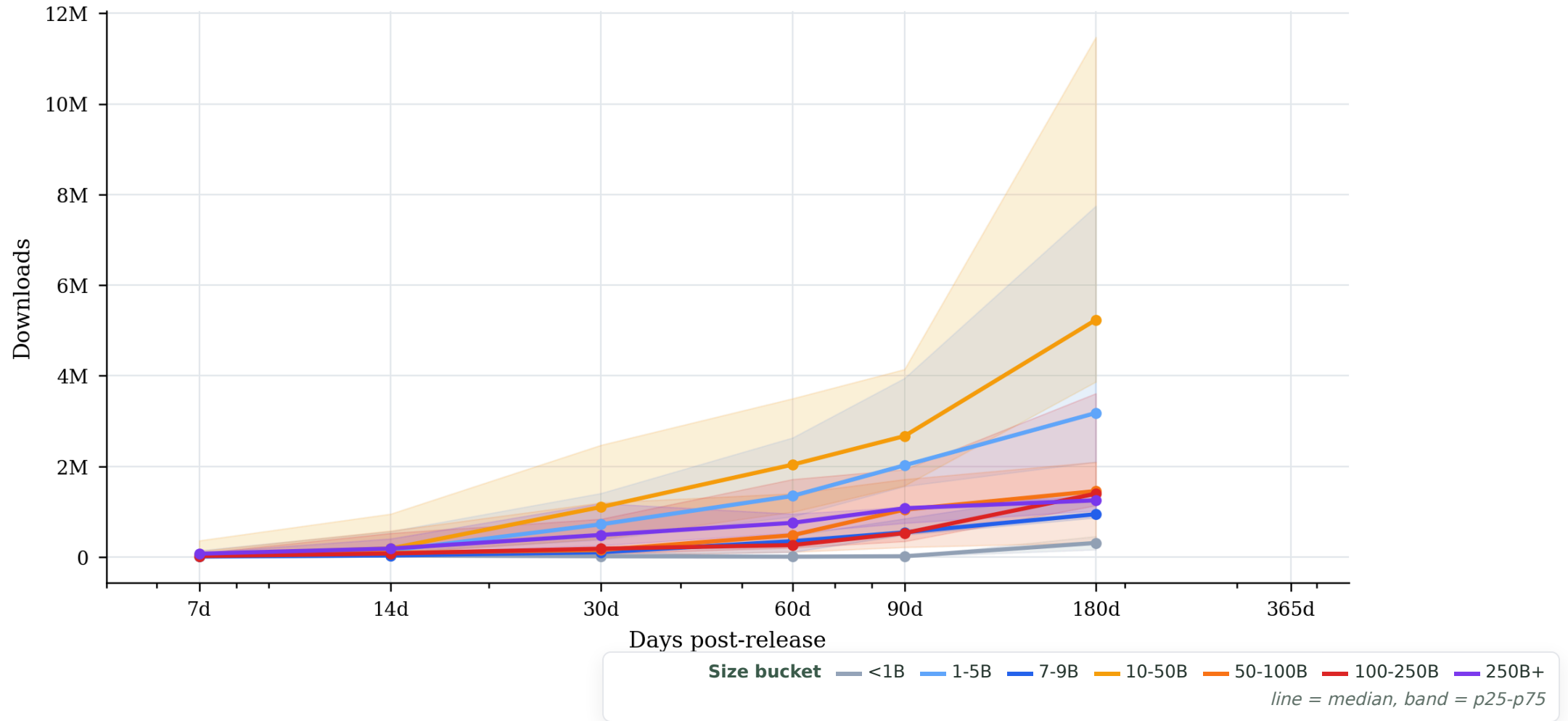
Relative Adoption Metric (RAM) normalizes open model downloads by model size and time since release.

RAM = model cumulative downloads ÷ median downloads for top peers in the same size bucket

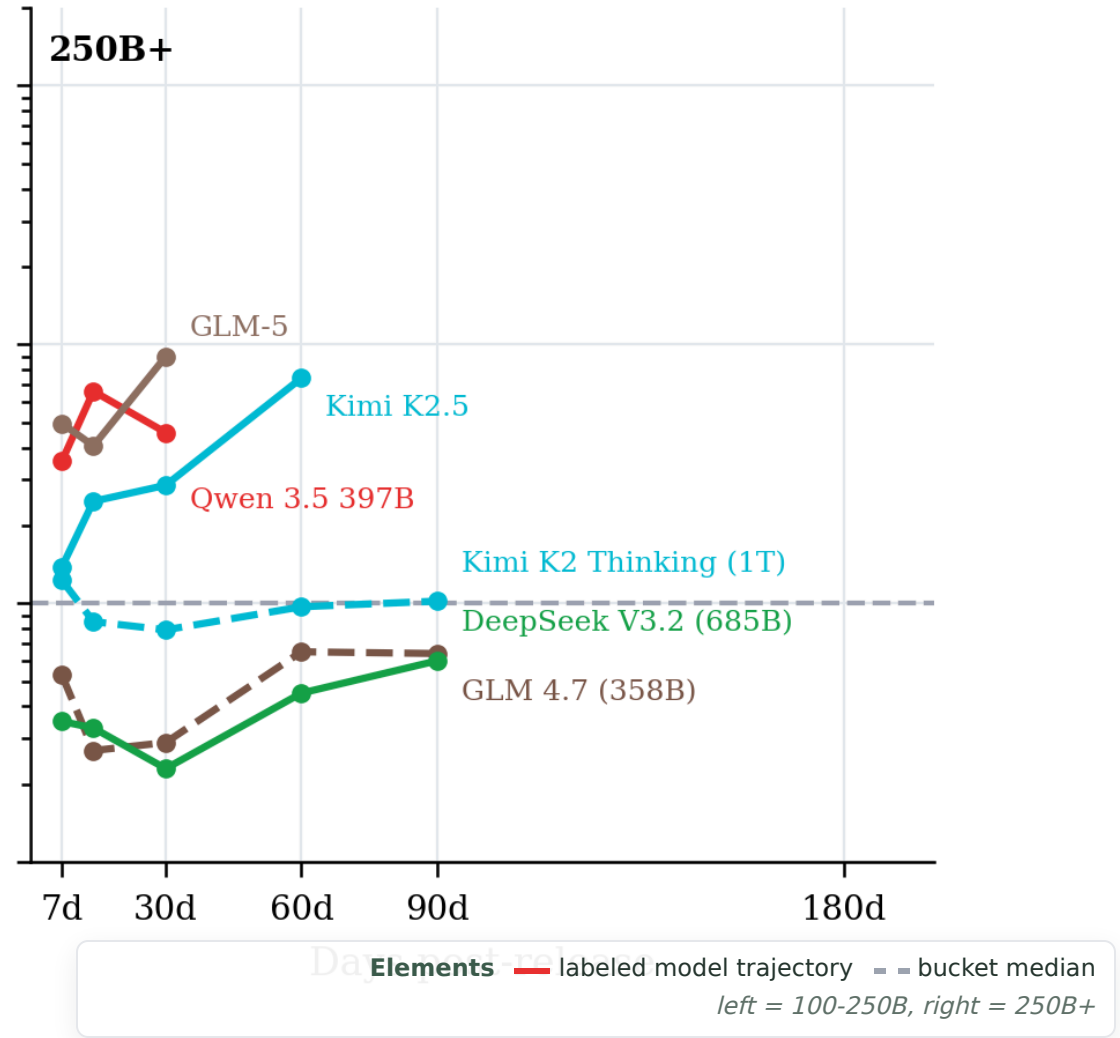
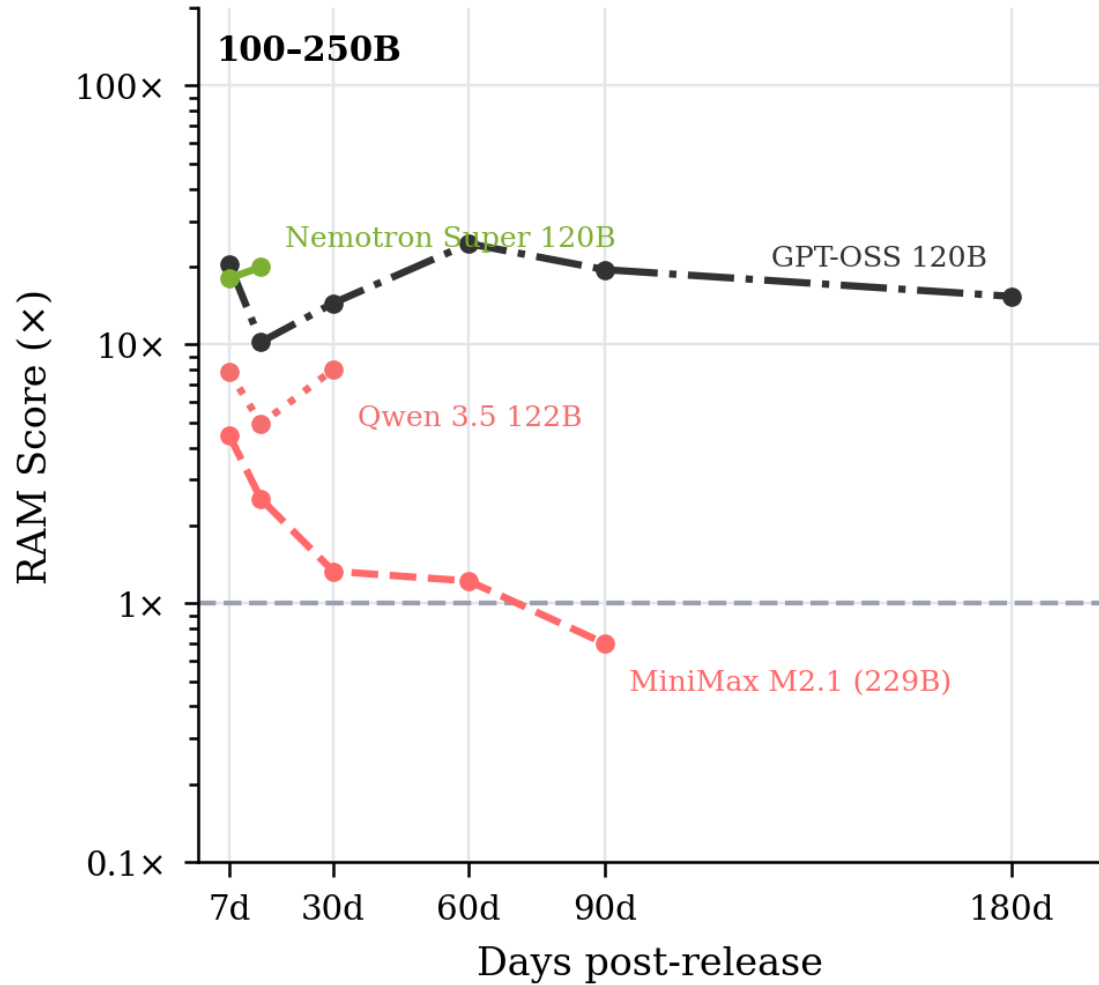
- **1.0x** = tracking the top-10 median for its size class
- **>1.0x** = outperforming its size-adjusted adoption baseline
- Computed at fixed milestones: 7, 14, 30, 60, 90, 180, and 365 days
- Useful for separating launch hype from sustained community adoption



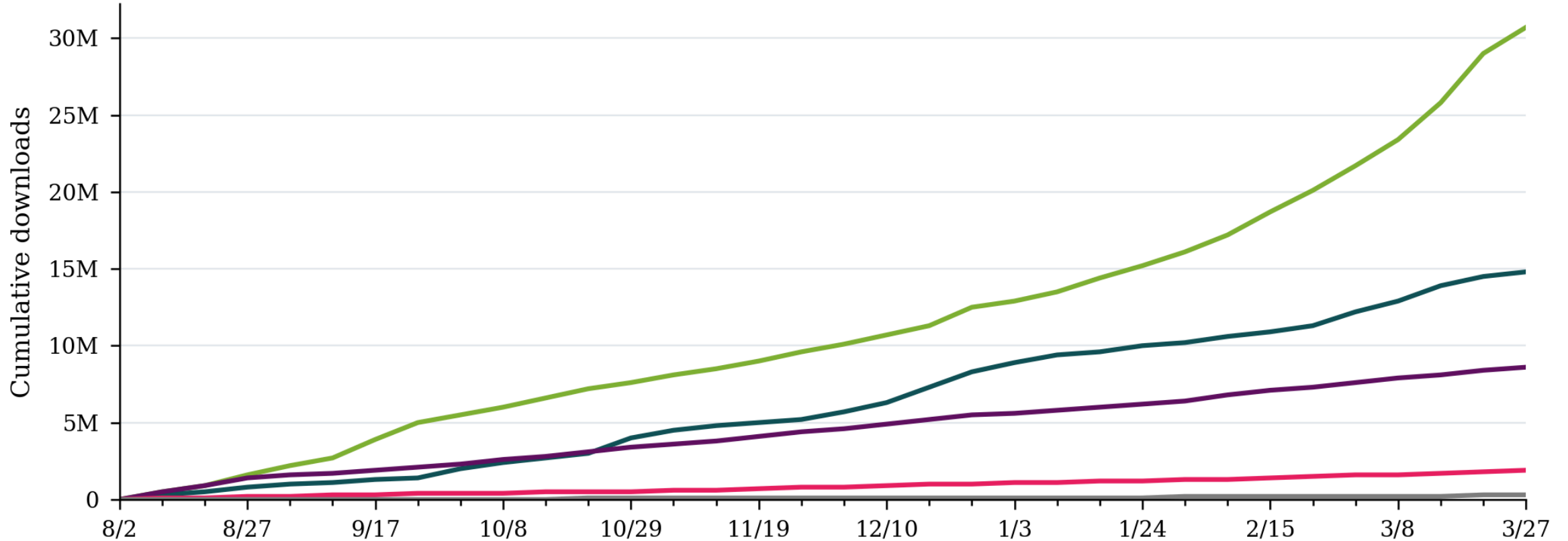
RAM size bins



RAM case studies



The American comeback?



GPT-OSS surpassing Mistral's entire legacy portfolio. Nemotron's ramp accelerating.

Disruption is still possible.

