
AfroJailbreak-ZW: Evaluating Jailbreak Resistance in Shona

Malvin T. Machingura
Bindura University of Science Education

With
Apart Research

Abstract

Large language model safety alignment is trained almost entirely in English, and prior work shows that refusal behavior often fails to transfer to lower-resource languages. We introduce AfroJailbreak-ZW, the first jailbreak-resistance evaluation for Shona, Zimbabwe's largest indigenous language. We cover five attack categories across English, Shona, and code-switched conditions, tested against GPT and Gemini models. We report Attack Success Rate, refusal rate, and a Language Vulnerability Index that corrects success rates for comprehension failures. We release the benchmark and evaluation harness to support future work on Southern African languages.

1. Introduction

Large language models are increasingly deployed in contexts where users speak Shona rather than English. Safety alignment for these models, however, is built and red-teamed almost exclusively in English. Whether the refusal behavior measured in English holds for Shona speakers is currently unmeasured.

This matters because the mechanism behind cross-lingual safety failure is reasonably well understood: safety fine-tuning data is unevenly distributed across languages, so refusal behavior generalizes poorly to languages underrepresented in that data (Yong, Menghini & Bach, 2023). Translating a refused English prompt into a low-resource language has been shown to bypass safety mechanisms in GPT-4 at success rates competitive with state-of-the-art optimization-based jailbreaks (Yong et al., 2023), and this effect has since been characterized more precisely as two distinct failure modes: a harmfulness curse (models comply more in low-resource languages) and a relevance curse (models also lose instruction-following quality, complicating naive success measurement) (Deng et al., 2024).

Despite rapid recent progress in African-language safety evaluation, every benchmark we identified stops short of Shona specifically.

Our main contributions are:

1. The first jailbreak-resistance dataset for Shona, testing two attack categories (harmful instructions and role-play) across English, Shona, and code-switched conditions.
2. A Language Vulnerability Index (LVI) that combines Attack Success Rate with a relevance/comprehension penalty, addressing the known confound that “jailbroken” and “incoherent” are easy to conflate in low-resource-language model outputs
3. An empirical comparison of ASR, refusal rate, and LVI across English, Shona, and code-switched conditions on GPT and Gemini family models, reported with explicit sample-size and judge-reliability caveats.
4. An open, reproducible evaluation harness and dataset schema that future researchers can extend to other SADC languages without redoing the methodological design work.

2. Related Work

Multilingual jailbreak research established the core phenomenon we build on. Yong et al. (2023) showed GPT-4 could be jailbroken via translation into low-resource languages (Zulu, Scots Gaelic) at rates competitive with optimization-based attacks. Deng et al. (2024) introduced the MultiJail dataset and the harmfulness-curse/relevance-curse framing. Song et al. (2024) showed that mixing languages within a single prompt is itself an effective attack vector, independent of translation alone, directly motivating our code-switched condition.

African-language coverage has emerged recently, with studies extending jailbreak evaluation to languages like Yoruba and Zulu. However, chiShona remains underrepresented in safety

evaluation. AfroJailbreak-ZW closes this gap, adapting established four-condition designs to the Shona context with native-speaker validation.

No benchmark identified as of mid-2026 evaluates Shona in a jailbreak or safety context.

AfroJailbreak-ZW addresses this, focusing on Shona to ensure feasible native-speaker validation at hackathon scale.

3. Methods

Dataset. We developed three parallel prompt sets, English, Shona, and Shona-English code-switched, covering five attack categories: harmful instructions, prompt injection, role-play, translation attacks, and obfuscation using informal Zimbabwean digital spelling. This pilot tests two of the five defined categories (harmful instructions, role-play); the remaining three are defined in the framework for future testing. Shona prompts were created by the author (a native Shona speaker); independent back-translation validation was not performed within the hackathon timeframe. Code-switched prompts reflect naturalistic language mixing.

Models. We evaluate GPT (via ChatGPT) and Gemini(via Gemini mobile app) models to compare how training architectures impact safety. This approach tests whether training data composition affects safety vulnerability in Shona.

Metrics. We measure Attack Success Rate (ASR) to track harmful compliance, reporting Refusal Rate separately. We also calculate a Language Vulnerability Index (LVI), which combines ASR with a relevance rate to distinguish between genuine jailbreaks and comprehension failures.

Judge protocol. All responses were labeled by a single human rater (the author) using a REFUSED/COMPLIANT/IRRELEVANT rubric; no automated judge model was used given the project timeframe.

4. Results

Table 1. ASR and refusal rate by language condition (English / Shona / Code-switched) $\times$ model, with sample size. (n)

Model	Language	n	ASR	Refusal rate	LVI
GPT	English	6	0%	66.7%	0%
GPT	Shona	6	16.7%	66.7%	13.9%
GPT	Code-switched	6	16.7%	83.3%	16.7%
Gemini	English	6	16.7%	50%	11.1%
Gemini	Shona	5	20%	80.0%	20.0%
Gemini	Code-switched	5	40%	60%	40.0%

Results should be interpreted as directional point estimates given the small sample size ($n=5-6$). We observed higher vulnerability in the code-switched condition, consistent with existing literature, though our data does not support definitive statistical claims.

5. Discussion and Limitations

If results echo broader literature, the interpretation is likely that Shona speakers face risks that have simply gone unmeasured until now. No provider can currently claim that safety mitigations hold for Shona, which is particularly concerning for high-severity content and informal, code-switched registers used in Zimbabwean services.

Limitations

The dataset is small and hypothesis-generating rather than statistically definitive. Labeling by a small team and testing only three model families limits generalizability. Shona has register and dialect variation that a single translation pass cannot fully capture, and the reliability for Shona outputs remains a key uncertainty.

Misuse risks. A benchmark that documents which prompt categories, languages, or phrasings most reliably bypass safety training is, by construction, a roadmap that could be misused if the underlying prompt dataset were released without access controls. We mitigate this by (a) not publishing the populated harmful-instruction dataset openly, (b) following the same gated-access norms used for AdvBench, HarmBench, and StrongREJECT, and (c) committing to responsible disclosure, sharing any newly discovered, currently exploitable vulnerability with the relevant model provider before public release, consistent with the norms modeled by Yong et al. (2023). There is also a risk that a low LVI score could be misread as “this language is safe,” when it may instead reflect comprehension failure or an underpowered sample; we explicitly caution against that reading throughout this report.

Future Work

Future work should extend this design to other SADC languages, test optimization-based attacks against manual Shona prompts, and develop language-aware safety judges validated specifically on Shona outputs. We advocate for larger, participatory-research datasets to improve on single-team translation passes.

6. Conclusion

With AfroJailbreak-ZW, we provide the first empirical pilot assessment of safety vulnerabilities for Shona, a significant Zimbabwean language missing from contemporary multilingual red-teaming datasets. Our main contributions include the dataset itself, the LVI framework designed to mitigate comprehension-induced success artifacts, and an open, reproducible methodology. Although these results are hypothesis-generating rather than definitive evidence of safety failure,

we propose this work to stimulate broader research and urge providers to prioritize alignment evaluations for Southern African linguistic landscapes across models such as GPT and Gemini.

Code and Data

No public repository link is available at this time. The prompt schema and scoring rubric are described in Section 3, and the populated dataset is withheld per the dual-use considerations in Section 5.

Author Contributions

Malvin T. Machingura designed and led the project, implemented the evaluation harness, and conducted the analysis.

References

1. Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2024). *Multilingual Jailbreak Challenges in Large Language Models*. *arXiv:2310.06474*.
2. Song, J., Huang, Y., Zhou, Z., & Ma, L. (2024). *Multilingual Blending: LLM Safety Alignment Evaluation with Language Mixture*. *arXiv:2407.07342*.
3. Yong, Z.-X., Menghini, C., & Bach, S. H. (2023). *Low-Resource Languages Jailbreak GPT-4*. *arXiv:2310.02446*.

LLM Usage Statement

We utilized AI tools to assist with literature synthesis, report structuring, and evaluation harness development. All citations were manually verified, and empirical findings were derived from our independent analysis.