DeepSeek-V3 Shared Expert Strategy and Token Throughput on Consumer GPUs

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the shared expert strategy in DeepSeek-V3's MoE architecture affect token generation throughput compared to standard sparse MoE models on consumer-grade GPUs. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures. Research question: How does the shared expert strategy in DeepSeek-V3's MoE architecture affect token generation throughput compared to standard sparse MoE models on consumer-grade GPUs?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

14 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The NVIDIA H800 GPU SXM architecture features reduced FP64 computational performance and NVLink bandwidth for regulatory	×	0.05
The NVLink bandwidth in H800 SXM nodes is reduced from 900 GB/s to 400 GB/s.	×	0.02
Each H800 SXM node is equipped with eight 400G Infiniband (IB) CX7 NICs.	×	0.01
Tensor Parallelism is avoided during training due to its inefficiency under limited NVLink bandwidth.	×	0.03
DualPipe is employed to overlap attention and MoE computation with MoE communication.	×	0.05
The system achieves all-to-all communication at speeds exceeding 40GB/s with eight 400Gbps Infiniband (IB) NICs.	×	0.02
DeepEP is open-sourced, enabling highly efficient expert parallelism.	×	0.05
DeepSeek-V3 employs the DeepSeek-MoE and Multi-head Latent Attention (MLA) architectures.	✓	0.19
DeepSeek-V3 incorporates FP8 mixed-precision training, significantly lowering computational costs.	×	0.13
DeepSeek-V3 integrates speculative decoding based on its Multi-Token Prediction Module, which significantly increases th	×	0.10
DeepSeek-V3 deploys a Multi-Plane two-layer Fat-Tree network to replace a traditional three-layer Fat-Tree topology, red	×	0.07
DeepSeek-V3 model has a KV Cache Per Token Multiplier of 1x.	×	0.09
Qwen-2.5 72B (GQA) model has a KV Cache Per Token Multiplier of 4.66x.	×	0.02
LLaMA-3.1 405B (GQA) model has a KV Cache Per Token Multiplier of 7.28x.	×	0.02
DeepSeek-V2 MoE model has a training cost of 155 GFLOPS/Token.	×	0.10
DeepSeek-V3 MoE model has a training cost of 250 GFLOPS/Token.	×	0.12
Qwen-72B Dense model has a training cost of 394 GFLOPS/Token.	×	0.04
LLaMa-405B Dense model has a training cost of 2448 GFLOPS/Token.	×	0.04

References

- http://arxiv.org/abs/2505.09343v2
- http://arxiv.org/abs/2412.19437v2
- http://arxiv.org/abs/2601.15021v1