Language Models in Formal Theorem Proving and Mathematical Verification Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do language models perform on formal theorem proving and mathematical verification tasks v12. 7 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: Exploring Length Generalization in Large Language Models. Research question: How do language models perform on formal theorem proving and mathematical verification tasks v12.

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 3.7/10.

3 Results

4 papers retrieved. 7 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
Fine-tune, Prompting, Fine-tune + Prompting, Fine-tune + Scratchpad, Prompting + Scratchpad, and Fine-tune + Prompting +	×	0.05
Models of different scales (244m, 422m, 1b, 64b) were used in the study.	×	0.03
Training configurations included different learning rates (2e-05, 2e-04, 2e-03) and batch sizes (32, 128).	×	0.02
The Boolean Variable Assignment task involves executing Python programs line by line to determine the value of the final	×	0.00
The diverse variable assignment split includes a wide range of boolean operators and maximally diverse programs.	×	0.02
The chain-like variable assignment split is a variant of the dataset.	×	0.02
Length generalization performance does not improve with model scale, as shown in Figure 3.	×	0.14

References

- https://www.semanticscholar.org/paper/a99474d3e90465a466b916a8b854b699537cf9e7
- https://www.semanticscholar.org/paper/87cc0acccfbc6011a47ee1b031213b0c72b9761a
- https://arxiv.org/abs/2207.04901