SECOND ORDER

AI, Capital & Work · A Second Order research briefing

ISSUE 04 · JUNE 30, 2026

The Stack Splits

In a single June fortnight Washington restricted its most capable models and Beijing released its best as open weights. The variable that sets the global AI standard has moved from capability, where the US still leads, to diffusion, where policy is working against it.

IN THIS BRIEFING

One question: who sets the standard the world builds on?

THE THROUGH-LINE

Since Issue 01 the series has read the buildout from inside the US, where the capital and the frontier sit. June forces a different question: who controls the model the world actually runs. Washington suspended its two most capable public models and held a third for review in the same fortnight China released a top model as open weights on its own silicon. The capability gap is the narrowest it has been; the deployment gap just widened the other way.

WHY NOW

Fable 5 and Mythos 5 pulled June 12. GLM-5.2 shipped open June 13. GPT-5.6 held June 25. The economics that govern model procurement changed this month, whether or not the roadmap has caught up.

WHAT YOU WILL LEARN

1. What happened: the June split (sec. 01)
2. Why the constraint is policy, not capability (sec. 02)
3. How the base layer flipped to China (sec. 03)
4. A full stack that routes around Nvidia (sec. 04)
5. Six operator moves, a dated watchlist, and what would change our mind

IF YOU REMEMBER ONE LINE: the standard is set at the layer the world builds on, not the one that scores highest, and that layer is moving east.

THE CLAIM

The controls meant to slow China are slowing the US

> *“At a time when access to frontier models is abruptly cut off for non-technical reasons, we are even more convinced.”*
>
> z.ai (Zhipu), on releasing GLM-5.2 as open weights, June 2026

51

GLM-5.2 on the Intelligence Index, the top open model; the leader overall, Fable 5 at 60, is suspended in the US

2.2x

Chinese open families' monthly Hugging Face downloads against US open families (Second Order)

0

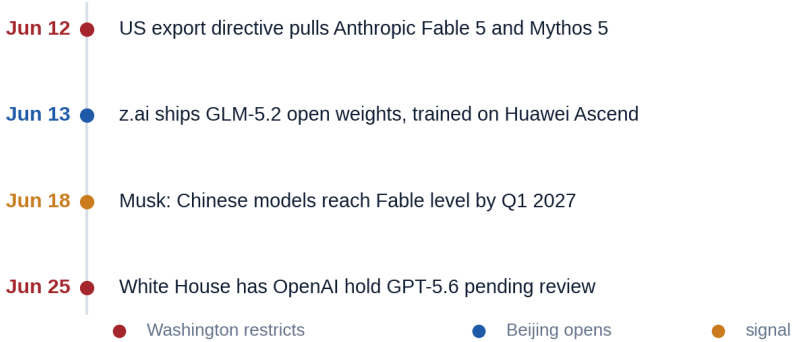
Nvidia GPUs in GLM-5's training; about 100,000 Huawei Ascend accelerators instead

Washington restricts as Beijing opens

EXHIBIT 1

Washington restricts as Beijing opens: the June 2026 split

Frontier-model access events, June 2026



Note: Fable 5 and Mythos 5 access suspended by US export-control directive; GPT-5.6 (Sol, Terra, Luna) held pending review.

Source: Anthropic statement (Jun 12); z.ai/Zhipu release (Jun 13); Musk on X (Jun 18); TechCrunch, CNN, Axios (Jun 25).

SO WHAT

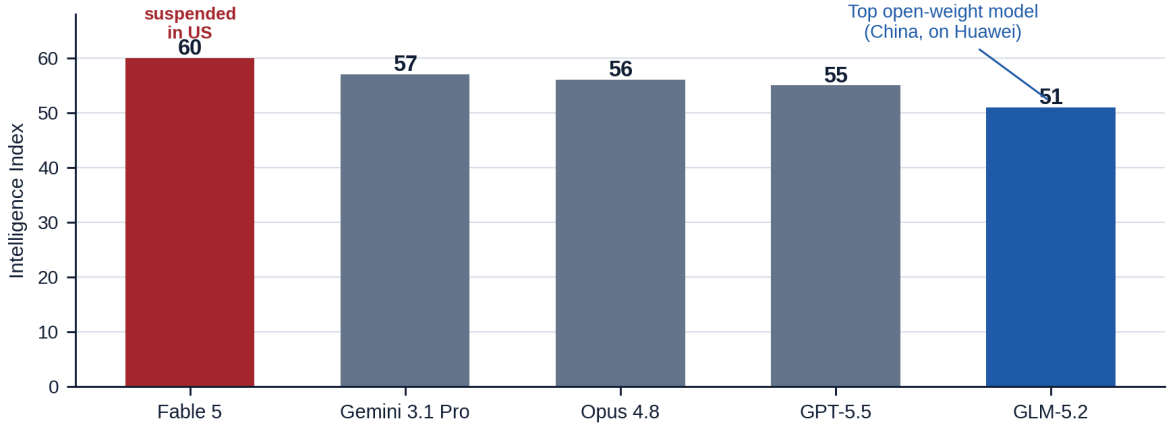
- Capability did not drive this. In one fortnight the US restricted three frontier models while China released one as open weights.
- The June 12 directive targeted foreign-national access to Fable 5 and Mythos 5; with no way to segment users, Anthropic shut both for everyone.
- GPT-5.6's 30-day pre-release review turns US frontier deployment into a licensing queue.

The constraint moved from capability to policy

EXHIBIT 2

The capability gap is small, and the #1 model is the one the US pulled

Artificial Analysis Intelligence Index, leading models, as of Jun 27 2026



Note: Fable 5 (60) leads but is suspended in the US; GLM-5.2 is the top open-weight model, on a Huawei stack.

Source: Artificial Analysis Intelligence Index (artificialanalysis.ai), retrieved Jun 29 2026.

SO WHAT

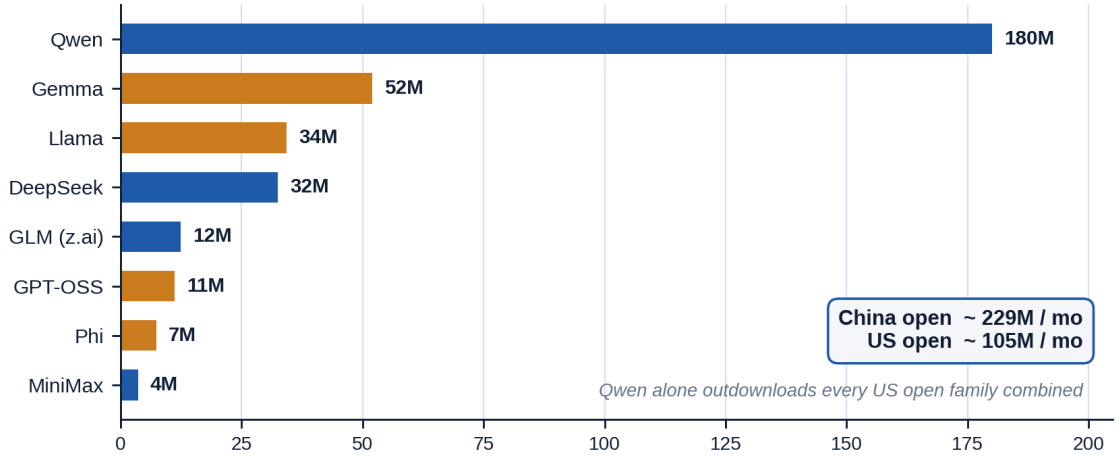
- The lead is real and narrow: the best deployable US model (Opus 4.8, 56) leads GLM-5.2 (51) by five points.
- The model at the top, Fable 5 at 60, is the one Washington suspended.
- A frontier that cannot be shipped does not set a standard.

Open weights compound, and adoption has flipped

EXHIBIT 3

The open-weight base has flipped to China

Monthly Hugging Face downloads, leading open LLM families, top models per family (millions)



Note: Sum of each family's top models by 30-day downloads; counts overstate real usage (automated and CI pulls).

Source: Second Order calculations from the Hugging Face Hub API (huggingface.co/api/models), retrieved Jun 29 2026.

SO WHAT

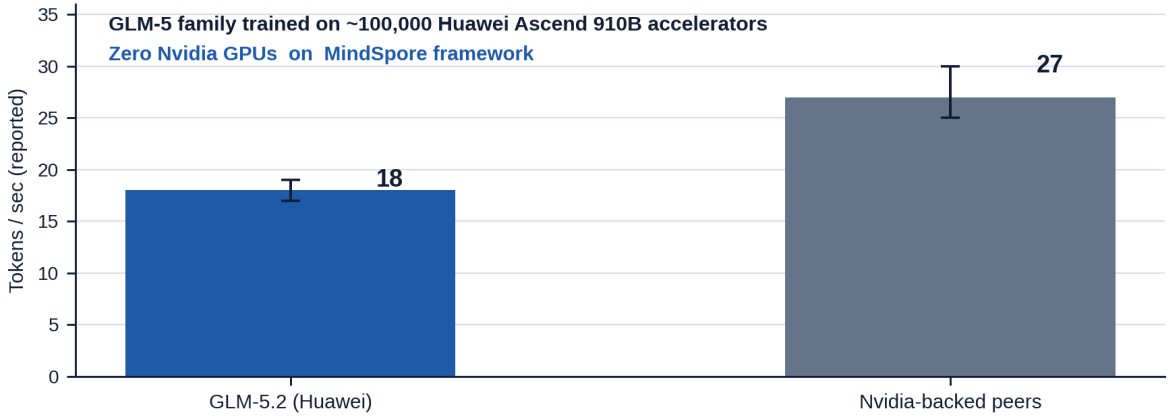
- Chinese open families draw about 229M Hugging Face downloads a month against about 105M for US open families (our count).
- Qwen alone outdraws every US open family combined and anchors 113,000+ derivatives, more than Google and Meta together.
- Open weights are about a third of OpenRouter tokens; the base layer is open, and increasingly Chinese.

China routes around Nvidia too

EXHIBIT 4

China's frontier runs on a domestic stack, at a speed tax

Reported GLM-5.2 inference throughput vs Nvidia-backed peers (tokens/sec)



Note: Inference rates are reported ranges (whiskers), not benchmarked here; the training-chip claim is vendor-originated.

Source: Training-chip claim: digitaltoday.co.kr, letsdatascience.com, techtimes.com (multi-outlet). Inference rates: reported, single-source.

SO WHAT

- GLM-5 was trained on about 100,000 Huawei Ascend accelerators, no Nvidia, on the MindSpore framework.
- The cost is a throughput penalty near a third: about 17-19 tokens/sec against 25-30 for Nvidia-backed peers.
- Open weights on domestic silicon is a complete stack beyond US export law.

SECOND-ORDER EFFECT

The standard is set by what people build on

The first-order reading is a US win: keep the best models at home, and deny them to rivals.

The second-order reading runs the other way, and it follows from how standards are won. They are won at the base layer, where availability matters more than peak capability, and they compound: each week a model is the one developers can hold and fine-tune, the tooling and the next layer of work settle onto it.

So the restrictions work against their intent. When the top model can be withdrawn by directive, the rational builder hedges toward weights that cannot be recalled, which today means the Chinese families. Washington is defending a capability lead by conceding the standard it was meant to secure.

WHAT WE WOULD DO

The operator's playbook

1. Access is a supply line

Do not single-source a closed model that can be suspended by directive. Name a fallback and the trigger that moves you to it.

2. Test an open fallback now

Keep a self-hostable open model in parallel to production, so the switching cost is a measured quantity, not a discovery under pressure.

3. Separate weights from jurisdiction

Run open weights on infrastructure you control rather than a China-hosted API, and keep regulated data off jurisdictions you would not choose.

4. Build model-agnostic

Standardize on interfaces that let you swap the underlying model in a day; the leaderboard and the policy regime will keep moving.

5. Map your silicon

Know your exposure to Nvidia against the alternatives, and to the memory bottleneck behind all of it. Next week's issue.

6. Assume open abroad

Non-US customers can and will build on Chinese open weights. Price and position for a default model that is not American.

WATCHLIST

What would change our mind

GPT-5.6 release: a broad release within weeks is a speed bump; continued delay is a standing license regime.

Fable 5 and Mythos 5: whether, and on what terms, they return; a quiet restore argues a fix, not policy.

Q1 2027: the Musk and z.ai horizon for an open model at Fable's level; watch the gap close or stall.

Open-weight share: Hugging Face and OpenRouter: do the US open efforts (GPT-OSS, Gemma, OLMo) recover momentum?

Falsifier: a fast US carve-out plus a plateau in Chinese open adoption would break the thesis.

NEXT ISSUE · THE CAPEX ECONOMY. Growth is increasingly the buildout itself, and a rising share of it is memory, not compute. Gavin Baker puts high-bandwidth memory at 30 to 40% of hyperscaler capex by 2027, supplied by a three-firm oligopoly. We separate the capex figure markets quote from the capacity it buys.