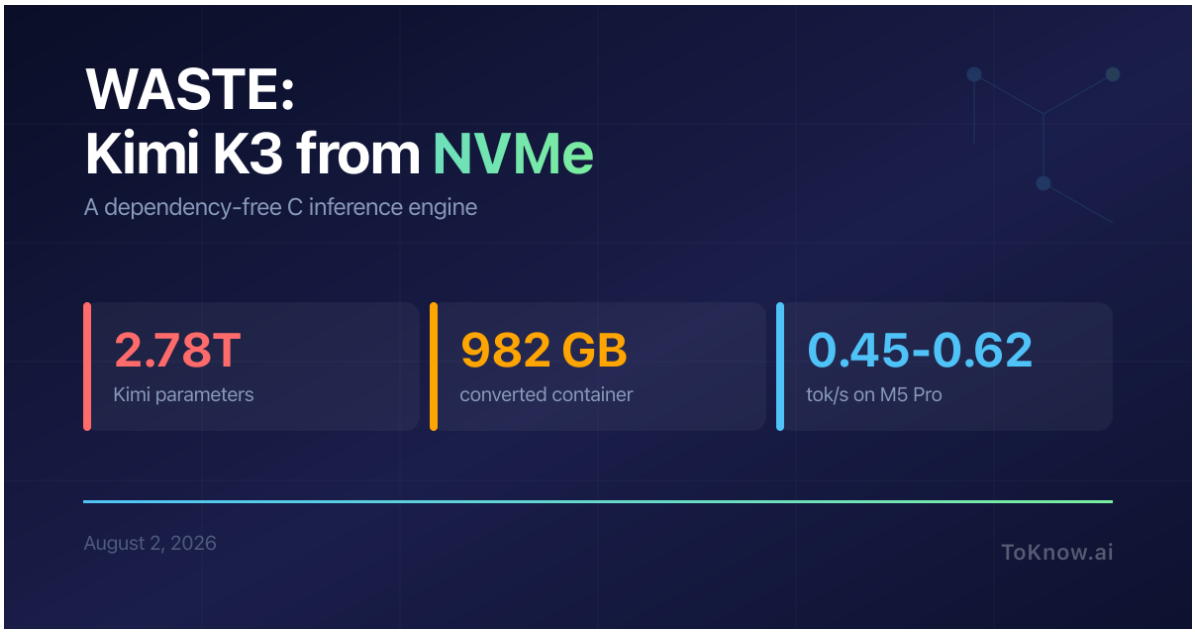
WASTE: Run Kimi K3 by Streaming Experts from NVMe

Kabui, Charles

2026-08-02

Read at ToKnow.ai



WASTE is an embeddable C11 inference engine that keeps a model's shared trunk in memory and streams only the selected experts from NVMe storage. It targets the full 2.78-trillion-parameter Kimi K3, not a pruned or distilled copy. The converted WASTE container is 982 GB, the model opens with 29.06 GB of RAM, and the project recommends 64 GB plus a fast internal SSD. WASTE's efficiency documentation reports 0.45 to 0.62 tokens per second on

a 64 GB M5 Pro. WASTE uses 3-bit residual vector quantization for experts, keeps more sensitive shared weights at 4 or 8 bits, and uses router lookahead to begin reads before the next layer needs them. At 4K context, K3's compressed KV cache is about 0.21 GB.

The useful difference is WASTE's small software boundary. Inference needs libc and pthreads, not Python, PyTorch, CUDA, or BLAS. The documentation also measures Kimi-Linear 48B at 10.65 tokens per second in a 19 GB container with a 1.28 GB minimum. K3 remains the main and best-tested target, and the format and API are not frozen. WASTE includes a CLI, an OpenAI-compatible server, a C API, and multimodal examples. Conversion takes about 4.7 hours with three workers and needs another 1.42 TB of temporary space. A USB enclosure measured 0.94 GB/s versus 12.78 GB/s for the internal SSD. WASTE is Apache 2.0. The efficiency notes explain why more allocated memory can make it eight times slower.

Sources:

- WASTE repository and documentation
- WASTE engine design
- WASTE efficiency measurements
- WASTE file format
- Kimi-Linear 48B model
- WASTE Apache 2.0 licence
