What is the impact of quantization techniques (e.g., 8-bit vs. 4-bit) on the speed/accuracy trade-off of OpenP

Assignee Research

June 10, 2026

Abstract

Quantization methods reduce the number of bits required to represent each parameter in a model, trading accuracy for smaller memory footprints and inference latencies. However, the final model size depends on both the number of parameters of the original model and the rate of compression. For example, a 30B 8-bit model and a 60B 4-bit model have the same number of bits but may have very different zero-shot accuracies. In this work, we study this trade-off by developing inference scaling laws of zero-shot performance in Large Language Models (LLMs) to determine the bit-precision and model size

1 Introduction

This paper examines: The case for 4-bit precision: k-bit Inference Scaling Laws. Research question: What is the impact of quantization techniques (e.g., 8-bit vs. 4-bit) on the speed/accuracy trade-off of OpenPangu-MLA 13B when benchmarked against MMSU and other multimodal evaluation suites?.

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 6.9/10.

3 Results

13 papers retrieved. 10 claims extracted; 6 independently verified. Quality review score: 6.9/10.

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
4-bit quantization is universally optimal across all cases tested for model scales ranging from 19M to 176B parameters.	✓	0.15
None of the tested quantization methods improve scaling behavior for 6 to 8-bit precision.	✓	0.21
For 4-bit precision, using floating point data types enhances bit-level scaling trends more effectively than integer or	×	0.09
For most 4-bit models, a quantization block size between 64 and 128 is optimal.	×	0.10
Outlier-dependent quantization (proxy quantization) stabilizes 3-bit OPT and Pythia models but does not improve bit-leve	✓	0.15
The study conducted more than 35,000 zero-shot quantization experiments.	✓	0.18
Experiments covered LLM families including BLOOM, OPT, NeoX/Pythia, and GPT-2.	✓	0.16
On modern hardware like GPUs, loading a number from main memory takes more than 100 times longer than performing an arit	×	0.03
Total model bits are strongly related to inference latency.	×	0.11
Scaling laws for zero-shot performance were analyzed for bit precisions ranging from 3 to 8 bits.	✓	0.23

References

- http://arxiv.org/abs/2009.06488v2
- http://arxiv.org/abs/2105.03536v1
- http://arxiv.org/abs/2212.09720v2