

# TurboFieldfare, Deltafin, and WASTE: Three Ways to Run MoE Models from Disk

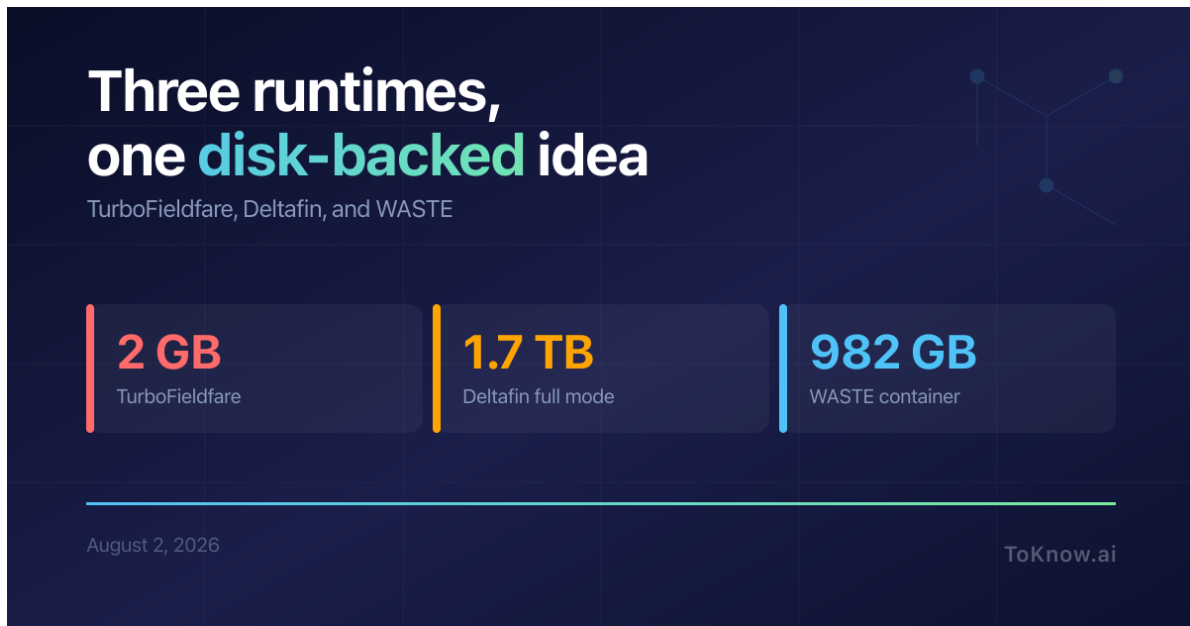
Kabui, Charles

2026-08-02

---

[Read at ToKnow.ai](#)

---



[TurboFieldfare](#), [Deltafin](#), and [WASTE](#) share a trick: each token in a mixture-of-experts model activates only part of the weights. They keep shared weights resident and fetch routed experts from storage. TurboFieldfare runs Gemma 4 26B-A4B with about 2 GB for weights, while Deltafin and WASTE target Kimi K3 at roughly 2.8 trillion parameters. Deltafin's install

is 1.7 TB and reports 0.2660 tokens per second on a 64 GB M1 Max. WASTE's converted container is 982 GB and reports 0.45 to 0.62 on a 64 GB M5 Pro.

Their engineering choices split three ways. TurboFieldfare uses custom Swift and Metal kernels for Apple Silicon, a native Mac app, and text-only Gemma inference. Deltafin spans Apple Silicon and Linux, exposes an OpenAI-compatible API, and uses draft models that Kimi K3 must verify. WASTE is a C11 library with a CLI, server, and multimodal path. Its documentation covers Kimi K3 and Kimi-Linear 48B, while the format and API remain unfinished. Its router lookahead changes when reads start, not which experts K3 selects.

Choose TurboFieldfare for an 8 GB Mac and a modest local model. Choose Deltafin for Kimi K3 through an existing API. Choose WASTE when embeddable C, Linux support, or direct NVMe control matters. All three trade memory for storage latency, setup time, and limited throughput. TurboFieldfare has no remote server security. Deltafin's repository is MIT-licensed, but its fetched Kimi code and weights have separate terms. WASTE needs nearly a terabyte of converted storage. None matches a hosted model for interactive speed.

Sources:

- [TurboFieldfare repository and benchmark](#)
- [Deltafin repository and current measurements](#)
- [WASTE repository and documentation](#)
- [WASTE efficiency measurements](#)
- [Official Kimi K3 model repository](#)

---

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