

The current balance of power in open models

The expanded form of a testimony I prepared for Congress.

NATHAN LAMBERT

SEP 21, 2026

I was recently invited to brief a group of Congressional members and staff on the state of open-weight models in the lens of U.S.-China competition. I'm sharing my prepared remarks as a state of the union on open models that is accessible to a broader audience.

Interconnects AI is a reader-supported publication. Consider becoming a subscriber.

✓ Subscribed

Recap: What is an open source v. open-weight vs. closed model?

Open language models are AI models where their weights are publicly available for inspection or downstream use. These are most often contrasted to so-called “closed” AI models. Closed models offer access only through Application Programming Interfaces (APIs) that developers can use to directly query a model, like GPT-4 or Claude Opus 4.5, or through products, like ChatGPT and Claude Code.

Open language models primarily are bucketed into two categories, open-weight and open-source models. Open-weight models are the most common form, such as popular models like Meta’s Llama, Alibaba’s Qwen, Google’s Gemma, or DeepSeek’s models. These models are

governed by licenses, governing documents dictating what is allowed with downstream use, and are often accompanied by inference code in libraries such as Transformers, VLLM, SGLANG, etc. Since about April 2025, Chinese AI companies have been the clear leader in open-weight models.

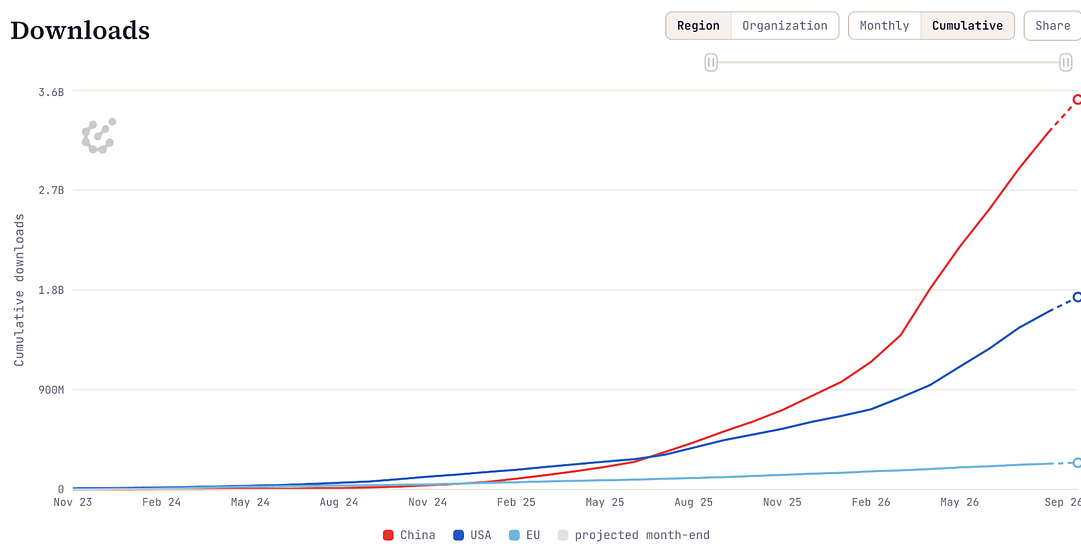
True “open-source” models are similar to these, as they include the weights, licenses, and inference code, but they also include the complete information needed to reproduce the model – the training code and training data. The most prominent open-source models have been built in the United States, led recently by the Allen Institute for AI’s Olmo models that I helped build in my recent 2.5 years there. The other prominent open-source models are also built by American non-profit organizations, including OpenAthena’s Marin models and EleutherAI’s Pythia models.

Open-weight, open-source, and every other label for a model – including closed models primarily offered via an API – exist on a spectrum. For example, Nvidia’s Nemotron models are far more open than most open-weight models, releasing large quantities of their training data under permissive licenses, but they’re not fully open-source because they do not release all of the data. Closed models also [exist on a spectrum](#) based on what information the API reveals and the terms of use.

The state of competition between American and Chinese open-weight models (unit economics, technical capabilities, etc.)

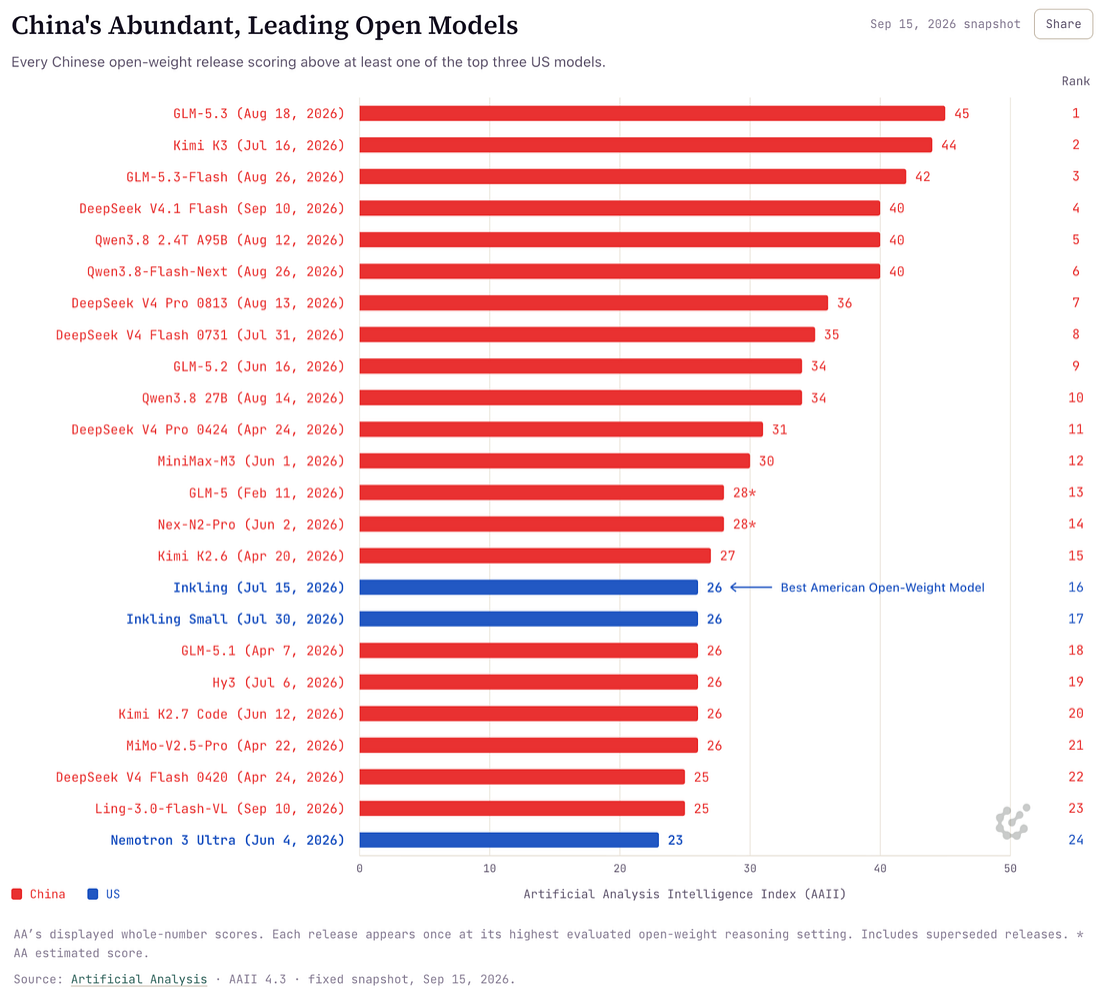
We are living in a world where [GLM-5.2](#) and [Kimi K3](#), some of the latest, leading Chinese models, have enacted a step change in the commercial viability of open models — crossing a similar threshold in agentic capabilities that Anthropic’s Claude Code crossed in December of 2025.

America was the early leader in open language models, primarily through Meta's Llama models, which were used extensively across research and commercial tasks. Chinese open-weight models surpassed American open-weight models in these two key areas about 18 months ago. The simple metric showing this is Hugging Face Downloads, where China took the lead in July of 2025 primarily through the success of Alibaba's Qwen models. I personally maintain tools to track this data, and since I first published the [American Truly Open Models \(ATOM\) Project](#) in August of 2025, China's [download lead has grown to about 1.6B](#) – with a total of 3.2B downloads, twice that of America's total.



On popular capabilities benchmarks, such as the Artificial Analysis Intelligence Index (AAII), the Chinese open-weight models have a clear lead over American counterparts. The top three Chinese models as of writing this on September 14, 2026 are Z.ai's GLM-5.3 and GLM-5.3-Flash and Moonshot AI's Kimi K3 with scores of 45, 42, and 44 respectively. By comparison, the leading American models are Thinking Machines' Inkling and Inkling Small, both with a score of 26, and Nvidia's Nemotron 3 Ultra, with a score of 23. The top American models were released in June and July of 2026, and are updated less frequently than their Chinese counterparts. For example, Chinese labs released models

with scores above these American models 2-6 months before the American companies got there (e.g. GLM-5 or DeepSeek V4 Pro). There is a trend of *more* American companies releasing models, including names like Arcee AI, Poolside and IBM, but they are not rapidly closing this performance gap. Other benchmarks tell a similar story.



The top American open models on the Artificial Analysis Index are behind 15 other Chinese made models.

Together, Chinese open-weight models are approximately 2-5 months behind the closed American frontier, with the open-weight American models being approximately 6-9 months behind the likes of OpenAI and Anthropic. The Chinese labs are closest in tasks with clear user demand,

such as agentic coding, and further behind on more open-ended scientific tasks, such as physics or biology.

The reasons why Chinese labs can produce these strong models, despite having fewer resources than American counterparts, is still an open debate and [heavily influenced by different work cultures](#), but is also influenced by a few key technical factors. The Chinese labs release their models faster and focus on a slightly narrower distribution of tasks, flattering them slightly on public benchmarks. Releasing faster helps them score higher because all the labs are making consistent progress, so once you “finish” a model to be released, it is a snapshot of performance at that given time — labs where that time is later tend to score higher. Still, the models built by the Chinese labs are genuinely strong and represent real competition to the American industry. This competition will not decrease meaningfully as the closed labs patch vulnerabilities in their API offerings which enable distillation.

Distillation is most impactful in new domains and does not make it trivial to create a universally strong final model. I estimate that if distillation was fully prevented, e.g. with know-your-customer (KYC) tools at Anthropic and OpenAI, the gap from the strongest American models to Chinese open-weight models would only increase by 1-2 months.

For example, the Chinese labs are rapidly changing their posture towards paying for training data in 2026. Earlier in the year, the top Chinese labs including Moonshot AI and [Z.ai](#) had a strong preference towards building data workflows in-house, but by the summer they had begun to buy the cutting edge data – challenging RL environments for agentic tasks – from both established American companies and new Chinese startups.

With the advance of open weight models in China towards the frontier of capabilities, and the recent documentation of growing risks around frontier models in areas such as cybersecurity (e.g. the [OpenAI-](#)

HuggingFace incident), there's growing regulatory uncertainty on *how continued releases can enable a safer ecosystem?*

A structural challenge in open-weight models is that there are few effective methods for stopping pieces of open software from reaching bad actors. If an attempt was made to restrict access to the strongest open-weight models from China because they amplify risks, the parties who would be set back are American businesses. We have an example of this – HuggingFace [used a Chinese open-weight model to understand the cyberattack](#) because closed models would not answer their requests. Thus, managing the risks of open-weight models [often comes down to ecosystem preparation](#).

Open-weight models are becoming an essential tool for AI diffusion, and the best path to get ahead of these risks and unbalanced relationships where American companies rely on models built in China is to continue to enable investment in open models in the US. Ownership of open models allows better coordination and preparation of risks that are global in their nature while accelerating diffusion of AI services throughout the domestic economy.

[Leave a comment](#)

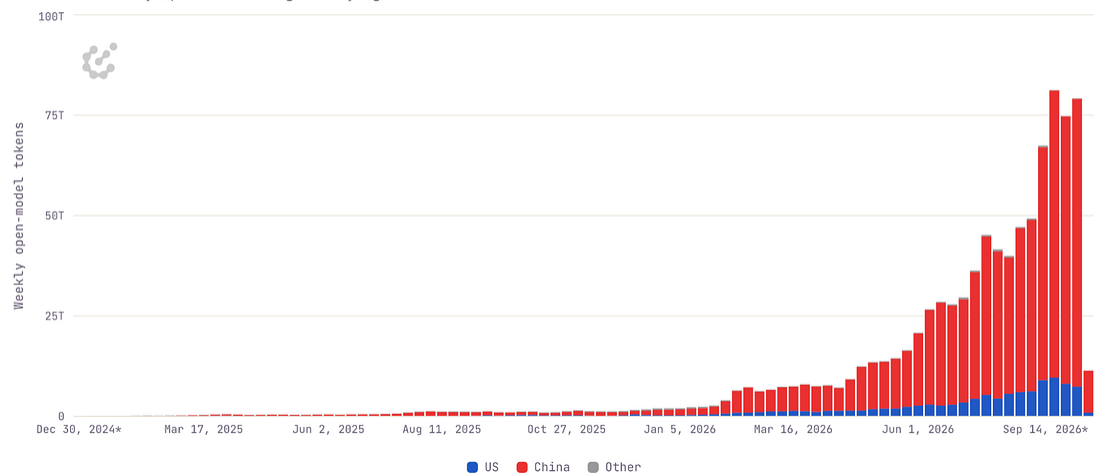
The state of open model adoption: How is open-source being used by academia, businesses, and other countries?

Open-weight language models have grown substantially in general interest and economic viability in 2026, allowing early glimpses of more direct ways to compare adoption of models from the US, China, or elsewhere on top of Hugging Face metrics. One example is OpenRouter usage. OpenRouter is a popular LLM inference platform that supplies a single interface to switch between models, open and closed, from the US and China. This platform is primarily known for trying different

open-weight models. The platform has shared usage data for the top models since Jan. 1, 2025, and shown growth in usage from ~1T tokens processed from open models in a week of September 2025 to ~80T tokens per week today. In that time, [Chinese models have grown from ~70% market share to over 80% of usage](#). Other platforms that are designed to commercialize open models show similar data, such as the open-source coding agent OpenCode, which shows an [inference volume of ~95% or higher with Chinese models](#).

OpenRouter Open Model Token Volume by Region

Token volume of daily top 50 models categorized by region.



These open platforms are the best *approximation* of open model usage we have – a large proportion of open model usage is on platforms that do not disclose per-model breakdowns, such as Together AI or Fireworks AI, and in private deployments for enterprise applications.

Many prominent technology companies and startups have been building on Chinese open-weight models for their AI features, such as [Harvey](#), the legal agent, [Cursor](#), the coding agent, and [DoorDash's](#) use of Kimi models, [Airbnb's](#) use of Qwen, or Perplexity's use of DeepSeek. These prominent companies are the tip of the iceberg, where a large swath of younger Silicon Valley startups are building on Chinese models in order to have low-cost, flexible options. There is a [growing trend of American startups and companies entering enterprise agreements](#) with Chinese

model labs in order to get permission to use their models in their products – a new form of cross-border technology collaboration I have not witnessed in my career.

The foundation of innovation on Chinese models extends further into the AI ecosystem. To a first order approximation, most of academic research is conducted on Alibaba's Qwen family of models. Having met multiple members of the Qwen leadership team [during my trip to China](#), they are very invested in and intentional about this type of adoption, which will not be easy to claw back to American models.

To quantify the adoption of open models across **academia**, I scanned every paper in the 5 most popular ML categories of arXiv (cs.AI, cs.CL, cs.CV, cs.LG, stat.ML), the preprint platform popular in AI research. The results clearly track my understanding of the evolving leadership in AI research, showing LLMs becoming a foundational layer of ML research – mentions of any open model were 2% in January of 2023 and 50% in September of 2026 – and the leading role shift from the U.S. to China in the same time period.

For example, in April to May of 2023, a few months after Meta's original Llama (a backronym, Large Language Model Meta AI, first released in Feb. of 2023), about 2,600 of 12,000 new AI/ML papers on arXiv mentioned at least one prominent open model family. Of all those scanned papers, ~5.5% mentioned Llama and ~1% mentioned a Chinese model. In the fall of 2024, during Llama's peak, about 23% of papers mentioned Llama with about 7.5% mentioning Qwen, the most direct Chinese competition. Today, Llama has lost its lead in academia, being mentioned in about 21% of papers still, which is remarkable longevity, but Qwen's share has risen to 30% of papers. Overall, any Chinese open weight model is mentioned in over 40% of papers, over the U.S.'s 30%, with China's share continuing to grow.

Model Mentions in AI Research

Share of scanned arXiv papers in cs.AI, cs.CL, cs.CV, cs.LG and stat.ML

Period	Papers which mention an LLM Mentioned / scanned (%)	Model families		Tracked open-weight groups	
		Llama	Qwen	U.S. group	China group
April-May 2023	2,640 / 11,892 (22.2%)	5.52%	0%	6.21%	1.03%
Sept.-Nov. 2024	9,243 / 23,258 (39.7%)	23.36%	7.46%	26.11%	11.19%
June-August 2026	17,843 / 32,002 (55.8%)	18.73%	31.70%	28.05%	37.59%

Papers may mention both U.S. and Chinese models -- can be counted in both.
Regional groups encompass the 48 most popular open model families.

Source: Interconnects Adoption Dashboard | data through Sept. 8, 2026
dashboard.interconnects.ai



This shows that we clearly have a lot of work to do in order to re-establish the U.S. as the home of AI research in the era of open-weight language models. There are signs of hope.

In our [research](#), we find that American models of comparable capabilities-to-size regions to their Chinese counterparts get adopted at disproportionate rates. In the last year we've seen OpenAI's first open-weight models since ChatGPT, [gpt-oss](#), become one of the most adopted open-weight models of all time. Since then, Google's [Gemma 4](#) models have been some of the only ones ever to show similar adoption numbers to Qwen's most popular small models, and Nvidia's [Nemotron](#) models have modest adoption despite numerous more capable models at the same size point.

Share

Summary

The story of open models in 2026 is one of establishing economic relevance. This is the convergence of many stories across the AI ecosystem, summarized as:

1. **The gap from open to closed models available to users has been *decreasing over the last 3 years*.** This varies by task, but can be estimated as a 2-5 month gap in capabilities. With capabilities overall progressing so fast, this has seen open-weight AI models unlock substantial markets in 2026 and points to more inflection points in the near future.
2. **Open model usage is exploding in high-value industries (e.g. software engineering, legal services, financial services), indicating an emergence of an alternative ecosystem to the best closed models.** Platforms offering inference primarily on open models, from Together, OpenRouter, Fireworks, Baseten, etc., are seeing incredible growth as the first winners of an open model post-training economy (other layers include finetuning APIs such as Thinking Machines' Tinker). This is combined with numerous anecdotes from technical staff in the AI industry that uses open-weight models such as GLM-5.3 as an alternative to Claude or GPT due to a combination of speed, lower prices, customizable offerings, and privacy.
3. **Chinese AI companies are the clear leaders in open weight models.** Relative to 2025, where Chinese models like DeepSeek R1 shook the AI world with surprise, the American AI labs have been recovering in their positions with open-weight models, but despite more substantial investment in the US, the Chinese labs regularly are producing notably stronger models adored by many types of users.

4. **Distillation of American AI models by Chinese labs does not explain the entire story of their success.** Distillation is an industry standard technique of training another AI model on the outputs from a usually stronger model. The technique is most prevalent in the Chinese AI industry, which has used basic exploits to extract reasoning traces and additional data from American companies' products that are not fully secured. The best estimates are that distillation helps reduce the performance gap of Chinese companies relative to the American frontier by 1-2 months.
5. **Chinese models, particularly Alibaba's Qwen family, are established as a foundational layer of research and development across academia and local model users.** In recent months, Chinese open weight models were mentioned in 38% of AI papers, above the U.S.'s 28% – and the Chinese share is growing much faster than its American counterparts. This, along with other political factors and the closed nature of leading American AI companies, is contributing to an accelerated decline in America's lead as the preeminent AI research hub in the world.
6. **Open weight models are entering the capability levels where new risks, e.g. cybersecurity, can be enabled by numerous open-weight models being available, necessitating an ecosystem level response in preparation.** This new era of risks is also enabling a period of political uncertainty, where there is regulatory attention on the strongest AI models, but massive uncertainty on how policy would be legally enacted. At the same time, [many researchers and engineers rely on open models due to more permissive safeguards](#), where the closed models such as Claude and GPT often refuse critical cybersecurity defensive work or biology research.

For more data, view the [Interconnects Dashboard](#).

Conclusions

In 2026 the Chinese labs are clearly maintaining their status as the leaders of the open-weight AI ecosystem. This comes as open-weight models have passed an inflection point in economic viability and in the face of increased activity from American labs as model competition. The leading Chinese labs do not appear to be meaningfully challenged, as they expand their enterprise and research adoption globally.

This landscape of open models comes at a crucial time in the broader AI ecosystem. We're seeing OpenAI and Anthropic take massive steps forward with their latest public models, and at the same time call for coordinated care on how we manage the next stage of AI progress. What is happening in the confines of a few AI labs today, especially with extreme talent and compute density, is a precursor to what will soon emerge in the open model ecosystem. Open models are going to be the substrate for everyone else in the world outside of the few true frontier AI labs, to harness an acceleration in software engineering and other computational practices. This represents a substantial source of soft power, influence, and potential for the organizations that enable this broad access to transformative intelligence.

With this future coming soon, we need to collectively stay humble about the exact path open models will take. There are a lot of unknowns with open models – e.g. we don't have good data on how they're used in [countries other than the U.S. and China](#). With the distribution of ML training expertise being broad, i.e. tens of organizations and thousands of people that are within a year of the frontier of capabilities, it is a matter of when, not if, open models cross the performance thresholds that enable new workflows. The collective approach should be to understand how to use this broadly accessible, open intelligence for good while proactively mitigating the potential harms.

Thank you to Florian Brand and Kevin Xu for feedback and/or suggestions for this work. For more research informing this post, see the [open-source AI](#)

reading list.