Psychometric-Based vs. Proof Pass Rate Metrics in Large-Scale Theorem Proving Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the psychometric-based evaluation method compare to traditional proof pass rate metrics in terms of accuracy and computational efficiency when applied to large-scale theorem proving. 9 claims were extracted from source literature; 2 were 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: Psychometric-Based Evaluation for Theorem Proving with Large Language Models. Research question: How does the psychometric-based evaluation method compare to traditional proof pass rate metrics in terms of accuracy and computational efficiency when applied to large-scale theorem proving benchmarks like miniF2F and MPTP Challenge?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 9 claims extracted; 2 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
The miniF2F-Graded dataset contains 488 theorems classified by type.	×	0.10
The scatter plot in Figure 2 shows the relationship between theorem difficulty and discrimination.	×	0.08
Theorems in the difficulty range of 0.4–0.6 generally demonstrate higher discrimination.	×	0.05
The evaluation method proposed reduces the evaluation cost for both Annotation LLMs and Evaluation LLMs by an average of	✓	0.16
The Ability Score and Pass@128 rankings are largely the same.	×	0.04
The miniF2F dataset is one of the most widely used datasets in the field of theorem proving with LLMs.	✓	0.17
The MiniF2F dataset contains theorems formalized in Metamath, Lean, Isabelle, and HOL Light.	×	0.05
260 theorems in the MiniF2F dataset are derived from the MATH dataset.	×	0.08
Theorems in MATH are classified into five difficulty levels.	×	0.08

References

- http://arxiv.org/abs/2502.00855v1
- http://arxiv.org/abs/2205.06203v1
- http://arxiv.org/abs/2504.13359v2