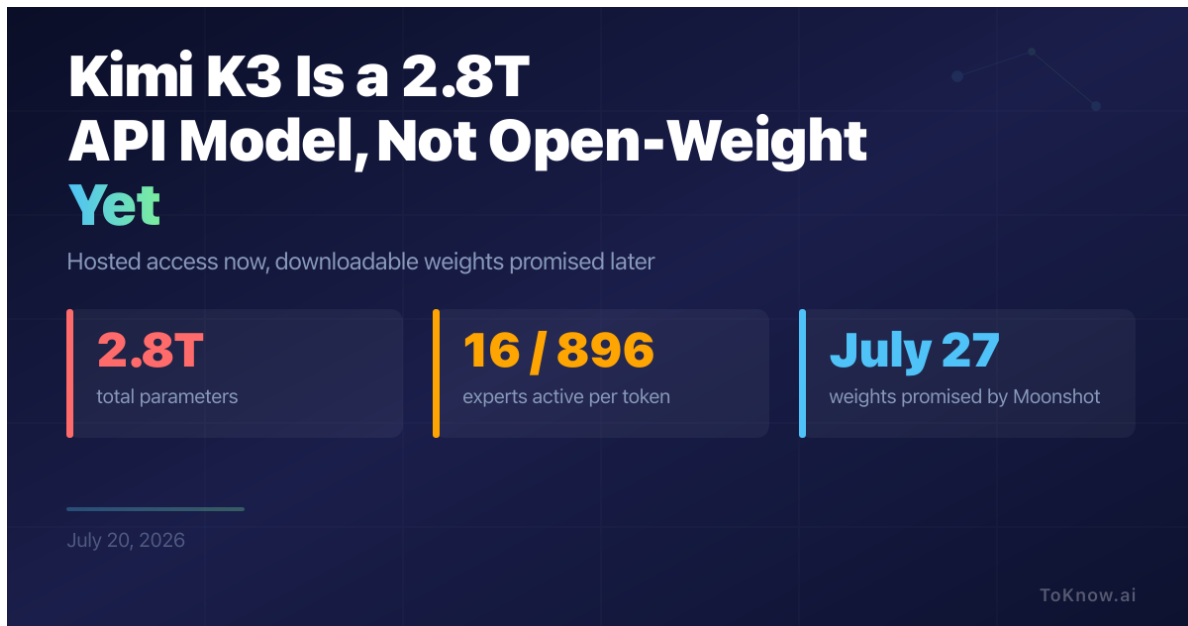


Kimi K3 Is a 2.8T API Model, Not an Open-Weight Release Yet

Kabui, Charles

2026-07-20

[Read at ToKnow.ai](#)



Moonshot AI launched [Kimi K3](#) with 2.8 trillion total parameters, 896 specialised parts called experts, and 16 experts active while processing each token. It accepts up to 1,048,576 tokens of context. Moonshot says its new architecture makes training 2.5 times more efficient than Kimi K2, but the full technical report is not available yet. The immediate correction is simpler: Moonshot calls K3 “open”, while its [API guide](#) says the full weights will arrive by July 27.

The company's [official model listing](#) had 18 models on July 20 and no K3 checkpoint. Today, K3 is available through Kimi's hosted products and API, not as a downloadable open-weight model.

Developers can use the API now. [Official pricing](#) is \$0.30 per million cached input tokens, \$3 for uncached input, and \$15 for output. K3 always reasons before answering and defaults to maximum effort, which can increase response time and output cost. Running it yourself is not possible until the weights, licence, and deployment files arrive. Even then, Moonshot recommends at least 64 accelerators, putting a full 2.8T deployment beyond most teams.

Model releases increasingly happen in stages: hosted access first, downloadable files later. That is useful, but the labels should follow the artefacts. Until people can inspect the checkpoint and licence, K3 is a proprietary service with a dated promise to become open-weight.

Read More: [Kimi K3 beats GLM-5.2 on capability, while GLM-5.2 is easier to host](#).

Sources:

- [Kimi K3 technical blog](#)
- [Kimi K3 API quickstart](#)
- [Kimi K3 pricing](#)
- [Moonshot AI model listing](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)