Dynamic Expert Sharing in Sparse MoE Models for Efficient Code Generation

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does increasing the number of experts in sparse MoE models improve inference efficiency (throughput) while maintaining pass@1 accuracy on self-invoking code generation tasks as benchmarked on. Among parallel decoding paradigms, diffusion large language models (dLLMs) have emerged as a promising candidate that balances generation quality and throughput. However, their integration with Mixture-of-Experts (MoE) architectures is constrained by an expert explosion: as the. 14 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: Dynamic Expert Sharing: Decoupling Memory from Parallelism in Mixture-of-Experts Diffusion LLMs. Research question: Does increasing the number of experts in sparse MoE models improve inference efficiency (throughput) while maintaining pass@1 accuracy on self-invoking code generation tasks as benchmarked on HumanEval Pro?.

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 4.4/10.

3 Results

12 papers retrieved. 14 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
MoE dLLMs are more memory-bound than their dense counterparts.	×	0.11
Dynamic Expert Sharing (DES) mitigates the memory bottleneck in MoE dLLMs.	✓	0.20
DES-Vote extends the Pareto frontier in the trade-off between activated experts and relative accuracy compared to Vanill	×	0.08
The average activated experts for LLaDA-MoE is approximately 96.0%.	×	0.04
The average activated experts for LLaDA2.0-mini is approximately 96.5%.	×	0.03
DES-Vote achieves a relative accuracy of approximately 99.5% with an average activated expert count near 98.5%.	×	0.06
On the MBPP benchmark, DES-Vote achieves a relative accuracy of approximately 100% with a MoE kernel latency of roughly	×	0.04
Vanilla MoE on LLaDA-MoE exhibits a MoE kernel latency of approximately 1.6 ms with a relative accuracy of roughly 97.5%	×	0.05
The roofline model indicates a peak performance of 2250 TFLOPs/s for the evaluated systems.	×	0.02
Existing methods designed for AR models fail to mitigate the global HBM traffic bottleneck because unique expert weights	×	0.09
DES-Seq strategy takes the union of experts chosen independently by each token in the block to allow for vectorized exec	×	0.05
DES-Vote is a mechanism where tokens collectively vote for experts based on their weighted router saliency.	×	0.11
Accuracy results in Figure 1(c) are averaged across HumanEval, MBPP, MATH500, and GSM8K benchmarks.	×	0.02
Figure 1(d) evaluates DES-Vote's performance specifically on the MBPP benchmark.	×	0.04

References

- http://arxiv.org/abs/2602.00879v1

- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2505.20225v1