Scaling Laws and Logical Reasoning in DeepSeek-V3 with MoE and MLA Architectures

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the effect of model size on language model performance on logical reasoning tasks v10. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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: What is the effect of model size on language model performance on logical reasoning tasks v10.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

10 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 4.7/10.

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.	×	0.01
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.	×	0.06
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.	×	0.03
DeepSeek-V3 employs the DeepSeek-MoE and Multi-head Latent Attention (MLA) architectures.	✓	0.18
DeepSeek-V3 incorporates FP8 mixed-precision training.	✓	0.17
DeepSeek-V3 integrates speculative decoding based on its Multi-Token Prediction Module.	×	0.05
DeepSeek-V3 deploys a Multi-Plane two-layer Fat-Tree network to replace a traditional three-layer Fat-Tree topology.	×	0.07
DeepSeek-V3 aims to address memory efficiency, cost-effectiveness, and inference speed challenges in scaling LLMs.	×	0.14
DeepSeek-V3's KV Cache Per Token Multiplier is 1x, compared to 4.66x for Qwen-2.5 72B (GQA) and 7.28x for LLaMA-3.1 405B	×	0.03
DeepSeek-V3 MoE has a training cost of 250 GFLOPS/Token, compared to 394 GFLOPS/Token for Qwen-72B Dense and 2448 GFLOPS	×	0.04

References

- http://arxiv.org/abs/2505.09343v2
- http://arxiv.org/abs/2603.09200v1
- http://arxiv.org/abs/2407.04973v1