Language Model Performance on Formal Theorem Proving and Mathematical Verification

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do language models perform on formal theorem proving and mathematical verification tasks v15. 17 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PutnamBench: Evaluating Neural Theorem-Provers on the Putnam Mathematical Competition. Research question: How do language models perform on formal theorem proving and mathematical verification tasks v15.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

15 papers retrieved. 17 claims extracted; 1 independently verified. Quality review score: 4.8/10.

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
PUTNAMBENCH exceeds prior benchmarks by providing support for all of Lean 4, Isabelle, and Coq, on a set of difficult co	✓	0.15
PUTNAMBENCH contains 640 problems.	×	0.07
PUTNAMBENCH includes problems from categories such as Algebra, Analysis, Number Theory, Geometry, and Linear Algebra.	×	0.04
PUTNAMBENCH contains 253 Algebra problems, 226 Analysis problems, 107 Number Theory problems, 68 Geometry problems, and	×	0.03
GPT-4 was prompted in a pass@10 setting with temperature T = 0.7 to generate proofs for each problem in PUTNAMBENCH.	×	0.02
GPT-4 successfully generated a proof for one problem (Putnam 1988 B1) in the Lean 4 formalizations of PUTNAMBENCH.	×	0.08
COPRA, with default hyperparameters for search and a pass@1, successfully generated a proof for one problem (1988 B1) in	×	0.04
ReProver, both with and without retrieval, failed to generate any successful proofs in the Lean 4 formalizations of PUTN	×	0.04
GPT-4 successfully generated a proof for one problem in the Isabelle formalizations of PUTNAMBENCH.	×	0.06
DSP successfully generated proofs for four problems in the Isabelle formalizations of PUTNAMBENCH.	×	0.07
Sledgehammer successfully generated proofs for three problems in the Isabelle formalizations of PUTNAMBENCH.	×	0.07
GPT-4 successfully generated a proof for one problem in the Coq formalizations of PUTNAMBENCH.	×	0.08
COPRA successfully generated a proof for one problem in the Coq formalizations of PUTNAMBENCH.	×	0.08
Tactician and CoqHammer failed to generate any successful proofs in the Coq formalizations of PUTNAMBENCH.	×	0.06
The only problem solved in both Lean and Coq is Putnam 1988 B1, which is not solved by any method in Isabelle.	×	0.05
The evaluation is based on the pass@n metric, which measures a prover's ability to produce a successful proof given a bu	×	0.03
GPT-4 was used for evaluations in each of the languages (Lean, Isabelle, Coq).	×	0.09

References

- http://arxiv.org/abs/2401.12061v1
- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2407.11214v2