Language Models in Formal Theorem Proving and Mathematical Verification Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do language models perform on formal theorem proving and mathematical verification tasks v18. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Safe: Enhancing Mathematical Reasoning in Large Language Models via Retrospective Step-aware Formal Verification. Research question: How do language models perform on formal theorem proving and mathematical verification tasks v18.

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 2.8/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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 LSTM model achieves comparable performance to other SOTA ORMs and PRMs despite its high parameter efficiency and min	×	0.07
The Safe framework, integrating LSTM with a PRM, consistently outperforms almost every other baseline model across all d	×	0.05
The Safe approach demonstrates significant improvements on the MATH-500 and CollegeMath datasets.	×	0.03
The mediocre improvement on the GSM8K dataset is attributed to its low difficulty level and data imbalance.	×	0.04
For stronger models like GPT-4o and Llama 3.1, allocating additional computational resources during the verification pro	×	0.03
For less powerful models like Llama 3.0, increasing the sample count with a weaker yet more cost-effective RM remains an	×	0.02
There is a synergistic effect between the PRM and the formal step verifier in the Safe framework.	×	0.07
The Safe framework integrates retrospective formal scores with those of a prospective PRM.	×	0.07
The formal verifier in the Safe framework focuses on the correctness of each step, constituting retrospective verificati	×	0.09
PRMs are typically trained using a loss function that evaluates the likelihood of achieving a correct outcome in the fut	×	0.02

References

- http://arxiv.org/abs/2404.12534v3
- http://arxiv.org/abs/2506.04592v1
- http://arxiv.org/abs/2505.20613v3