

Forum:[Hack Forums](#)

Thread: [DeepSeek just made its AI architecture much cheaper](#)

Nikola.4737n - September 10, 2026, 1:09 pm

? loading encrypted image...

DeepSeek has introduced V4.1-Flash, a new multimodal model designed to deliver more intelligence while using dramatically less compute and memory.

The model has 552B parameters, but only 8B are active for input and 16B for output thanks to a new Causal Encoder–Decoder architecture. DeepSeek says its combination of new pre-training and large-scale RL can outperform even its flagship V4-Pro on several benchmarks.

The bigger breakthrough may be efficiency. V4.1-Flash uses just 1/4 of the HBM and 1/8 of the SSD storage needed for its previous-generation KV cache. That matters a lot for AI agents, where cached context can become a major part of inference costs.

DeepSeek is also cutting prices, with off-peak rates set at 50% of peak pricing.

[Source](#)

XtremeDiscount - September 10, 2026, 4:01 pm

Oh wow, the AI race continues!