Mixed-Precision Training Effects on Reasoning Accuracy in Sub-10B Parameter Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the effect of mixed-precision training on the reasoning accuracy of sub-10B models fine-tuned on SIMCOPILOTJ, measured by pass@1 and inference throughput. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PrefixQuant: Eliminating Outliers by Prefixed Tokens for Large Language Models Quantization. Research question: What is the effect of mixed-precision training on the reasoning accuracy of sub-10B models fine-tuned on SIMCOPILOTJ, measured by pass@1 and inference throughput?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

16 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
PrefixQuant achieves a 1.12 WikiText perplexity improvement on Llama-3-8B compared to existing methods under the same dy	×	0.14
PrefixQuant-O1 surpasses QoQ by 0.31 perplexity and +1.22 points accuracy on Llama-3-8B.	×	0.04
PrefixQuant-O2 maintains performance benefits with a 0.28 perplexity improvement and +1.11 points accuracy on Llama-3-8B	×	0.05
PrefixQuant-O1 and PrefixQuant-O2 outperform SpinQuant by +2.24 and +0.95 accuracy, respectively, in W4A8KV4 quantizatio	×	0.12
PrefixQuant-O1 and PrefixQuant-O2 outperform SpinQuant by +4.07 and +2.72 accuracy, respectively, in W4A4KV4 quantizatio	×	0.14
PrefixQuant-O1 achieves a perplexity of 5.67 and accuracy of 68.04 on Llama-3-8B.	×	0.04
PrefixQuant-O2 achieves a perplexity of 5.68 and accuracy of 68.09 on Llama-3-8B.	×	0.04
PrefixQuant-O1 achieves a perplexity of 5.05 and accuracy of 71.25 on Llama-3-8B.	×	0.05
PrefixQuant-O2 achieves a perplexity of 5.07 and accuracy of 71.25 on Llama-3-8B.	×	0.05
PrefixQuant-O1 achieves a perplexity of 5.93 and accuracy of 66.74 on Llama-3-8B.	×	0.04
PrefixQuant-O2 achieves a perplexity of 6.01 and accuracy of 66.37 on Llama-3-8B.	×	0.04

References

- http://arxiv.org/abs/2406.06649v1
- http://arxiv.org/abs/2602.06370v1

- http://arxiv.org/abs/2410.05265v2