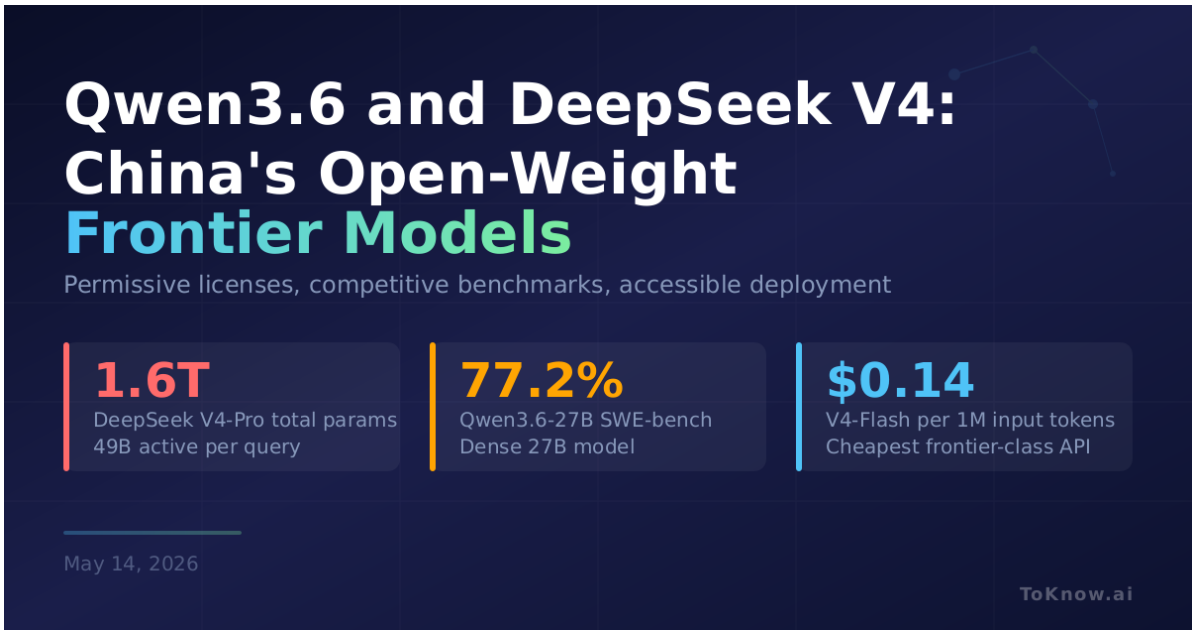
Qwen3.6 and DeepSeek V4: China's Open-Weight Models Now Match Frontier Competitors

Kabui, Charles

2026-05-14



Alibaba open-sourced Qwen3.6-27B on April 22, a dense 27-billion-parameter multimodal model that handles text, images, and video natively. It replaces standard transformer attention with Gated Delta Networks, supports 262K context (extensible to 1M), and includes

both thinking and non-thinking modes so developers can trade latency for reasoning depth. On coding benchmarks it scores 77.2% on SWE-bench Verified, 87.8% on GPQA Diamond, and 94.1% on AIME 2026. Weeks earlier, DeepSeek released V4-Pro and V4-Flash, both Mixture-of-Experts models with 1M-token context. V4-Pro packs 1.6 trillion total parameters with 49 billion active per query, trained on over 32 trillion tokens. Its hybrid attention architecture cuts KV cache to 10% of V3's at the same context length. V4-Flash uses 284 billion total (13 billion active) and approaches V4-Pro's reasoning scores at a fraction of the compute. Both families ship under permissive open licenses: Apache 2.0 for Qwen, MIT for DeepSeek.

What makes these releases practically significant is accessibility. Qwen3.6-27B fits on a single high-end GPU while outperforming many larger models on coding and math, scoring higher than Qwen3.5's 35B MoE flagship on SWE-bench despite being a smaller dense model. DeepSeek V4-Flash, at $0.14 per million input tokens, undercuts every major US API provider. For developers and companies evaluating model options, self-hosting these open-weight models now delivers frontier-competitive performance without API lock-in.

The gap between Chinese open-weight releases and US closed-source frontiers has compressed to single-digit percentages on most standard benchmarks. The competitive pressure runs both ways: cheaper open models from China push pricing down, while US labs respond with tighter vertical integration.

Read More: A deeper look at DeepSeek V4's Huawei chip strategy is in DeepSeek V4: 1.6 Trillion Parameters on Huawei Chips. Qwen3.5's initial multimodal MoE architecture is covered in Qwen3.5: One Model for Text, Images, Video, and Agent Tasks.

Sources:

- Qwen3.6-27B Model Card on Hugging Face
- DeepSeek V4-Pro Model Card on Hugging Face
- DeepSeek V4-Flash Model Card on Hugging Face
- Qwen3.6-27B Blog Post by Qwen Team
