Quantization Trade-offs in Fine-Tuned Secure Language Models on Resource-Constrained Hardware

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of quantization on the throughput-accuracy trade-off for fine-tuned SecLM models deployed on resource-constrained hardware. As the rapid scaling of large language models (LLMs) poses significant challenges for deployment on resource-constrained devices, there is growing interest in extremely low-bit quantization, such as 2-bit. Although prior works have shown that 2-bit large models are. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Unifying Block-wise PTQ and Distillation-based QAT for Progressive Quantization toward 2-bit Instruction-Tuned LLMs. Research question: What is the impact of quantization on the throughput-accuracy trade-off for fine-tuned SecLM models deployed on resource-constrained hardware?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.2/10.

3 Results

12 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 6.2/10.

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
NTP-QAT with SEQ (ParetoQ) substantially outperforms existing QAT techniques such as BitDistiller and EfficientQAT at 2-	×	0.06
UPQ is evaluated on Llama 3.2 1B Instruct, Llama 3.2 3B Instruct, and Llama 3.1 8B Instruct.	×	0.04
Llama 3.2 1B Instruct is trained with a schedule of 30B tokens.	×	0.05
Llama 3.2 3B Instruct and Llama 3.1 8B Instruct are trained with a shorter schedule of 5B tokens due to resource constra	×	0.02
The pre-training dataset used is DCLM-Edu, filtered from DCLM by applying an educational quality classifier.	×	0.03
All training texts were packed with a context length of 1024 tokens.	×	0.03
UPQ achieves state-of-the-art performance on INT2 quantization of instruction-tuned LLMs.	✓	0.21
FlexRound slightly outperforms OmniQuant on most benchmarks across PTQ, NTP-QAT, and Distill-QAT.	×	0.08
UPQ progressively quantizes FP16 instruction-tuned LLMs in two stages: first to INT4 using block-wise PTQ, then to INT2	✓	0.32
Block-wise PTQ is effective in preserving the intrinsic capabilities of instruction-tuned LLMs at performance levels com	✓	0.19
The initial INT4 quantization step significantly reduces quantization error in the subsequent INT2 quantization, leading	×	0.11
The IFEval score remains low after INT4 quantization, indicating that instruction-following ability has yet to be recove	×	0.06

References

- http://arxiv.org/abs/2508.03332v2
- http://arxiv.org/abs/2506.09104v1
- http://arxiv.org/abs/2306.01076v2