SOVEREIGN: What is the trade-off between inference latency and quality-of-service (e.g., response accuracy) when deployin

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enh

1 Introduction

Analysis of: Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures. Research goal: What is the trade-off between inference latency and quality-of-service (e.g., response accuracy) when deploying Llama, Mistral, Qwen, and DeepSeek under varying hardware constraints (e.g., GPU vs. CPU inference)?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 8.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
DeepSeek-V3 was trained on 2,048 NVIDIA H800 GPUs.	✓	0.21
DeepSeek-V3 demonstrates hardware-aware model co-design for cost-efficient training and inference at scale.	✓	0.27
DeepSeek-V3/R1 model architecture includes Multi-head Latent Attention (MLA) for enhanced memory efficiency.	✓	0.26
DeepSeek-V3/R1 model architecture includes Mixture of Experts (MoE) architectures for optimized computation-communicatio	✓	0.27
DeepSeek-V3/R1 model architecture includes FP8 mixed-precision training to unlock the full potential of hardware capabil	✓	0.27
DeepSeek-V3/R1 model architecture includes a Multi-Plane Network Topology to minimize cluster-level network overhead.	✓	0.28
The paper discusses potential future hardware directions, including precise low-precision computation units, scale-up an	✓	0.37
The paper underscores the critical role of hardware and model co-design in meeting the escalating demands of AI workload	✓	0.26

References

- https://doi.org/10.2139/ssrn.5133368
- https://doi.org/10.48550/arxiv.2412.19437

- https://doi.org/10.1145/3695053.3731412