Quantization-Induced False Positive Rates in LLaMA 3.2 and Mistral on BugsInPy

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does 4-bit quantization impact the false positive rate of LLaMA 3.2 versus Mistral on the BugsInPy dataset compared to FP16 baselines. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MuonQ: Enhancing Low-Bit Muon Quantization via Directional Fidelity Optimization. Research question: How does 4-bit quantization impact the false positive rate of LLaMA 3.2 versus Mistral on the BugsInPy dataset compared to FP16 baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.9/10.

3 Results

4 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.9/10.

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
MuonQ4 closely tracks the full-precision Muon32 baseline in training loss curves across all model scales.	×	0.06
Naive Muon4 exhibits a persistent loss gap that does not diminish over the course of training.	×	0.03
The advantage of MuonQ4 over naive quantization becomes more pronounced as model scale increases.	×	0.03
MuonQ4 consistently matches the Muon32 baseline on various benchmarks, including ARC-Challenge, ARC-Easy, OpenBookQA, Bo	×	0.02
MuonQ4 reduces optimizer memory by 7.3$\times$ on average compared to Muon32.	×	0.03
MuonQ4 incurs only a modest memory increase compared to Muon4 but recovers the majority of the full-precision performanc	×	0.01
Pre-quantization normalization prevents non-uniform error accumulation in MuonQ.	×	0.10
Structural decomposition ensures momentum orthogonalization stability in MuonQ.	×	0.09
μ-law companding quantization improves resolution in the dense near-zero region in MuonQ.	×	0.11
MuonQ uses relative error (RE) and cosine similarity (CS) to evaluate quantization quality.	×	0.04

References

- http://arxiv.org/abs/2605.11396v1
- http://arxiv.org/abs/nucl-ex/0511007v1
- http://arxiv.org/abs/2310.06825v1