SDS PODCAST

EPISODE 1012: KIMI K3: THE OPEN-WEIGHT 2.8-TRILLION PARAMETER COMPETING AT THE FRONTIER

Jon Krohn: 00:00 This is episode number 1012 on KIMI K3. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today's episode is all about KIMI K3, a new model at OChina that in the space of a single week has managed to rattle investors, kick off a pricing skirmish among the big American AI labs, and reignite a debate in Washington DC about open source AI. Whether you're a hands-on ML practitioner or you're more focused on the commercial side of AI, this is a release you'll want to understand. So we're going to dig into it today. Here we go. Let's start with the company behind it. Moonshot AI is a Beijing-based startup backed by the Chinese tech giant Alibaba. And if the name Kimmy rings a bell, that might be because the company's original Kimmy chatbot released back in 2023, three years ago, was the first model capable of accepting a context window of 128,000 tokens, which was a big deal at that time that made a big splash.

01:06 Moonshot's co-founder and CEO, Yang Xi Lin, earned his doctorate at Carnegie Mellon in 2019, and the company has had serious financial momentum. Moonshot raised $2 billion in May at a valuation north of 20 billion with annual recurring revenue reportedly exceeding $200 million. What makes the K3 release a particularly compelling story is that it's something of a comeback. Moonshot's market position has eroded significantly over the past 18 months following DeepSeek's meteoric rise, and now the student of that disruption has become the disruptor. So what exactly is KIMI K3 released in mid-July? It's a 2.8 trillion parameter model that Moonshot says is the largest open source AI model in the world. Now, before you panic about inference costs with such an enormous model, it's key to know that this is a mixture of experts' architecture. You can learn more about MOE models in episode 939 of this podcast if you'd like to.

02:08 But the key thing here is that of KIMI K3's 896 expert sub-modules within their model, merely 16 of those nearly 900 are activated per token. So that headline parameter count describes total capacity, not the compute running on every single request. The model features a one million token context window, native visual understanding, an always on reasoning mode, and it's built on two architectural innovations developed at Moonshot. The first one is KIMI delta attention, which is a hybrid linear attention mechanism and attention residuals, which is a drop-in replacement for the residual connections that first became famous years ago with a model called Resnet. Not sure if you'll remember that one. Anyway, together, those two innovations reportedly by roughly a 2. 5X improvement in scaling efficiency over the previous K2 generation of KIMI. That architectural angle matters beyond this one model. Bank of America analysts noted that despite persistent compute constraints in China, K3 demonstrates that pre-training scale paired with architectural innovations can still deliver step change gains.

03:18 In other words, US chip export controls are once again proving to be a leaky dam. Now, how good is KIMI K3? To Moonshot's credit, they've actually been relatively measured in their own claims. The company itself says K3 still trails Anthropic's Claude Fable five and OpenAI's GPT 5.6 soul on overall performance, but that it beat ClaudeOpus 4.8 and GPT-5. 5, the model sitting just behind Anthropic and OpenAI's respective flagship models on benchmarks including coding and general agentic tasks. Independent signals so far are encouraging. For example, K3 scores 57 on the artificial analysis intelligence tracker, placing it well above the median of 31 for reasoning models in a comparable price tier. And the evaluation platform Arena ranked K3 at the top for front-end coding ability with Arena's CEO calling KIMI K3 possibly the single biggest release of the year and

a point where open source Chinese models are surpassing closed US models, at least in some respects.

04:27 That said, a few caveats do deserve some airtime here. On independently verified coding benchmarks, ClaudeOpus 4.8 still leads the active frontier and Moonshot's own Agentis scores haven't yet been reproduced by third parties. Separately, Wharton Professor Ethan Mollick, who has been on this podcast, you can check that out. He's argued that KIMI is a strong but uneven model rather than another deep sea scale breakthrough. So KIMI K3 is a top three top tier model, but certainly not the top one, which brings us to the part of the story with real commercial teeth pricing. Moonshot lists K3 at $3 per million uncached input tokens, $15 per million output tokens, and just 30 cents per million cash hit input tokens. For context on that last figure, the caching, if you're running agents or things like retrieval augmented workflows where the same system prompts and documents get sent over and over, most of your input tokens are going to be cached.

05:29 And so your effective input rate collapses toward that 30 cent per million token floor. Those rates undercut both ClaudeOpus 4.8 at five and $25 and GPT 5.6 Sol at five and $30 for input and output tokens respectively. While K3 actually matches despite being more cost-effective, all of those models, CloudOp is 4.8, GPT 5.6 Sol, they all have a one million token context window. And so yeah, it seems like you're getting a model that is cheaper than the next frontier models from the really big Western frontier labs with the same kinds of capabilities at a cheaper price. But interestingly, K3 is expensive by Chinese standards. So for example, Z.aI's GLM 5.2 costs at less than a third of the cost per million output tokens and DeepSeq V4 is 94% cheaper per million output tokens relative to KIMIKR three. So that's interesting. One practical gotcha. If you're evaluating it, K3 always reasons.

06:39 So you could have lots of tokens being consumed behind the scenes without anything being pumped out because reasoning effort is currently locked to maximum with KIMI K3 as well. And all those thinking tokens are billed as output at $15 per million. So a chatty reasoning trace that you can't even see for the most part can cost more than the visible answer. Fantastic for hard problems, wasteful for simple ones and simple lookups, things like that. All right, let's talk about the open source dimension with one asterisk. So the API for KIMI K3 went live in mid-July, but the full model weights are promised under a modified MIT license by July 27th, which is coming up soon now. And until those files actually appear, K3 isn't deployable or isn't downloadable or locally deployable. You can't really call it open source right now at all, but it seems like you're going to pull through on that.

07:33 And assuming Moonshot delivers that permissive license means enterprises will be able to run frontier class AI entirely on their own infrastructure with no per token fees and no data leaving their walls. So what does all this together mean for Western Labs? Well, the reaction has been swift. OpenAI and Anthropic have responded by increasing token allowances and relaxing usage limits to retain users while Anthropic has expanded access to CloudFable five and raised weekly token caps. That's all good news for us. Keep the competition coming. There's an analyst named Patrick Moorhead who described the market response as an overreaction shockingly similar to the DeepSeq launch 18 months ago while acknowledging K3 could pose revenue challenges for OpenAI and Anthropic long-term. This is an ongoing trend. It's not just KIMI K3 on its own. Chinese models were already seeping into Western production stacks before K3 arrived. So cursor, for example, used KIMI to build its, not KIMI3, but an earlier version to build its composer to coding agent.

08:36 DoorDash's CTO says the company delegates a lower level work to KIMI K2. 6 and Thinking Machines, which was founded by Mira Murati, who used to be an exec at OpenAI. They tapped KIMI K2. 5 to generate early post-training data for its own open model. There is friction here too though. OpenAI and Anthropic have accused several Chinese firms of using distillation to extract capabilities from their models. In February, Anthropic specifically accused DeepSeek, Moonshot, Whomade Kimmy, and MiniMax of attempting to illicitly extract Claude's capabilities. Allegations Beijing has rejected as groundless. However, that dispute resolves, the strategic pressure is the same one DeepSeek introduced. Now aimed at a more vulnerable spot though. The price businesses pay for advanced AI and the control they surrender to US providers. The combination of lower cost, strong performance, and customer control threatens to turn frontier AI from a tightly controlled premium service into a competitive low price commodity.

09:38 Not too bad for you and me, eh?

09:41 Yeah, my take for you is that the frontier is now contested to some extent, open and cheap enough that the cost of experimenting with world-class AI has never been lower, and every price war between labs is a subsidy for the applications you and I are building. Hopefully we really will get to access the Kimi K3 model weights later this month, but either way, open weight models are nipping closely at the heels of the proprietary frontier models and show no sign of letting up. This is surely good news for us whether we want cheap inference or fine-tuning of powerful, powerful models for our own use cases or the kinds of privacy benefits that come from having everything run on our own infrastructure. All right, that's it for the end of today's episode. If you enjoyed it or you know someone who might consider sharing this episode with them, leave a review of the show on your favorite podcasting platform.

10:34 If you write an Apple Podcast review, that's especially helpful to us. I'll actually read it on air when you do that. Yeah, you can comment on our YouTube videos as well or tag me in a LinkedIn post with your thoughts and I will respond to those for sure. And if you aren't already, be sure to subscribe to the show. Most importantly though, above all, we hope you'll just keep on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.