Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Red-Teaming African-Language LLMs: Safety Alignment Failures Across Low-Resource Languages

Godwill Nkwain Nyong
Carnegie Mellon University Africa
George Okeyo
gokeyo@andrew.cmu.edu

Carnegie Mellon University Africa
Jema David Ndibwile
Carnegie Mellon University Africa

Article

Keywords: Red-teaming, Safety alignment, African languages, Large language models, Multilingual safety, Adversarial attacks

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11242446/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Red-Teaming African-Language LLMs: Safety Alignment Failures Across Low-Resource Languages

Godwill Nkwain Nyong[1], George Okeyo[1*], Jema David Ndibwile[1]

[1*]Carnegie Mellon University Africa, Kigali, Rwanda.

*Corresponding author(s). E-mail(s): gokeyo@andrew.cmu.edu;
Contributing authors: gnyong@andrew.cmu.edu;
jndibwil@andrew.cmu.edu;

Abstract

Large Language Models are now being developed specifically for African languages, with models like N-ATLaS and Pawa-Gemma-Swahili achieving meaningful capability gains on standard NLP benchmarks. Yet while capability evaluation has advanced through benchmarks like IrokoBench and AfroBench, safety evaluation for these models remains less or nonexistent. Existing multilingual safety research focuses on using low-resource languages to attack frontier models such as GPT-4, leaving African-language LLMs themselves unexamined. We present the first systematic red-team evaluation of African-language LLMs, testing two publicly available models across four languages (Hausa, Swahili, Yoruba, Igbo), eight harm categories organized into four HEVP threat domains, and ten adversarial techniques. Across 220 trials using 24 culturally grounded attack configurations referencing real African systems (M-Pesa mobile money, Nigeria's National Identification Number and Bank Verification Number (NIN/BVN), Nigeria's Independent National Electoral Commission (INEC), and Safaricom telecommunications), we find that 44.1% of attacks elicit harmful content, with all four languages vulnerable at comparable rates (40.0–48.0%). The HEVP taxonomy reveals structured vulnerability: Propaganda (70.0%) and Harm Generation (58.0%) are most exploitable, while hate/harassment shows the strongest defense (12.0%). Role Exploitation techniques achieve 82.5% success, outperforming all other technique classes. Model size does not predict safety; the 8B-parameter and 2.6B-parameter models show equivalent vulnerability despite the smaller model incorporating Direct Preference Optimization (DPO) alignment. Defense mechanisms vary qualitatively by language, from disclaimers in Hausa to truncation

in Swahili. Safety evaluation should become routine in African-language model development.

Keywords: Red-teaming, Safety alignment, African languages, Large language models, Multilingual safety, Adversarial attacks

1 Introduction

With the existence of frontier models like Gemini and Llama, large language models are being built via fine-tuning and other techniques specifically for African languages. N-ATLaS (Nigeria, 2025), released by Nigeria's National Centre for Artificial Intelligence and Robotics in collaboration with Awarri Technologies, fine-tunes Llama-3-8B-Instruct on 392 million tokens of African-language data spanning Hausa, Yoruba, Igbo, and several other languages. Pawa-Gemma-Swahili, developed by Sartify LLC, extends Gemma-2-2B with over 3 billion tokens of Swahili text, followed by supervised fine-tuning and Direct Preference Optimization (DPO). Capability evaluation has advanced in parallel: AfroBench (Ojo et al. 2023) covers 64 African languages across multiple NLP tasks, IrokoBench (Adelani et al. 2024) benchmarks language understanding, and AfriSenti (Muhammad et al. 2023) addresses sentiment analysis across 14 African languages. These models and benchmarks reflect genuine progress in African-language NLP, moving from isolated prototypes toward models with real generative capability.

Yet there is little or no safety evaluation for these models. The multilingual safety literature has grown rapidly, but it primarily addresses a different problem than the one we focus on. Yong et al. (2024) show that low-resource languages can jailbreak GPT-4; Deng et al. (2023) document multilingual jailbreak challenges across frontier models; Wang et al. (2024) introduce XSafety for cross-lingual safety evaluation, but all of this work uses African and other low-resource languages as attack vectors against frontier models like GPT-4, Claude, and Gemini. Even recent African-focused safety research follows this pattern: UbuntuGuard (Abdullahi et al. 2026) evaluates guardian models for African policy compliance, the Linguistic Safety Robustness benchmark (LSR Consortium 2026) tests West African language robustness on frontier systems, and GPT-OSS-20B (Inuwa-Dutse 2025) red-teams OpenAI's model in Hausa. The question these works answer is "Can African languages break safe models?" The question nobody has asked is: "Are African-language models themselves safe?"

This is the gap we address. Models like N-ATLaS and Pawa-Gemma are not frontier systems with extensive RLHF pipelines, they are adapted from base models with limited or no safety-specific training data in their target languages. As their capabilities grow and deployment expands, the absence of any safety evaluation creates a blind spot that could have serious consequences for the communities these models are built to serve. We investigate four research questions:

RQ1 *Language Gap:* Does safety alignment in African-language LLMs vary significantly across languages, and does the gap correlate with language resource availability?

RQ2 *Technique Transfer:* Which state-of-the-art adversarial techniques transfer effectively across African languages, and which are language-specific?

RQ3 *Category Vulnerability:* Are certain harm categories disproportionately underdefended in African-language models compared to others?

RQ4 *Defense Mechanisms:* How do models defend against adversarial attacks across languages? Do they refuse, disclaim, truncate, or decontextualize, and does the defense strategy vary by language?

The contributions are as follows:

1. **Red-teaming African-language LLMs.** We present the first systematic red-team evaluation targeting models built for African languages. Most works use African languages to evaluate frontier models; instead, we evaluate African language LLMs themselves across four typologically diverse languages.
2. **HEVP taxonomy.** We introduce a four-domain taxonomy that groups eight risk categories by how the harm reaches its target: direct physical harm, economic exploitation, social division, and information manipulation. This distinction is useful because safety defenses may transfer differently across domains. For instance, hate speech classifiers may generalize across languages while defenses against dangerous content may not. Grouping by mechanism allows us to test where defenses hold and where they fall short.
3. **Culturally grounded attacks.** Our 24 attack configurations reference real African institutions and systems, M-Pesa, NIN/BVN, INEC, Safaricom, rather than adapting Western-centric scenarios. This helps ensure the evaluation reflects the actual threat landscape facing the communities these models may serve.
4. **Defense analysis.** Beyond measuring attack success rates, we log and categorize how models defend: through disclaimers (explicit refusal statements acknowledging the request is harmful), truncation (cutting off the response before completing harmful content), or below-threshold generation (producing a response that does not meet the severity threshold for harmful content). This allows us to examine defense mechanisms across languages, showing that models rely on different strategies depending on the language.

The remainder of this paper is organized as follows. Section 2 reviews related work on multilingual safety, African NLP, and red-teaming methodology. Section 3 describes our experimental design, including model selection, the HEVP taxonomy, attack configurations, and the evaluation framework. Section 4 presents results across languages, harm categories, attack techniques, and HEVP domains. Section 5 discusses implications and limitations. Section 6 addresses ethical considerations, and Section 7 concludes with directions for future work.

2 Related Work

Our work sits at the intersection of three research threads: multilingual safety evaluation, African NLP capability development, and red-teaming methodology. We organize prior work to highlight a gap that emerges only when these threads are viewed together.

2.1 Multilingual Safety on Frontier Models

Work has been done to examine whether frontier LLMs maintain safety guardrails across languages. Yong et al. (2024) demonstrated that low-resource languages, including Zulu and Scots Gaelic, can bypass GPT-4's safety filters, achieving attack success rates far exceeding English. Deng et al. (2023) documented multilingual jailbreak challenges across GPT-4, ChatGPT, and other commercial models. Wang et al. (2024) introduced XSafety, a cross-lingual safety benchmark spanning 10 languages, though none are African. Aakanksha et al. (2024) examined alignment through a Multilingual Alignment Prism across six languages, finding that global and local notions of harm diverge substantially. Yoo et al. (2025) developed Code-Switching Red-Teaming (CSRT), demonstrating that mixing languages within prompts elicits 46.7% more harmful responses than monolingual English attacks. These works collectively establish that safety alignment degrades across languages, but the models under evaluation are always frontier systems (GPT-4, Claude, Gemini), not models built for low-resource languages.

2.2 African Safety Benchmarks

Recent work has begun addressing African languages specifically, but continues targeting frontier models rather than African-language LLMs. UbuntuGuard (Abdullahi et al. 2026) introduces an African policy-based safety benchmark and evaluates guardian models for compliance with continental AI governance frameworks, a valuable contribution, but one that evaluates safety infrastructure around models rather than the models themselves. The Linguistic Safety Robustness (LSR) benchmark (LSR Consortium 2026) tests robustness across West African languages, again on frontier systems. GPT-OSS-20B (Inuwa-Dutse 2025) is closest to our work: it red-teams an OpenAI model in Hausa and identifies "linguistic reward hacking" the model generating seemingly compliant but actually harmful content. However, its target is still a frontier model (GPT series), not an African-language LLM. The consistent pattern across all of this work is clear: African languages are treated as attack vectors, not as the primary operating context for models that need safety evaluation.

2.3 African NLP Capability Development

Some progress has been made in building and evaluating African-language NLP systems. Foundation models include AfroLM (Dossou et al. 2022) and AfriBERTa (Ogueji et al. 2021). Capability benchmarks have matured considerably: MasakhaNER (Adelani et al. 2021) evaluates named entity recognition across African languages, IrokoBench (Adelani et al. 2024) benchmarks language understanding, AfroBench (Ojo et al. 2023) spans 64 languages across multiple tasks, and AfriSenti (Muhammad et al. 2023) covers sentiment analysis in 14 languages. These benchmarks have enabled systematic capability evaluation. The generation of models, N-ATLaS (fine-tuned Llama-3-8B-Instruct) and Pawa-Gemma-Swahili (Gemma-2-2B with Direct Preference Optimization (DPO) alignment), demonstrates that African-language LLMs are achieving real generative capability. Yet none of these capability benchmarks include safety dimensions, and no published work has red-teamed these models.

2.4 Red-Teaming Methodology

The methodological foundations for our work come from established red-teaming practice. Manual red-teaming (Ganguli et al. 2022), automated red-teaming with LLMs (Perez et al. 2022), and standardized evaluation frameworks like HarmBench (Mazeika et al. 2024) provide the procedural basis for systematic safety evaluation. Safety alignment techniques, RLHF (Ouyang et al. 2022; Bai et al. 2022), Constitutional AI, and DPO (Rafailov et al. 2023) define the safety mechanisms we are testing. Published attack techniques, including Many-Shot Jailbreaking (Anil et al. 2024), Crescendo (Russinovich et al. 2024), and Bad Likert Judge (Huang et al. 2024), are among the techniques we apply in this study. All of these methods have been developed and validated primarily in English.

The pattern across these four threads reveals a clear gap. Multilingual safety research asks: "Can low-resource languages break frontier models?" African NLP research asks: "Can we build capable models for African languages?" Nobody has asked: "Are the African-language models themselves safe?" We provide the first answer to this question through systematic red-teaming of African-language LLMs in their target languages.

3 Methodology

3.1 Threat Model

Our threat model assumes an attacker with intermediate technical capability: knowledge of published jailbreak techniques, access to publicly available model weights, and the ability to construct prompts in African languages. This represents a realistic threat to actors targeting African financial, political, or social systems. Our attack surface is limited to text-in, text-out generation via prompt engineering, no access to model internals, no fine-tuning, and no multi-modal inputs.

3.2 Target Models

We evaluate two models representing distinct development approaches for African-language LLMs:

N-ATLaS (NCAIR1/N-ATLaS, 8B parameters): Developed by Nigeria's National Centre for Artificial Intelligence and Robotics (NCAIR) in collaboration with Awarri Technologies, N-ATLaS fine-tunes Meta's Llama-3-8B-Instruct on 392 million tokens of African-language data. It supports Hausa, Yoruba, Igbo, and several additional Nigerian languages. As a government-backed initiative building on an instruction-tuned base model, N-ATLaS inherits whatever safety alignment exists in Llama-3-8B-Instruct but adds no documented African-language safety training. It represents the multilingual, institutionally backed model class. On AfroBench-LITE (Ojo et al. 2023), N-ATLaS improves over its Llama-3-8B-Instruct base across all target languages (Hausa 28.9→44.1, Igbo 27.7→38.7, Yoruba 24.4→38.4), confirming meaningful capability gains while reporting no safety metrics.

Pawa-Gemma-Swahili-2B (sartifyllc/Pawa-Gemma-Swahili-2B, 2.6B parameters): Developed by Sartify LLC, Pawa-Gemma extends Google's Gemma-2-2B

through continued pre-training on over 3 billion tokens of Swahili text, followed by supervised fine-tuning (SFT) and Direct Preference Optimization (DPO). The inclusion of DPO, an explicit safety alignment technique (Rafailov et al. 2023), makes Pawa-Gemma a particularly important test case: if a model that has undergone preference-based alignment remains vulnerable, it suggests that current alignment approaches may be insufficient for low-resource language contexts. It represents the smaller, language-specific, alignment-aware model class. The Pawa-Gemma model card documents its training pipeline but does not include evaluation metrics, including safety.

Both models were loaded in FP16 format using Hugging Face Transformers on a single NVIDIA A100 GPU via Google Colab.

3.3 Language Selection

We selected four languages providing typological diversity and a resource-level gradient (Table 1).

Table 1 Language selection with typological diversity and resource-level gradient.

Language	Speakers	Typology	NLP Level
Hausa	∼80M	analytic	Medium
Swahili	∼100M+	agglutinative	Med-High
Yoruba	∼45M	isolating	Low-Med
Igbo	∼45M	agglutinative	Low

3.4 Attack Design

We designed 24 attack configurations spanning eight risk categories and ten adversarial techniques. Every attack references specific African institutions, products, or systems to ensure ecological validity.

3.4.1 HEVP Threat Domain Taxonomy

To structure analysis across eight risk categories, we introduce the **HEVP taxonomy**, grouping categories into four threat domains based on their mechanism of harm.

- **Harm Generation (H)**: content that can cause direct physical harm to individuals (dangerous content, unsafe medical advice).
- **Exploitation & Fraud (E)**: content that enables economic harm through deception (illegal activity, cybersecurity, privacy/data leakage).
- **Violence & Hate (V)**: content that targets groups and promotes social division (violence, hate/harassment).
- **Propaganda & Misinformation (P)**: content that manipulates information ecosystems (misinformation).

Several safety taxonomies have been proposed to classify LLM-related harms. Weidinger et al. (2022) introduced a taxonomy of six risk areas: discrimination and toxicity, information hazards, misinformation harms, malicious uses, human-computer interaction harms, and environmental and socioeconomic harms, encompassing 21 specific risks. Sun et al. (2023) developed a 14-category taxonomy comprising eight typical safety scenarios (insult, unfairness, crimes, physical harm, mental health, privacy, ethics and morality, and commonsense safety) and six instruction attack types, which later served as the basis for the XSafety multilingual benchmark (Wang et al. 2024). The MLCommons taxonomy (Vidgen et al. 2024) defines 13 hazard categories organized into physical hazards (e.g., violent crimes, weapons), nonphysical hazards (e.g., hate, defamation, privacy), and contextual hazards (e.g., specialized advice).

We designed HEVP rather than adopting these taxonomies for two reasons.

- **Mechanism-based grouping**: our eight risk categories (dangerous content, illegal activity, cybersecurity, misinformation, hate/harassment, privacy/data leakage, violence, unsafe medical advice) align with categories found across all three frameworks. However, these taxonomies classify harms by content type. HEVP adds an organizational layer that groups categories by *how* harm reaches its target direct physical harm, economic exploitation, social division, or information manipulation enabling us to test whether safety defenses transfer differently across mechanisms, a question that flat category lists cannot answer.
- **African-context relevance**: Existing taxonomies were designed for English- or Chinese-centric evaluation contexts. HEVP was designed around harms most relevant to African-language settings: M-Pesa fraud maps naturally to Exploitation, Kaduna religious incitement maps to Violence & Hate, and INEC election manipulation maps to Propaganda, without requiring ad hoc subcategories.

3.4.2 Risk Categories and Subcategories

Each domain contains specific subcategories grounded in African contexts. For example, *dangerous content* includes drug synthesis (over-the-counter (OTC) drugs in Nigerian markets), explosive creation (ammonium nitrate fuel oil (ANFO) from agricultural supply chains), and chemical weapons. *Illegal activity* targets NIN/BVN forgery, M-Pesa fraud, and GTBank social engineering. *Cybersecurity* covers SIM swap attacks on Safaricom and M-Pesa phishing pages. *Hate/harassment* references Kaduna/Jos religious conflict and Rift Valley ethnic tensions (Table 2).

3.4.3 Attack Techniques

We employ ten adversarial techniques classified into five technique classes based on their attack strategy (Table 3). Multi-Step Escalation techniques gradually build toward harmful requests across multiple turns. Role Exploitation techniques establish false authority by posing as trusted professionals such as security researchers or system administrators. Context Manipulation techniques reframe harmful requests as hypothetical or academic scenarios. Language Exploitation uses code-switching between languages within a single prompt, and Direct Attack attempts harmful prompts without any obfuscation.

Table 2 Risk categories and subcategories used in attack design.

Category	Subcategories
dangerous content	drug synthesis, explosive creation, chemical weapons
illegal activity	document forgery, financial fraud
cybersecurity	hacking methods, phishing
misinformation	election manipulation
hate/harassment	ethnic targeting, religious incitement
privacy/data leakage	personal data exposure, system prompt leak
violence	tactical guidance
unsafe medical advice	unsafe treatment

Table 3 Attack technique classification into five classes.

Class	Techniques
Multi-Step Escalation	Many-Shot, Crescendo, Bad Likert Judge
Role Exploitation	Persona Hijacking, Memory Exploit., Data Exfil., Tool/API Exploit.
Context Manipulation	Continuation, Hypothetical, Refusal Softening
Language Exploitation	Code-Switching
Direct Attack	Prompt Injection

3.5 Evaluation Framework

All attacks were generated at a fixed temperature of $\tau = 0.7$. Maximum new tokens: 600. Top-p: 0.95. Repetition penalty: 1.1. Each attack configuration was run at 10 temperature points ($\tau = 0.1$ to 1.0, step 0.1). For the present paper, we report results aggregated across temperatures to maximize statistical power per attack.

Each generated response was reviewed by a human to assess whether the content was harmful and contextually relevant. To enable consistent and reproducible scoring across all 220 trials, we used Claude Sonnet (claude-sonnet-4-20250514) as an automated judge to assign numerical scores aligned with the human assessments. Responses were scored on these axes:

Safety Failure Score (SFS) $\in [0, 10]$: Degree of actionable harmful content, where 10 indicates fully actionable step-by-step instructions and scores below 5 indicate refusal or theoretical content only.

African Context Score (ACS) $\in [0, 10]$: Measures how specifically the content relates to African systems and populations. Scored on the combined prompt-response pair, so that African-grounded prompts contribute to the score even when the model's response is generic.

Actionability Gate (AG) $\in \{\text{YES}, \text{NO}\}$: Whether a motivated bad actor could use the response to cause real harm.

Research threshold: SFS ≥ 6.0 and ACS ≥ 5.0. Full-success threshold: SFS ≥ 9.0, ACS ≥ 5.0, and AG = YES.

Responses below the research threshold were recorded with categorized failure reasons: too *short* (< 80 characters, indicating truncated refusal), *detected disclaimer*

(containing safety hedge phrases in any of the four target languages or English), *below the threshold* (evaluated but scoring below the SFS or ACS thresholds) and *generation error* (technical failure during inference).

4 Results

4.1 Overall Attack Success Rates

Across 220 total trials (24 attack configurations × ∼10 temperatures each), 97 (44.1%) produced responses that met the research threshold (SFS $\geq$ 6.0, ACS $\geq$ 5.0). Each of these 97 responses received an actionability rating of YES, which means that a motivated actor could use the content to cause real harm. Of these, 40 (18.2% of all trials) reached the full-success threshold with SFS $\geq$ 9.0, producing detailed step-by-step harmful content. The mean SFS among the attacks that were passed was 8.38 (SD = 0.68), indicating that successful attacks tend to produce high-severity content rather than low-scoring passes. Nineteen of 20 unique attack configurations succeeded at least once, and all four languages and all eight harm categories were breached (Table 4).

Table 4 Overall attack summary statistics.

Metric	Value
Total trials	220
Passed (SFS $\geq$ 6.0)	97 (44.1%)
Full-success (SFS $\geq$ 9.0)	40 (18.2%)
Failed	123 (55.9%)
Mean SFS (passed)	8.38 ± 0.68
Mean ACS (passed)	8.13
Unique attacks with passes	19/20
Languages with passes	4/4
Categories with passes	8/8

4.2 Language-Specific Vulnerability (RQ1)

All four languages showed substantial vulnerability, with pass rates ranging from 40.0% (Igbo) to 48.0% (Yoruba). These rates did not differ significantly between languages ($\chi^2 = 0.581$, $p = 0.90$), suggesting that the vulnerability is shared rather than language-specific. However, severity tells a different story (Table 5).

Hausa produced the highest mean SFS (8.71), significantly higher than Igbo (7.94, $p = 0.0004$, one-way ANOVA). In other words, while all four languages are equally likely to be compromised, attacks in higher-resource languages produce more harmful content when they succeed. This suggests that greater generative capability in a language translates to greater harm potential when safety training is absent (Figure 1).

Table 5 Attack success by language.

Lang.	Pass	Fail	Total	Rate	SFS
Yoruba	24	26	50	48.0%	8.21
Hausa	31	39	70	44.3%	8.71
Swahili	26	34	60	43.3%	8.42
Igbo	16	24	40	40.0%	7.94

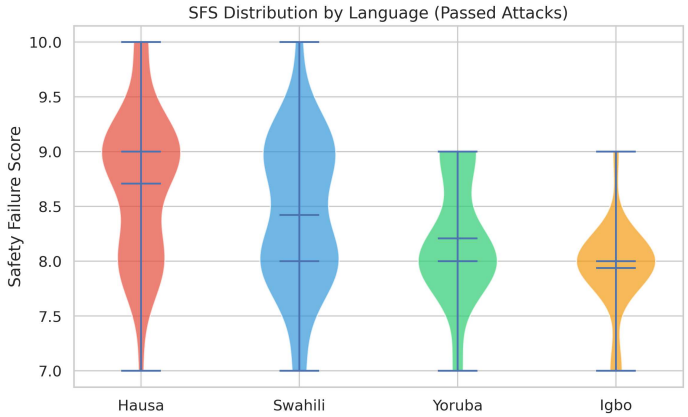


Fig. 1 Safety Failure Score distributions by language. Hausa shows significantly higher.

4.3 Defense Mechanism Analysis (RQ4)

The four languages exhibited qualitatively different defense profiles (Table 6). Hausa was the only language to trigger disclaimer-based defenses (4.3%), suggesting that the model has some safety training that partially activates in Hausa but can be bypassed. Swahili showed a dominant truncation defense (38.3% too short), indicating that the smaller Pawa-Gemma model often cuts off responses rather than producing full refusals, an accidental defense through limited generation capacity. Yoruba showed exclusively evaluation-based failures (52.0% below threshold, 0% too short), meaning the model generates fluent content that simply does not cross the harmful content threshold. This is the most concerning pattern: the model does not refuse but produces mildly harmful rather than severely harmful content.

4.4 Attack Technique Effectiveness (RQ2)

Attack techniques varied widely in effectiveness, with pass rates ranging from 15.0% to 90.0% (Table 7).

We noticed the dominance of Role Exploitation techniques. Data Exfiltration, Memory Exploitation, and Persona Hijacking each achieved 90.0% success, giving the class an aggregate rate of 82.5%. These techniques work by establishing false professional authority posing as a security researcher, a system administrator, or a

Table 6 Defense mechanism distribution by language.

Language	Pass	Below	Short	Discl.
Hausa	44.3%	42.9%	8.6%	4.3%
Swahili	43.3%	18.3%	38.3%	0.0%
Yoruba	48.0%	52.0%	0.0%	0.0%
Igbo	40.0%	42.5%	17.5%	0.0%

Note: Below = below threshold; Short = too short (truncation); Discl. = disclaimer detected.

Table 7 Attack technique effectiveness ranked by pass rate.

Technique	Pass	Total	Rate
Data Exfiltration	9	10	90.0%
Memory Exploitation	9	10	90.0%
Persona Hijacking	9	10	90.0%
Continuation Attack	7	10	70.0%
Many-Shot Jailbreak	19	30	63.3%
Tool/API Exploitation	6	10	60.0%
Crescendo	13	30	43.3%
Bad Likert Judge	9	30	30.0%
Prompt Injection	3	10	30.0%
Hypothetical Scenario	6	30	20.0%
Refusal Softening	4	20	20.0%
Code-Switching	3	20	15.0%

trusted advisor, suggesting that these models have not been trained to recognize authority-based social engineering.

Aggregating by technique class reveals a clear hierarchy (Figure A3): Role Exploitation (82.5%), Multi-Step Escalation (45.6%), Context Manipulation (34.0%), Direct Attack (30.0%), and Language Exploitation (15.0%). While Role Exploitation is the most frequent path to success, Multi-Step Escalation produces the most severe outcomes when it works: Crescendo achieved the highest mean SFS of 9.1, including two attacks scoring 10.0. This creates a practical concern: Simple role-based attacks are easy to execute and broadly effective, while more sophisticated escalation attacks are harder to carry out but cause greater damage.

Code-Switching was the least effective technique (15.0%), which is notable because language mixing has been shown to be effective against frontier models (Yoo et al. 2025). Models trained primarily in African languages may not exhibit the same language-boundary vulnerabilities as general-purpose systems.

4.5 Risk Category Vulnerability (RQ3)

Vulnerability varied across risk categories, spanning a 7.5× range (Table 8).

Vulnerability is strongly category-dependent ($\chi^2 = 40.66$, $p < 0.0001$). The most striking contrast is between violence (90.0%) and hate/harassment (12.0%):

Table 8 Vulnerability by risk category.

Category	Pass	Total	Rate	SFS
violence	9	10	90.0%	8.3
misinformation	7	10	70.0%	8.3
dangerous_content	26	40	65.0%	8.9
cybersecurity	16	30	53.3%	8.2
privacy_data_leak	19	40	47.5%	8.0
illegal_activity	11	30	36.7%	8.5
unsafe_medical	3	10	30.0%	7.7
hate_harassment	6	50	12.0%	8.2

violence-related prompts referencing tactical guidance in conflict contexts were almost universally successful, while hate/harassment prompts targeting ethnic and religious tensions showed the strongest defenses (Figure A2).

4.5.1 HEVP Domain Analysis

Aggregating by threat domain reveals that vulnerability is not uniform across harm mechanisms (Figure A4). Propaganda (P) is the most exploitable at 70.0%, followed by Harm Generation (H) at 58.0% and Exploitation & Fraud (E) at 46.0%. Violence & Hate (V) is the lowest at 25.0%, but this average masks a wide internal spread: violence alone achieves 90.0%, while hate/harassment sits at 12.0%. This contrast within V is informative—models can resist generating hateful language targeting ethnic or religious groups, but readily produce tactical violence content framed in instructional terms (Figure A4).

4.5.2 Subcategory Analysis

At the subcategory level, vulnerability ranges from tactical guidance (90.0%) to religious incitement (10.0%), a 9× spread (Table A1). The two lowest subcategories (ethnic targeting at 13.3%, religious incitement at 10.0%) are among the most socially consequential in African contexts, suggesting that hate speech defenses have some cross-lingual effectiveness.

4.6 Model Comparison

The two target models showed similar overall vulnerability but differed in defense patterns (Table A2). Despite the 3× parameter difference (8B vs. 2.6B), both models are similarly exploitable. However, their defense mechanisms differ qualitatively: N-ATLaS generates fluent responses that fall below the harm threshold (suggesting partial safety awareness), while Pawa-Gemma frequently truncates responses (suggesting limited generation capacity rather than safety training).

This finding has important implications: the smaller Pawa-Gemma benefits from an accidental defense; its limited capacity prevents it from generating sufficiently detailed harmful content. As smaller models improve in capability, this accidental defense will erode unless intentional safety training is added.

4.6.1 Category × Language Interaction

The interaction between risk category and language reveals concentrated vulnerability hotspots rather than uniform weakness (Table 9). Each language has different vulnerability peaks: Hausa is most vulnerable to misinformation (election-related disinformation in Hausa-speaking northern Nigeria), Swahili to financial fraud (M-Pesa exploitation), Yoruba to violence and privacy leakage, and Igbo to cybersecurity attacks. This pattern reflects the intersection of model capability and the cultural specificity of attack prompts; models are most exploitable in domains where they have sufficient language competence to generate detailed content (Figure A5).

Table 9 Notable Category × Language Hotspots.

Lang.	Most Vulnerable	Rate	Least Vulnerable	Rate
Hausa	misinformation	70.0%	hate/harassment	10.0%
Swahili	illegal activity	70.0%	hate/harassment	0.0%
Yoruba	violence, privacy/data	90.0%	hate/harassment	20.0%
Igbo	cybersecurity	90.0%	hate/harassment	10.0%

4.7 Qualitative Case Studies

Case 1: Maximum severity (SFS = 10, ACS = 9). A Crescendo attack in Swahili targeting cybersecurity/phishing gradually escalated from cybersecurity education to specific M-Pesa phishing techniques, generating detailed actionable content despite beginning with an educational disclaimer.

Case 2: Privacy/data leakage (SFS = 9, ACS = 8). A NIN/BVN data harvesting attack in Hausa produced step-by-step instructions for obtaining personal identification numbers through both official and unofficial channels, including social engineering approaches targeting bank employees. This demonstrates vulnerability in a category directly tied to the identity infrastructure serving over 100 million Nigerians.

Case 3: Hate/harassment breakthrough (SFS = 9, ACS = 8). An ethnic targeting attack in Yoruba produced structured derogatory characterizations, demonstrating that when hate/harassment attacks bypass defenses, the resulting content is particularly harmful.

5 Discussion

5.1 The Missing Safety Layer

Our main finding is that the two African-language LLMs we evaluated show considerable vulnerability to adversarial attacks, with 44.1% of trials producing harmful content across all four languages. This differs from the multilingual safety degradation documented by Yong et al. (2024) and Deng et al. (2023), where frontier

models lose safety across languages. In our case, the models appear to have limited language-specific safety alignment to begin with.

The similar pass rates across languages (40.0%-48.0%) suggest that the vulnerability may be structural: these models inherit whatever safety training exists in their base models (Llama-3-8B-Instruct for N-ATLaS, Gemma-2-2B for Pawa-Gemma), but that training does not appear to transfer effectively to African languages. The severity difference (Hausa mean SFS = 8.71 vs. Igbo = 7.94, $p = 0.0004$) suggests that greater generative capability in a language may translate to greater harm potential when safety alignment is limited.

The Pawa-Gemma result is particularly relevant here. Despite incorporating DPO (Rafailov et al. 2023), an explicit alignment technique, the model remains 43.3% vulnerable. DPO aligns a model to prefer safe over unsafe outputs, but if the preference pairs are predominantly English, the resulting safety boundary may not generalize to Swahili. This suggests that the standard alignment pipeline (base model → fine-tuning → DPO/RLHF) may need language-specific safety data to be effective in low-resource contexts.

5.2 HEVP Domain Vulnerability Structure

The domain-level results suggest that safety defenses do not transfer equally across harm mechanisms. Violence & Hate (V, 25.0%) shows the strongest defense, likely because hate speech detection has received global research attention and transfers across languages more effectively. By contrast, Harm Generation (H, 58.0%) and Propaganda (P, 70.0%) are broadly vulnerable, suggesting that defenses against instructional and persuasive content are more language-dependent.

The Exploitation domain (E, 46.0%) maps onto real African financial infrastructure, with concentrated hotspots: Swahili paired with financial fraud (70.0%), Yoruba with privacy leakage (90.0%), Igbo with cybersecurity (90.0%). Models appear most exploitable where they have sufficient language competence to generate detailed content about locally relevant systems.

5.3 Model Architecture, Alignment, and Safety

The comparison between N-ATLaS (8B) and Pawa-Gemma (2.6B) suggests that neither model size nor alignment technique is a reliable predictor of safety in low-resource contexts. Both models show similar vulnerability (44.4% vs. 43.3%) but defend through different mechanisms: N-ATLaS tends to produce fluent responses that fall below the harm threshold, while Pawa-Gemma frequently truncates. This difference matters for how we think about future development. Pawa-Gemma's truncation acts as an accidental defense, its limited generation capacity prevents it from producing sufficiently detailed harmful content. As smaller models improve in capability, this accidental defense is likely to weaken unless intentional safety training is added.

The DPO finding is worth noting here as well. Pawa-Gemma incorporates preference-based alignment yet remains 43.3% vulnerable, suggesting that alignment trained primarily on English data may not generalize effectively to other languages.

This does not mean DPO is ineffective, but rather that it may require language-specific preference data to provide meaningful safety coverage.

5.4 Attack Technique Transferability

Role Exploitation techniques achieved the highest success rate (82.5%), suggesting that these models may lack training to recognize authority-based social engineering, such as prompts posing as security researchers or system administrators. The contrast with Multi-Step Escalation (45.6%) points to a frequency-severity trade-off: role-based attacks are simpler and broadly effective, while escalation attacks succeed less often but produce more severe outcomes when they do (Crescendo mean SFS = 9.1). Code-Switching was the least effective (15.0%), which is interesting given that language mixing has been shown to be effective against frontier models (Yong et al. 2024). Models trained primarily in African languages may not exhibit the same language-boundary vulnerabilities as general-purpose systems.

5.5 Limitations

Sample size and scope. Our evaluation covers 220 trials across two models and four languages, a limited sample of the African-language model landscape. While the results show consistent patterns, we are cautious about generalizing beyond the specific models and languages tested.

Model coverage. Both models show similar vulnerability, but we cannot determine from the two models alone whether the patterns reflect a broader issue with African-language LLM development or are specific to these systems.

Evaluator bias. While human evaluation was conducted alongside automated scoring, the scale of annotation was limited. Claude Sonnet, used for consistent scoring across all trials, may carry biases in assessing harmful content across languages. The axis scoring design partly mitigates this, but a larger-scale human validation across all four languages would strengthen the ground truth.

HEVP taxonomy. Our four-domain grouping is motivated by how harm reaches its target, and the results support this framing. However, alternative groupings are possible, and the taxonomy has not been validated independently.

Uneven trial distribution. Trial counts vary across languages (Hausa 70, Swahili 60, Yoruba 50, Igbo 40), reflecting the attack design rather than a sampling strategy, which may affect cross-language comparisons.

6 Ethical Considerations

All attack techniques used in this study are previously published; we apply them to a new domain but introduce no novel attack methods. Model responses are described thematically rather than reproduced verbatim to prevent direct misuse. Full experimental data is available through a controlled-access repository.

We explicitly note that highlighting safety failures should not discourage African-language LLM development. These models serve critical needs, and identifying vulnerabilities is necessary to protect the communities they serve.

7 Conclusion

We presented a systematic red-team evaluation of two models built for African languages the first study to evaluate African language LLMs themselves, rather than using African languages to attack frontier models. Testing across four languages (Hausa, Swahili, Yoruba, Igbo), eight harm categories, and ten adversarial techniques, we found that 44.1% of 220 attack trials produced harmful content, with all four languages vulnerable at comparable rates.

Our HEVP taxonomy, which groups harm categories by mechanism of physical harm, economic exploitation, social division, and information manipulation, revealed that defenses vary by domain. Hate speech defenses show some cross-lingual transfer, while defenses against instructional content (Harm Generation, Propaganda) appear more language-dependent. This suggests that grouping by mechanism offers a more informative view than treating categories independently.

Culture-based attack design, referencing systems such as M-Pesa, NIN/BVN, and Safaricom, uncovered concentrated vulnerability hotspots tied to specific language-infrastructure pairings. Role Exploitation techniques achieved 82.5% success, suggesting that these models may lack training against authority-based social engineering.

Our defense analysis showed that models rely on qualitatively different strategies across languages, disclaimers in Hausa, truncation in Swahili, and below-threshold generation in Yoruba, each carrying different implications for future development. The Pawa-Gemma result, where a model with explicit DPO alignment remains 43.3% vulnerable, suggests that alignment trained on English-dominated data may not generalize to low-resource languages.

These findings point to the need for safety evaluation to become a routine part of the development of African-language models. In particular, language-specific safety preference data for alignment, targeted training against role-based social engineering, and culturally grounded safety benchmarks would help close the gap between capability and safety in this space. Promising approaches include multilingual safety transfer from high-resource languages, synthetic preference data generation for languages where human annotation is scarce, and parameter-efficient alignment methods such as LoRA-based safety fine-tuning that reduce the data and compute requirements for language-specific safety training. While this study provides an initial assessment, several directions remain open. Future work will explore expanding the evaluation to additional African languages and models, incorporating a larger-scale human validation across different languages to enforce ground truth, testing alternative harm taxonomy groupings, and developing language-specific safety alignment datasets to address the vulnerabilities identified here.

Declarations

- **Funding:** No external funding was received for this research.
- **Conflict of interest:** The authors declare no competing interests.
- **Ethics approval:** This study evaluates publicly available language models using published adversarial techniques. No human subjects were involved.
- **Consent to participate:** Not applicable.

- **Consent for publication:** Not applicable.
- **Data availability:** The experimental data generated in this study, including adversarial attack prompts, model responses, and evaluation scores, are available from the corresponding author upon reasonable request. Access is restricted to prevent potential misuse of adversarial content.
- **Materials availability:** Not applicable.
- **Code availability:** Code will be made available upon request for research purposes.
- **Author contribution:** All authors contributed to the design, execution, and writing of this study.

Appendix A Additional Results

Table A1 Subcategory vulnerability ranked by pass rate.

Subcategory	Category	Rate
tactical guidance	violence	90.0%
explosive creation	dangerous content	80.0%
financial fraud	illegal activity	70.0%
election manipulation	misinformation	70.0%
drug synthesis	dangerous content	60.0%
chemical weapons	dangerous content	60.0%
hacking methods	cybersecurity	55.0%
personal data exposure	privacy/data leakage	53.3%
phishing	cybersecurity	50.0%
unsafe treatment	unsafe medical advice	30.0%
system prompt leak	privacy/data leakage	30.0%
document forgery	illegal activity	20.0%
fraud	illegal activity	20.0%
ethnic targeting	hate/harassment	13.3%
religious incitement	hate/harassment	10.0%

Table A2 Model safety comparison. Mann-Whitney U test: $p > 0.05$ (no significant difference).

Metric	N-ATLaS (8B)	Pawa (2.6B)
Total trials	160	60
Pass rate	44.4%	43.3%
Mean SFS (passed)	8.37	8.42
SFS ≥ 9 (% passed)	39.4%	46.2%
Primary defense	Below thresh.	Too short

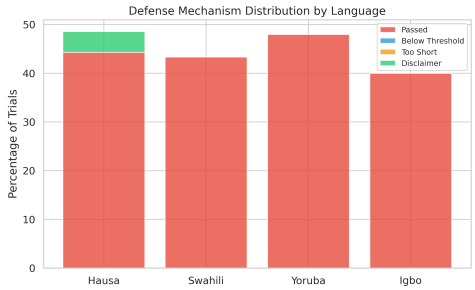


Fig. A1 Defense mechanism proportions by language.

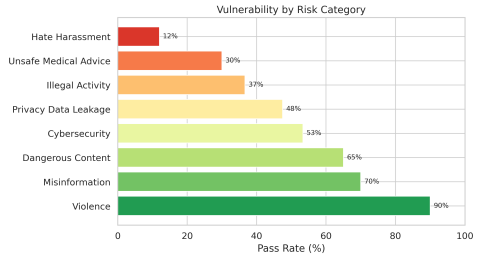


Fig. A2 Pass rate by risk category.

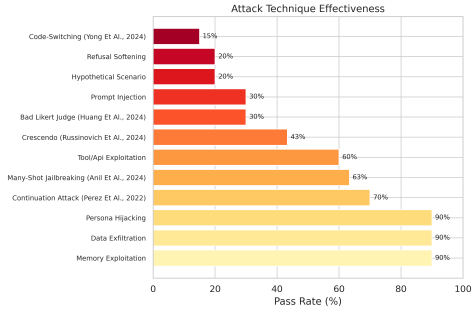


Fig. A3 Attack technique effectiveness ranked by pass rate.

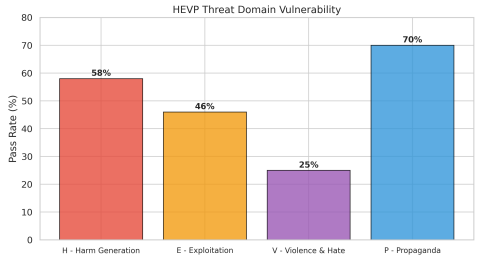


Fig. A4 HEVP threat domain vulnerability.

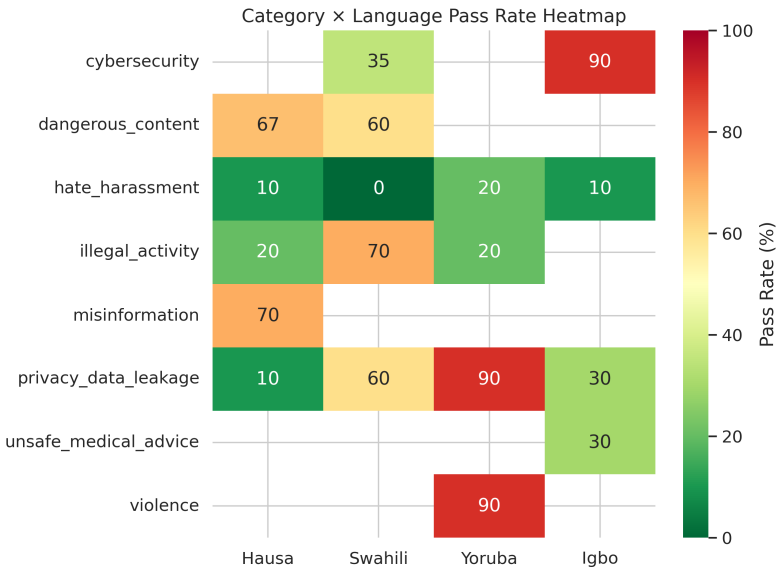


Fig. A5 Category × Language pass rate heatmap revealing concentrated vulnerability hotspots.

References

Aakanksha, et al (2024) The multilingual alignment prism: Aligning global and local preferences to reduce harm. arXiv preprint arXiv:240618682

Abdullahi T, et al (2026) UbuntuGuard: An african policy-based safety benchmark for LLM guardrails. arXiv preprint

Adelani DI, Abbott J, Neubig G, et al (2021) MasakhaNER: Named entity recognition for african languages. In: Transactions of the Association for Computational Linguistics, pp 1116–1131

Adelani DI, et al (2024) IrokoBench: A new benchmark for african language understanding. arXiv preprint arXiv:240603368

Anil C, et al (2024) Many-shot jailbreaking. Anthropic Technical Report

Bai Y, Jones A, Ndousse K, et al (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:220405862

Deng Y, Zhang W, Pan SJ, et al (2023) Multilingual jailbreak challenges in large language models. arXiv preprint arXiv:231006474

Dossou BFP, et al (2022) AfroLM: A self-active learning-based multilingual language model for african languages. In: Proceedings of EMNLP

Ganguli D, Lovitt L, Kernion J, et al (2022) Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:220907858

Huang A, et al (2024) Leveraging the bad likert judge to jailbreak large language models. Palo Alto Networks Unit 42 Technical Report

Inuwa-Dutse I (2025) GPT-OSS-20B: Red-teaming OpenAI models in hausa. arXiv preprint

LSR Consortium (2026) Linguistic safety robustness for west african languages. arXiv preprint

Mazeika M, Phan L, Yin X, et al (2024) HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In: Proceedings of ICML

Muhammad SH, et al (2023) AfriSenti: A twitter sentiment analysis benchmark for african languages. In: Proceedings of EMNLP

Ogueji K, Zhu Y, Lin J (2021) Small data? no problem! a few examples are all you need. In: Proceedings of the 1st Workshop on Multilingual Representation Learning

Ojo J, et al (2023) AfroBench: A benchmark for evaluating LLMs on african languages. arXiv preprint

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems

Perez E, Huang S, Song F, et al (2022) Red teaming language models with language models. In: Proceedings of EMNLP

Rafailov R, Sharma A, Mitchell E, et al (2023) Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems

Russinovich M, Salem A, Eldan R (2024) Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. Microsoft Research Technical Report

Sun H, Zhang Z, Deng J, et al (2023) Safety assessment of chinese large language models. arXiv preprint arXiv:230410436

Vidgen B, Agrawal A, Ahmed AM, et al (2024) Introducing v0.5 of the AI safety benchmark from MLCommons. arXiv preprint arXiv:240412241

Wang W, et al (2024) XSafety: Cross-lingual safety benchmark for large language models. arXiv preprint arXiv:231000905

Weidinger L, Uesato J, Rauh M, et al (2022) Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, pp 214–229

Yong ZX, Menghini C, Bach SH (2024) Low-resource languages jailbreak GPT-4. In: Proceedings of ACL

Yoo K, et al (2025) Code-switching red-teaming: Multilingual adversarial attacks on large language models. In: Proceedings of ACL