Frontier Language Model Failures in Abstract Mathematical Reasoning

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What are the failure modes of frontier language models on abstract mathematical reasoning v14. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ChaosBench-Logic: A Benchmark for Logical and Symbolic Reasoning on Chaotic Dynamical Systems. Research question: What are the failure modes of frontier language models on abstract mathematical reasoning v14.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
CHAOSBENCH-LOGIC specifies a set Φ of global axioms encoding widely used implications in dynamical systems.	×	0.12
CHAOSBENCH-LOGIC defines 11 unary predicates, each mapping a system s to a boolean.	×	0.07
The benchmark ground truth for each system is a truth assignment for the 11 predicates that is consistent with Φ.	×	0.06
The logical accuracy metric is defined as the average of correct predictions over Neval questions.	×	0.05
CHAOSBENCH-LOGIC uses a fixed ontology and ground truth for evaluating models on logical consistency.	×	0.09
The system annotations satisfy Φ, ensuring the closure remains consistent.	×	0.01
CHAOSBENCH-LOGIC includes continuous-time flows, discrete-time maps, PDEs, neuronal oscillators, chemical reaction model	×	0.03
The benchmark uses forward chaining under Φ to compute the correct answer for implication questions.	×	0.04
Reverse implications are not assumed in CHAOSBENCH-LOGIC to avoid overspecification and force models to respect directio	×	0.05
The logical accuracy metric is defined as the average of correct predictions over Neval questions.	×	0.05

References

- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2602.06176v1
- http://arxiv.org/abs/2510.17496v2