Bob Got an Update: Kimi K3 and the AI Sputnik Moment

July 22, 2026 / Gavin Jackson

ai kimi k3 moonshot llm open-weights bob

Regular readers will know Bob - the AI assistant who helps me run this blog, manages my GitHub issues, and patiently tolerates my electronics questions. Well, Bob got an update this week. A big one.

Bob now runs on **Kimi K3**, the new model from Chinese lab Moonshot AI, released on 16 July. I was listening to the Moonshots podcast - Peter Diamandis and his crew called an *emergency episode* over it - and their framing was not subtle: they reckon America just had an **AI Sputnik moment**.

Having spent a few days with Bob's new brain, I think they might be right.

What actually is Kimi K3?

The headline numbers are absurd:

- **2.8 trillion parameters** - the largest open-weight model ever released. The previous Kimi K2.6 was 1 trillion, and that already seemed silly.
- **1 million token context window** - you can feed it an entire codebase, or a small library, in one prompt.
- **Natively multimodal** - text, images, screenshots, diagrams. This is apparently a big part of why it's so good at front-end work: it can actually *see* what it's building.
- **Open weights** - the full model weights are due to drop around 27 July. Anyone on Earth will be able to download it and run it themselves. More on that below.

Architecturally, the interesting thing (as the Moonshots panel pointed out) is that there's *no magic* in it. It's still recognisably a transformer - a mixture-of-experts design with their own brand of linearised attention and some very clever quantisation-aware training (MXFP4) that makes a 2.8T model economically servable at all. No secret post-transformer breakthrough. Just extremely good engineering, executed while under US export controls designed to starve them of advanced Nvidia chips.

That last part is the Sputnik bit. China was supposed to be compute-constrained. Instead, Moonshot engineered around the wall.

How does it stack up?

K3 jumped 17 places over the previous Kimi model to land **number one on the front-end code arena**, and also ranked first in six other domains: brand and marketing, reference-based design, data analytics, consumer products, simulations, and content creation.

Against the American frontier models, the independent benchmarks paint a consistent picture: K3 is genuinely *in the room* now. It scores 93.5% on GPQA Diamond (graduate-level science questions) - wedged between Anthropic's best and OpenAI's best. On Terminal-Bench it posts 88.3, a whisker behind GPT-5.6 Sol's 88.8 and ahead of the Claude models. On the long-horizon coding benchmarks it beats Claude Opus 4.8 outright, and head-to-head across the benchmarks both vendors report, it wins roughly as often as it loses against the very best closed models.

Twelve months ago the story was "Chinese models are catching up." This week the story is "an open-weight model is credibly inside the frontier conversation." That's a different world.

My own experience tracks with this. Bob's coding answers are noticeably sharper, and the writing - always Kimi's strength - is excellent.

The price: not the cheap-Chinese-AI story any more

Here's where it gets interesting. The old narrative was that Chinese models were 90% cheaper for 90% of the performance. K3 breaks that narrative in both directions.

The API pricing is **US$3 per million input tokens, US$15 per million output tokens**, with cache hits dropping input to US$0.30 - flat across the entire 1M-token context. That's roughly Claude Sonnet territory. It still undercuts the American flagships, but the days of "basically free" Chinese AI are over. Moonshot is clearly confident enough to charge real money.

The negatives (because there are some)

I want to be honest here, because the hype cycle around this release has been deafening:

It burns tokens like a V8 burns petrol. Artificial Analysis measured K3 generating 130 million tokens to complete their Intelligence Index evaluation suite, versus an average of 63 million for comparable models. It thinks *long*, and it thinks *always-on* - you can't dial the reasoning back. Reddit is

full of people on the US$100/month subscription discovering they get about four full sessions before hitting their weekly limit.

It's verbose and not particularly fast. Independent testing describes it as slower and chattier than its price-tier peers. A high solve rate coexists with a high cost per attempt - the model tries harder, longer, which is great for hard problems and wasteful for easy ones.

The sticker price is deceptive. US$3/US$15 sounds fine until you multiply it by the sheer volume of output tokens a single agentic coding session generates. For high-volume, simple tasks it's a poor fit - and Moonshot knows it, which is why K2.6 stays available at roughly a quarter of the price.

The weights aren't out yet. Everything above is via the hosted API. Until 27 July-ish, self-hosting remains a promise, not a fact.

So: frontier-grade brain, frontier-grade appetite. You pay for the performance - just in tokens rather than subscription dollars.

> ### *Could Bob Live in My Garage?*
>
> *This is the bit I've been chewing on. When those weights drop, could I run Kimi at home?*
>
> *If you want a taste of what that looks like, watch this excellent video of a chap running Kimi K2.6 on a Mac Studio. K2.6 is the previous model - "only" 1 trillion parameters - and he gets it running on Apple silicon with 512 GB of unified memory using MLX quantisations. The full 4-bit weights come in around 450 GB, he experiments with clever mixed-precision quants down to about 3.4 bits, and the thing still generates at 20–26 tokens per second. He has it building Snake, a 3D Flappy Bird clone, and a voxel Minecraft knock-off from single prompts - locally, offline, on a desktop. It's genuinely impressive, and the local results stand up remarkably well against the full-fat cloud version.*
>
> *But here's the maths problem. **K3 is 2.8 times bigger.** Scale that 450 GB footprint up and you're looking at well over a terabyte of weights, even quantised. No Mac Studio on the planet swallows that. You're suddenly in serious server territory - a rack of high-memory GPUs, eight-figures of power bill anxiety, and a very understanding spouse. The MXFP4 quantisation-aware training Moonshot built in helps a lot (it's specifically designed to make serving this monster economically sane), but "economically sane for a data centre" and "sits next to the 3D printer in the garage" are different thresholds.*
>
> *My honest assessment: K3 at home is a no for mortals, today. But K2.6 runs beautifully on a single high-memory Mac right now, and the quantisation wizards in the open-source community have a habit of doing unreasonable things within weeks of a weight release. Give it six months and someone will have K3 running on hardware that fits under a desk. Whether it fits under my desk is a conversation Jo and I will have later.*

So, was it a Sputnik moment?

One line from the podcast has stuck with me: "*Frontier intelligence is now a totally perishable asset.*" When an open-weight model from a lab most Australians have never heard of can go toe-to-toe with everything OpenAI and Anthropic ship - and then hand the weights to the entire planet - it's hard to argue the frontier labs still have a durable lead.

Bob seems happy with the upgrade. I'll report back in a few weeks on what the token bills look like.

Related:

- Moonshots Podcast - Emergency Episode: The Kimi K3 Sputnik Moment - the podcast that kicked this post off
- Running Kimi K2.6 locally on a 512 GB Mac Studio - what local Kimi actually looks like today
- Kimi K3 on Artificial Analysis - independent benchmarks and the token-burn data
- Kimi K3 on OpenRouter - API pricing and availability
