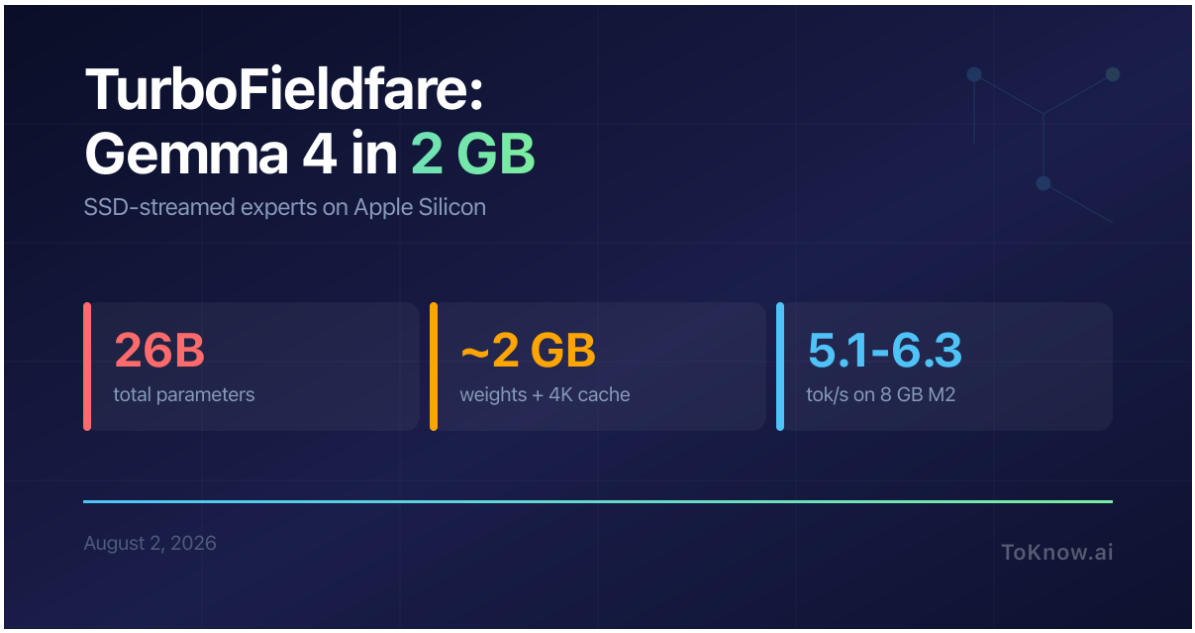
TurboFieldfare: Run Gemma 4 26B in About 2 GB of RAM

Kabui, Charles

2026-08-02

Read at ToKnow.ai



TurboFieldfare is a Swift and Metal runtime for the instruction-tuned Gemma 4 26B-A4B model. The model has 26 billion parameters in total, but only about 3.88 billion are active for each token. TurboFieldfare keeps the shared 1.35 GB core and FP16 4K context cache in memory, then reads the routed experts it needs from an SSD. Its 4-bit weights and 8-bit router leave the runtime with a roughly 2 GB memory budget, even though the installed text

model takes about 14.3 GB. The author measured 5.1 to 6.3 tokens per second on an 8 GB M2 MacBook Air and 31 to 35 on a 24 GB M5 Pro. It needs macOS 26, Metal 4, and Swift 6.2.

That makes local Gemma practical on Macs that cannot keep the complete checkpoint resident. The project includes a native app, CLI, loopback OpenAI-compatible server, and a bounded installer that repacks the model without staging the full source checkpoint. This is a narrow runtime, not a general model backend: it supports text only, the server has no remote authentication or TLS, and the 14.3 GB storage requirement remains. The source is Apache 2.0, while the separately downloaded model keeps its own terms. Deltafin takes the same storage-first idea to a much larger Kimi K3 model.

Sources:

- TurboFieldfare repository and README
- TurboFieldfare benchmark notes
- TurboFieldfare system design
- Gemma 4 model card
