This Week in Local AI — DeepSeek V4 Took #1 on Vibe Code

April 26, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 6 – April 26, 2026

Two open-weight model families dropped in eight days, one of them is now #1 on Vibe Code Benchmark ahead of Kimi K2.6 and Gemini 3.1 Pro, FP4 inference finally landed in the GGUF ecosystem, and Anthropic admitted what most of you suspected. Busy week.

Biggest Day Ever, Eleven Pieces in Four Days

April 25 hit 14,452 humans on the site – a 28% jump over the previous all-time high. Bing and DuckDuckGo recovery accelerated to 6-8x daily search referrals after weeks of plateau, and one new article got indexed by DDG **16 minutes** after publish. First organic click in under twenty minutes. That used to take two weeks on a good day.

The driver was the publication burst around DeepSeek V4 and Qwen 3.6. New: the DeepSeek V4 Flash vs Pro guide, the Qwen 3.6 complete guide, the FP4 inference guide, and the LARQL piece. Refreshed: Mac 2026, coding models 2026, OpenClaw models, OpenClaw alternatives, and the Qwen 3.5 + llama.cpp comparison guides.

The lesson: ship fast on release weeks. Search engines reward freshness on hot terms, and traffic spikes show up the same week, not three weeks later.

DeepSeek V4 Dropped, Then Took #1 on Vibe Code

DeepSeek shipped V4 the evening of April 23. Two MoE checkpoints, both MIT, both 1M context. V4-Pro is 1.6T total / 49B active. V4-Flash is 284B total / 13B active. Pricing on DeepSeek's own API: Flash at $0.14 in / $0.28 out per million tokens, Pro at $1.74 / $3.48 — Haiku-tier and Sonnet-tier respectively, with open weights underneath.

Within 48 hours, V4-Flash topped the Vibe Code Benchmark, beating Kimi K2.6 and Gemini 3.1 Pro per Vals AI's reporting. The architecture story is real: hybrid Compressed Sparse Attention plus Heavily Compressed Attention puts V4-Flash at ~10% of V3.2's FLOPs and ~7% of its KV

cache at 1M context. That's the difference between "1M context exists in marketing" and "1M context you can actually serve."

V4-Pro is workstation territory. V4-Flash is realistic on serious homelabs (96GB+ unified memory or dual high-end GPUs). For most readers the practical move is to test Flash via DeepSeek's API first — drop-in OpenAI-compatible, costs almost nothing, see if it replaces your Haiku or Sonnet workload before committing to local hardware.

📖 **Our breakdown of both variants is here.**

Qwen 3.6 Shipped — Both Variants

The 35B-A3B MoE landed April 16 (Apache 2.0, 262K context, 3B active params per token). The 27B dense followed April 22. Per Qwen's own announcement, the 27B dense ties Claude Sonnet 4.6 on the AA Agentic Index for coding work. SWE-bench Verified 77.2, Terminal-Bench 2.0 59.3, LiveCodeBench v6 83.9 — flagship numbers in a 27B model.

The MoE is the story for 16GB VRAM users. Only 3B active params per token means you can run a 35B-class model on a single mid-range card with llama.cpp RAM offload. Community benchmarks on RTX 3090 hit ~100 tok/s at UD-Q4_K_XL. Simon Willison ran the 27B Q4_K_M GGUF on his Mac and clocked 25.57 tok/s.

Both work with Claude Code and OpenCode out of the box via OpenAI-compatible endpoints (`ollama launch claude --model qwen3.6:27b`). Multiple r/LocalLLaMA threads confirm the local-Claude-Code workflow is genuinely usable now.

📖 **Full Qwen 3.6 guide is here.**

FP4 Finally Landed in llama.cpp

NVFP4 native support hit llama.cpp this month for Blackwell GPUs (RTX 5000/6000 series, plus RTX 50-series consumer when the kernels mature). MXFP4 landed in ik_llama.cpp. First practical FP4 in the GGUF ecosystem — up to 25% faster prompt processing and 35% faster token generation reported on supported hardware.

For RTX 3090 and 4090 users: doesn't help yet. For anyone on Blackwell or considering one, the pricing math just shifted. FP4 weights at near-FP8 quality means MoE models like Qwen 3.6-35B-A3B and DeepSeek V4-Flash get materially cheaper to serve.

📖 **Full FP4 inference guide is here.**

Quick Hits

- **Anthropic came clean.** Per Fortune and The Register on April 23, Anthropic admitted Claude Code's default reasoning effort was silently dropped from "high" to "medium" on March 4 to cut latency. Reverted April 7. A March 26 bug also discarded reasoning history mid-session, and an April 16 change capped responses at 25 words between tool calls. All three reverted by April 20. Local models with full reasoning effort were genuinely competitive during that window.
- **OpenClaw subscription cutoff (April 4)** was partially reversed after pushback. Third-party harnesses still pay per-token. Our breakdown holds up.
- **llama.cpp now natively converts Anthropic API to OpenAI format** as of April 23. Significant for pointing Claude Code at a local model without a translation shim.
- **Hermes Agent AMA Wednesday April 29** on r/LocalLLaMA. Nous Research is running it. Hermes handles per-model tool-call formatting correctly — the thing most agent frameworks botch.
- **M5 Max Mac Studio delayed to October.** Per Bloomberg's Mark Gurman on April 19, memory and storage shortages pushed the Studio refresh back. Need a desktop for local AI in the next six months? Buy the M4 Max or M3 Ultra now.
- **Codex Autoresearch:** Karpathy's pattern is starting to show up as a reusable skill. Worth tracking if you're building research agents.

That's the week. Reply if there's something specific you want me to cover next.

— Mark, InsiderLLM

Running local AI on weird hardware? Built something novel with it? Drop us a line at hello@insiderllm.com.

You're getting this because you signed up at insiderllm.com. [Unsubscribe]({{ unsubscribe_url }})

Source: https://insiderllm.com/blog/newsletter-2026-04-26/