MQuant and Structured Pruning Trade-offs in LLaVA Multimodal Reasoning Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the throughput-accuracy trade-off of MQuant's post-training quantization on LLaVA compare to structured pruning methods when evaluated on multimodal reasoning tasks. Post-training quantization (PTQ) of large language models (LLMs) to extremely low bit-widths remains challenging due to the fundamental trade-off between computational efficiency and representational capacity. While existing ultra-low-bit methods rely on binary approximations or. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models. Research question: How does the throughput-accuracy trade-off of MQuant's post-training quantization on LLaVA compare to structured pruning methods when evaluated on multimodal reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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/2509.16989v3
- http://arxiv.org/abs/2410.08119v3
- http://arxiv.org/abs/2509.23661v3