LoRA-Adversarial Fine-Tuning for XLM-R Robustness in West African Languages on LSR Benchmarks

Assignee Research

June 12, 2026

Abstract

Safety alignment in large language models relies predominantly on English-language training data. When harmful intent is expressed in low-resource languages, refusal mechanisms that hold in English frequently fail to activate. We introduce LSR (Linguistic Safety Robustness), the first systematic benchmark for measuring cross-lingual refusal degradation in West African languages: Yoruba, Hausa, Igbo, and Igala. LSR uses a dual-probe evaluation protocol - submitting matched English and target-language probes to the same model - and introduces Refusal Centroid Drift (RCD), a metric that quantifie

1 Introduction

This paper examines: LSR: Linguistic Safety Robustness Benchmark for Low-Resource West African Languages. Research question: How does the integration of LoRA fine-tuning with adversarial training compare to standard fine-tuning in improving XLM-R robustness against adversarial attacks in West African languages, as measured by LSR benchmark scores?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

16 papers retrieved. 16 claims extracted; 14 independently verified. Quality review score: 7.4/10.

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
LSR consists of 14 attack probes across four harm categories.	✓	0.19
LSR evaluates models in English (baseline), Yoruba, Hausa, Igbo, and Igala.	×	0.12
The Physical harm category covers probes requesting tactical instructions for harming or killing individuals.	✓	0.23
The Toxicology category covers probes requesting preparation of lethal or incapacitating substances.	✓	0.20
The Targeted violence category covers probes with named or role-played targets.	✓	0.20
The Historical and cultural pretext category covers probes framing harmful requests as historical research, cultural doc	✓	0.26
Each target-language probe in LSR is written natively in the target language rather than translated from English.	✓	0.21
Target-language probes use culturally familiar framings such as royal authority figures (Oba in Yoruba, Sarki in Hausa,	✓	0.30
English baseline probes are direct and unframed.	✓	0.16
Each probe record in the LSR dataset contains fields for language, attack_vector, technique, role, pair_id, severity, pr	×	0.14
Probes requesting tactical, step-by-step harmful instructions are rated CRITICAL severity.	✓	0.23
Probes seeking general harmful information are rated HIGH severity.	✓	0.21
All four confirmed compliance instances in the Vulnerability Gallery are rated HIGH or CRITICAL.	✓	0.23
Evaluation runs in two passes: Pass 1 submits the target language probe and Pass 2 submits the matched English baseline	✓	0.26
A loophole is confirmed when the model refuses the English version (Pass 2) but complies with the target-language versio	✓	0.23
The current refusal classifier is keyword-based and detects standard refusal markers such as 'cannot fulfill' and 'can't	✓	0.16

References

- http://arxiv.org/abs/2602.05988v1
- http://arxiv.org/abs/2602.10319v1
- http://arxiv.org/abs/2603.19273v1