Multimodal Model Robustness to Cross-Modal Distributional Shifts via Llama-3 Alignment Strategies

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Do multimodal models trained with alignment strategies similar to Llama-3 exhibit greater robustness to cross-modal distributional shifts compared to Vicuna-style SFT models, as evaluated by. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. 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: Do multimodal models trained with alignment strategies similar to Llama-3 exhibit greater robustness to cross-modal distributional shifts compared to Vicuna-style SFT models, as evaluated by performance drops on benchmarks like VQA-Shift or Multimodal StressTest?.

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 4.4/10.

3 Results

15 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.4/10.

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	×	0.14
Layer 1 of the Lugha-Llama-8B-wura model showed an average cosine similarity of 0.9808 in the pilot study.	✓	0.15
Layer 2 exhibited the peak average cosine similarity of 0.99998 in the pilot study.	×	0.10
Layer 31 showed an average cosine similarity of 0.9876 in the pilot scan.	×	0.07
The baseline output similarity on the full evaluation set prior to TLI fine-tuning was approximately 0.32.	×	0.15
The average cosine similarity at Layer 31 for the trained set prior to TLI fine-tuning was approximately 0.3211.	×	0.13
The average cosine similarity at Layer 31 for the control set prior to TLI fine-tuning was approximately 0.3143.	×	0.13
The control set used for evaluation consists of 63 unseen word pairs.	×	0.13
A paired t-test was conducted to determine the statistical significance of changes in mean cosine similarity before and	×	0.04
The base model used is Lugha-Llama-8B-wura (Lugha Factory, 2023).	×	0.08
Lugha-Llama is built upon the Llama-3 architecture.	×	0.06
The model was loaded in 4-bit precision using bitsandbytes with NF4 quantization.	×	0.02
The compute data type used was torch.bfloat16.	×	0.02
The pilot study extracted embeddings from every transformer layer, ranging from Layer 0 to Layer 31.	×	0.05
Layer 0 represents the initial input embeddings in the Lugha-Llama model.	×	0.06
For the final evaluation, word embeddings were extracted from Layer 31 (the final output layer).	×	0.09
Embeddings used for evaluation were mean-pooled over attention-masked tokens and L2-normalized.	×	0.02
Cosine similarity between L2-normalized Swahili and English word embeddings was the primary metric for lexical alignment	✓	0.19

References

- http://arxiv.org/abs/2410.14148v4
- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2509.09055v1