# W4A4 vs. INT8 Quantization Impact on Llama-2-7B HumanEval Performance and H100 Latency

Assignee Research

June 1, 2026

**Abstract**

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does W4A4 quantization affect the HumanEval pass@1 score of Llama-2-7B compared to INT8 quantization while maintaining real-time inference latency on NVIDIA H100 GPUs. Reducing the latency and model size has always been a significant research problem for live Automatic Speech Recognition (ASR) application scenarios. Along this direction, model quantization has become an increasingly popular approach to compress neural networks and reduce. 11 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: 4-bit Conformer with Native Quantization Aware Training for Speech Recognition. Research question: How does W4A4 quantization affect the HumanEval pass@1 score of Llama-2-7B compared to INT8 quantization while maintaining real-time inference latency on NVIDIA H100 GPUs?.

## 2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 11 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.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| The most lightweight configuration without WER loss of Large model is to have int4 weights and int8 activations. | × | 0.04 |
| For Small model, the most lightweight configuration without WER loss is to use int8 for both weights and activations. | × | 0.05 |
| The I4WI8A Conformer (S) model has significantly outperformed all the baselines while having a more aggressive quantizat | × | 0.07 |
| 'Fake' QAT with 4-bit weights has the same accuracy as the float model on test-clean and little accuracy reduction on te | × | 0.06 |
| Training with native QAT is 7% slower than float model on TPU. | × | 0.05 |
| Training with Fake4W is 6% slower than native QAT. | × | 0.05 |
| Int8 models (E0 and E1) have similar results compared to float32 models (B0). | × | 0.07 |
| Int4 models with PTQ (E2) cannot produce reasonable performance. | × | 0.02 |
| Native QAT (E3) has additional 0.6/0.4 degradations compared to int8 models. | × | 0.06 |
| LibriSpeech models are highly overparameterized (LibriSpeech training set: 960 hours; large-scale training set: 400k ho | × | 0.06 |
| The production model is trained with an extremely large amount of data, so the model is not overfitting anymore. | × | 0.04 |

## References


- http://arxiv.org/abs/2505.14302v1
- http://arxiv.org/abs/2203.15952v4
- http://arxiv.org/abs/2601.04719v1