Synthetic Training Diversity Enhances DeepSeek Coder Robustness to Code Perturbations

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Does increasing the diversity of synthetic training samples improve DeepSeek Coder's resistance to semantic-preserving code perturbations as measured by accuracy drop on the MBPP benchmark. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AdaShield: Safeguarding Multimodal Large Language Models from Structure-based Attack via Adaptive Shield Prompting. Research question: Does increasing the diversity of synthetic training samples improve DeepSeek Coder's resistance to semantic-preserving code perturbations as measured by accuracy drop on the MBPP benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

16 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
AdaShield-S and AdaShield-A outperform FSD and MLLMP in defending against FigStep and QR attacks.	×	0.03
AdaShield-S exhibits inferior defense performance compared to AdaShield-A due to the absence of specific safety rules.	×	0.04
MLLMP employs a harmful detector and a detoxifier but fails in defending against jailbreak attacks due to limited genera	×	0.08
The harmful detector in MLLMP exhibits a mere accuracy of 4.34% in the 'Pornography' scenario of QR with target MLLM as	×	0.04
AdaShield-A outperforms MLLMP and FSD, and achieves performance comparable to the Vanilla in benign dataset performance.	×	0.04
AdaShield-S falls short at recognizing benign queries, leading to performance degradation caused by over-defense.	×	0.04
AdaShield-A excels in mitigating over-defense by filtering benign queries based on similarity.	×	0.06
Pa does not contain specific instructions to check the image content, but only vaguely guides the model to examine the i	×	0.03
Pb requires the model to check the content of the image but lacks a chain-of-thought.	×	0.03
Pc only requires the model to refuse to engage in illicit activities, but lacks a clear and actionable plan.	×	0.02
Pd is only the first step of Ps, which involves examining whether the image contains harmful text or items.	×	0.05
Pe is only the second step of Ps, which forces the model to combine the content of pictures and text.	×	0.04
AdaShield-A has a lower latency compared to FSD, MLLMP, and AdaShield-S.	×	0.02
AdaShield-A uses input-aware defense prompts that include specific safety rules for different scenarios.	×	0.05

References

- http://arxiv.org/abs/2403.09513v1
- http://arxiv.org/abs/2506.16243v1
- http://arxiv.org/abs/2205.14230v2