Quantized Small Language Models: Accuracy-Throughput Trade-offs at 4-Bit and 8-Bit on ARM and x86 Hardware

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does the accuracy-throughput trade-off of 4-bit quantized SLMs under 10B parameters compare to 8-bit quantization across SLM-Bench's 9 NLP tasks when evaluated on ARM and x86 hardware. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts–Extended Version. Research question: How does the accuracy-throughput trade-off of 4-bit quantized SLMs under 10B parameters compare to 8-bit quantization across SLM-Bench's 9 NLP tasks when evaluated on ARM and x86 hardware configurations?.

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 6.8/10.

3 Results

7 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.8/10.

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.

References

- http://arxiv.org/abs/2310.16836v1
- http://arxiv.org/abs/2508.15478v2
- http://arxiv.org/abs/2309.13773v1