Targeted Lexical Injection versus Full Fine-Tuning for Confidence Calibration in Lugha-Llama on Adversarial Swahili-English Text

Assignee Research

June 12, 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: What is the effect of Targeted Lexical Injection on the calibration of confidence scores for Lugha-Llama when processing adversarial Swahili-English mixed text compared to standard full fine-tuning?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.0/10.

3 Results

15 papers retrieved. 18 claims extracted; 13 independently verified. Quality review score: 7.0/10.

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) of the Lugha-Llama-8B-wura model showed an average cosine similarity of approximately 0.3153	✓	0.25
Layer 1 of the Lugha-Llama-8B-wura model showed an average cosine similarity of 0.9808 in the pilot study.	✓	0.22
Layer 2 exhibited the peak average cosine similarity of 0.99998 in the pilot study.	✓	0.20
Layer 31 showed an average cosine similarity of 0.9876 in the pilot scan.	✓	0.19
The baseline output similarity on the full evaluation set prior to TLI fine-tuning was approximately 0.32.	×	0.14
The average cosine similarity at Layer 31 for the trained set prior to TLI fine-tuning was approximately 0.3211.	✓	0.21
The average cosine similarity at Layer 31 for the control set prior to TLI fine-tuning was approximately 0.3143.	✓	0.19
The control set used for evaluation consisted of 63 unseen word pairs.	×	0.14
A paired t-test was conducted to determine the statistical significance of changes in mean cosine similarity before and	✓	0.25
The base model used is Lugha-Llama-8B-wura (Lugha Factory, 2023).	✓	0.20
Lugha-Llama is built upon the Llama-3 architecture.	×	0.13
The model was loaded in 4-bit precision using bitsandbytes with NF4 quantization.	✓	0.19
The compute data type used was torch.bfloat16.	×	0.12
The pilot study extracted embeddings from every transformer layer, ranging from Layer 0 to Layer 31.	×	0.10
Layer 0 represents the initial input embeddings in the Lugha-Llama model.	✓	0.20
For the final evaluation, word embeddings were extracted from Layer 31 (the final output layer).	✓	0.15
Embeddings used for evaluation were mean-pooled over attention-masked tokens and L2-normalized.	✓	0.20
Cosine similarity between L2-normalized Swahili and English word embeddings was the primary metric for lexical alignment	✓	0.26

References

- http://arxiv.org/abs/2604.09529v1
- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2506.15415v1