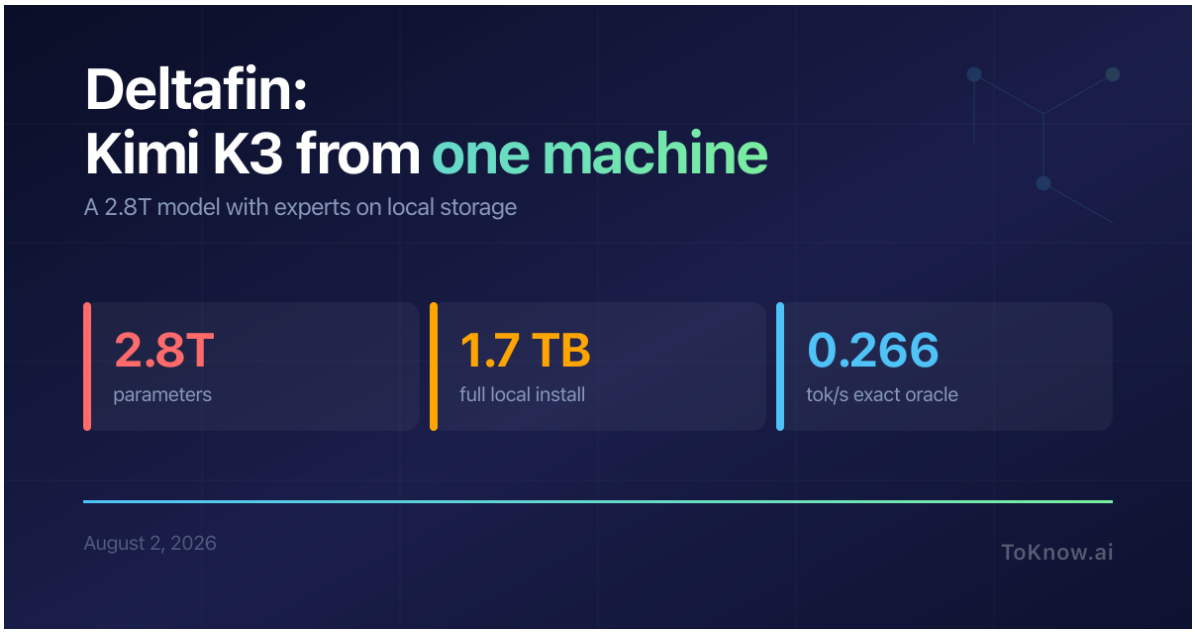
Deltafin: Run the Full Kimi K3 from One Machine

Kabui, Charles

2026-08-02

Read at ToKnow.ai



Deltafin is a research runtime for running the 2.8-trillion-parameter Kimi K3 on one Apple Silicon or Linux workstation. Its recommended full mode stores about 1.7 TB of weights locally and reads the selected mixture-of-experts from disk. A streaming mode needs about 215 GB, but fetches missing experts over the network and can take more than three minutes per uncached token. A single target pass reads 16 experts across 92 layers, about 25.8 GB of

expert data. The project reports an exact-oracle median of 0.2660 tokens per second, or 3.76 seconds per token, on a 64 GB M1 Max after converting the resident spine to int8.

This is an existence proof, not a fast chat service. Deltafin supports MPS, CUDA, and CPU paths, plus an OpenAI-compatible API for local clients and coding agents. Its draft models can suggest tokens, but K3 verifies every emitted token and remains the authority. The trade-offs are blunt: inference is greedy, one request runs at a time, long chats rebuild their full prompt cache, and the streaming mode is meant for trying the idea rather than daily use. The repository includes an MIT licence, while the Kimi modeling code and weights fetched during setup follow Moonshot AI's separate terms. WASTE offers a C runtime with a different storage and memory trade-off.

Sources:

- Deltafin repository and current measurements
- Deltafin optimization notes
- Official Kimi K3 model repository
- Deltafin licence file
