Neuro-Symbolic vs. Neural Proof Generation Robustness to Adversarial Perturbations

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do neuro-symbolic proof generation methods perform in terms of robustness against adversarial perturbations in theorem statements compared to end-to-end neural approaches on formal mathematics. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: NeuroShield: A Neuro-Symbolic Framework for Adversarial Robustness. Research question: How do neuro-symbolic proof generation methods perform in terms of robustness against adversarial perturbations in theorem statements compared to end-to-end neural approaches on formal mathematics benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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 proposed NeuroShield is evaluated on the GTSRB dataset and achieves over 98% accuracy on clean (non-adversarial) inp	×	0.08
The model is trained from scratch for only 10 epochs, without any pre-training.	×	0.07
Under FGSM attack with ε = 8/255, the FGSM-Neuro-Symbolic model achieves 63.10% accuracy.	×	0.11
Under PGD attack with ε = 8/255, the PGD-Neuro-Symbolic model achieves 56.00% accuracy.	×	0.10
The FGSM-Neuro-Symbolic model improves FGSM accuracy from 45.0% (FGSM adversarial training) to 63.10%, an improvement of	×	0.13
The PGD-Neuro-Symbolic model improves PGD accuracy from 38.65% (PGD adversarial training) to 56.00%, a gain of +17.35%.	×	0.15
The pure Neuro-Symbolic model (without any adversarial training) maintains high clean accuracy ($\sim$97.90%) but drops under	×	0.10
The PGD-Neuro-Symbolic model achieves 56.00% PGD accuracy, slightly outperforming LNL-MoEx-Ti (55.4%) and approaching th	×	0.12
Under FGSM attack with ε = 8/255, the Neuro-Symbolic model reaches 63.10% accuracy, which is competitive with LNL-S (64.	×	0.07
LNL-MoEx-S achieves the highest FGSM accuracy (77.8%).	×	0.06

References

- http://arxiv.org/abs/2601.13162v1
- http://arxiv.org/abs/1801.04693v1
- http://arxiv.org/abs/2404.12534v3