Quantization and Hardware Effects on Small Language Model Throughput in SLM-Bench

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the inference throughput of small language models on SLM-Bench tasks vary across different quantization levels and hardware accelerators. Edge computing enables real-time data processing closer to its source, thus improving the latency and performance of edge-enabled AI applications. However, predictive AI models often fall short when dealing with complex, dynamic tasks that require advanced reasoning and. 11 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Toward Edge General Intelligence With Multiple-Large Language Model (Multi-LLM): Architecture, Trust, and Orchestration. Research question: How does the inference throughput of small language models on SLM-Bench tasks vary across different quantization levels and hardware accelerators?.

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 8.1/10.

3 Results

4 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.1/10.

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
Edge computing enables real-time data processing closer to its source.	✓	0.26
Edge computing improves the latency and performance of edge-enabled AI applications.	✓	0.20
Predictive AI models often fall short when dealing with complex, dynamic tasks that require advanced reasoning and multi	✓	0.33
The survey explores the integration of multi-LLMs to address challenges in edge computing where multiple specialized LLM	✓	0.29
The survey reviews the transition from conventional edge AI models to single LLM deployment and ultimately to multi-LLM	✓	0.30
Dynamic orchestration, resource scheduling, and cross-domain knowledge transfer are enabling technologies key for multi-	✓	0.29
Trusted multi-LLM systems ensure robust decision-making in environments where reliability and privacy are crucial.	✓	0.26
Multimodal multi-LLM architectures involve multiple LLMs specializing in handling different data modalities such as text	✓	0.24
Multimodal multi-LLM architectures integrate outputs from specialized LLMs for comprehensive analysis.	✓	0.17
Future directions for multi-LLM systems include improving resource efficiency and establishing trustworthy governance.	✓	0.21
Future directions for multi-LLM systems include addressing privacy, trust, and robustness concerns.	✓	0.24

References

- https://doi.org/10.1109/tccn.2025.3612760
- https://doi.org/10.1109/access.2025.3610994
- https://doi.org/10.48550/arxiv.2409.00088