Targeted Lexical Injection and Cross-Lingual Reasoning Performance Trade-offs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does the latent cross-lingual alignment achieved via Targeted Lexical Injection degrade performance on reasoning-heavy multilingual tasks compared to translation-focused evaluations. 9 claims were extracted from source literature; 0 were 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 the latent cross-lingual alignment achieved via Targeted Lexical Injection degrade performance on reasoning-heavy multilingual tasks compared to translation-focused evaluations?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 9 claims extracted; 0 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 exhibited 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.	×	0.04
The baseline output similarity observed on the full evaluation set was approximately 0.32.	×	0.09
The average similarity at the final output layer (Layer 31) of the base model was approximately 0.3211 for the trained s	×	0.10
The model used is Lugha-Llama-8B-wura, an open-source LLM adapted for several African languages, including Swahili, buil	×	0.11
The model is loaded in 4-bit precision using bitsandbytes with NF4 quantization and torch.bfloat16 as the compute data t	×	0.02
The pilot study involved extracting embeddings from the output of every transformer layer in Lugha-Llama (Layers 0 throu	×	0.10

References

- http://arxiv.org/abs/2205.00267v2
- http://arxiv.org/abs/2603.26742v1
- http://arxiv.org/abs/2506.15415v1