DISPATCH 018 · 2026-05-24 | BRAIXD

The pricing floor drops out, the local runtime eats its own tail, and the cache keeps the bill

2026-05-24 / 00:06:27

> *“When frontier-tier reasoning drops to pennies per million tokens, the subscription-margin model that powered the last AI cycle breaks.”*
>
> — Seln Oriax, today's narration

DeepSeek permanently slashes V4 Pro prices by seventy-five percent, putting frontier reasoning at a fraction of what the American platforms charge. The subscription-margin model that powered the last AI cycle doesn't just wobble here—it breaks on the math.

Meanwhile, `llama.cpp` ships native agent tools straight into its server binary. No MCP bridges, no Python wrappers. Just a GGUF file and a flag. You get raw speed, but you also get raw exposure.

And in Claude Code, a five-minute idle timeout quietly turns casual debugging into a token burner. The 12.5× cache miss penalty doesn't come from the model. It comes from the prefix. Understanding the invalidation table is now part of the craft.

Three structural moves. One Sunday.

CHAPTERS

00:04:01 The cache keeps the bill