

Sustainability Metrics of Quantized Small Language Models Across GPU and Edge Devices

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the relationship between model size (1B-10B parameters) and the sustainability metrics (e.g., energy efficiency) of quantized SLMs on SLM-Bench when deployed on GPU vs. edge devices. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.7/10. 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: What is the relationship between model size (1B-10B parameters) and the sustainability metrics (e.g., energy efficiency) of quantized SLMs on SLM-Bench when deployed on GPU vs. edge devices?.

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 5.7/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.7/10.

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.

References

- <http://arxiv.org/abs/2501.16273v2>
- <http://arxiv.org/abs/2508.15478v2>
- <http://arxiv.org/abs/astro-ph/0510346v1>