Frontier Language Model Failures in Abstract Mathematical Reasoning

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the failure modes of frontier language models on abstract mathematical reasoning v17. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FineGRAIN: Evaluating Failure Modes of Text-to-Image Models with Vision Language Model Judges. Research question: What are the failure modes of frontier language models on abstract mathematical reasoning v17.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

14 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
Flux achieves a success rate of 0.655 for generating single-object prompts.	×	0.04
Flux achieves a success rate of 0.103 for generating two-object prompts.	×	0.04
Flux achieves a success rate of 0.034 for generating three-object prompts.	×	0.04
SD 3.5 Large achieves a success rate of 0.483 for generating single-object prompts.	×	0.04
SDXL achieves a success rate of 0.138 for generating single-object prompts.	×	0.03
All evaluated models experience a decline in success rates as the number of objects in the prompt increases from one to	×	0.05
For the prompt 'A person hitting a hard drum that has sand on the drum', FineGRAIN identified a failure mode in Flux, SD	×	0.09
For the prompt 'A person hitting a hard drum that has sand on the drum', the VQAScore for Flux was 0.893.	×	0.01
For the prompt 'A person hitting a hard drum that has sand on the drum', the VQAScore for SD35 was 0.909.	×	0.01
For the prompt 'A person hitting a hard drum that has sand on the drum', the CLIP Score for Flux was 0.316.	×	0.01
For the prompt 'A person hitting a hard drum that has sand on the drum', the CLIP Score for SD35 was 0.266.	×	0.01
FineGRAIN outputs 1 if an image contains a failure mode and 0 otherwise.	×	0.09
The 'Counts or Multiple Objects' failure mode is defined as the model struggling to generate a precise number of distinc	×	0.05
The 'Color attribute binding' failure mode is defined as the model having difficulty correctly associating colors with s	×	0.09
In a token limit experiment, Flux achieved an average score of 0.32.	×	0.02
In a token limit experiment, SDXL achieved an average score of 0.00 across all token limits (3, 10, 20, 50).	×	0.02

References

- http://arxiv.org/abs/2512.02161v1
- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2601.01982v1