Batch Size Scaling and Latency Degradation in Codestral Static Code Classification

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the correlation between batch size scaling and latency degradation for Codestral models when performing static analysis code classification. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Impact of White-box Deployment Strategies for Edge AI on Latency and Model Performance. Research question: What is the correlation between batch size scaling and latency degradation for Codestral models when performing static analysis code classification?.

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 8.2/10.

3 Results

7 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.2/10.

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 empirically assesses the accuracy vs latency trade-off of white-box and black-box operators and their combinat	✓	0.31
The study includes inference experiments with 3 white-box operators (QAT, Pruning, Knowledge Distillation), 2 black-box	✓	0.35
The experiments are conducted across 3 tiers (Mobile, Edge, Cloud) on 4 commonly-used Computer Vision and Natural Langua	✓	0.29
The combination of Distillation and SPTQ operators (DSPTQ) should be preferred over non-hybrid operators when lower late	✓	0.41
Among the non-hybrid operators, the Distilled operator is a better alternative in both mobile and edge tiers for lower l	✓	0.46
Operators involving distillation show lower latency in resource-constrained tiers (Mobile, Edge) compared to the operato	✓	0.39
For textual subject models, which have low input data size requirements, the Cloud tier is a better alternative for the	✓	0.39

References

- http://arxiv.org/abs/2410.21676v4
- http://arxiv.org/abs/2507.07101v4
- http://arxiv.org/abs/2411.00907v3