Quantization Effects on DeepSeek R1 and Codestral Code Generation Efficiency

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the impact of quantization techniques (e.g., 4-bit, 8-bit) on the inference efficiency (throughput, latency) of DeepSeek R1 when generating optimized Python code suggestions compared to. The growing demand for Large Language Models (LLMs) in applications such as content generation, intelligent chatbots, and sentiment analysis poses considerable challenges for LLM service providers. To efficiently use GPU resources and boost throughput, batching multiple requests. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Atom: Low-bit Quantization for Efficient and Accurate LLM Serving. Research question: What is the impact of quantization techniques (e.g., 4-bit, 8-bit) on the inference efficiency (throughput, latency) of DeepSeek R1 when generating optimized Python code suggestions compared to Codestral?.

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 3.7/10.

3 Results

7 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2310.19102v3
- http://arxiv.org/abs/2507.07145v1
- http://arxiv.org/abs/2505.02390v2