

Early-Layer LoRA Fine-Tuning with Targeted Lexical Injection for Robustness in Low-Resource Lugha-Llama

Assignee Research

June 14, 2026

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities, yet their performance in low-resource languages (LRLs), such as Swahili, often lags due to data scarcity and underrepresentation in pre-training. A key challenge is achieving robust cross-lingual lexical alignment, crucial for tasks like translation and cross-lingual information retrieval. This paper introduces Targeted Lexical Injection (TLI), a novel and efficient fine-tuning approach. We first demonstrate that Lugha-Llama-8B-wura, a Swahili-centric LLM, exhibits strong, near-perfect lexical alignment for Swahili-English

1 Introduction

This paper examines: Targeted Lexical Injection: Unlocking Latent Cross-Lingual Alignment in Lugha-Llama via Early-Layer LoRA Fine-Tuning. Research question: How does early-layer LoRA fine-tuning with Targeted Lexical Injection affect the robustness of Lugha-Llama against adversarial perturbations in low-resource language settings?.

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 7.5/10.

3 Results

12 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 7.5/10.

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.	✓	0.26
Layer 1 showed an average cosine similarity of 0.9808.	×	0.15
Layer 2 exhibited the peak average cosine similarity, reaching 0.99998.	✓	0.22
Layer 31 showed an average similarity of 0.9876 in the pilot scan.	✓	0.19
The baseline output similarity observed on the full evaluation set was approximately 0.32.	×	0.13
The base Lugha-Llama-8B-wura model showed an average similarity of approximately 0.3211 for the trained set at the final	✓	0.24
The base Lugha-Llama-8B-wura model showed an average similarity of approximately 0.3143 for the control set at the final	✓	0.23
The model uses Lugha-Llama-8B-wura as the base model.	×	0.13
Lugha-Llama-8B-wura is an open-source LLM specifically adapted for several African languages, including Swahili, built u	✓	0.28
The model is loaded in 4-bit precision using bitsandbytes with NF4 quantization and torch.bfloat16 as the compute data t	✓	0.25
A pilot study was conducted to determine which layer of the Lugha-Llama-8B-wura model exhibits the highest degree of inh	✓	0.33

References

- <http://arxiv.org/abs/2504.15610v3>

- <http://arxiv.org/abs/2506.15415v1>
- <http://arxiv.org/abs/2602.05988v1>