Robustness of Quantized LLaMA 3.2 and Mistral to Syntactic Bug Description Variations

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the robustness of LLaMA 3.2 and Mistral to syntactic variations in bug descriptions change under 4-bit quantization compared to FP16 precision. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can We Enhance Bug Report Quality Using LLMs?: An Empirical Study of LLM-Based Bug Report Generation. Research question: How does the robustness of LLaMA 3.2 and Mistral to syntactic variations in bug descriptions change under 4-bit quantization compared to FP16 precision?.

2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 papers retrieved. 12 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.

5 Extracted Claims

Claim	Verified	Confidence
The study filtered bug reports to include only those containing descriptions, Steps to Reproduce (S2Rs), Expected Behavi	×	0.12
Bug reports containing stack traces or code snippets were excluded from the training dataset to avoid noise and complexi	×	0.03
Bug reports with a CTQRS score greater than 14 were retained based on the threshold defined by Zheng et al. [74].	×	0.07
After filtration, the dataset contained 3,966 bugs with all required information.	×	0.03
200 bug reports were manually reviewed to verify quality after the initial filtration steps.	×	0.06
The Llama3 model was used to generate unstructured bug report summaries from the 3,966 well-structured reports.	×	0.09
Each unstructured report was generated three times during the synthetic data generation process.	×	0.04
Generated reports were retained only if they achieved an SBERT similarity score exceeding 85% and a cosine similarity sc	×	0.03
The final synthetic dataset comprised 3,903 well-structured bug reports paired with their generated summaries.	×	0.06
The study employs both qualitative and quantitative evaluation metrics, whereas GIRT, ChatBR, and BugBlitz employ only q	×	0.02
ChatBR utilizes a Few-Shot adaptation technique, while GIRT, BugBlitz, and the current study utilize Fine-tuning.	×	0.05
ChatBR uses the ChatGPT model, while GIRT, BugBlitz, and the current study use Open Source models.	×	0.12

References

- http://arxiv.org/abs/2504.18804v1
- http://arxiv.org/abs/2512.22671v2

- http://arxiv.org/abs/2102.07660v2