What is the impact of model size scaling on code repair accuracy for LLaMA 3.2 and Mistral when quantized to 4

Assignee Research

June 10, 2026

Abstract

Large language models (LLMs) have demonstrated strong performance on a wide range of software engineering tasks, including code generation and analysis. However, most prior work relies on cloud-based models or specialized hardware, limiting practical applicability in privacy-sensitive or resource-constrained environments. In this paper, we present a systematic empirical evaluation of two locally deployed LLMs, LLaMA 3.2 and Mistral, for real-world Python bug detection using the BugsInPy benchmark. We evaluate 349 bugs across 17 projects using a zero-shot prompting approach at the function leve

1 Introduction

This paper examines: An Empirical Evaluation of Locally Deployed LLMs for Bug Detection in Python Code. Research question: What is the impact of model size scaling on code repair accuracy for LLaMA 3.2 and Mistral when quantized to 4-bit versus FP16 on the BugsInPy dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2306.09896v5
- http://arxiv.org/abs/2411.07586v1
- http://arxiv.org/abs/2604.23361v1