Quantized vs. Full-Precision Waveform Models: GPU Latency Gains on Clinical Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the relative gain in inference speed when using quantized versions of waveform-integrated models compared to full-precision models, as measured by GPU latency on standardized clinical. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: wa-hls4ml: A Benchmark and Surrogate Models for hls4ml Resource and Latency Estimation. Research question: What is the relative gain in inference speed when using quantized versions of waveform-integrated models compared to full-precision models, as measured by GPU latency on standardized clinical benchmark datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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.

References

- http://arxiv.org/abs/2504.17803v1
- http://arxiv.org/abs/2511.05615v1
- http://arxiv.org/abs/2403.00868v3