DeepSeek V4 gets deployable, a July 24 trap, and a quiet price cut

June 15, 2026 · by Mark Bartlett

Download this post as PDF

Mostly a DeepSeek week, and not for the reason you'd expect — the news isn't a new model, it's the unglamorous work of the tooling catching up so you can actually run V4. Plus a deprecation date that'll quietly break your code if you're not watching it, and a price cut worth re-running your math on.

DeepSeek V4 Is Going From "Announced" to "Deployable"

When V4 shipped in April, the headline was the model — 1.6T-param Pro, 284B Flash, MIT-licensed, 1M context. The part that doesn't make the launch post is that "the weights exist" and "you can serve this in production" are two different milestones, often weeks apart.

This week closed some of that gap. vLLM's latest release (v0.23.0) gave V4 a serious hardening pass. V4 first landed in vLLM back in v0.22.0; this round was the optimization follow-up — a faster attention kernel, expert-parallelism support for the big MoE, better long-context KV-cache handling, and decoupling V4's attention metadata from V3.2's so the two stop interfering. None of that is exciting. All of it is the difference between "technically supported" and "fast enough to bother with."

The takeaway if you're eyeing V4 locally: if you tried it at launch and it felt rough, it's worth another look — the backends are visibly maturing release over release. Flash is still the one most people should care about. At 13B active params it's in the same hot-path class as the mid-size MoEs you're probably already running, so if your rig handles those, Flash is at least in the conversation. Pro is still datacenter hardware.

One to Act On: The July 24 Deprecation

This is the item worth thirty seconds of your time today, because it's the kind of thing that silently breaks a working pipeline.

If you have code calling DeepSeek's API with `model="deepseek-chat"` or `model="deepseek-reasoner"`, those names retire **July 24, 2026.** They still work right now because DeepSeek is routing them to V4 Flash as a compatibility shim — but after the cutoff, calls using those strings fail.

The change is trivial: swap the model string to `deepseek-v4-flash` (or `deepseek-v4-pro` for the heavy tier). The trap is just not knowing to do it. If you have V3.2-era DeepSeek code in a side project or a cron job somewhere, it's worth changing now while it's in front of you rather than debugging a dead endpoint in late July.

(V3.2 itself is no longer separately callable — DeepSeek folded everything into V4 in April. If you run the R1-Distill models locally, those are untouched and still the best budget reasoning on a single consumer card. It's only the API names changing.)

📖 **Our updated DeepSeek V3.2 / R1-Distill guide is here.**

V4 Pro Quietly Got 4× Cheaper

Filed under good news you may have missed: DeepSeek made V4 Pro's launch discount permanent at the end of May. The rate is now **$0.435 per million input tokens / $0.87 output** — down from the $1.74 / $3.48 it launched at, and it's the standing price, not a promo that snaps back.

It changes the Flash-vs-Pro math. Flash is still cheaper, but the output gap narrowed from roughly 12× to about 3×. If you wrote Pro off as too expensive at launch, it's worth re-pricing — for agent workloads at volume where the quality step-up actually matters, Pro at a third of its old cost is a different calculation than it was six weeks ago.

📖 **Our DeepSeek V4 Flash vs Pro guide is here.**

Quick Hits

- **Check your DeepSeek model strings.** Same theme as above, but worth its own line: anything still calling `deepseek-chat` has a clock on it now. A one-line change today versus a broken job in July.
- **vLLM v0.23.0** is the release with the V4 work above — that's the current mark if you're pinning versions for a local setup. llama.cpp is rolling along in the b9600s with routine point builds; nothing there you need to chase.

- **Flash is the V4 most people should test.** If you've got hardware that runs the mid-size MoEs, V4 Flash is in range — and now that the vLLM path is maturing, it's a better-supported test than it was a month ago.

Heading to AI Engineer World's Fair

Personal note to close. I'll be at the AI Engineer World's Fair at Moscone, June 29 – July 2 — about two weeks out and I'm locking in plans. If you're going to be there, reply and let me know. Meeting a few of you in person is one of the better reasons I do this.

I'm planning to work the floor — the inference platforms, the GPU-cloud companies, the agent-tooling crews — and bring back firsthand takes on what's real versus what's just slideware. Firsthand and no-hype is the whole point of this site, so I'm looking forward to doing it in person for once.

That's the week. Next edition lands next Monday, and as the Fair gets closer I'll have a "who's worth your attention on the floor" rundown.

— Mark, InsiderLLM

Source: https://insiderllm.com/blog/newsletter-2026-06-15/

Free guides for running AI locally