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

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do sparse mixture-of-experts models compare to dense transformers on mathematical reasoning v16. 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: Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection. Research question: How do sparse mixture-of-experts models compare to dense transformers on mathematical reasoning v16.

2 Methodology

Systematic literature search across multiple databases yielded 15 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

15 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
The threshold T=50 is used for labeling queries as 'Sparse Retriever' or 'Dense Retriever' in the results of this paper.	×	0.05
Similar results are obtained for threshold values of 100, 150, and 200.	×	0.03
BERT is fine-tuned using labels derived from the presence and rank of the top relevant passage Fq returned by the sparse	×	0.04
BERT has been extensively used with success on various tasks in Natural Language Processing and Information Retrieval, i	×	0.06
BERT utilizes a cross-encoder framework, which jointly encodes the inputs and candidates in a single transformer.	×	0.03
The model performs full-cross self-attention over the given input and label to attain higher accuracy.	×	0.05
A linear classification layer with binary cross entropy loss is used on top of the first vector produced by the transfor	×	0.03
The MS MARCO passage collection consists of 8.8 million passages accompanied by more than 500k pairs of query and judged	×	0.02
For over 90% of queries in the MS MARCO passage collection, there is a single judged relevant passage per query.	×	0.03
The MS MARCO Development Set contains 6,980 queries for development and validation.	×	0.03
BM25 is employed as the sparse retriever, implemented by the open-source Anserini system from the University of Waterloo	×	0.03

References

- http://arxiv.org/abs/2605.10933v3
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2109.10739v1