Sparse Mixture-of-Experts vs. Dense Transformers in Mathematical Reasoning Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do sparse mixture-of-experts models compare to dense transformers on mathematical reasoning v10. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: DECO: Sparse Mixture-of-Experts with Dense-Comparable Performance on End-Side Devices. Research question: How do sparse mixture-of-experts models compare to dense transformers on mathematical reasoning v10.

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 3.8/10.

3 Results

12 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
DECO incorporates ReLU in the router to determine expert activation.	×	0.07
ReLU is fully differentiable, inherently induces sparsity, and supports token-dependent activation ratios.	×	0.04
DECO applies a scaling operator to the routing scores before they are multiplied by the expert outputs.	×	0.04
DECO uses a learnable vectorized scaling factor instead of a fixed scalar scaling factor.	×	0.03
Non-gated experts exhibit more favorable properties within the specific context of ReLU-based routing.	×	0.12
In a ReLU-activated MoE, non-gated experts obtain a more stable trend of activation ratio compared to gated variants.	×	0.13
DECO introduces NormSiLU as an expert activation function.	×	0.11
DECO achieves a PPL of 34.36 on the Small model, 27.74 on the Medium model, 21.87 on the Large model, and 18.38 on the X	×	0.03
Dense models achieve a PPL of 34.40 on the Small model, 27.85 on the Medium model, 21.93 on the Large model, and 18.41 o	×	0.04
DECO achieves a PPL of 36.90, 39.18, 42.81, and 47.38 on different benchmarks.	×	0.02
Dense models achieve a PPL of 36.80, 39.01, 42.80, and 46.98 on different benchmarks.	×	0.03

References

- http://arxiv.org/abs/2605.10933v3

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2603.11114v1