

Frontier Language Model Failures in Abstract Mathematical Reasoning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the failure modes of frontier language models on abstract mathematical reasoning v20. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ARS: Adaptive Reasoning Suppression for Efficient Large Reasoning Language Models. Research question: What are the failure modes of frontier language models on abstract mathematical reasoning v20.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

4 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 study evaluates Qwen2.5-Math-1.5B-Instruct, Qwen2.5-Math-7B-Instruct, and DeepSeek-R1-Distill-Qwen-7B models.	×	0.01
The ARS algorithm utilizes difficulty thresholds d1 and d2, and confidence thresholds c1, c2, and c3.	×	0.03
In FAST mode, the ARS policy is configured with 2 drafts and 10 tokens per draft.	×	0.01
In MOD mode, the ARS policy uses a budget of 64 tokens.	×	0.01
In the default mode, the ARS policy uses a sc_k parameter of 3.	×	0.02
The generation process terminates if the text length reaches 1200 tokens.	×	0.03
Confidence scores in ARS are computed using entropy confidence on tentative answers.	×	0.05
Token suppression occurs if the next token is in the trigger set and the suppression probability exceeds a random value.	×	0.05
The maximum token limit per response is set to 1200 tokens to prevent excessive generation.	×	0.04
Reasoning reflection behaviors are triggered by specific keywords including 'Wait', 'But', and 'Alternatively'.	×	0.03
The objective of ARS is to minimize expected output length $E[T]$ subject to an accuracy degradation threshold ϵ .	×	0.03
ARS operates through three core components: Multi-checkpoint certainty estimation, Progressive threshold adaptation, and	×	0.15
ARS establishes multiple checkpoints at regular intervals during generation, unlike previous methods that rely on single	×	0.06
Table 1 presents performance comparison results on the GSM8K dataset.	×	0.05
Table 2 presents performance comparison results on the MATH500 dataset.	×	0.04
ARS consistently achieves superior length reduction while maintaining competitive accuracy across all tested model scale	×	0.10

References

- <http://arxiv.org/abs/2510.00071v2>
- <http://arxiv.org/abs/2508.04848v1>
- <http://arxiv.org/abs/2309.02144v1>