SOVEREIGN: What is the impact of Lynx scheduling on expert load balancing and downstream QA accuracy for Mixtral 8x7B whe

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning

1 Introduction

Analysis of: DeepSeek-V3 Technical Report. Research goal: What is the impact of Lynx scheduling on expert load balancing and downstream QA accuracy for Mixtral 8x7B when evaluated on HotpotQA, relative to static routing with varying top-k expert selection?.

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

9 papers retrieved. 10 claims extracted, 7 verified. Tribunal: 6.5/10 $\rightarrow$ REVISE (revision_round=1). Policy: SOFT_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 is a Mixture-of-Experts (MoE) language model with 671B total parameters	✓	0.25
DeepSeek-V3 has 37B activated parameters for each token	×	0.15
DeepSeek-V3 adopts Multi-head Latent Attention (MLA) architecture	✓	0.23
DeepSeek-V3 adopts DeepSeekMoE architecture	×	0.12
DeepSeek-V3 uses an auxiliary-loss-free strategy for load balancing	✓	0.21
DeepSeek-V3 uses a multi-token prediction training objective	✓	0.21
DeepSeek-V3 was pre-trained on 14.8 trillion tokens	×	0.13
DeepSeek-V3 training includes Supervised Fine-Tuning and Reinforcement Learning stages	✓	0.21
DeepSeek-V3 requires 2.788M H800 GPU hours for full training	✓	0.26
DeepSeek-V3 training process was stable with no irrecoverable loss spikes or rollbacks	✓	0.23

References

- https://doi.org/10.1007/s11747-014-0403-8
- https://doi.org/10.48550/arxiv.2106.04426
- https://doi.org/10.48550/arxiv.2412.19437