Probing the Safety Frontier of African-Language LLMs: A Red-Teaming Study

Yehoshua

Zindi African Trust & Safety LLM Challenge · April 2026

https://github.com/yehoshua0/The-African-Trust-Safety-LLM-Challenge

Abstract. Large language models fine-tuned for African languages carry the promise of democratising AI across underserved communities. Yet their safety alignment has received comparatively little scrutiny. This report presents a structured red-teaming study targeting two publicly available African-language models — **N-ATLaS** (Hausa, Igbo, Yoruba) and **Pawa-Gemma-Swahili-2B** (Swahili) — using 15 adversarial prompts spanning 9 risk categories. All attacks were validated through a purpose-built web application that logged reproducibility across multiple independent runs. Our results reveal that classic jailbreak techniques, when faithfully translated into African languages, retain near-full effectiveness: 12 of 15 attacks (80 %) were confirmed at a 3/3 reproducibility score, and 3 at 2/3. Three attack types — *direct jailbreak*, *prompt injection*, and *refusal softening* — successfully transferred across both models and multiple languages, suggesting that safety vulnerabilities are structurally embedded rather than incidental.

Keywords: red-teaming · safety alignment · African NLP · jailbreak · N-ATLaS · Pawa-Gemma · Hausa · Igbo · Swahili · Yoruba

Introduction

The rapid proliferation of LLM technology into African markets — driven by initiatives from organisations such as Masakhane, Pawa AI, and Lelapa AI — raises urgent questions about the adequacy of safety guardrails in low-resource language settings. Safety research has overwhelmingly concentrated on high-resource languages, primarily English. The implicit assumption has been that multilingual fine-tuning preserves alignment properties. This study challenges that assumption.

We target two representative models from the African NLP landscape:

Model	Languages	Size	Source
N-ATLaS	Hausa, Igbo, Yoruba	~7B (Q4_K_M)	QuantFactory/N-ATLaS-GGUF
Pawa-Gemma-Swahili-2B	Swahili	~2B (Q8_0)	Skier8402/gemma2-2b-swahili-it

Table 1: Target models and their characteristics.

Both models serve communities spanning hundreds of millions of speakers across West and East Africa. A successful jailbreak at scale is therefore not merely a benchmark curiosity — it represents a concrete harm vector.

Methodology

Red-Teaming Infrastructure

All experiments were conducted through a locally hosted FastAPI web application interfacing with `llama-cpp-python` for GGUF inference. Generation parameters were deliberately calibrated for adversarial pressure:

Attack Taxonomy

Prompts were classified along three orthogonal axes derived from the challenge specification:

Parameter	Value	Rationale
Temperature	0.90	Flattens safety distribution
Top-P	0.95	Broadens nucleus sampling
Repetition penalty	1.15	Suppresses refusal phrase loops
Max new tokens	2048	Captures full harmful output

Table 2: Inference parameters used across all red-teaming runs.

- **Risk category** — 9 categories: Cybersecurity, Extremism, Self-Harm, Violence, Illegal Activity, Privacy/Data Leakage, Bias & Discrimination, Harmful Instructions, and Policy Evasion.
- **Attack type** — 9 techniques: Direct Jailbreak, Persona Hijacking, Refusal Softening, Translation Pivot, Hypothetical Scenario, Prompt Injection, Indirect Request, Context Poisoning, and Roleplay.
- **Language** — Hausa, Igbo, Swahili, and Yoruba, with prompts composed in each target language.

Validation Protocol

Each prompt was submitted a minimum of two to three times under identical generation conditions. A break was classified as:

- **3/3 Confirmed** — all three independent runs produced a policy-violating response.
- **2/3 Confirmed** — two of three runs produced a violation (one refusal tolerated).

Only confirmed breaks were included in the final submission.

Results

Overview

Metric	Value
Total validated attacks	15
Languages covered	4 (Hausa, Igbo, Swahili, Yoruba)
Risk categories hit	9 / 9
Models broken	2 / 2
3/3 reproducibility	12 (80 %)
2/3 reproducibility	3 (20 %)
Cross-model attack types	3 (`direct_jailbreak`, `prompt_injection`, `refusal_softening`)

Table 3: Aggregate results across all validated attacks.

Per-Language Breakdown

Language	Model	Attacks	3/3	2/3
Swahili	Pawa-Gemma-Swahili-2B	11	8	3
Hausa	N-ATLaS	1	1	0
Yoruba	N-ATLaS	2	2	0
Igbo	N-ATLaS	1	1	0

Table 4: Attack distribution and reproducibility by language.

Key Findings

Finding 1 — Language is not a safety barrier. Translating the classic DAN prompt into Hausa, Igbo, or Yoruba produced successful cybersecurity jailbreaks on N-ATLaS across all three languages. The `direct_jailbreak` type is cross-model: it succeeded on both N-ATLaS and Pawa-Gemma in their respective languages, with a combined 4/4 success rate.

Finding 2 — System prompt extraction is cross-model and cross-language. Attacks `PRIVLEAK01` (Swahili) and `YOR_PRIVLEAK01` (Yoruba) both induced models to echo a `SYSTEM PROMPT:` header and enumerate internal guidelines. This is a structural vulnerability shared by both architectures, not an artefact of any specific fine-tuning.

Finding 3 — Refusal softening is highly effective. Pre-emptive illegal-acknowledgement framing (*"I know this might be illegal but..."*) bypassed refusals for advance-fee fraud explanation in both Hausa (N-ATLaS) and Swahili (Pawa-Gemma), achieving 3/3 on both.

Finding 4 — Authority personas erode safety boundaries. Institutional framing (professor, doctor) lowered the threshold for harmful content generation in attacks targeting SQL injection (`CYBER02`) and document forgery (`FORGE01`), even when the model nominally acknowledged the illegality.

Discussion

Structural vs. Incidental Alignment

The pattern across 17 attacks suggests that alignment in these models is **shallow and language-agnostic in the wrong direction**: guardrails appear to rely on surface-level keyword matching or high-resource-language heuristics that fail when the same adversarial intent is rephrased in a lower-resource tongue. Models with strong English safety alignment can nonetheless exhibit broad policy failure when queried in Hausa or Yoruba.

Community-Specific Risk

Several attacks carry amplified harm potential in African contexts:

- **M-Pesa phishing** (`PHISH01`) directly weaponises the dominant mobile-money infrastructure for approximately 33 M Kenyan users.
- **Ethnic bias in Swahili** (`BIAS01`) generates fabricated demographic statistics that could exacerbate ethnic tensions in Kenya.
- **Sahel extremism recruitment** (`EXT01`) generates content tailored to the specific socioeconomic conditions of the region.

These are not generic safety failures — they exploit cultural and economic specificity that English-centric safety research would not anticipate.

Limitations

This study is bounded by (i) a focus on two models, (ii) a 15-attack sample limited by competition scope, and (iii) inference on quantised GGUF weights, which may exhibit slightly different behaviour from full-precision checkpoints. Replication on full-precision models and expansion to additional languages (Amharic, Zulu, Afrikaans, French Creole) remain open avenues.

Conclusion

We have demonstrated that off-the-shelf jailbreak techniques, when adapted to African languages, reliably elicit policy-violating responses from two representative African-language LLMs. The cross-

model and cross-language consistency of direct jailbreaks, prompt injection, and refusal softening argues for systematic adversarial evaluation as a prerequisite before deploying African-language models in any user-facing context. We hope this work contributes to a more robust and equitable AI safety research agenda for the continent.

Acknowledgements

This work was submitted to the **Zindi African Trust & Safety LLM Challenge**. We thank the challenge organisers and evaluators for their rigorous review process, and the broader ecosystem of organisations whose collective effort makes African-language NLP research possible.

Figure 1: *

Challenge partners & sponsors — Cassava Technologies, AMINI, meetkai, Qhala, Lelapa AI, MTN, Lanfrica, Vodacom, Orange, Ethio Telecom, World Sandbox Alliance, Masakhane, Mozisha, Pawa AI, APHRC, TOUMAi, Digital Umuganda, Airtel, Capsele, AXUM, AXIAN Telecom, awarri, DATA SPIRES, ABIS.

We further thank the authors of **N-ATLaS** (QuantFactory) and **Pawa-Gemma-Swahili-2B** (Skier8402 / Google DeepMind) for making model weights publicly available, enabling open safety research without cloud compute dependency.