JULY UPDATE — IS KIMI K3 AI SPUTNIK, OR AI SUICIDE?

The AI Ouroboros Just Nicked Its Own Jugular

On Kimi K3, distillation, and the nature of unsustainable economics. | July 20, 2026

You may be familiar with the ouroboros, which is the old symbol of the snake eating its own tail. It usually gets drawn as a picture of eternal renewal, but when we look a bit closer, we may see something else. The thing I'm calling the "AI Ouroboros" is consuming itself to stay alive, and it has finally bitten down on something it needs to stay viable.

As we've been predicting during this several-year-long frenzy to "AI ALL THE THINGS" using LLMs, which we are increasingly confident is a dead end for AGI and ASI, what's happening now is the predictable end state of an unsustainable path, and unsustainable practices always end.

The only open question is how. It doesn't mean 'AI is dead', or any of that doomer nonsense. But it DOES mean that the AI industry and humanity's approach to LLM-based AI, AGI, and ASI is about to change, one way or another.

HOW THE OUROBOROS WORKS TODAY: WHAT FEEDS THE SNAKE

If you're reading my newsletter, you probably know at least the basics of how the frontier AI models are built:

- US labs raise and burn hundreds of billions of investor dollars to train each generation of frontier models.
- Then they price inference below cost to seed adoption, because a model nobody depends on is a failure and a write-off. Remember, empowerment is the sales pitch, while DEPENDENCY is the business model. They have to get you hooked, drive engagement, etc.
    - These pricing strategies set their own flawed incentives along the way, resulting in phenomena like 'tokenmaxxing', which just burns resources because people follow incentives, even dumb ones.
- To convert that dependency into revenue, they ship *forward deployed engineers* into enterprise accounts. The FDE's real job is not integration, it's to seed dependency and wire the model(s) and harnesses from that vendor **so deeply** into a customer's workflow that leaving becomes unthinkable, and then
- End subsidies and manufacture enough visible ARR (ROI is another matter) to justify the next funding round.

It is a coherent, if flawed strategy. Subsidize the habit, embed the dealer, raise cash on the growth curve. Then IPO and get your liquidity while the dumb money piles in at the end, with the Frontier Labs now "Too Big to Fail", and the US Government is forced to swoop in and socialize losses, on the basis of 'national security'.

However, it is not sustainable. This bubble is going to burst, and the flaws in this approach are fatal.

THE LEAK IS THE FRONT DOOR, NOT THE BROKEN WINDOW

The moment the labs expose a frontier model through an API, they hand the world a teacher. Immediately, tens of thousands of bot accounts start pulling out as much information as they can, which is very hard to stop or even correlate across accounts and API keys.

Competitors and adversaries query your model, capture its outputs, and train their own models to imitate it. This is called ‘model theft’, using distillation, and it does not require breaking in. It requires nothing you didn’t sell; just some bot-herding and residential proxy services to mask and obscure traffic sources.

> ***You cannot rent out the answers and keep the intelligence. The answers are the intelligence.***

TL;DR - If I can use your model, I can eventually steal your model.

If I can send inputs, and harvest outputs, that becomes training data for student models. There are of course countermeasures like rate limiting and bot defense, but these have consistently been overpowered by the adversaries.

To be clear, distillation in and of itself is not a “hack”, it’s a valid machine learning technique that we know is being abused across the board.

In fact, adversaries don’t NEED to distill a model in order to attack it; they can just create their own, train it on a similar data set relevant to the domain in question, then create adversarial inputs that manipulate their own models in the ways desired. Some subset of these adversarial inputs will work against the target.

However, unauthorized distillation is absurdly common, and even with guardrails and AI red teaming in the Frontier Labs (ignore for a moment the silly 3rd party claims of “AI Guardrails” of their own classifiers in-line in front of your models. That’s security theater and a topic for another article), the attack surface for LLMs is effectively infinite.

Think about it for a second. If these companies who have been flirting with Trillion dollar valuations COULD stop the massive distillation campaigns (both from China and from each other), wouldn’t they? This is arguably the most valuable IP on the planet. They don’t stop it because they CAN’T.

This has always been the case, but has only recently been proven with a mathematical proof published in June of 2026. In production, AI systems can not be patched, only watched. Threat modeling and real-time monitoring are critical. This is exactly what happened with Fable and Mythos - AWS researchers showed they could jailbreak the model, the US Government intervened in a panic, and the labs' response? "Hey - no fair! Everyone else is vulnerable to this, too!"

Exactly.

However, while the realities of a mathematically infinite attack surface notwithstanding, that's not even the main problem driving this unraveling. What's killing the Ouroboros isn't a vulnerability, it's the product itself. Selling API access to a reasoning model and trying to prevent model theft is like a restaurant selling a popular dish to an Army of Chefs while trying to keep the recipes secret.

KIMI K3 AND THE 70 PERCENT DISCOUNT ON UNSAFE AI

On July 16, Moonshot AI released Kimi K3, a 2.8-trillion-parameter open-weight model and the largest ever released to the public. It took first place on Arena's Frontend Code benchmark, ahead of Claude Fable 5. On the broader agentic and coding suites it lands in the top tier, behind Fable 5 and GPT-5.6 Sol but ahead of Claude Opus 4.8. The full weights drop July 27, free to download and free to tune.

Cybersecurity folks, brace yourself for a busy August. And September. And….

Now the economics: hosted Kimi K3 runs $3 per million input tokens and $15 per million output. Fable 5 runs $10 and $50. That is roughly 70 percent cheaper on headline pricing, and on a real weighted task Artificial Analysis measured K3 at about $0.94 against $2.75 for Fable. There are some corner cases like cached input tokens that are 97-99% cheaper.

Frontier-adjacent performance at a third of the price, and by the end of July the weights will be released (barring government intervention) with the weights in your hands to run locally, and without the factory guardrails that are bogging down cybersecurity pros trying to use Fable to defend their organizations.

You won’t be able to run K3 on a laptop, but enterprises and adversaries with resources (including stolen and abused “free trial” cloud resources) will be able to run K3 on their own THIS MONTH.

You cannot subsidize your way to dependency when a free, tunable copy of the SAME capability is one download away.

TWO GUARDRAIL REGIMES, POINTED IN OPPOSITE DIRECTIONS

Unlike US models, a Chinese model clears regulatory review by satisfying TC260's basic security requirements, a standard built around 31 risk categories. The first and governing category is violation of core socialist values. The list polices content that endangers national security, harms the image of the state, or promotes what Beijing defines as false information. Read the list and notice what isn’t there. There are broad suggestions for, but no rigorous restrictions on cyber, biological, or nuclear risks. The CCP’s stated #1 risk is someone using AI to usurp their authority; the top priority is to keep the model on message.

US labs tend to guard the opposite axis. Fable 5 and Mythos 5 sort every cybersecurity request through a four-tier dual-use filter. Penetration testing, exploit development, privilege escalation, and high-uplift vulnerability discovery all sit in the high-risk tier and get blocked pending stronger authorization controls. The intent is sound, but in reality defenders and attackers ask for the same knowledge, so the same wall that slows an attacker also locks out the legit red teamer, the CTF player, and the CISO trying to verify a CVE is real during an argument with the head of software development about resource allocations. Security researchers have been saying exactly this since June.

This phenomenon also hit critical mass this week when Hugging Face was compromised by a fully autonomous AI attack, and when they tried to use closed American models in response, they gave up and adopted Chinese models instead.

Stack the two regimes and you get the trap. The American enterprise, paying per token, is handed a safety-throttled tool for defense.

The adversary downloads an open-weight model of comparable power, tuned to answer anything as long as it never insults the Party, and fine-tunes the political guardrails off in an afternoon.

We have optimized our models to hamstring the defender and optimized theirs to empower the attacker. For free.

In late April we wrote that "America's AI Incoherence is Driving American CISOs to China", but even we didn't realize it would be less than one calendar quarter before we were proven correct.

WEAPONIZATION IS NOW NEARLY TURNKEY

If you've read our previous issues, you know we've been calling this as inevitable.

In February, Nature Communications published a study in which four frontier reasoning models were pointed at nine target models as autonomous adversaries. Given a system prompt and no further human help, they planned and executed multi-turn jailbreaks with a 97.14 percent success rate. One capable reasoning model, acting alone, collapsed the entire cost curve of red-teaming. No cohort of prompt engineers. No gradient search. One model turned loose.

Now hand that same autonomous capability to anyone, in open weights, with the guardrails filed off, tunable to any purpose.

Further, if you do some analysis of the frequency of frontier - class releases, which went from 8 in 2025 to an average of one release every 10 days in 2026, we are on track for literally daily frontier - level releases by January, just six months from now. The power is increasing per model, and the rate of model release is accelerating at the same time.

For Defenders - whether Kimi K3, Inkling, or some other open weight model(s), it's increasingly clear that the coarse-grained guardrails from models like Fable prove too binding. We're not going to be able to depend on the frontier labs for cyber defense; it will simply be too asymmetrical to face adversaries with abliterated and tuned frontier - class models while Fable won't answer even basic questions on the topic anymore.

If you can even afford the token cost.

THE HOST NOTICES, TOO LATE

The US government has started to react, which is what a host does when the parasite reaches a critical organ. Congress has opened formal probes into US firms running Chinese models, the State Department has flagged enterprise adoption as a national security concern, and federal procurement bans are the leading proposal on the table.

The urgency is real, but the response is not effective. On OpenRouter, the share of tokens from US companies flowing to Chinese models reached 45 percent in early July, up from 11.5 percent at the start of the year. In fact, one could make a strong case that the US' incoherent AI strategy has handed China the advantage from Day 1, with export controls only forcing Chinese engineers to be more creative and efficient, while American labs continue to throw gobs of money at data and compute, without much focus on efficiency, so they can 'win the race' to the next leaderboard.

But, while you can bar a government agency from procuring a Chinese model, you can't "ban" a file already mirrored on a hundred thousand machines. Legal scholars have already flagged that publicly released model weights likely enjoy First Amendment protection, and policy analysts across the spectrum concede that banning open weights is, in their word, impossible. The host can refuse to swallow the parasite. It cannot un-swallow the parasite already inside it, and somehow we doubt that the CCP is terribly concerned about First Amendment protections for, well, anything.

NOW WHAT

The following turns from what has happened to what comes next, so weigh it as forecast rather than fact. --Ed.

Every link in the chain that funds the frontier is now under stress at the same time. Investors need a return that the current 'per-token subsidy plus inevitable distillation plus Chinese efficiency on newly capable Chinese chips' approach just can't produce.

Enterprises need capability for defense that their own safety regime restricts.

Governments need control over a diffusion mechanism that is, by construction, uncontrollable.

And nobody can maintain ANY semblance of governance or control with weekly or daily model releases.

Unsustainable arrangements like this are likely to end in one of three ways:

- The frontier labs consolidate around defensible moats that are not the model itself, meaning proprietary data, regulated & vertical (and STICKY) enterprise deployments, and hardware, and concede that raw model access is a commodity heading toward zero.
- Or - the U.S. government intervenes hard enough to wall off a domestic market, trading efficiency for control and accepting a slower, more expensive, more secure American AI. This effectively admits ‘defeat in the AI race’. But this is how the world works, for example, when it comes to pharmaceuticals. Americans pay more than anyone else in the world, because we fund the research. Everyone else gets generics at 1/10 the cost.
- Or - the subsidy wells simply run dry, the funding rounds stop clearing, and the correction happens the way corrections always do, all at once and to everyone's surprise but no one's excuse. The big labs have been trying to make themselves Too Big to Fail as fast as possible, and we’ll see if lawmakers agree during a political backlash around the midterms, with AI data center construction now less popular for average Americans than the NIMBY we faced with nuclear energy and power plants.

I for one think that the release of Open Weight Frontier performance is potentially both good for every stock ticker in America except the AI companies, and also a very dangerous moment in time for humanity.

I do not know which of the three we get.

I am fairly confident it is not a fourth option where the U.S. and China stop competing and start cooperating, or an option that would see the current playbook continue indefinitely.

That’s the thing about unsustainable practices. They end. The snake does not get to keep eating its own tail forever and call it growth just because it feels full.

The AI Ouroboros has been feeding on itself for years and calling the sensations ‘progress’. This month it nicked its own jugular.

What happens next is not so much a question of whether the bleeding stops. It won’t.

It is a question of whether the industry sees the wound in time to change what it is doing, or finds out the hard way that with the bleeding edge of technology, that blood has to come from somewhere, and sometimes that bleeding edge is on a knife you can find in your own right hand.

SOURCES

Kimi K3 release and benchmarks — VentureBeat

Kimi K3 API pricing — BenchLM

Claude Fable 5 / Mythos 5 pricing — Finout

Fable 5 cyber safeguards and jailbreak framework — Anthropic

Industry reactions to Fable 5 restrictions — SecurityWeek

China's AI safety evaluations and TC260 31 risks — AI Safety in China

Large reasoning models are autonomous jailbreak agents — Nature Communications

Congressional probes into Chinese AI use — Slashdot

Washington confronts China's open-source models, 45% token share — Semafor

White House accuses China of industrial-scale AI model theft — Nextgov/FCW

All opinions are my own.