Targeted Lexical Injection for Adversarial Code-Switching Robustness in Low-Resource African Languages

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent does Targeted Lexical Injection improve robustness against adversarial code-switching inputs in low-resource African languages compared to adapter-based tuning on multilingual NLI tasks. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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: To what extent does Targeted Lexical Injection improve robustness against adversarial code-switching inputs in low-resource African languages compared to adapter-based tuning on multilingual NLI tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

14 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 5.2/10.

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
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.	×	0.09
Layer 31 showed an average similarity of 0.9876 in the pilot scan.	×	0.05
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	×	0.14
The base Lugha-Llama-8B-wura model showed an average similarity of approximately 0.3143 for the control set at the final	×	0.14
The model uses Lugha-Llama-8B-wura as the base model, which is an open-source LLM adapted for several African languages,	×	0.09
The model is loaded in 4-bit precision using bitsandbytes with NF4 quantization and torch.bfloat16 as the compute data t	×	0.03
The pilot study identified Layer 2 as exhibiting the highest degree of inherent cross-lingual lexical alignment for Swah	✓	0.27

References

- http://arxiv.org/abs/2605.17152v1
- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2303.03750v2