Impact of Sampling Strategies on F1-Score Stability in Code Vulnerability Detection

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of different sampling strategies (e.g., stratified, random) on the stability of F1-scores for Llama3, Codestral, and Deepseek R1 when evaluated on code vulnerability detection. Large language models (LLMs) are increasingly used in software development, but their level of software security expertise remains unclear. This work systematically evaluates the security comprehension of five leading LLMs: GPT-4o-Mini, GPT-5-Mini, Gemini-2.5-Flash, Llama-3.1. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Assessing the Software Security Comprehension of Large Language Models. Research question: What is the impact of different sampling strategies (e.g., stratified, random) on the stability of F1-scores for Llama3, Codestral, and Deepseek R1 when evaluated on code vulnerability detection benchmarks with varying contamination rates?.

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 5.0/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.0/10.

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.

References

- http://arxiv.org/abs/2506.11022v2
- http://arxiv.org/abs/2512.21238v1
- http://arxiv.org/abs/2504.16584v1