Failure Modes of Frontier Language Models in Abstract Mathematical Reasoning

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the failure modes of frontier language models on abstract mathematical reasoning v15. 16 claims were extracted from source literature; 2 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: Achieving >97% on GSM8K: Deeply Understanding the Problems Makes LLMs Better Solvers for Math Word Problems. Research question: What are the failure modes of frontier language models on abstract mathematical reasoning v15.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 16 claims extracted; 2 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
DUP achieves a new SOTA result on the GSM8K benchmark, with an accuracy of 97.1% under the zero-shot setting.	✓	0.28
Extensive experiments on 10 diverse reasoning benchmarks show that the DUP method consistently outperforms the other cou	✓	0.29
The DUP method is a zero-shot prompting method.	×	0.12
The DUP method is evaluated on six Arithmetic Reasoning benchmarks: GSM8K, SVAMP, MultiArith, AddSub, AQuA, and SingleEq	×	0.07
The DUP method is also evaluated on two Commonsense Reasoning benchmarks: CommonsenseQA and StrategyQA.	×	0.07
The DUP method is evaluated on two Symbolic Reasoning benchmarks: Last Letter and Coin Flip.	×	0.07
The GSM8K dataset contains 1319 samples with number answers.	×	0.06
The MultiArith dataset contains 600 samples with number answers.	×	0.01
The AddSub dataset contains 395 samples with number answers.	×	0.01
The SVAMP dataset contains 1000 samples with number answers.	×	0.01
The SingleEq dataset contains 508 samples with number answers.	×	0.01
The AQuA dataset contains 254 samples with option answers.	×	0.01
The Last Letters dataset contains 500 samples with string answers.	×	0.01
The Coin Flip dataset contains 500 samples with Yes/No answers.	×	0.01
The StrategyQA dataset contains 2290 samples with Yes/No answers.	×	0.01
The CommonsenseQA dataset contains 1221 samples with option answers.	×	0.01

References

- http://arxiv.org/abs/2404.14963v5
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2508.04848v1