Llama3 and Codestral Safety Filter Performance in Adversarial Code Generation

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do Llama3 and Codestral differ in false positive rates and latency overhead when enforcing safety constraints on adversarial code generation tasks. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Stress Testing of SPARK Humanoid Safety Filters. Research question: How do Llama3 and Codestral differ in false positive rates and latency overhead when enforcing safety constraints on adversarial code generation tasks?.

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 5.3/10.

3 Results

16 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 5.3/10.

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 study replicates and analyzes the SPARK case G1SportMode_D1_WG_SO_v1.	×	0.09
The study evaluates six safety filters: RSSA, RSSS, SSA, CBF, PFM, and SMA.	✓	0.22
The evaluation uses a shared benchmark setting, fixed time horizon, and controlled seeds.	×	0.04
A post-processing pipeline was built to convert high-dimensional .npz logs into goal-tracking, minimum-distance, and col	✓	0.28
In baseline results, the PFM filter tracks the goal closely but has more collision steps.	×	0.12
In baseline results, the SMA filter yields the lowest average environment-collision count.	×	0.04
In baseline results, SSA, RSSA, and RSSS filters show more balanced behavior compared to PFM and SMA.	×	0.11
The experiments observed long runtimes and repeated 'No Solution' outputs, suggesting feasibility limits when constraint	×	0.02
The SPARK humanoid safety benchmark was replicated in MuJoCo.	✓	0.20
The study stress-tests filters using obstacle crowding, perception noise, and sensor latency.	×	0.14
The benchmark case G1SportMode_D1_WG_SO_v1 uses the Unitree G1 humanoid in SportMode with first-order dynamics, a whole-	×	0.09
For baseline replication, the study used random seeds 20, 21, and 22.	×	0.06
Each baseline replication run consisted of 5000 simulation steps.	×	0.03
Artificial potential fields use repulsive terms to push robots away from obstacles.	×	0.02
Control Barrier Functions (CBFs) provide an optimization-based framework for enforcing forward invariance of safe sets i	×	0.04
Wei and Liu organized filters including PFM, SSA, CBF, SMA, and SSS under a unified framework to study safety–efficiency	×	0.08

References

- http://arxiv.org/abs/1804.09888v1
- http://arxiv.org/abs/2605.19009v1
- http://arxiv.org/abs/2507.20135v1