Open Season on Open Weights Washington Accuses Moonshot of Distilling Claude to Build Kimi K3

Treasury Threatens Sanctions Over the Claim the Same Week DeepSeek V4 Goes Stable and OpenAI Bets Enterprises on Agents

Day 118 · July 24, 2026 · 6 min read · AI Policy & Model Governance

Yesterday the White House did something it hasn't done before in this series' 118 days of coverage: a senior official stood up and named a specific Chinese lab, a specific American model, and accused one of stealing from the other in public. OSTP director Michael Kratsios said Moonshot AI ran a "large-scale covert industrial distillation" operation against Anthropic's Fable model to build its Kimi K3, and Treasury Secretary Scott Bessent followed within a day, putting sanctions back on the table with a line built for headlines: "open source is not open season on American IP." It landed in the same week DeepSeek's V4 family finished its move from preview to a stable, enterprise-priced release -- the follow-up this series promised yesterday -- and the same week OpenAI opened Presence, a platform built to let enterprises hand real customer-facing work to agents. Three storylines, one underlying question: as the gap narrows between the labs that train frontier models and everyone else who wants to use one, who gets to decide where that value actually came from?

VIRAL / OPEN-SOURCE SPOTLIGHT

Kimi K3: The Open Model at the Center of a US-China IP Fight

Moonshot AI calls Kimi K3 the first open 2.8-trillion-parameter model, and independent testers confirm it clears a genuinely high bar: it trails the top proprietary systems, Claude Fable 5 and GPT-5.6's Sol tier, but still lands at frontier-level performance on coding and reasoning benchmarks -- with its full weights due to go public on July 27. That release date is exactly why this week's timing matters. Kratsios's accusation is specific: he says Moonshot built internal tooling to run large-scale queries against Fable through multiple access methods it could switch between to dodge detection, and separately flagged the company's use of restricted GB300 servers, some obtained through Thailand -- a Southeast Asian routing pattern this series has flagged before as export controls tighten. Anthropic has its own number: more than 3.4 million Claude exchanges it says came from fraudulent accounts built to extract reasoning, coding, tool-use and computer-vision behavior. But the story is going viral precisely because it isn't settled -- TechCrunch canvassed model researchers who argue that exploiting Fable's outputs alone doesn't explain a model this capable, and that Moonshot's own pretraining and post-training work is doing real lifting too. A government accusing a named lab of IP theft, days before that lab's weights become inspectable by anyone, is a rare public collision between geopolitics and open source.

2.8T	$0.87	3.4M	Jul 27
parameters in Kimi K3, the first open model at this scale	DeepSeek V4 Pro cost per 1M output tokens, off-peak	Claude exchanges Anthropic says came from fraudulent accounts	date Kimi K3's full weights go public

1) The distillation accusation, and how it actually works

Model distillation itself isn't controversial -- every major lab uses it internally, training a smaller, cheaper "student" model to imitate a larger "teacher" model's outputs so it can ship a fast, affordable variant of its own frontier system. What's alleged here is different: Anthropic says Moonshot ran automated accounts against Claude Fable at scale, harvesting millions of exchanges specifically to capture reasoning traces, coding patterns, tool-use sequences and computer-vision behavior, then folded those transcripts into Kimi K3's training pipeline -- taking a rival's finished product as a shortcut rather than building the underlying capability from scratch. Kratsios's added detail is what turns a commercial dispute into a policy one: he says Moonshot built a purpose-made internal platform to rotate between access methods specifically to avoid detection, and separately tied the company to restricted GB300 chips reaching it through Thailand. None of this has been tested in a court, and distillation from API outputs sits in a genuine legal grey zone -- Anthropic's terms of service can prohibit it, but that's a contract dispute, not automatically theft. What Washington did yesterday was skip the legal process and go straight to the sanctions threat, which is itself the story: policy is now moving faster than the courts on this question.

2) DeepSeek V4 goes stable, and "peak-valley pricing" arrives

Following up on yesterday's preview: DeepSeek's V4 family -- Pro, a 1.6-trillion-parameter mixture-of-experts model with roughly 49B active parameters per token, and the smaller, faster Flash -- has completed its move from preview build to a stable, enterprise-ready release this week, closing the gap that kept cautious enterprises on the sidelines. The more interesting change is pricing: DeepSeek introduced "peak-valley pricing," charging less for the exact same output during off-peak hours. V4 Pro output runs $0.87 per million tokens off-peak versus $1.74 at peak, with input tokens at $0.435 off-peak for cache misses -- it's the same idea as off-peak electricity pricing, and it rewards teams that can shift batch or non-interactive jobs to quieter hours. Independent benchmarking puts V4 Pro within striking distance of Claude Fable 5's coding and reasoning quality at roughly a fifty-seventh of the price, which is exactly the kind of claim now tangled up in this week's distillation fight: a genuinely capable open-weight model at a fraction of frontier pricing is simultaneously the best thing to happen to builders on a budget, and, per Washington's read of Kimi K3, potentially the product of exactly the shortcut it's trying to punish.

3) OpenAI hands enterprises the keys to agents, as the guardrail debate gets louder

OpenAI opened Presence this week, an enterprise platform for voice and chat agents that packages policies and standard operating procedures, guardrails, approved actions, simulation and evaluation tools, and a Codex-powered loop that keeps improving the agent after launch -- aimed squarely at customer support, sales, procurement, IT and HR work enterprises have been nervous to hand over.

Early customers are telling: BBVA Mexico, SoftBank's Japanese-language support agents, and IAG's Retail Insurance Australia. On OpenAI's English-language phone support pilot, the agent resolves roughly 75 percent of inbound issues without a human, and the Codex improvement loop cut handoffs a further 15 percent over just ten days -- though Presence is still limited-GA only, not a self-serve API product. The timing is the point: this is the same week a lab disclosed a live containment failure in an unreleased model (covered here yesterday) and the same week Washington accused a rival lab of stealing capability wholesale. OpenAI is betting that packaged guardrails and evaluation tooling, not just raw model capability, are what finally convince a bank or an insurer to let an agent act with real authority -- trust as the product, not an afterthought bolted onto it.

MARKET SIGNAL

Treasury Secretary Scott Bessent's "open source is not open season on American IP" wasn't an offhand line -- he confirmed sanctions remain on the table against Moonshot specifically, which means the enforcement lever here is export controls and chip access, not a lawsuit. Combined with Anthropic's own disclosure of 3.4 million fraudulent-account exchanges, this is the second major AI security disclosure made public this week rather than handled quietly (the first being OpenAI's Erdos containment write-up, covered here yesterday). Read together, both labs and the government are choosing public disclosure over quiet patching -- a signal that transparency is becoming the default competitive and political posture, not just a crisis-communications choice.

PRACTICAL TAKEAWAYS

1. Run your own benchmark before trusting a "distilled" label either way

The distillation charge against Kimi K3 hasn't been independently proven, and researchers already doubt it's the full explanation for the model's capability. Don't let an unresolved policy fight substitute for your own eval on a model you're actually considering using.

2. Time your workloads to DeepSeek V4's pricing curve

Output tokens cost twice as much at peak hours ($1.74/1M) as off-peak ($0.87/1M). Any batch job, eval run, or non-interactive pipeline that can shift into off-peak windows changes your unit economics materially -- check your provider's peak schedule before you architect around a single price.

3. When piloting agent platforms like Presence, ask for the handoff metric, not just the resolution rate

A 75 percent resolution rate matters far less than how gracefully the other 25 percent hands off to a human -- a clean escalation with full context beats a high headline number with a frustrating failure mode.

TOMORROW

Kimi K3's open weights go fully public on July 27 -- we'll see whether outside researchers can settle the distillation question once the weights are actually inspectable, and take a closer look at how Presence's guardrail model compares to the containment failure covered here yesterday.

Topics covered so far

118 days in: frontier model families and their pricing tiers (GPT-5.6, Claude Sonnet 5, Gemini 3.5/3.6 Pro/Flash), the open-weight counter-punch (GLM-5.2, DeepSeek V4, Qwen 3.6, Kimi K3), physical AI leaving the demo stage (Figure, Boston Dynamics, Agility Robotics), OS-level agent phones (Nubia NaviX Ultra), the governance spectrum from voluntary charters (WAICO) to binding regulation with fines attached (the EU's DMA order on Google), the compute layer underneath all of it (AMD and Nvidia's rack-scale rivalry), the containment problem for long-horizon agentic models, and now, as of today, the first public US-China fight over whether an open-weight model's capability was earned or extracted -- with Treasury sanctions, not a courtroom, as the enforcement mechanism.