Codestral-7B and Codestral-70B Throughput-Latency Trade-offs in Vulnerability Classification

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 2 peer-reviewed papers addressing the following research question: What is the throughput and latency trade-off between Codestral-7B and Codestral-70B when classifying vulnerabilities in Big-Vul under varying levels of parallelized inference and model quantization. 5 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Least squares quantization in PCM. Research question: What is the throughput and latency trade-off between Codestral-7B and Codestral-70B when classifying vulnerabilities in Big-Vul under varying levels of parallelized inference and model quantization?.

2 Methodology

Systematic literature search across multiple databases yielded 2 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

2 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 8.7/10.

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
In pulse-code modulation (PCM), with a given ensemble of signals to handle, the quantum values should be spaced more clo	✓	0.33
In the limit as the number of quanta becomes infinite, the asymptotic fractional density of quanta per unit voltage shou	✓	0.36
The optimization criterion used is that the average quantization noise power be a minimum.	✓	0.22
The result obtained here goes over into the Panter and Dite result as the number of quanta become large.	✓	0.28
The optimum quantization schemes for 2^b quanta, b=1,2,\ldots{},7, are given numerically for Gaussian and for Laplacian distrib	✓	0.23

References

- https://doi.org/10.1109/tcom.1980.1094577
- https://doi.org/10.1109/tit.1982.1056489