Training Data Overlap and Performance Inflation in LLaMA 3.2 and Mistral Code Repair

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: To what extent does training data overlap with the BugsInPy dataset inflate the reported code repair performance of LLaMA 3.2 and Mistral. 10 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the role of context in AI-driven program repair and test generation. Research question: To what extent does training data overlap with the BugsInPy dataset inflate the reported code repair performance of LLaMA 3.2 and Mistral?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

1 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.2/10.

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
Learning-based techniques show promise for automating software development tasks.	✓	0.25
Current approaches treat context in an ad hoc manner.	✓	0.20
Existing techniques select context through arbitrary heuristics, such as fixed token windows, enclosing methods, or enti	✓	0.28
The goal of the work presented in this dissertation is to systematically leverage different forms of context to improve	✓	0.33
A graph-to-sequence learning approach that captures semantic context through program analysis outperforms state-of-the-a	✓	0.34
A retrieval-based technique for selecting demonstration examples during few-shot prompting outperforms task-specific and	✓	0.38
An automated technique for generating issue-reproducing tests from natural language bug reports successfully generates r	✓	0.32
The technique generates reproducing tests for cases uniquely solved by the approach that were missed by all prior work.	✓	0.25
The complexity of multi-hunk patches is characterized through empirical analysis of real-world bugs.	✓	0.22
Hunk divergence is introduced as a metric.	×	0.06

References

- `https://openalex.org/W7159897974`