RV Pro

Steno Signals

with Andreas Steno

Steno Research

July 20, 2026

Steno Signals: Say Hello to Kimi. She is going to increase the GPU and memory demand

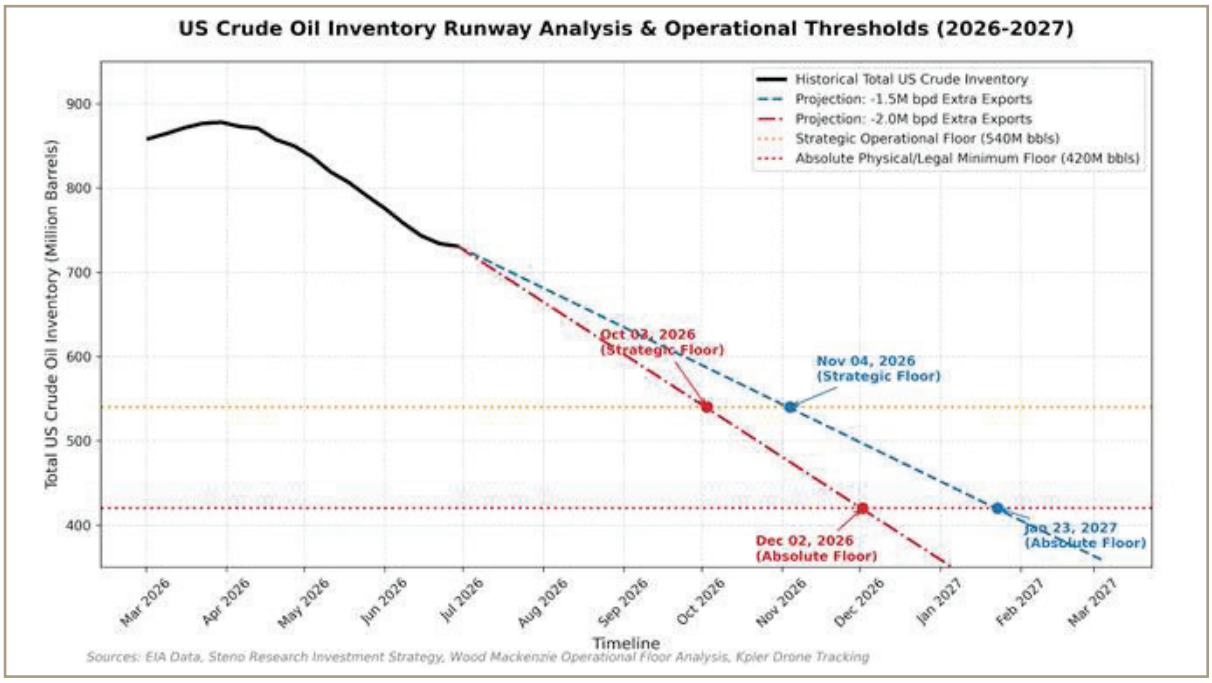
First, let's spend a few seconds on what actually worries me, and then we will get to our new friend Kimi from China, which does not worry me at all. At least if you are not a holder of Anthropic or OpenAI stock, where it is a potential issue for them, but not for the hardware trade.

After a week of escalating attacks, the weekend saw two U.S. servicemen killed and heavy attacks on Iranian Republican Guard (IRGC) headquarters in Tehran as well as widening attacks on Gulf nations. We are clearly in an escalatory cycle where both sides are upping the ante bit by bit. However, we still see no signs of the U.S. preparing for the big escalation, which would be a ground invasion. Such an undertaking would take half a million troops and 2-3 months to prepare, so it's not something they just decide to do.

We are always just one phone call from Pakistani mediators and one Truth post away from a complete de-escalation. However, both sides seem set on fighting it out for at least a few more weeks right now. The IRGC banked more than $2 billion in the 4-week ceasefire before they decided to disrupt the peace and restart the war. They have a war chest for another 2-3 months. Meanwhile, the U.S. doesn't have any way of forcing compliance. Few worthwhile targets are remaining in Iran to bomb, and despite claims to the contrary, they have evidently not managed to destroy Iran's military capabilities.

So the solution is, once again, negotiations, which are very tricky because the IRGC is quite content with this conflict going on for months. I only truly see a solution if China gets involved, or if Trump completely TACOs again. And there is still time to be patient on both sides, as the oil math looks better for BOTH the West and IRGC after that window of opportunity was used by both sides during the ceasefire. The crude oil inventory situation is deteriorating faster than most realize, and the clock is ticking toward a forced resolution.

Chart 1: U.S. crude oil inventory is running down fast, and the clock is ticking

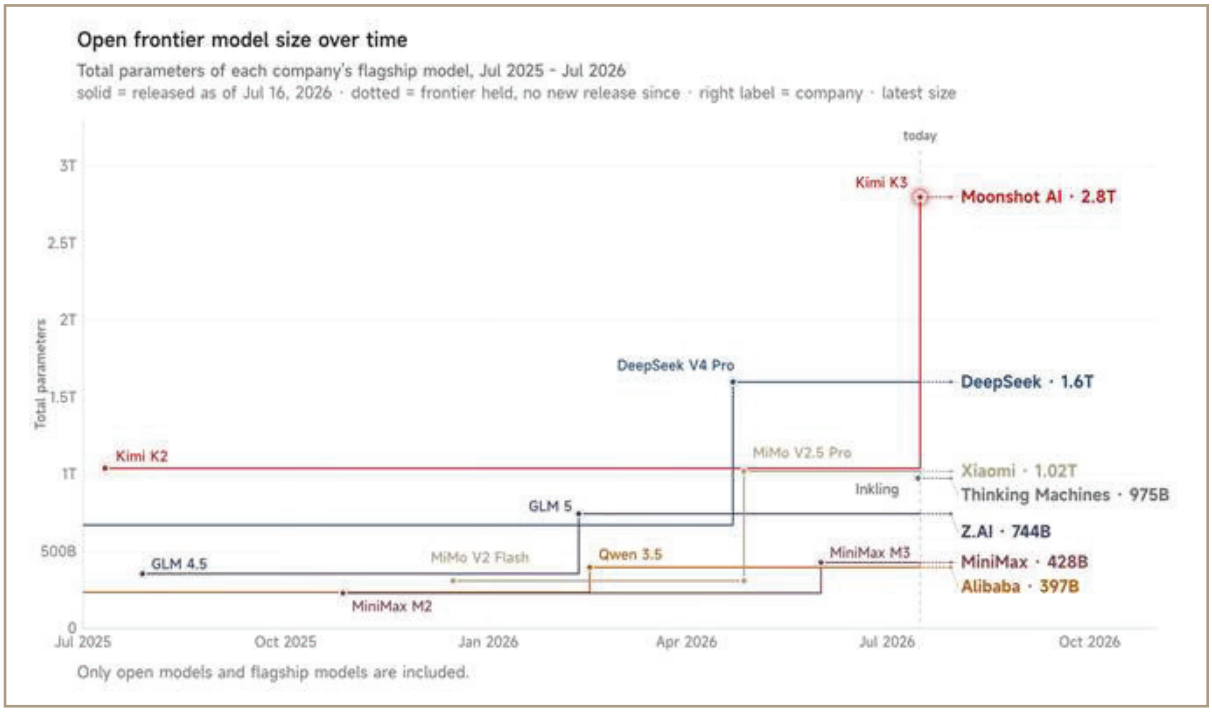


Now to Kimi, and why nothing structurally new happened last week.

Moonshot AI dropped Kimi K3 on July 17, and markets immediately lost their minds. The Philadelphia Semiconductor Index shed 12.5% on the week, its worst week in 15 months. Nvidia fell. TSMC dropped 7%. SoftBank fell 9%. The "DeepSeek moment" narrative was dusted off and served up hot. But here is the thing: if you look at the actual data, nothing structurally new happened. At all.

Yes, K3 is the largest open-weight model ever released at 2.8 trillion parameters. Yes, it runs competitive benchmarks. Yes, it is cheaper to access via API than Claude Fable 5. But the underlying dynamic is identical to every prior Chinese AI release: a well-funded startup replicates, fine-tunes, and scales up techniques pioneered at U.S. labs and delivers a model that, yes, may be good, but is ultimately neither innovative nor improved. The chart below shows the model size race over the past year. Moonshot went from Kimi K2 at 1 trillion parameters to K3 at 2.8 trillion. Impressive. But DeepSeek V4 Pro was already at 1.6 trillion before this. The parameter arms race is real, but it is not a surprise.

Chart 2: Open frontier model size over time. Kimi K3 is now the biggest, but the U.S. is still funding the next generation

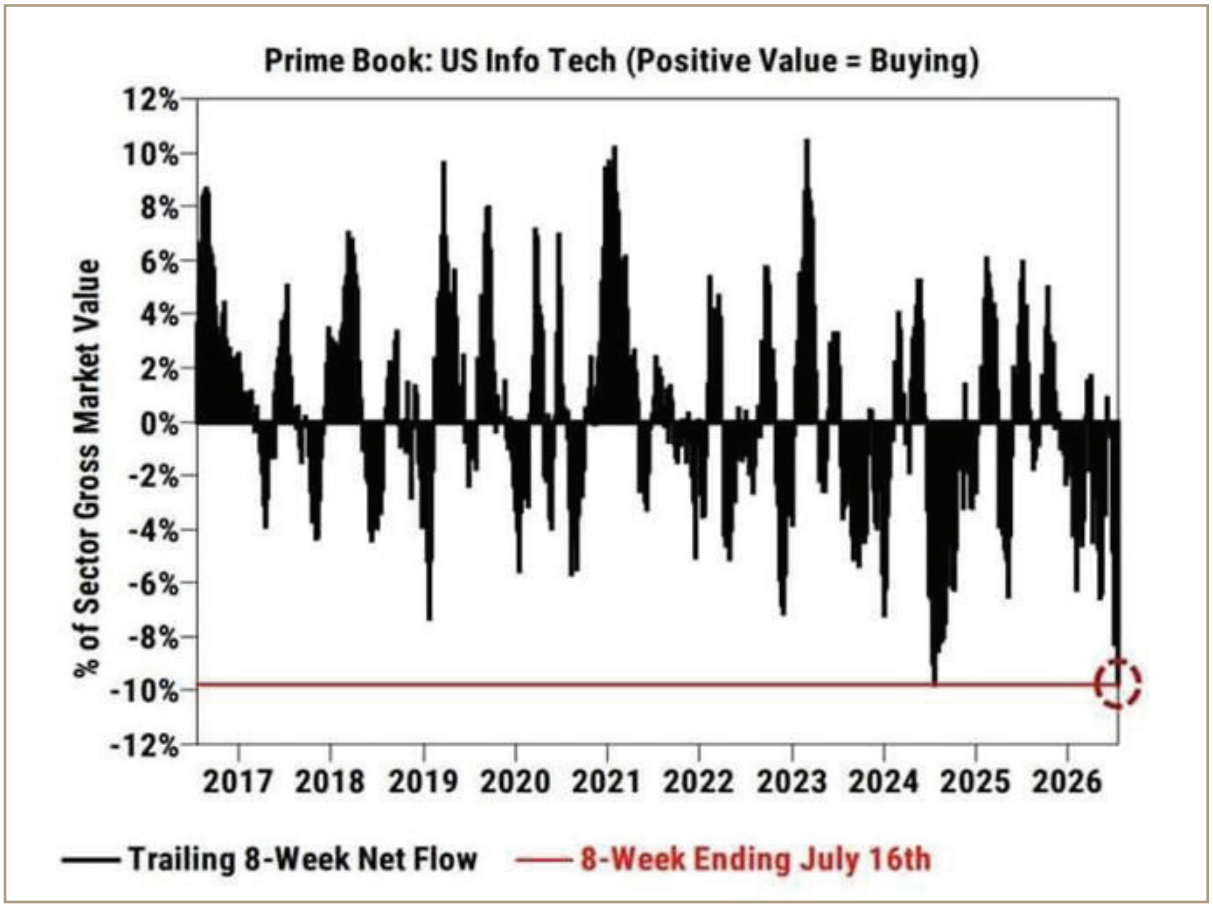


The more important context is investment. The U.S. AI ecosystem (Anthropic, OpenAI, Google, Meta, and Microsoft) is spending north of $300 billion on AI infrastructure in 2026 alone. Moonshot's total funding is a rounding error by comparison. China's aggregate AI investment is substantial in absolute terms but represents a fraction of the U.S. effort at the frontier. You cannot close a 10x funding gap with clever engineering alone, and the benchmarks reflect exactly that.

Hedge funds were already positioning for exactly this kind of panic. The Goldman prime book had been running the largest 8-week net outflow from U.S. Info Tech since their data series begins, with the 8-week flow ending July 16th breaching the -10% gross market value threshold, a level only seen twice in nearly a decade. In other words, the professional money was already short tech in size before K3 dropped. The model release gave them a narrative to drive it further, not a new reason to be bearish.

The irony is that the smart short-side argument against tech right now has nothing to do with Chinese AI capability. It has to do with the YoY earnings wall of worry for semiconductor companies and positioning crowding. K3 is a convenient story, not the actual thesis. When the short covering comes, it will come fast, but it will take some time, as we need volatility to soften first (Warsh needs to pivot).

Chart 3: Hedge funds were already the most short U.S. Info Tech in a decade. The K3 release was a catalyst, not a cause

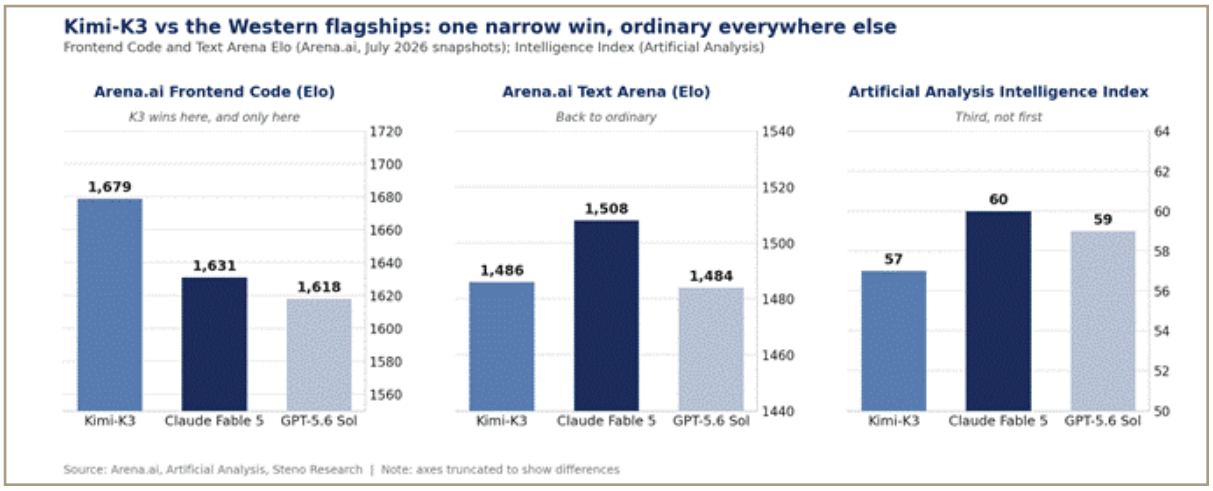


What the benchmarks actually show: one narrow win, ordinary everywhere else

Let's look at what K3 actually achieved, because the details are far less impressive than the headlines. On Arena.ai's Frontend Code Arena, K3 takes first place with an Elo of 1,679 versus Claude Fable 5's 1,631. That is a real win, and I will give Moonshot credit for it. But then look at the other two panels.

On Arena.ai's Text Arena, which reflects general-purpose quality as judged by millions of blind human pairwise votes, K3 scores 1,486 Elo, finishing behind Claude Fable 5 at 1,508 and level with GPT-5.6 Sol at 1,484. And on the Artificial Analysis Intelligence Index, which aggregates performance across the broadest set of tasks, K3 sits at 57, third place, behind Fable 5 at 60 and GPT-5.6 Sol at 59. This is a model that leads in one specific domain and finishes third overall. That is not a paradigm shift. That is incremental progress at the tail end of the gap.

Chart 4: Kimi K3 vs the Western flagships: one narrow win, ordinary everywhere else



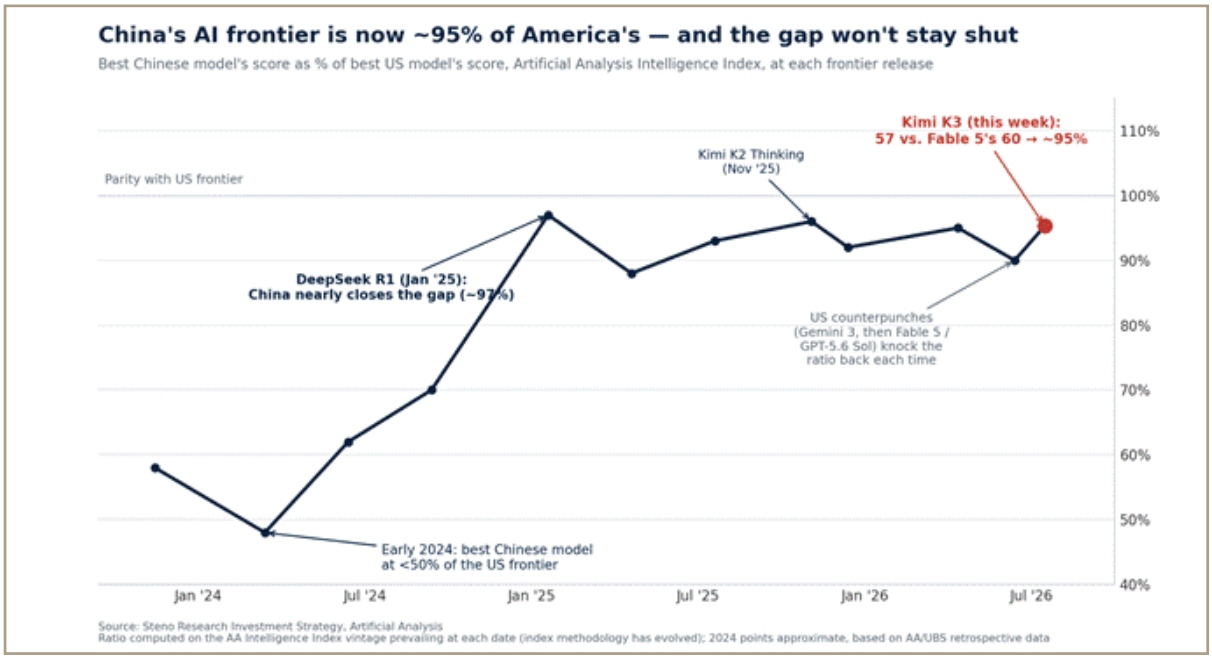
China is still fighting for the last 5-7%, and the gap keeps snapping back

The longer historical view is the most important chart in this piece. In early 2024, the best Chinese model sat below 50% of the U.S. frontier on the Artificial Analysis Intelligence Index. DeepSeek R1 in January 2025 brought that ratio to nearly 97%, triggering the first great "AI panic" selloff. Since then, the U.S. has counterpunched with Gemini 3, Claude Fable 5, and GPT-5.6 Sol, knocking the ratio back each time. Kimi K3 brings China back to roughly 95%.

Notice the pattern. China closes the gap. The U.S. releases the next generation. The gap reopens. China closes again. This is not convergence to parity. It is a treadmill. And the reason is simple: you cannot replicate what has not been published yet. Chinese labs are extraordinarily good at absorbing, adapting, and scaling published research from U.S. labs. DeepSeek's MoE architecture, the attention mechanisms, the training recipes all draw heavily from work done at Google, Meta, and the U.S. academic ecosystem. When the U.S. frontier moves, China follows it. When the U.S. frontier stays still, China catches up. The frontier has not been still.

There is also the compute constraint. Despite enormous domestic investment, China's access to leading-edge GPU silicon remains restricted. CXMT, China's HBM producer, is still at 25-45% yield on HBM3, well behind SK Hynix and Micron. The training runs that will define the next generation of frontier models require hardware that China simply cannot procure in the same quantities or quality. The gap at 5-7% on the intelligence index may look small. In practice, for the most demanding reasoning and agentic tasks, that 5-7% is enormous.

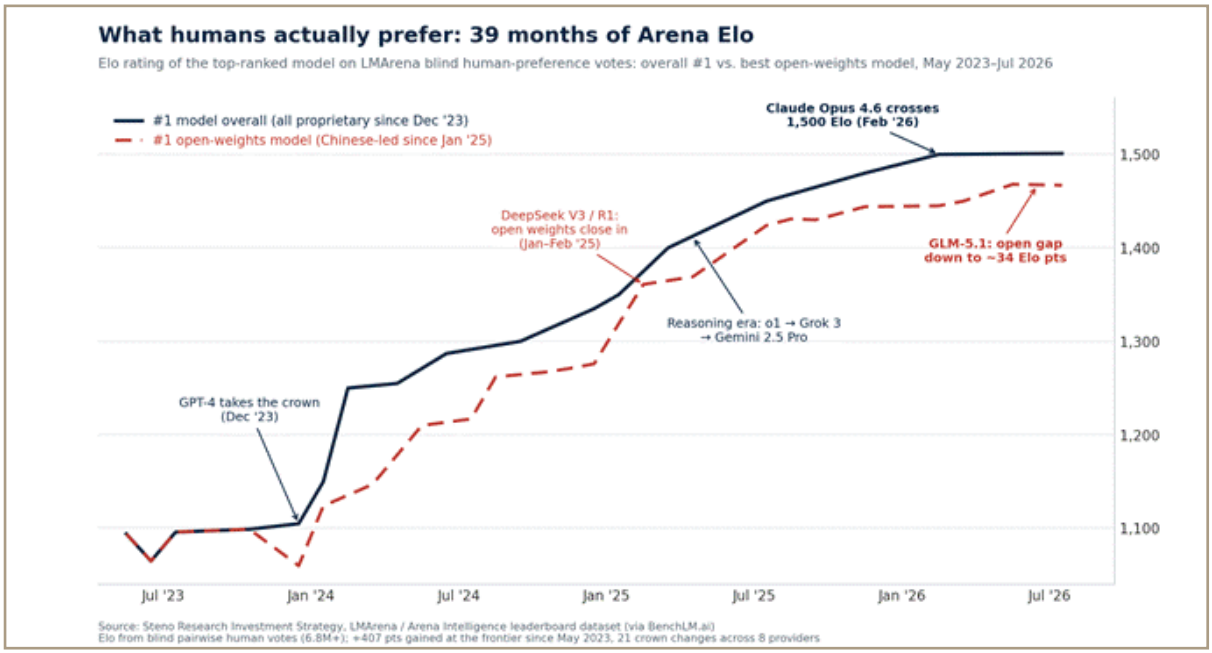
Chart 5: China's AI frontier is now ~95% of America's, but the gap keeps snapping back every time the U.S. counterpunches



The human preference data tells the same story over 39 months. Every model that has held the #1 spot on LMArena's blind pairwise vote since December 2023 has been a proprietary U.S. model. The open-weights gap, Chinese-led since January 2025, has closed from roughly 300 Elo points to about 34 Elo points today with GLM-5.1. K3 is not at the top of the open-weights leaderboard. GLM-5.1 is. And even GLM-5.1 is 34 points behind the proprietary frontier. That gap sounds small. On a 1,500-point scale where individual model quality differences are measured in single digits, 34 points is actually significant.

The key point is that Chinese open-weights models have been the best open-weights models for 18 months, and they still have not cracked the overall #1 position. The reason is that the compute and data advantages of closed-source training runs are not replicable through fine-tuning and architecture tricks alone. K3 is a fantastic open-weights model. It is not a threat to the closed-source frontier.

Chart 6: What humans actually prefer: 39 months of Arena Elo, and every #1 is still proprietary

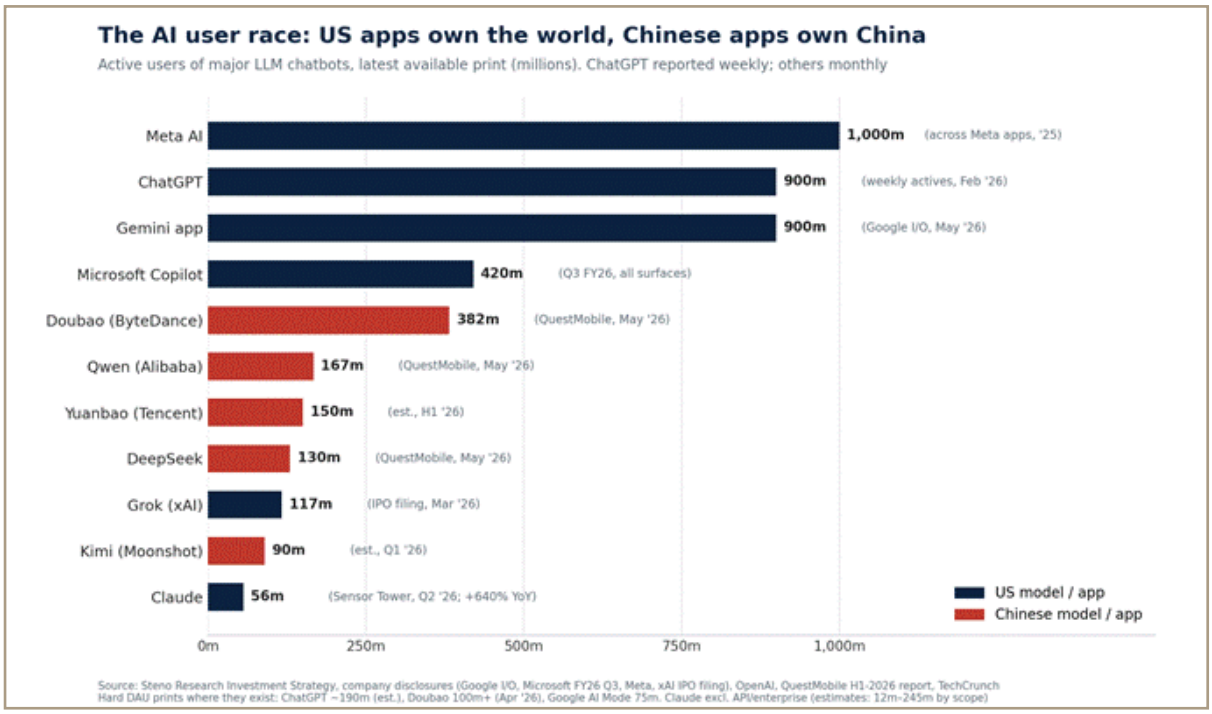


The user race: Chinese apps dominate China, and nowhere else

The distribution reality is important context for anyone trying to assess the commercial threat from Kimi. ChatGPT has 900 million weekly active users. Gemini has 900 million. Meta AI has 1 billion across its app surfaces. Microsoft Copilot has 420 million. These are structural distribution advantages that no open-weight model release changes overnight. Kimi (Moonshot) has an estimated 90 million users, almost entirely in China. Claude has 56 million, up 640% year-on-year but still well behind the major consumer platforms.

The Chinese AI apps (Doubao, Qwen, Yuanbao, DeepSeek, Kimi) dominate within China because they have access to the Chinese market through WeChat integrations, local app stores, and regulatory environments that effectively block the U.S. competitors. Outside China, they have minimal consumer footprint. This is the same pattern seen in social media, e-commerce, and search. China builds world-class products for the Chinese market. The global consumer AI market remains almost entirely U.S.-controlled.

Chart 7: The AI user race: U.S. apps own the world, Chinese apps own China

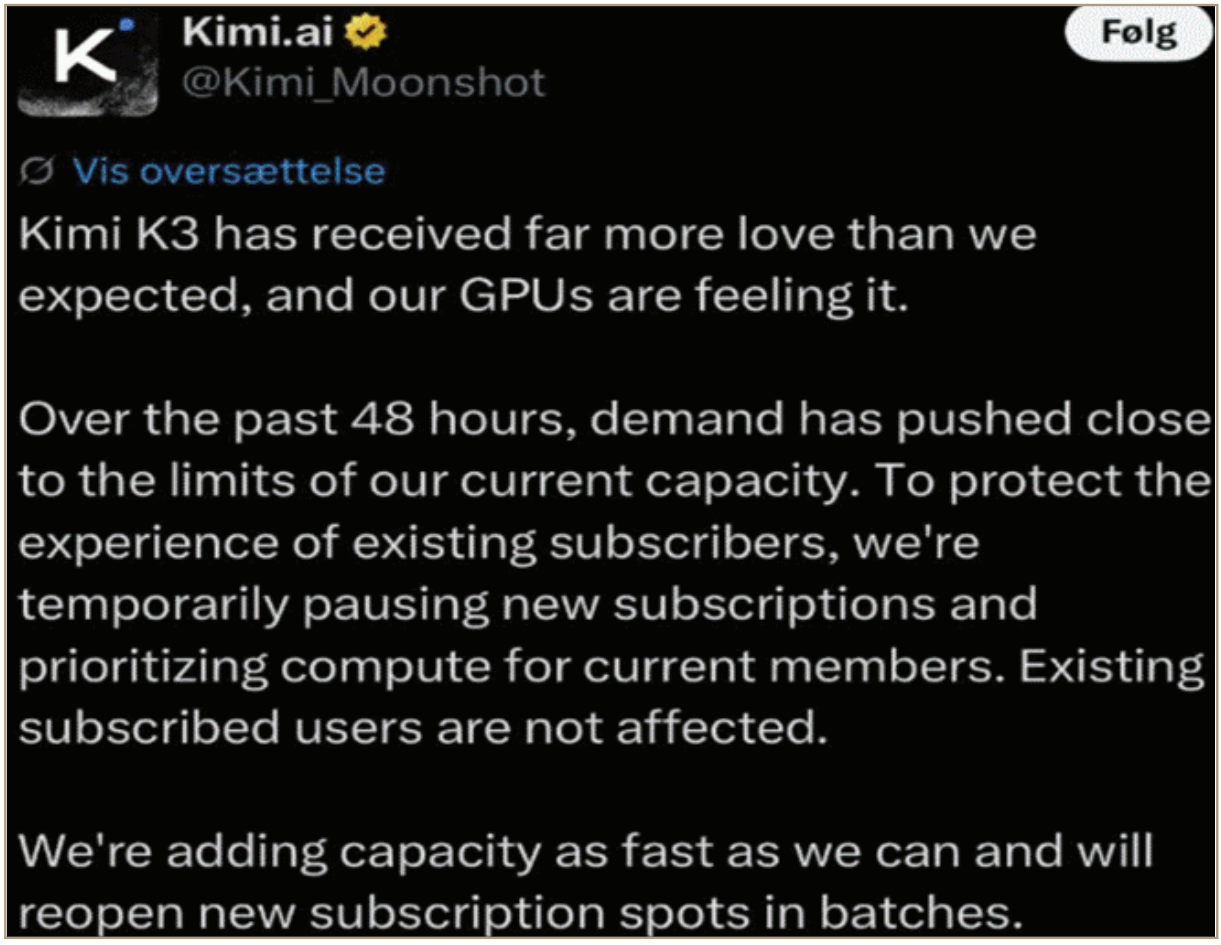


The hardware bull case: Moonshot ran out of GPUs in 48 hours

Now we get to the part that the market completely got backwards. Kimi K3 launched on July 17. By July 19, less than 48 hours later, Moonshot had to pause new subscriptions entirely. The tweet below is the only data point you need to understand why the semi-selloff was a mistake. Read it carefully: “our GPUs are feeling it… demand has pushed close to the limits of our current capacity… we’re temporarily pausing new subscriptions.”

This is not the behavior of a company that has figured out how to do AI with less compute. This is the behavior of a company whose new model is so popular it immediately overwhelmed their entire GPU fleet. A cheaper, more capable model does not reduce hardware demand. It explodes it. This is Jevons Paradox in its most literal form: efficiency gains lower the cost of consumption, which drives up total consumption. Every time a new cheap frontier model ships, the addressable market for inference compute expands. Every time it expands, the hyperscalers and AI startups buy more GPUs. This cycle has been running for three years and shows no sign of stopping.

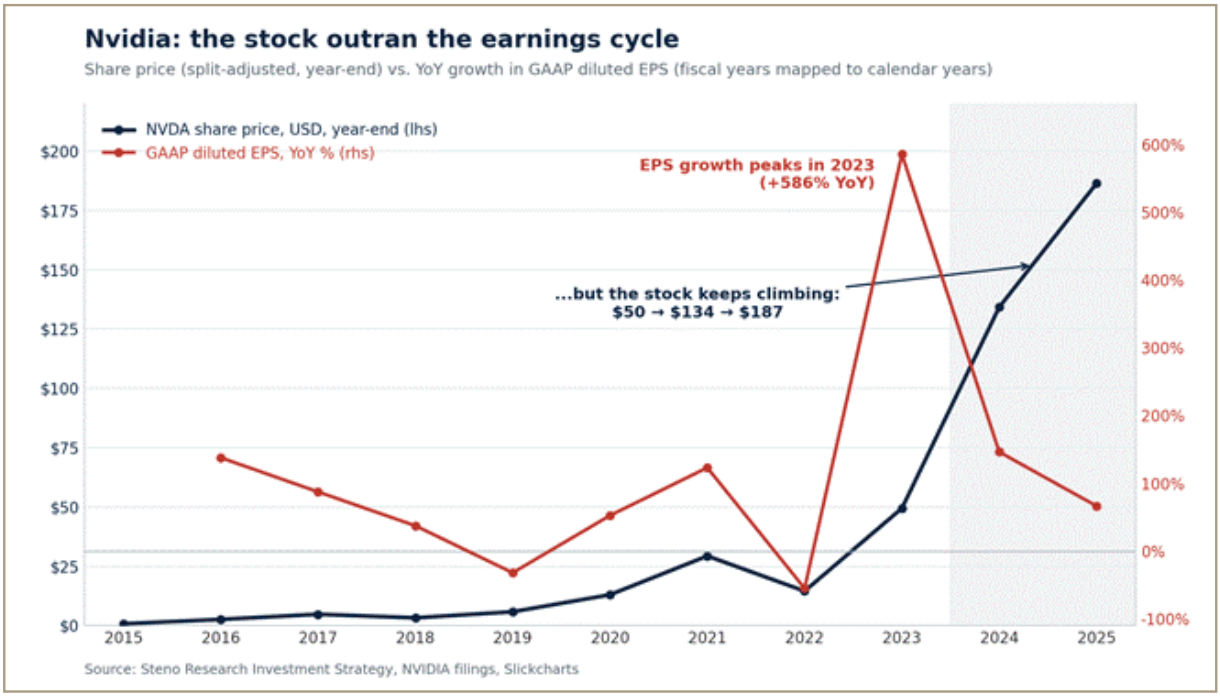
Chart 8: Moonshot's own tweet: GPU capacity full in 48 hours. This is bullish for hardware, not bearish



Moonshot's ARR hit $300 million in June 2026, and the company is in active talks to push its valuation past $30 billion. At that kind of growth trajectory, they are not looking for ways to buy fewer GPUs. They are desperately trying to buy more. Every analyst who looked at K3 and sold Nvidia should look at this tweet and ask whether they had the causality backwards. The same playbook played out with DeepSeek R1 in January 2025: Nvidia fell 17% on the day of the release, recovered fully within weeks, and then went on to new highs as it became obvious that cheap inference drives volume, not restraint.

Nvidia's story is not about any individual model release. It is about a structural capex supercycle driven by the fact that every new application of AI, whether it comes from San Francisco or Beijing, requires more compute at scale. K3 is not the end of that cycle. It is another lap of it.

Chart 9: Nvidia: the stock keeps outrunning the earnings cycle, and the K3 selloff is just another entry point

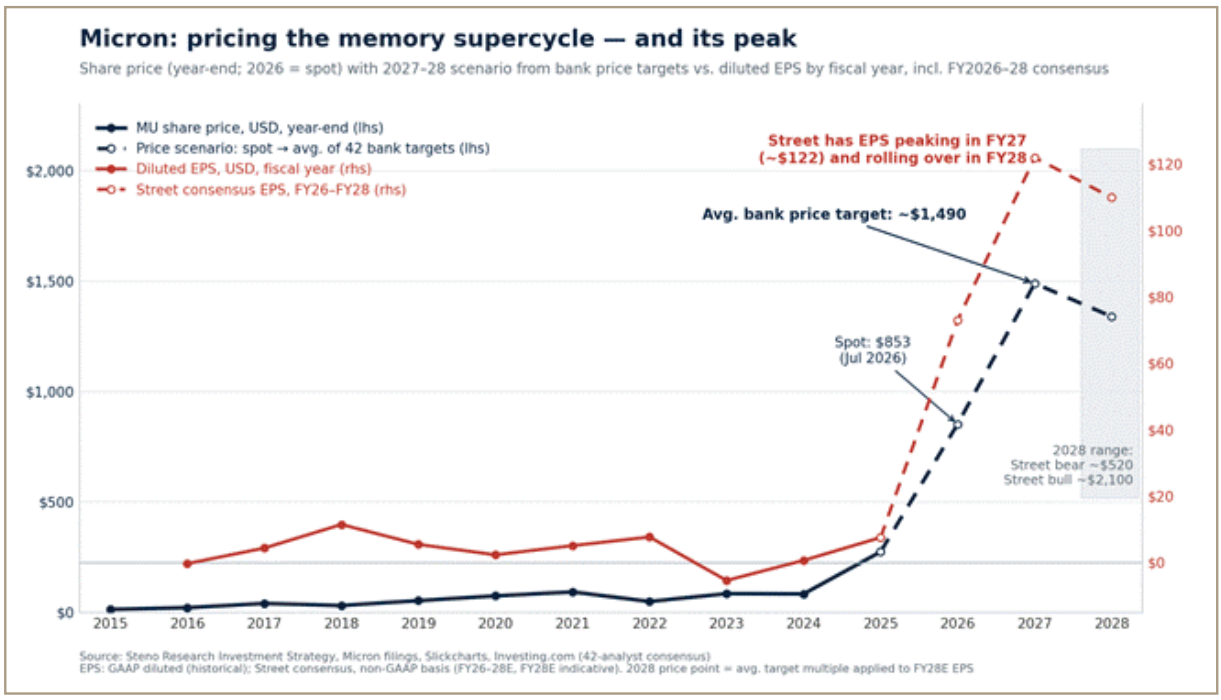


The memory angle: this is as much about HBM as it is about GPUs

Bloomberg's headline on the same day as the tech rout read: "Moonshot's Kimi K3 May Be More About Memory Than Compute." This is the correct framing, and it deserves more attention than it got. A 2.8 trillion parameter MoE model with a 1 million token context window has enormous High Bandwidth Memory requirements at inference. You cannot run K3 efficiently without stacking HBM. You cannot serve K3 at scale, the kind of scale that runs out of subscriptions in 48 hours, without buying very large quantities of SK Hynix HBM3E and Micron HBM4.

The memory supercycle thesis rests on exactly this dynamic. Every new large language model that ships with a massive context window and MoE architecture is an HBM demand event. The street has Micron EPS peaking in FY27 at around $122 and rolling over in FY28. With an average bank price target of approximately $1,490 versus a spot price around $853 in July 2026, the memory supercycle trade still has substantial upside embedded in consensus, and that is before accounting for the possibility that models like K3 extend the demand runway beyond what consensus currently models.

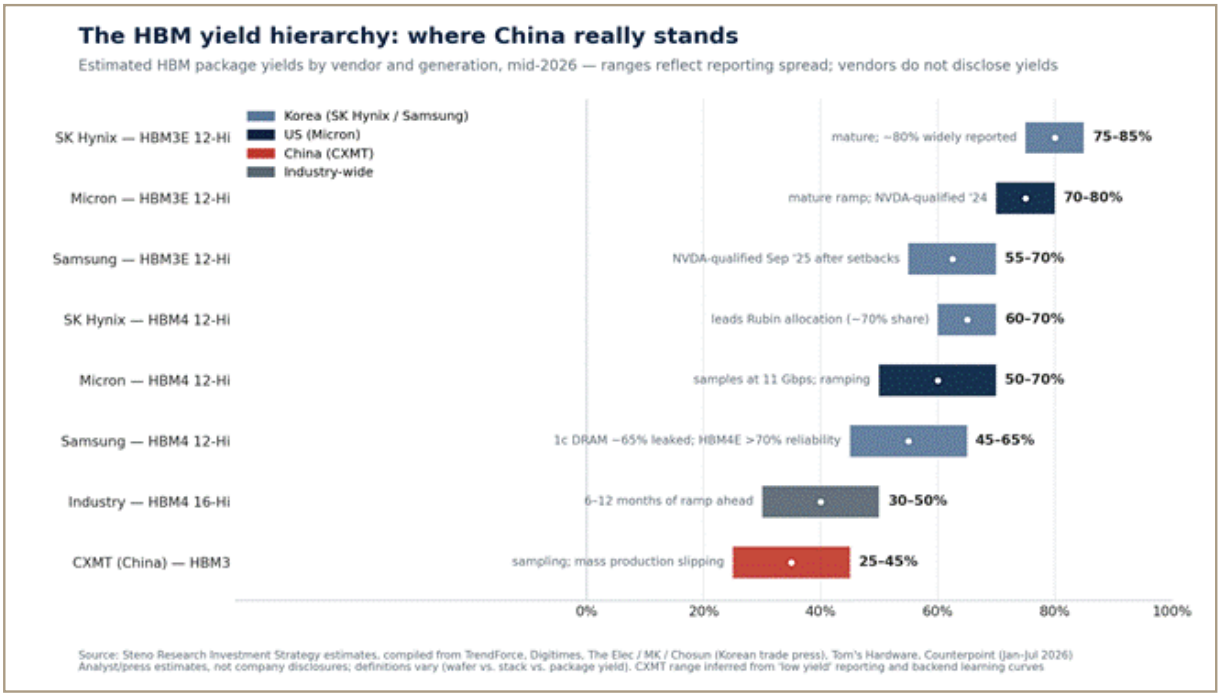
Chart 10: Micron: pricing the memory supercycle, and the consensus has it peaking in FY27



China's own attempt to break into the HBM supply chain is instructive here. CXMT, the primary Chinese HBM producer, is still achieving only 25-45% package yield on HBM3, an older generation that SK Hynix has been shipping at 75-85% yield for years. Mass production timelines keep slipping. Analysts who initially expected CXMT to be sampling HBM4 by late 2026 have pushed those estimates out repeatedly. The yield gap between CXMT and the Korean/U.S. producers is not a minor lag. It is a structural capability gap that reflects the difference between having export-controlled EUV equipment and not having it.

This matters because it means the bottleneck in the HBM supply chain, the component that Kimi K3 and every model like it needs most, remains entirely in the hands of SK Hynix, Micron, and Samsung. Chinese AI progress does not change that. If anything, it tightens it further, because every large open-weight model that gets released and goes viral puts additional demand pressure on HBM supply chains that China cannot easily supplement from domestic sources.

Chart 11: The HBM yield hierarchy: CXMT is nowhere near the frontier, and that bottleneck is your investment thesis



Where this IS a problem: Anthropic is most at risk, and Claude is the reason why

Let me be direct about the competitive dynamic that actually matters here, because I think the market is focused on the wrong victim. The conventional wisdom after every Chinese model release is that OpenAI is the primary target. But OpenAI has 900 million weekly ChatGPT users, a deeply embedded consumer brand, a Microsoft enterprise distribution machine, and a model portfolio that spans from GPT-4o mini to o3. They have pricing power, distribution, and brand loyalty that no open-weight model threatens in the near term.

Anthropic is in a structurally different position. Claude does not have a 900 million user consumer app. Claude's dominant use case, by a wide margin, is coding. Cursor, which runs Claude as its primary model, has become the most popular AI coding tool in the world. The entire Claude Fable 5 value proposition for enterprise developers is: best-in-class reasoning applied to coding tasks. That is exactly the one benchmark category where K3 beats Fable 5 outright. And K3 does it at a third of the price.

Think about the decision framework for an enterprise developer or a startup building a coding workflow in July 2026. Claude Fable 5 costs $10 per million input tokens and $50 per million output tokens. Kimi K3 costs $3 input and $15 output. ChatGPT sits at $2.50 input and $15 output. Fable 5 is 3-4x more expensive than every serious alternative on input, and more than 3x pricier on output.

For a company running millions of coding agent turns per day, that price differential is not a minor consideration. It is the difference between a viable unit economics model and an unsustainable one.

The counter-argument is that Fable 5's quality advantage justifies the premium. And on most benchmarks, that holds: Fable 5 leads on SWE-bench Verified at 95% versus K3's 60.4%, and dominates on hard problem categories. But enterprise developers are not running the hardest possible problems every turn. The majority of coding agent workloads (autocomplete, boilerplate generation, test writing, debugging straightforward bugs) are well within K3's capability envelope. "Good enough at a third of the price" is a winning pitch in every enterprise sales conversation I have ever seen.

The Anthropic IPO narrative, which was already being stress-tested by expectations of OpenAI's own public debut, gets substantially more complicated every time a frontier-quality model appears at this kind of pricing differential in Claude's core use case. DeepSeek was the dress rehearsal for this conversation. Kimi K3 is the main event. Not because K3 is definitively better than Claude. It is not, overall. But because it is close enough, in the right domain, at the wrong price for Anthropic to ignore.

The bottom line

The tech rout last week was a fear-driven overreaction to a development that, on close inspection, confirms rather than undermines the AI capex supercycle. Nothing structurally new happened with Kimi. China is still fighting for the last 5-7% of the intelligence gap, with materially less investment, on hardware that remains export-controlled and domestically constrained. They copy, they iterate, they ship impressive products, and they reliably trigger panics that reliably become buying opportunities.

For hardware investors, the trade is clear: buy the dip. Moonshot ran out of GPUs in 48 hours. Micron and SK Hynix supply the HBM that every large model needs at scale. The memory supercycle has legs that no amount of Chinese model progress shortens.

For those watching the model providers, the picture is more nuanced. OpenAI is insulated by distribution. Google is insulated by integration. Anthropic's insulation comes from coding quality, and that is precisely what K3 has now partially eroded. Claude remains a world-class product. But Kimi just made the pricing conversation a lot harder for Anthropic's enterprise sales team. Welcome to Kimi. She is not going to destroy the GPU cycle. She is going to feed it.

Andreas Steno

Andreas is the core macro consultant in Steno Research. Building on years of experience as Global Chief Strategist at Nordea Bank, he is one of the most quoted and sought-after macro analysts out there. Andreas' expertise is the FX/rates, energy, real estate and equity spaces, but doesn't shy away from hot takes on other topics if the underlying analysis is strong enough. Andreas anchors the weekly editorial 'Steno Signals'.

RV® Pro