DeepSeek V4

Highly Efficient Million-Token Context LLMs

1. Executive Summary

DeepSeek V4 is a massive generational leap in Large Language Model efficiency. With a focus on sovereign AI and hardware independence, V4 achieves flagship performance while drastically reducing the computational and memory requirements for long-context tasks.

2. Model Specifications

Model Version	Total Parameters	Activated Params	Context Window
DeepSeek-V4-Pro	1.6 Trillion	49 Billion	1,000,000 Tokens
DeepSeek-V4-Flash	284 Billion	13 Billion	1,000,000 Tokens

3. Key Innovations

3.1 Manifold-Constrained Hyper-Connections (mHC)

Ensures training stability for trillion-parameter models by constraining the layer mixing matrix to the Birkhoff polytope (doubly stochastic matrices). This bounds the spectral norm and ensures smooth gradient flow across 1000+ layers.

3.2 H

Alternates between Compressed Sparse Attention (CSA) for fine-grained retrieval and Heavily Compressed Attention (HCA) for global signal summarization. KV cache is reduced by 90% compared to V3.2, making 1M token context economically viable.

3.3 M

A brand new optimizer replacing AdamW for core parameters. Uses orthogonalization-based Newton-Schulz iterations to accelerate convergence and ensure stability in large MoE training.

4. Hardware Independence

V4 is designed for 'Sovereign AI.' It uses TileLang (a custom DSL) and MXFP4 quantization to run efficiently on non-Nvidia hardware like Huawei Ascend and Cambricon chips. This breaks the dependency on proprietary CUDA ecosystems.

5. B

- SimpleQA-Verified: 57.9 (Ranked #1 in open source)

- Codeforces Rating: 3206 (Rivals top-tier closed models)

- Inference Cost: Reduced by 70%+ compared to V3.2