Your RTX 3090 Doesn't Send Policy Change Emails

April 6, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 5 – April 5, 2026

Anthropic just proved why owning your inference stack matters. And Google shipped a model that makes it easier to do.

Anthropic Cuts OpenClaw Off From Claude Subscriptions

Starting April 4, Claude Pro and Max subscriptions no longer cover third-party agent harnesses. If you run OpenClaw, PI Agent, or any non-Anthropic tool against Claude's API, you pay per-token. Claude Code – Anthropic's own agent – stays on the flat rate.

This was inevitable. One OpenClaw power session can burn through a week's worth of Claude quota in a couple of hours. Agent workloads don't fit the "casual chat" pricing model that subscriptions were built for. Anthropic is the first major provider to draw that line explicitly, but OpenAI and Google will follow.

The framing matters: Anthropic isn't blocking third-party tools. They're making them expensive. A heavy OpenClaw session that cost $0 yesterday now costs $15-40 in API tokens. For teams running agents daily, this is a real budget hit.

The local AI angle writes itself. A used RTX 3090 running Qwen 3.5 27B doesn't care about policy changes. It doesn't throttle your agent sessions. It doesn't distinguish between first-party and third-party tools. It costs $900 once and runs forever. Structural risk isn't about whether the cloud is good today – it's about what changes tomorrow without your consent.

📖 **We wrote a full breakdown here.**

Gemma 4: One Week In, Here's What We Know

Google dropped Gemma 4 on April 2. After a week of community testing, the benchmarks held up but the practical story got more complicated.

The good: the 26B-A4B MoE hits 119 tok/s on a single RTX 3090 and 150 tok/s on a 4090. The 31B dense scored #3 among all open models on LMArena (ELO 1452). Apache 2.0 license. Vision, video, and audio input on the edge models. On FoodTruck Bench, the 31B beat GLM 5, Qwen 3.5 397B, and every Claude Sonnet – at $0.20 per run. Community leaderboard testing showed it "destroyed every model except Opus 4.6 and GPT-5.2."

The bad: KV cache VRAM was a disaster at launch. The 31B uses 0.85 MB per context token – 2-3x more than comparable models. Users with 40GB couldn't fit it at reasonable context lengths. llama.cpp patched the worst of it, and adding `-np 1` to your launch command cuts the SWA cache allocation by 3x for single users. Someone ran the 31B at full 256K context on a single RTX 5090 using TurboQuant KV compression.

The verdict so far: Qwen 3.5 still leads on coding and multilingual. Gemma 4 wins on context length (256K), multimodal breadth, and the Apache 2.0 license – which may be the biggest practical differentiator for anyone shipping products. Vision is mixed: benchmarks say great, users report inconsistent OCR. Run both and pick per task.

📖 **Our full Gemma 4 guide with VRAM tables and the -np 1 fix is here.**

The Claude Code Architecture, Ranked for Local AI

The Claude Code source leak was three weeks ago. Most coverage focused on the drama. We focused on the engineering.

Nate B Jones identified 12 core architecture patterns in the 512K-line TypeScript codebase. We ranked them by payoff for people running agents on local hardware. The key finding: the top 6 patterns solve problems that local users hit harder than cloud users – smaller context makes compaction more critical, weaker models make output verification more critical, consumer hardware makes crash recovery more critical.

The harness matters as much as the model. ForgeCode + Opus beats raw Claude Code by 16 points on Terminal-Bench 2.0. The patterns – diff-based editing, LSP integration, git awareness, tool registries with metadata – work regardless of what model sits behind them. OpenClaw, PI Agent, and Aider each implement some of these. None implement all of them.

📖 **Full ranking of all 12 patterns with a tool comparison scorecard here.**

Quick Hits

- **OpenClaw security, March report:** 33 vulnerabilities, CVSS 9.4 sandbox escape, privilege escalation, instances sold on BreachForums. If you're running OpenClaw, update immediately.
- **Qwen 3.6 open weights teased.** Chujie Zheng confirmed medium-sized models coming. No date. We'll have a day-one guide.
- **NanoCoder:** 950-line Python coding agent inspired by Claude Code, trending on GitHub (58 to 242 stars this week). Works with any LLM including local Ollama models.
- **llama.cpp b8671:** Gemma 4 KV cache fix, chat template fix, rotated KV cache implementation. Pull and rebuild.
- **text-generation-webui v4.3:** Gemma 4 support, ik_llama.cpp backend integration.
- **TurboQuant on Gemma 4:** Per-layer outlier-aware K quantization now beating public fork results on Qwen PPL benchmarks.
- **Hermes Agent by Nous Research:** Best open-source agent for local models right now per community consensus. Handles per-model tool call formatting – the thing most agent frameworks get wrong.

That's the week. Next edition drops next Sunday.

Running local AI on weird hardware? Built something novel with it? We're always looking for real benchmarks and creative local AI applications. Drop us a line at hello@insiderllm.com.

Source: https://insiderllm.com/blog/newsletter-2026-04-06/