Quantization Impact on CodeLlama-7B Performance in Binary Analysis Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How do different quantization techniques (e.g., 8-bit, 4-bit) affect the performance of codellama-7b-hf on binary analysis tasks compared to the float16 baseline. 9 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Enriching Location Representation with Detailed Semantic Information. Research question: How do different quantization techniques (e.g., 8-bit, 4-bit) affect the performance of codellama-7b-hf on binary analysis tasks compared to the float16 baseline?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

7 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 5.3/10.

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
Cyber-physical systems (CPS) are critical to modern infrastructure.	✓	0.22
Cyber-physical systems (CPS) are vulnerable to faults and anomalies that threaten their operational safety.	✓	0.26
The work evaluates the use of open-source Large Language Models (LLMs), such as Mistral 7B, Llama3.1:8b-instruct-fp16, a	✓	0.43
The methodology utilises retrieval-augmented generation (RAG) techniques, incorporating a novel two-step process where L	✓	0.41
The original prompt design yielded strong results for the battery dataset but required modification for the powertrain d	✓	0.37
The adjusted prompt, which emphasises rule inference, significantly improved anomaly detection for the powertrain datase	✓	0.34
Models like Mistral 7B achieved F1-scores up to 0.99.	✓	0.29
Llama3.1:8b-instruct-fp16 and Gemma 2 reached perfect F1-scores of 1.0 in complex scenarios.	✓	0.35
These findings demonstrate the impact of effective prompt design and rule inference in improving LLM-based fault detecti	✓	0.41

References

- https://doi.org/10.1109/access.2019.2909490
- https://doi.org/10.1186/s42400-025-00361-w
- https://doi.org/10.4230/lipics.giscience.2025.3