Quantization Impact on HumanEval Pass@1 Scores in Code Generation Models

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does 4-bit versus 8-bit quantization impact the HumanEval pass@1 scores of code generation models when evaluated on different programming languages. Democratization of AI is an important topic within the broader topic of the digital divide. This issue is relevant to LLMs, which are becoming popular as AI co-pilots but suffer from a lack of accessibility due to high computational demand. 11 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Quantized Large Language Models for Code Generation on Low-Resource Language Benchmarks. Research question: How does 4-bit versus 8-bit quantization impact the HumanEval pass@1 scores of code generation models when evaluated on different programming languages?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

1 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.0/10.

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 study evaluates the performance of five quantized code LLMs in Lua code generation tasks.	✓	0.27
Models with 7B parameters were tested on a consumer laptop at 2-, 4-, and 8-bit integer precisions.	✓	0.26
The quantized 7B models were compared to non-quantized code LLMs with 1.3, 2, and 3 billion parameters.	✓	0.27
Lua was chosen as the benchmark language to avoid model biases related to high-resource languages.	✓	0.22
Models quantized at 4-bit integer precision offer the best trade-off between performance and model size.	✓	0.29
4-bit quantized models can be deployed on an average laptop without a dedicated GPU.	✓	0.19
Model performance significantly drops at 2-bit integer precision.	✓	0.22
Models at 8-bit integer precision require more inference time than lower precision models.	✓	0.22
The increased inference time of 8-bit models does not effectively translate to better performance compared to 4-bit mode	✓	0.18
4-bit models with 7 billion parameters outperform non-quantized models with 1.3, 2, and 3 billion parameters.	✓	0.23
4-bit 7B models have comparable storage and memory demands to non-quantized models with 1.3, 2, and 3 billion parameters	✓	0.18

References

- https://doi.org/10.48550/arxiv.2410.14766