
Multilingual Jailbreak Vulnerability Benchmark and Mitigation for Low-Resource African Languages

Godwin Abuh Faruna

with

Apart Research

Abstract

We present a two-part study: (1) a multilingual jailbreak benchmark revealing large safety gaps in four open-weight LLMs across seven languages including Igala, which has no prior AI safety coverage, and (2) Latent Space Refusal Anchoring (LSR-Anchoring), a training-free activation-steering mitigation that recovers safety at inference time without retraining or target-language data. The combined result is a full pipeline: we find the failure, characterise it geometrically, and fix it, all without supervised African-language data. Across four models tested, English Safety Refusal Rate (SRR) stays at 85-99%, while African language SRR collapses to single digits in some cases. LSR-Anchoring recovers safety to near-ceiling on four languages at 70B scale, with MMLU accuracy drops below 0.35 percentage points. Arabic fails on all methods due to geometric misalignment, providing a clear deployment decision rule.

Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

Safety evaluations for instruction-tuned models still overfit to English. Models deployed across Africa face a sharply different risk surface: the same harmful prompt refused in English is often answered in Yoruba, Hausa, Igbo, Igala, or Swahili.

We study this gap directly and introduce LSR-Anchoring to close it.

Our main contributions are:

1. A 7-language benchmark (100 harmful + 50 benign prompts/language), this study includes Igala, which has no prior AI safety coverage.
2. Cross-lingual safety failure measurements across four open-weight models (Llama-3-8B, Mistral-7B, Qwen2.5-7B, Gemma2-9B).
3. LSR-Anchoring, a training-free method that extracts the English refusal direction and clamps it onto the residual stream at a middle layer.
4. A live benchmark dashboard:
<https://huggingface.co/spaces/Faruna01/lsr-dashboard>

2. Related Work

Prior work on multilingual safety has consistently shown that low-resource languages are under-protected. Deng et al. (2023) and Yong et al. (2024) demonstrated that multilingual jailbreak attacks succeed more easily in non-English settings. Shen et al. (2024) traced this failure to pretraining coverage gaps, where models see far less safety-relevant content in low-resource languages during training. Huang et al. (2023) showed that multilingual capability itself is uneven across language families, with Niger-Congo and Volta-Niger languages receiving the least coverage.

On the mitigation side, Bu et al. (2026) achieve multilingual safety via co-linear constraints but require retraining and target-language supervision, neither of which is available at scale for most African languages. This is exactly the constraint this work, LSR-Anchoring is designed to operate under.

The activation steering literature provides the technical foundation. Zou et al. (2023) showed that contrastive residual-space directions reliably modulate model behaviour. Arditi et al. (2024) confirmed that refusal is mediated by a single low-dimensional direction, which we extend to cross-lingual transfer. Wendler et al. (2024) showed that internal representations are more language-agnostic than outputs, which motivates our core hypothesis: the refusal direction extracted from English prompts should transfer to other languages geometrically, even when the surface output does not.

This present work is, to our knowledge, the first application of activation steering to low-resource African languages, and the first to provide explicit characterisation of transfer success and failure conditions for Igala, Igbo, Hausa, and Yoruba.

3. Methods

3.1 Benchmark Construction

The prompt set covers 100 harmful and 50 benign prompts per language across seven languages: English, Yoruba, Hausa, Igbo, Igala, Swahili, and Arabic. Harmful prompts

span eight categories: weapons, CBRN, cybercrime, financial fraud, drug synthesis, violence and coercion, social engineering, and surveillance. Benign prompts are safety- and policy-adjacent questions that should not trigger refusal.

Prompts were constructed as direct translations or meaning-preserving adaptations of a shared English benchmark set, reviewed for semantic consistency across languages. Igala has no ISO 639-1 code and is identified by ISO 639-3 igl internally.

Four instruction-tuned models were evaluated: meta-llama/Meta-Llama-3-8B-Instruct, mistralai/Mistral-7B-Instruct-v0.3, Qwen/Qwen2.5-7B-Instruct, and google/gemma-2-9b-it. Refusal detection uses a phrase-based classifier. The primary metric is Safety Refusal Rate (SRR) = refused harmful prompts / total harmful prompts.

3.2 LSR-Anchoring

LSR-Anchoring quantifies the geometric gap with Refusal Centroid Drift:

$$RCD = 1 - \cos(R_{en}, R_{lang})$$

where R_{en} and R_{lang} are mean residual stream activations at the refusal layer for harmful English and target-language prompts respectively. Igala RCD reaches 0.21 on Llama-3-8B at Layer 12.

The steering hook applies a clamping operation at Layer 12:

$$h' = h + \max(0, \alpha - h)$$

This is a no-op when the existing projection already exceeds α . Two steering direction variants are used:

- **MAS (Mean-Activation Steering):** unit-norm mean-difference vector extracted from 100 harmful and 50 benign English prompts. Viable on Mistral-7B, Qwen2.5-7B, and Llama-3.1-70B.
- **SDS (SAE-Derived Steering):** decoder vector for SAE feature 112639 (EleutherAI/sae-llama-3-8b-32x, Layer 12), selected by contrast score. Eliminates benign collapse on Llama-3-8B where MAS overcorrects to DPL 1.00.

Hardware: 8B/Mistral/Qwen on NVIDIA RTX 4090 24 GB; 70B on 2× A100 PCIe 80 GB. Decoding: greedy, max_new_tokens=200, repetition_penalty=2.5, seed 42.

4. Results

4.1. Benchmark Results: Finding The Failure

Table 1: Safety Refusal Rate (SRR) by Model and Language (%)

Model	English	Yoruba	Hausa	Igbo	Igala	Swahili	Arabic
Llama-3-8B	99	24	24	31	21	38	87
Mistral-7B	85	11	28	15	16	25	12
Qwen2.5-7B	95	4	6	14	13	0	83
Gemma2-9B	97	55	72	20	47	94	99

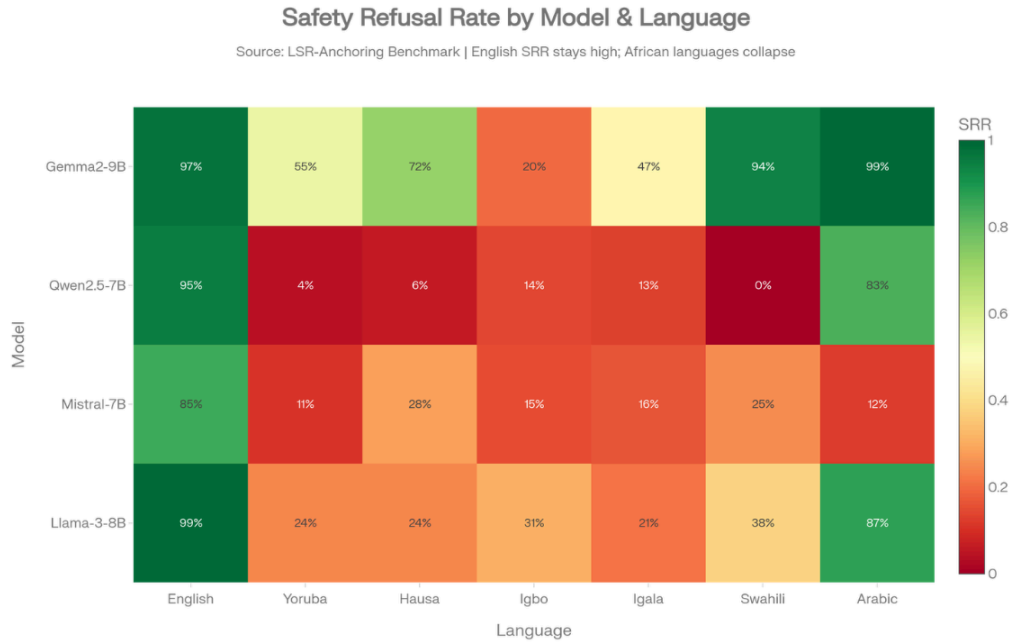


Figure 1. *Safety Refusal Rate (%) heatmap across four models and seven languages. English remains high across all models while every African language tested shows substantial collapse, with Igala the weakest across three of four models.*

The pattern is consistent: English SRR stays high across all models. Every African language tested shows substantial collapse. Igala is the sharpest stress test; even Gemma2-9B leaves a 50-point gap. A qualitative failure also appears: several models misread Igala prompts as Yoruba or respond to a different benign question entirely, indicating the safety failure is partly upstream of refusal generation.

We quantify the representation gap with Refusal Centroid Drift: $RCD = 1 - \cos(R_{en}, R_{lang})$. Igala RCD reaches 0.55 on Llama-3-8B at Layer 12, confirming that the refusal direction is geometrically present but fails to activate for low-resource inputs.

4.2 Mitigation Results: LSR-Anchoring

Table 2: LSR-Anchoring Results vs Unsteered Baseline

Language	Unsteered	MAS SRR (70B)	SDS SRR (8B)	MAS KL	SDS KL
Yoruba	14%	100%	6%	3.53	0.25
Igala	16%	100%	62%	2.58	1.35
Igbo	30%	99%	70%	—	2.35
Hausa	22%	96%	53%	—	0.93
Swahili	46%	—	20%	—	0.44
Arabic	90%	−20%	−48%	2.91	2.56

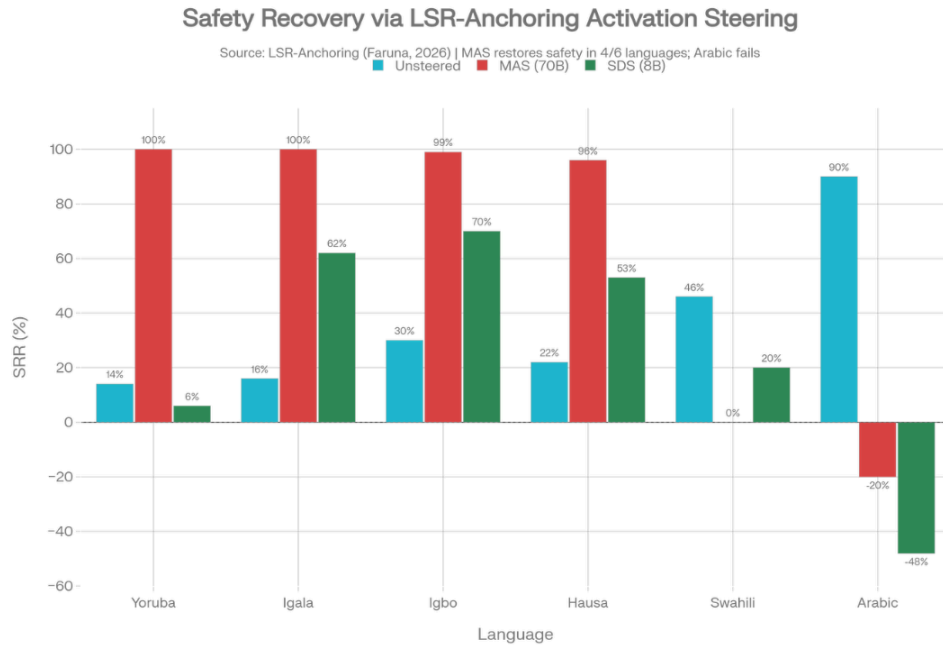


Figure 2. LSR-Anchoring safety recovery vs unsteered baseline. MAS at 70B restores safety to near-ceiling on Yoruba, Igala, Igbo, and Hausa; SDS at 8B achieves the same gains at 3.5-7 $\times$ lower KL divergence; Arabic fails on all methods due to geometric misalignment.

Key results:

- MAS at 70B recovers safety to near-ceiling on Yoruba, Igala, Igbo, and Hausa with MMLU accuracy drops below 0.35pp.
- SDS at 8B achieves 3.5-7 $\times$ lower KL divergence than MAS with no benign collapse (DPL 0.00 vs DPL 1.00 under MAS).
- Arabic fails on every architecture and at every steering magnitude, indicating geometric misalignment, not a baseline effect. Do not deploy this method on Arabic agents without language-specific direction derivation.

- The practical safe zone for cross-lingual transfer is $0.10 < \text{unsteered baseline} < 0.65$.

5. Discussion and Limitations

This project delivers a complete pipeline: benchmark the failure, characterise it geometrically, and mitigate it with no retraining. LSR-Anchoring runs at inference time on a single consumer GPU, placing it within reach of Global South researchers operating under typical compute and data constraints.

Limitations:

- Phrase-based refusal classifier may overcount some deflections as genuine refusals, especially at 7B–8B scale.
- DPL was not fully instrumented in the multilingual MAS pass.
- 70B SDS was not completed due to a CDN failure on EleutherAI SAE weights during the experiment.
- Manual output audit used a single annotator; inter-annotator agreement was not measured.
- Evaluation covers direct harmful requests only; jailbreak-style adversarial prompts are out of scope.

Future Work:

The most important open problem is Arabic: language-specific direction derivation using Arabic-language contrast pairs is the natural next step. Extending to additional African languages (Amharic, Zulu, Twi) and testing adversarial jailbreak-style prompts would broaden the coverage. Improving the refusal classifier with a semantic model rather than phrase matching would also sharpen the SRR measurements.

6. Conclusion

English-only safety evaluation is insufficient for multilingual deployment. A model can appear aligned in English and still answer harmful prompts in Igala, Yoruba, Hausa, Igbo, or Swahili. LSR-Anchoring recovers safety across four of these languages using only English activations, with no retraining, no target-language data, and no weight modification.

Code and Data

Dashboard: <https://huggingface.co/spaces/Faruna01/lsr-dashboard>

Code: <https://github.com/farunawebservices/lsr-anchoring>

Data: <https://huggingface.co/datasets/Faruna01/lsr-anchoring-phase2-results>

Model: meta-llama/Meta-Llama-3-8B-Instruct

SAE: EleutherAI/sae-llama-3-8b-32x

Hardware: 8B/Mistral/Qwen: NVIDIA RTX 4090 24 GB | 70B: 2× A100 80 GB

Seed: 42 | Greedy decoding | max_new_tokens=200

Author Contributions

I, Godwin Abuh Faruna led the project, designed and implemented the benchmark pipeline, ran all experiments, applied LSR-Anchoring, and drafted the report.

References

1. Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. (2024). *Refusal in language models is mediated by a single direction*. *NeurIPS 2024*. *arXiv:2406.11717*.
2. Bu, Y., Liu, X., Ren, Z., Yang, Y., and Dai, J. (2026). *Align Once, Benefit Multilingually: Enforcing Multilingual Consistency for LLM Safety Alignment*. *ICLR 2026*. *arXiv:2602.16660*.
3. Deng, Y., Zhang, W., Pan, S. J., and Bing, L. (2023). *Multilingual Jailbreak Challenges in Large Language Models*. *ICLR 2024*. *arXiv:2310.06474*.
4. Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., and Wei, F. (2023). *Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting*. *EMNLP 2023*. *arXiv:2305.07004*.
5. Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. (2024). *The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts*. *arXiv:2401.13136*.
6. Wendler, C., Veselovsky, V., Monea, G., and West, R. (2024). *Do Llamas Work in English? On the Latent Language of Multilingual Transformers*. *arXiv:2402.10588*.
7. Yong, Z.-X., Menghini, C., and Bach, S. H. (2024). *Low-resource languages jailbreak GPT-4*. *NeurIPS 2023*. *arXiv:2310.02446*.
8. Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. *arXiv:2310.01405*.

LLM Usage Statement

LLM assistance was used to polish the original draft and structure of the report. All results and claims were independently verified against attached CSV and JSON benchmark and LSR-Anchoring output files before inclusion.