Mixed-Precision Quantization Effects on Token Throughput in Llama-3.1-8B for CWE Detection

Assignee Research

June 13, 2026

Abstract

Large Language Models (LLMs) have demonstrated significant capabilities in understanding and analyzing code for security vulnerabilities, such as Common Weakness Enumerations (CWEs). However, their reliance on cloud infrastructure and substantial computational requirements pose challenges for analyzing sensitive or proprietary codebases due to privacy concerns and inference costs. This work explores the potential of Small Language Models (SLMs) as a viable alternative for accurate, on-premise vulnerability detection. We investigated whether a 350-million parameter pre-trained code model (codeg

1 Introduction

This paper examines: Case Study: Fine-tuning Small Language Models for Accurate and Private CWE Detection in Python Code. Research question: What is the impact of using mixed-precision (FP8/INT8) quantization during inference on token-per-second throughput while maintaining a minimum F1-score threshold for CWE detection in Llama-3.1-8B fine-tuned on Python datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.7/10.

3 Results

15 papers retrieved. 12 claims extracted; 12 independently verified. Quality review score: 6.7/10.

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
Farasat & Posegga [15] achieved an average accuracy of 98.6%, F1 score of 94.7%, precision of 96.2%, recall of 93.3%, an	✓	0.16
Bagheri et al. [31] reported an F1 score of 99% using a hybrid ML model (Self-attention + CNN - Conformer).	✓	0.21
Singh et al. [32] achieved an accuracy of 0.66, precision of 0.65, recall of 0.66, and F1 score of 0.64 for CWE predicti	✓	0.16
Dozono et al. [6] reported Python F1 scores for various LLMs: GPT-4o=0.80, GPT-4T=0.76, Gemini 1.5 Pro=0.75, CodeLlama-7	✓	0.31
Steenhoek et al. [34] evaluated LLMs on C/C++ and found a low balanced accuracy of 54.5%.	✓	0.17
Shestov et al. [35] achieved the best F1 score of 0.86 for binary classification using WizardCoder for JAVA CWE detectio	✓	0.23
Li et al. [36] reported a 5-6% improvement over the base model using LoRa and IA3 fine-tuning approach for LLMs.	✓	0.23
Jiang et al. [37] achieved the best F1 score of 87% using Llama 2-7b model with LoRa fine-tuning approach.	✓	0.22
The un-tuned codegen-mono model failed to detect a single CWE within any of the code snippets presented in a baseline ev	✓	0.30
The fine-tuned codegen-mono model achieved an accuracy of 99%, precision of $\approx$98.08%, recall of 100%, and F1-score of $\approx$99	✓	0.43
The MITRE Top 25 Most Dangerous Software Weaknesses list [30] was selected as the target for the detection model.	✓	0.21
Google's gemini-2.0-flash-thinking-exp-01-21 model was used via its API for generating synthetic data.	✓	0.17

References

- http://arxiv.org/abs/2508.03332v2

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