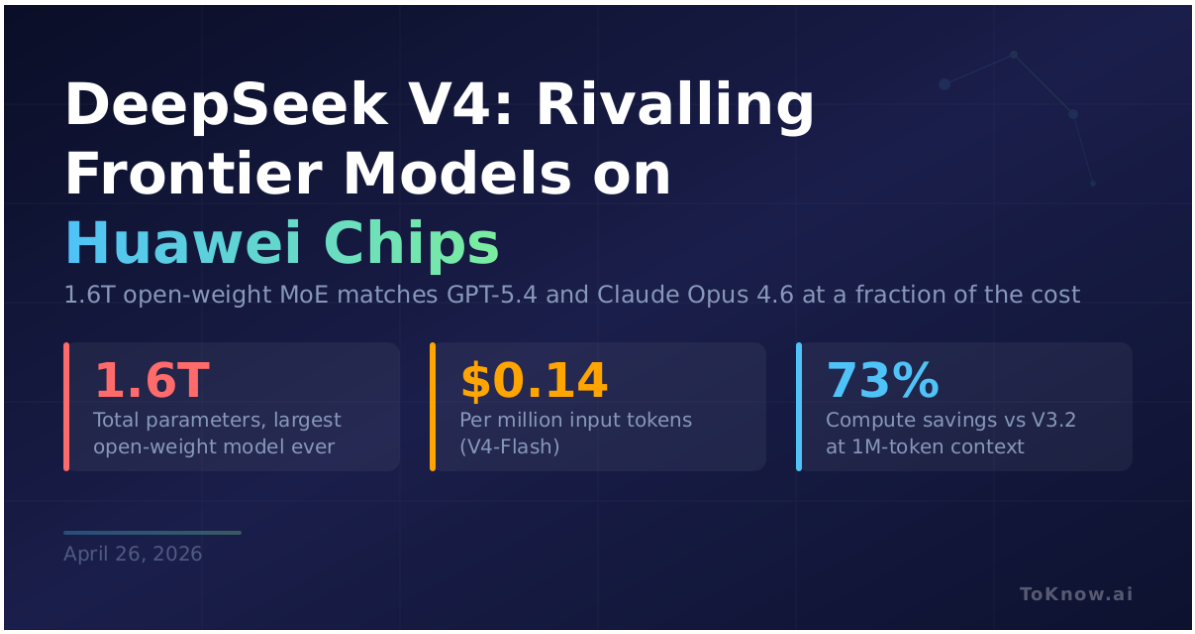
DeepSeek V4: 1.6 Trillion Parameters on Huawei Chips, Rivalling the Best Closed Models

Kabui, Charles

2026-04-26



DeepSeek released a preview of V4 on April 24, 2026, its biggest model update since R1 shook the AI industry in January 2025. V4-Pro packs 1.6 trillion total parameters (49 billion active per query) and a 1-million-token context window, up from V3's 128K. The smaller

V4-Flash runs 284 billion total parameters with just 13 billion active. Both use a hybrid attention architecture that compresses older context and focuses on nearby text, cutting V4-Pro's compute to 27% and its memory to 10% of what V3.2 needed at the same context length. V4-Flash is even leaner: 10% of the compute and 7% of the memory. On benchmarks, V4-Pro-Max matches Anthropic's Claude Opus 4.6 and OpenAI's GPT-5.4 on coding tasks like SWE-bench Verified (80.6% resolved) and scores 90.1% on MMLU-Pro. It leads all open-source models and trails leading closed models by only a few percentage points on reasoning. Both models were pre-trained on over 32 trillion tokens and released under the MIT License.

The practical shift is pricing. V4-Flash costs $0.14 per million input tokens and $0.28 per million output tokens, undercutting GPT-5.4 Nano, Gemini 3.1 Flash, and Claude Haiku 4.5. V4-Pro runs $1.74 per million input tokens, still a fraction of what OpenAI and Anthropic charge for comparable performance. For a developer building a coding assistant that reads an entire codebase, or a research agent processing a long document archive, V4 offers near-frontier capability at open-source prices. DeepSeek also optimized V4 for popular agent frameworks like Claude Code and CodeBuddy, making it a drop-in option for agentic workflows.

The bigger story is hardware. V4 is DeepSeek's first model optimized for Huawei's Ascend chips, a deliberate move away from Nvidia. Huawei confirmed its Ascend 950 supernodes will support V4, and DeepSeek gave Chinese chipmakers, not Nvidia or AMD, early access to the model. DeepSeek uses Huawei hardware for inference, though training still appears partially reliant on Nvidia GPUs. The market noticed: domestic chipmakers Hua Hong Semiconductor jumped 15% and SMIC rose roughly 9% on the announcement. Nvidia CEO Jensen Huang put the stakes bluntly on the Dwarkesh Podcast: "The day that DeepSeek comes out on Huawei first, that is a horrible outcome for our nation." US export controls have cut China off from Nvidia's best chips since 2022, and Beijing has been pushing firms toward domestic alternatives. V4 is the clearest signal yet that China's parallel AI stack, from chips to models, is becoming viable. Replacing Nvidia requires not just competitive silicon but a software ecosystem that developers will actually use. DeepSeek's open-source approach, MIT licensing, and agent-framework compatibility are a direct play for that developer adoption. NVIDIA's advantage was never just hardware; it was the lock-in from CUDA and the ecosystem built around it. Whether Huawei's Ascend platform can match that remains the open question, but V4 makes the possibility much harder to dismiss.

Read More: NVIDIA's own efficiency play with only 3B active parameters is covered in Nemotron Cascade 2.

Sources:

- MIT Technology Review: Three Reasons Why DeepSeek's V4 Matters
- Quartz: DeepSeek Is Back with a New Open-Source AI Model Built on Chinese Chips
- TechCrunch: DeepSeek Previews New AI Model That Closes the Gap with Frontier Models
- DeepSeek V4-Pro Technical Report on Hugging Face
- Wikipedia: DeepSeek

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*