Vector Society Brief

Open Weights at the Frontier

DeepSeek-V4 lands with a 1M-token default, two tiers, and drop-in OpenAI / Anthropic compatibility

---

The release

On **April 24, 2026**, DeepSeek shipped its V4 preview: a two-model family with open weights on Hugging Face, a 1M-token context window as the default (not a paid tier), and a sparse-attention architecture that is the actual story under the headline numbers. The previous generation (*deepseek-chat* and *deepseek-reasoner*) is being routed to V4-Flash during transition and fully retires **Jul 24, 2026, 15:59 UTC**.

The lineup

	V4-Pro	V4-Flash
Total parameters	1.6T	284B
Active per token	49B	13B
Default context	1M tokens	1M tokens
Positioning	Frontier-tier reasoning, agentic coding, world knowledge	Cost / latency optimized; reasoning “closely approaches” Pro

What's actually new

DeepSeek's headline architectural claim is **token-wise compression + DeepSeek Sparse Attention (DSA)**. The practical effect: 1M-token context without the quadratic compute and memory cost that has historically gated long-context inference. This is why 1M is the default, not the premium SKU. Long-context retrieval-style work (whole-codebase reasoning, multi-document synthesis, transcript analysis) becomes routine rather than a budgeting decision.

Drop-in compatibility

The API exposes both **OpenAI ChatCompletions** and **Anthropic Messages** surfaces. Existing code written against either SDK can swap the base URL and a model name; no client rewrite. DeepSeek explicitly calls out integration with **Claude Code, OpenClaw & OpenCode**. For shops standardized on Claude or OpenAI clients, V4 is the cheapest possible A/B test against frontier closed-source models — including against open weights you can pull and run yourself.

Why it matters

Vector	Implication
Open weights	Frontier-tier reasoning is downloadable. Self-hosting is a real option, not a research project.
1M default context	Whole-repo, whole-codebase, and long-transcript prompting stop being cost-prohibitive. Workflows that needed RAG can sometimes drop it.
Two tiers, same family	Cheap-tier and frontier-tier route to compatible APIs. Cost-optimize per task without changing client code or prompts.

OpenAI & Anthropic-compatible	Migration cost from incumbents is hours, not weeks. Vendor lock-in weakens.
DSA architecture	Long-context inference economics shift industry-wide. Expect competitor responses inside a quarter.

Caveats

- **Preview, not GA.** Behavior, pricing, and SLAs can shift before general availability.
- **Data residency.** The hosted API ships your tokens to DeepSeek infrastructure. For sensitive workloads, the open-weights path on your own GPU is the privacy-preserving option — that's the point of open weights.
- **Benchmarks vs production.** “Rivals top closed-source” is a benchmark claim. Validate on your own evals before committing a live workload.

Takeaway

- **The frontier is no longer behind a closed API.** Open weights at this tier change the build-vs-buy math for every team.
- **Long context is becoming free.** Architectures, not just hardware, are the lever. Plan products around the assumption that 1M-token prompts are routine.
- **API surfaces are converging.** When the same client speaks to OpenAI, Anthropic, and DeepSeek, model choice is a runtime decision, not an architectural one.
