Language Model Performance on Formal Theorem Proving and Mathematical Verification

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 v19. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Neural Theorem Proving: Generating and Structuring Proofs for Formal Verification. Research question: How do language models perform on formal theorem proving and mathematical verification tasks v19.

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 3.5/10.

3 Results

16 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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 ProofSeek framework uses 8 Isabelle proof methods (auto, simp, auto, blast, fastforce, eval, sos, arith, simp:field	×	0.03
The timeout for any proof step and Sledgehammer is set to 10 seconds and 40 seconds, respectively.	×	0.03
The experiments were run on two machines: one with an AMD EPYC 7763 64-Core Processor CPU @ 2.49GHz and a NVIDIA A40-48Q	×	0.02
The miniF2F-Test dataset includes 488 formal mathematical problems, split into a validation set and a test set, each con	×	0.06
The Isabelle part of the miniF2F-test dataset contains an additional informal statement and informal draft for each prob	×	0.06
Access is granted in AWS S3 Bucket Policies if and only if there exists at least one statement in the policy that allows	×	0.09
The ProofSeek framework consists of two core components: a component for fine-tuning a language model using SFT and RL,	×	0.08
The framework is generalizable across domains where the input is a mathematical statement, policy code, or natural langu	×	0.08
The fine-tuning process involves both supervised and reinforcement learning stages.	×	0.05

References

- http://arxiv.org/abs/2504.17017v2
- http://arxiv.org/abs/2506.04592v1
- http://arxiv.org/abs/2605.02790v2