DeepSeek-V3 Auxiliary-Loss-Free Routing for Long-Context Load Balancing Efficiency

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the load balancing efficiency of DeepSeek-V3's auxiliary-loss-free policy compare to traditional routing methods during long-context inference tasks. 13 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Auxiliary-Loss-Free Load Balancing Strategy for Mixture-of-Experts. Research question: How does the load balancing efficiency of DeepSeek-V3's auxiliary-loss-free policy compare to traditional routing methods during long-context inference tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

12 papers retrieved. 13 claims extracted; 4 independently verified. Quality review score: 5.3/10.

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
DeepSeekMoE outperforms conventional MoE architectures like GShard significantly.	×	0.03
DeepSeekMoE segments experts into finer granularity and isolates some experts as shared ones.	×	0.04
Sigmoid baseline performs better than the softmax baseline in the main experiments.	×	0.02
Experiments are based on two model sizes of 1B and 3B total parameters.	×	0.03
The 1B model is trained on 100B tokens and the 3B model on 200B tokens.	×	0.08
Cosine learning rate scheduler is applied for the 1B model and multi-step learning rate scheduler for the 3B model.	×	0.03
Auxiliary loss coefficient α is set to 0.001 for the baseline.	×	0.10
Loss-Free Balancing achieves lower perplexity and better load balance on both 1B and 3B models compared to the auxiliary	✓	0.28
Loss-Free Balancing achieves a perplexity of 9.50 and MaxVioglobal of 0.04 for the 1B model.	✓	0.15
Loss-Free Balancing achieves a perplexity of 7.92 and MaxVioglobal of 0.04 for the 3B model.	✓	0.16
Loss-Free Balancing maintains a better load balance throughout most of the training time as shown in Figure 3.	✓	0.22
MaxVioglobal reflects the degree of balanced expert utilization and efficiency upper bound when the batch size approache	×	0.07
MaxViobatch is more related to the training efficiency.	×	0.03

References

- http://arxiv.org/abs/2602.11688v1
- http://arxiv.org/abs/2409.04896v1
- http://arxiv.org/abs/2408.15664v1