Quantization Impact on Reasoning Capabilities in Large Language Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does model quantization affect reasoning capability in large language models v13. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency. Research question: How does model quantization affect reasoning capability in large language models v13.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

13 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The Llama 3.2 1B Instruct model in FP16 precision has a parameter size of 1.2B and a model size of 2364.74 MB.	×	0.02
The Llama 3.2 1B Instruct model quantized to Q8_0 has a model size of 1259.90 MB.	×	0.03
The Llama 3.2 1B Instruct model quantized to Q4_K_M has a model size of 770.29 MB.	×	0.03
The Llama 3.2 1B Instruct model quantized to Q3_K_S has a model size of 611.98 MB.	×	0.03
The Qwen 2.5 1.5B Instruct model quantized to Q4_1 has a parameter size of 1.5B and a model size of 969.75 MB.	×	0.03
The average score for the Llama 3.2 1B model in FP16 precision is 159.42.	×	0.03
The average score for the Llama 3.2 1B model with Q8_0 quantization is 75.85.	×	0.05
The average score across Q4 variants for the Llama 3.2 1B model is 83.95.	×	0.04
The average score across Q3 variants for the Gemma 2 2B model is 255.79.	×	0.03
The overall average score for the Qwen 2.5 1.5B model across all tested configurations is 110.36.	×	0.02
Post-training quantization (PTQ) applies quantization to a fully trained model without modifying the original training p	×	0.09
PTQ converts high-precision weights, typically 32-bit floating-point (FP32), to lower-bit formats such as INT8, INT4, or	×	0.07
Weight-Only Quantization quantizes only the weight tensor W of each linear layer.	×	0.09
Weight-Only Quantization accelerates memory-bounded General Matrix-Vector Multiply (GEMV) operations during the decoding	×	0.07
Weight-Activation Quantization quantizes both the input activation X and the weight tensor W of each linear layer.	×	0.09
Most LLM quantization techniques use Post-Training Quantization (PTQ) because it enables efficient inference without req	×	0.12

References

- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2402.12065v2
- http://arxiv.org/abs/2504.03360v1