LoRA Adapter Layer Scaling and Performance Trade-offs in Low-Resource African Languages

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does varying the number of LoRA adapter layers affect the performance trade-off between accuracy and inference latency in low-resource African language tasks within XTREME-R. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Analyzing Quality-Latency-Resource Trade-offs in a Technical Documentation RAG Assistant Using LoRA Adaptation. Research question: How does varying the number of LoRA adapter layers affect the performance trade-off between accuracy and inference latency in low-resource African language tasks within XTREME-R?.

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 2.7/10.

3 Results

13 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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 primary quality metric used is token-level F1 between the generated and the gold answer.	×	0.08
Headline results and per-regime tables and plots in Section 7 and Appendix I are computed on a held-out test split of n	×	0.02
The eval split consists of n = 745 samples and is used for configuration selection during the experimental loop.	×	0.01
A non-parametric bootstrap 95% confidence interval is reported based on 1,000 resamples on the test split.	×	0.08
Paired comparisons of ΔF1 between configurations use the paired bootstrap on the test split.	×	0.03
Embedding-based semantic metrics such as BERTScore are not reported in this study.	×	0.04
Two judge-based quality axes, correctness and groundedness, are computed by an external LLM judge identified as gpt-5.4-	×	0.04
The LLM judge receives only the triple (question, context, answer) and is blind to the generator identity, model_id, or	×	0.04
Correctness is defined as the content-level accuracy of the answer with respect to the provided context.	×	0.04
Groundedness is defined as the degree to which the answer is supported by the context with no unsupported additions.	×	0.02
The study aggregates judge ratings into correctness_pass@4 and groundedness_pass@4 scores, representing the fraction of	×	0.03
Four cost quantities are measured: mean inference latency (Linf), peak inference VRAM (Minf), total training time (Ttrai	×	0.04
The four cost metrics form the cost vector used in the Pareto analysis.	×	0.05
A configuration is defined as Pareto-optimal if no other configuration is at least as good on quality and every cost dim	×	0.02
In the two-dimensional Pareto fronts presented in Section 7, the quality coordinate is F1.	×	0.03
In the two-dimensional Pareto fronts, the cost coordinate alternates between mean inference latency, peak inference VRAM	×	0.06

References

- http://arxiv.org/abs/2602.05988v1
- http://arxiv.org/abs/2605.28222v1
- http://arxiv.org/abs/2106.09685v2