SOVEREIGN: How does the auxiliary-loss-free load balancing strategy in DeepSeek-V3 affect expert utilization diversity an

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: How does the auxiliary-loss-free load balancing strategy in DeepSeek-V3 affect expert utilization diversity and downstream code generation accuracy on HumanEval compared to traditional auxiliary-loss methods?.

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

8 papers retrieved. 15 claims extracted, 9 verified. Tribunal: 7.2/10 $\rightarrow$ 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 is a Mixture-of-Experts (MoE) language model.	✓	0.20
DeepSeek-V3 has 671 billion total parameters.	×	0.09
DeepSeek-V3 activates 37 billion parameters for each token.	×	0.07
DeepSeek-V3 adopts Multi-head Latent Attention (MLA) architecture.	✓	0.23
DeepSeek-V3 adopts DeepSeekMoE architecture.	×	0.11
DeepSeek-V3 uses an auxiliary-loss-free strategy for load balancing.	✓	0.21
DeepSeek-V3 sets a multi-token prediction training objective.	✓	0.24
DeepSeek-V3 was pre-trained on 14.8 trillion tokens.	×	0.12
DeepSeek-V3 underwent Supervised Fine-Tuning and Reinforcement Learning stages after pre-training.	✓	0.18
DeepSeek-V3 outperforms other open-source models in comprehensive evaluations.	✓	0.23
DeepSeek-V3 achieves performance comparable to leading closed-source models.	✓	0.26
DeepSeek-V3 required 2.788 million H800 GPU hours for its full training.	×	0.14
The DeepSeek-V3 training process experienced no irrecoverable loss spikes.	✓	0.19
The DeepSeek-V3 training process required no rollbacks.	×	0.13
DeepSeek-V3 model checkpoints are available at https://github.com/deepseek-ai/DeepSeek-V3.	✓	0.26

References

- https://doi.org/10.48550/arxiv.2412.19437

- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.4230/oasics.icpec.2025.4