Language Models in Formal Theorem Proving and Mathematical Verification

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do language models perform on formal theorem proving and mathematical verification tasks v11. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FVEL: Interactive Formal Verification Environment with Large Language Models via Theorem Proving. Research question: How do language models perform on formal theorem proving and mathematical verification tasks v11.

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 4.3/10.

3 Results

14 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
The Code2Inv dataset contains 133 programs in C.	×	0.03
The SV-COMP dataset contains over 23k C programs.	×	0.08
C-parser supports only part of the C99 standard.	×	0.02
The FVELER lemma distribution over step intervals is adjusted by setting the y-axis to a logarithmic scale.	×	0.04
LORA is used to fine-tune two advanced open-source large language models: Llama-3-8B-instruct4 and Mistral-7B-Instruct-v	×	0.08
The training data is converted into the alpaca format for fine-tuning.	×	0.05
During inference, the language model generates a lemma specification to verify that it satisfies the specifications.	×	0.05
The language model interacts with PISA for proof verification.	×	0.05
UAUTOMIZER is the overall champion of the 12th Competition on Software Verification (SV-COMP 2023).	×	0.06
ESBMC is based on K-induction, which is particularly useful for verifying the properties of loops and recursive function	×	0.04
Lemur presents a set of derivation rules and makes proposals using a language model to approximate the boundary conditio	×	0.04
The evaluation follows the settings of Lemur.	×	0.03
The methods compared include symbolic solvers: Uautomizer and ESBMC, and the LLM-based method: Lemur.	×	0.03
Formal verification for C code in the Isabelle environment is a great challenge.	✓	0.16

References

- http://arxiv.org/abs/2406.09757v2
- http://arxiv.org/abs/2506.04592v1
- http://arxiv.org/abs/2406.14408v2