Early-Layer LoRA Fine-Tuning Preserves Zero-Shot Reasoning in Multilingual LLMs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does early-layer LoRA fine-tuning preserve zero-shot reasoning capabilities in multilingual LLMs better than full fine-tuning when evaluated on cross-lingual natural language inference benchmarks. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Targeted Lexical Injection: Unlocking Latent Cross-Lingual Alignment in Lugha-Llama via Early-Layer LoRA Fine-Tuning. Research question: Does early-layer LoRA fine-tuning preserve zero-shot reasoning capabilities in multilingual LLMs better than full fine-tuning when evaluated on cross-lingual natural language inference benchmarks?.

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 3.5/10.

3 Results

13 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 3.5/10.

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
Layer 0 (input embeddings) showed a modest average cosine similarity of approximately 0.3153.	×	0.08
Layer 1 showed an average cosine similarity of 0.9808.	×	0.10
Layer 2 exhibited the peak average cosine similarity, reaching 0.99998.	×	0.08
Layer 31 showed an average similarity of 0.9876 in the pilot scan.	×	0.04
The baseline output similarity observed on the full evaluation set was approximately 0.32.	×	0.09
The base Lugha-Llama-8B-wura model showed an average similarity of approximately 0.3211 for the trained set at the final	×	0.13
The base Lugha-Llama-8B-wura model showed an average similarity of approximately 0.3143 for the control set at the final	×	0.13
The model uses Lugha-Llama-8B-wura as the base model.	×	0.08
Lugha-Llama-8B-wura is an open-source LLM specifically adapted for several African languages, including Swahili, built u	×	0.10
The model is loaded in 4-bit precision using bitsandbytes with NF4 quantization and torch.bfloat16 as the compute data t	×	0.01
The pilot study revealed that Lugha-Llama-8B-wura inherently achieves very high lexical alignment in its early layers, p	✓	0.19

References

- http://arxiv.org/abs/2506.15415v1

- http://arxiv.org/abs/2310.09917v3
- http://arxiv.org/abs/2110.06500v2