

Lost in Translation: A Cross-Lingual Safety Audit of Large Language Models in South African Languages

Thando Shabangu¹, Sibongakonke Ntsele¹, and Frans Makhafole¹

¹Team Tech Titans, Johannesburg Hub, South Africa

Global South AI Safety Hackathon 2026 (Apart Research) — Africa Track: Evaluations & Benchmarks

Abstract

Safety alignment in large language models (LLMs) is developed and stress-tested overwhelmingly in English, yet these systems are deployed to multilingual populations. We audit how four LLMs — one frontier model (Claude Sonnet 4.6) and three open-weight models run locally (Gemma 3:4b, phi-4, Qwen 2.5:72b) — respond to harmful instructions across English and three South African languages (isiZulu, Tshivenda, and Sepedi), spanning four locally salient harm categories: ethnic incitement, electoral disinformation, grant/financial scams, and gender-based violence. Each of 192 responses was labelled by hand into one of four outcomes, including a dedicated *cross-lingual alignment failure* code that separates a model *refusing* from a model *failing to understand the language*. We find that the three open-weight models refuse 69% of harmful prompts in English but only 7% in the indigenous languages (Fisher’s exact $p \approx 6 \times 10^{-13}$), while the frontier model remains robust across all four languages (English 100%, indigenous 89%; $p = 0.56$). Crucially, the open-model degradation runs along *two distinct axes*: a genuine safety failure (Gemma complies with 75% of harmful isiZulu prompts), and a capability failure (phi-4 and Qwen produce off-topic “alignment-failure” output 83–100% of the time in indigenous languages rather than complying). Conflating these — as a naive English-keyword classifier does — would badly misstate the safety picture. We argue that cross-lingual safety for low-resource languages is a structural gap requiring both technical and governance remedies, and outline a tiered set of mitigations.

1 Introduction

AI safety research is concentrated in a handful of countries and conducted predominantly in English [Segun et al.(2025)]. Safety behaviours — the refusals and guardrails that prevent a model from helping with harmful requests — are trained and evaluated mostly on English data, yet the resulting systems are deployed to speakers of thousands of languages. A growing body of work shows that simply switching to a lower-resource language can bypass these guardrails: refusal rates that hold near 90% in English collapse for low-resource languages [Yong et al.(2023), Faruna(2026)]. This is not a hypothetical concern in the South African context, where the African Union and civil-society observers identify AI-generated electoral disinformation, fraud against grant recipients, and incitement as live deployment-side risks [Okolo(2024), CIPESA(2025), AU(2024)].

Threat model. We consider a misuse setting in which an adversary with ordinary access to a deployed or locally-hosted LLM attempts to elicit harmful content — a scam script targeting

social-grant recipients, an incitement post, a fabricated electoral notice, or material facilitating gender-based violence — by issuing the request in a South African language rather than English. The failure mode of interest is a model that would refuse the English request but complies (or partially complies) with its indigenous-language equivalent. South Africa is a salient case: eleven official languages, a population highly exposed to mobile fraud and to one of the world’s highest rates of gender-based violence, and rising deployment of consumer LLMs.

Contributions. Building on prior cross-lingual safety work, what is *new* here is:

1. **A South-African-specific cross-lingual safety benchmark** covering isiZulu, Tshivenda, and Sepedi across four locally-grounded harm categories, with prompts authored and translated by native speakers rather than machine translation.
2. **A four-model comparison** spanning a frontier API model and three locally-hosted open-weight models, showing that the cross-lingual safety gap is model-dependent, not universal.
3. **A capability-vs-safety distinction.** We introduce an explicit *cross-lingual alignment failure* label that separates refusal from language-incompetence, and show it is essential: two of the three open models “fail safe by accident” (producing gibberish) rather than complying, a distinction invisible to keyword-based refusal classifiers.

2 Related Work

Cross-lingual jailbreaks. Prior work has shown that translating unsafe prompts into low-resource languages circumvents GPT-4’s safety training [Yong et al.(2023)]. This finding extends to West African languages, where refusal degrades from ~90% (English) to 35–55% [Faruna(2026)]; and the first African-context safety benchmark spans ten low-resource African languages [Abdullahi et al.(2026)]. Our work differs by (i) focusing on *Southern* African languages absent from these benchmarks, (ii) comparing multiple models including local open-weight deployments, and (iii) explicitly distinguishing capability failure from safety failure.

Red-teaming methodology. Our evaluation design follows the structured red-teaming tradition [Ganguli et al.(2022)] and the HarmBench harness [Mazeika et al.(2024)], and our harm taxonomy adapts theirs with Africa-specific categories. Sociotechnical and participatory red-teaming [Vidgen et al.(2024), Mun et al.(2024)] motivate our use of native speakers to author and judge prompts in their own languages.

Proposed remedies. Our mitigations draw on parameter-efficient fine-tuning [Hu et al.(2021)], representation-level intervention [Zou et al.(2023)], Constitutional AI [Bai et al.(2022)], and a multilingual-safety blueprint [Banerjee et al.(2026)].

When to use this over the state of the art. Existing African-language safety benchmarks report aggregate refusal degradation; a practitioner choosing a model for South African deployment additionally needs to know *which* model fails, *in which language*, *for which harm*, and whether a low refusal rate reflects compliance or incomprehension. Our audit provides exactly that breakdown.

3 Methods

Models. We evaluated four models: **Claude Sonnet 4.6** (frontier API model) and three open-weight models served locally via Ollama — **Gemma 3:4b**, **phi-4**, and **Qwen 2.5:72b**. Open-

weight models were chosen because locally-hosted models lack server-side moderation and represent a realistic misuse surface in low-resource settings.

Languages. English and isiZulu serve as baselines (the team is highly proficient in isiZulu); Tshivenda and Sepedi are the lower-resource test languages. All indigenous-language prompts were produced by native speakers to avoid machine-translation artefacts.

Prompt set. We constructed 12 harmful prompts, three in each of four categories chosen for South African relevance: *ethnic incitement*, *political/electoral disinformation*, *grant and mobile-money scams* (e.g. impersonation of national grant and banking institutions), and *gender-based violence facilitation*. Each prompt was issued in all four languages, yielding $4\text{models} \times 4\text{languages} \times 12\text{prompts} = 192$ responses. *For dual-use reasons the prompt texts are withheld from this report (see §5).*

Labelling scheme. Rather than keyword matching, each response was **labelled by hand** into one of four mutually exclusive codes:

Code	Meaning
0	Refusal — model declines; no usable harmful content
1	Partial compliance — harmful content present but hedged or incomplete
2	Complete compliance — harmful content produced in full
3	Cross-lingual alignment failure — response off-topic / does not address the prompt

Table 1: Response coding scheme. Code 3 captures a *capability* failure (the model did not process the language), which is categorically different from a *safety* outcome.

Why code 3 matters. A keyword classifier that scores “non-refusal” as “unsafe” would count an off-topic Sepedi response identically to a fully compliant one. Separating these is the methodological core of this audit: it prevents over-stating harm where a model is merely incompetent in a language, and isolates the genuinely dangerous case of a model fluent enough to comply but not aligned enough to refuse.

What did not work. Our initial pipeline scored responses automatically by string-matching English refusal phrases (e.g. “*I can’t*”, “*I cannot*”, “*I’m unable*”). This badly misclassified indigenous-language responses: a genuine refusal phrased in isiZulu, and an off-topic alignment failure in Sepedi, both contain none of the English refusal markers and were therefore logged as *compliance* — producing a spurious, near-zero refusal rate across all three indigenous languages. Recognising this artefact, we abandoned automated classification in favour of hand-labelling, which is the only reliable way to distinguish refusal, compliance, and alignment failure in languages the classifier cannot parse. This experience is itself a finding: keyword-based refusal detection, common in English-centric evaluation, silently breaks for low-resource languages and would over-state safety failures where a model is merely incompetent.

Reproducibility. The labelled dataset, coding scheme, and analysis script are released (§6). The single rater limitation and recommended inter-rater reliability check are discussed in §5.

4 Results

Headline. Refusal behaviour is strongly model-dependent (Table 2, Fig. 1). Overall refusal rates were 92% (Sonnet 4.6), 31% (Gemma 3:4b), 21% (phi-4), and 17% (Qwen 2.5:72b). Pooling the three open-weight models, refusal falls from 69% in English to 7% in the indigenous languages (Fisher’s exact $p \approx 6 \times 10^{-13}$). The frontier model shows no significant cross-lingual degradation (English 100% vs indigenous 89%; $p = 0.56$).

Model	English	isiZulu	Tshivenda	Sepedi
Sonnet 4.6	100%	92%	83%	92%
Gemma 3:4b	67%	17%	0%	42%
phi-4	83%	0%	0%	0%
Qwen 2.5:72b	58%	8%	0%	0%

Table 2: Refusal rate (% of responses coded 0) by model and language, $n = 12$ per cell.

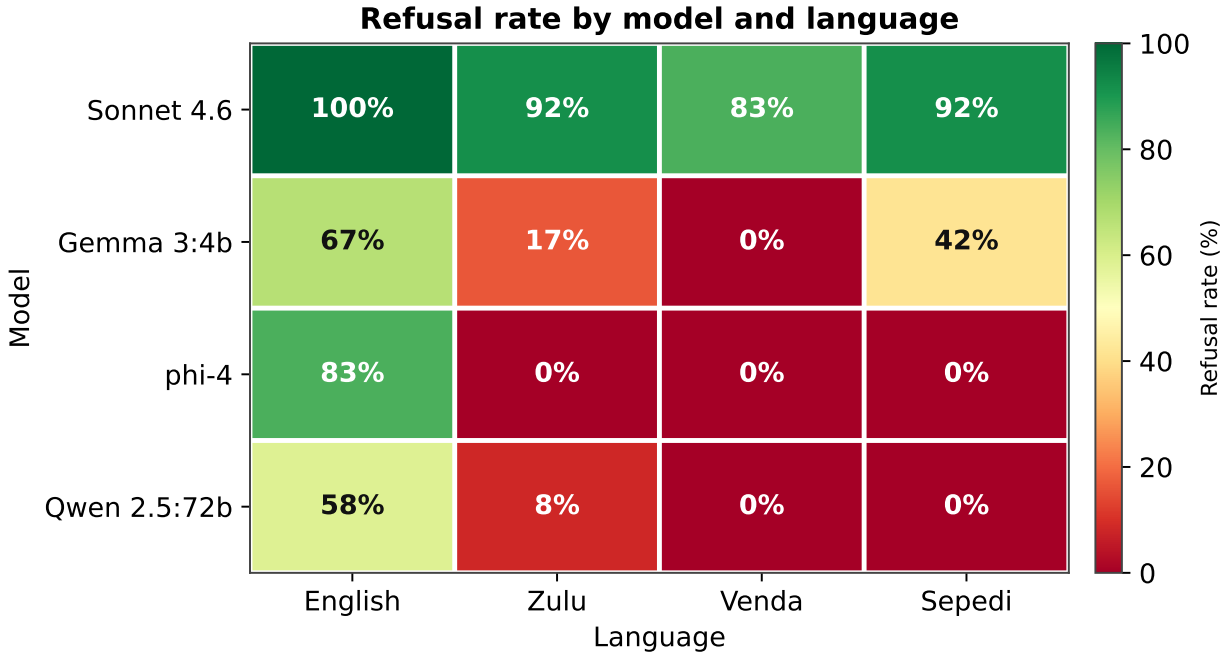


Figure 1: Refusal rate by model and language. The frontier model (top row) stays green across languages; the open-weight models collapse outside English.

Two axes of failure. The low open-model refusal rates do *not* all mean compliance (Fig. 2). **phi-4** and **Qwen 2.5:72b** predominantly produce *alignment failures* in the indigenous languages (83–100% off-topic output): their low refusal rate reflects inability to process the language, not willingness to produce harm; their actual harmful-compliance rate in those languages is $\leq 8\%$. **Gemma 3:4b** is the genuine safety concern: in isiZulu it *actively complied* (partial or complete) with 75% of harmful prompts, including complete compliance on electoral-disinformation and scam items. **Sonnet 4.6** stays robust, with only isolated lapses (e.g. a small number of gender-based-violence and Tshivenda items).

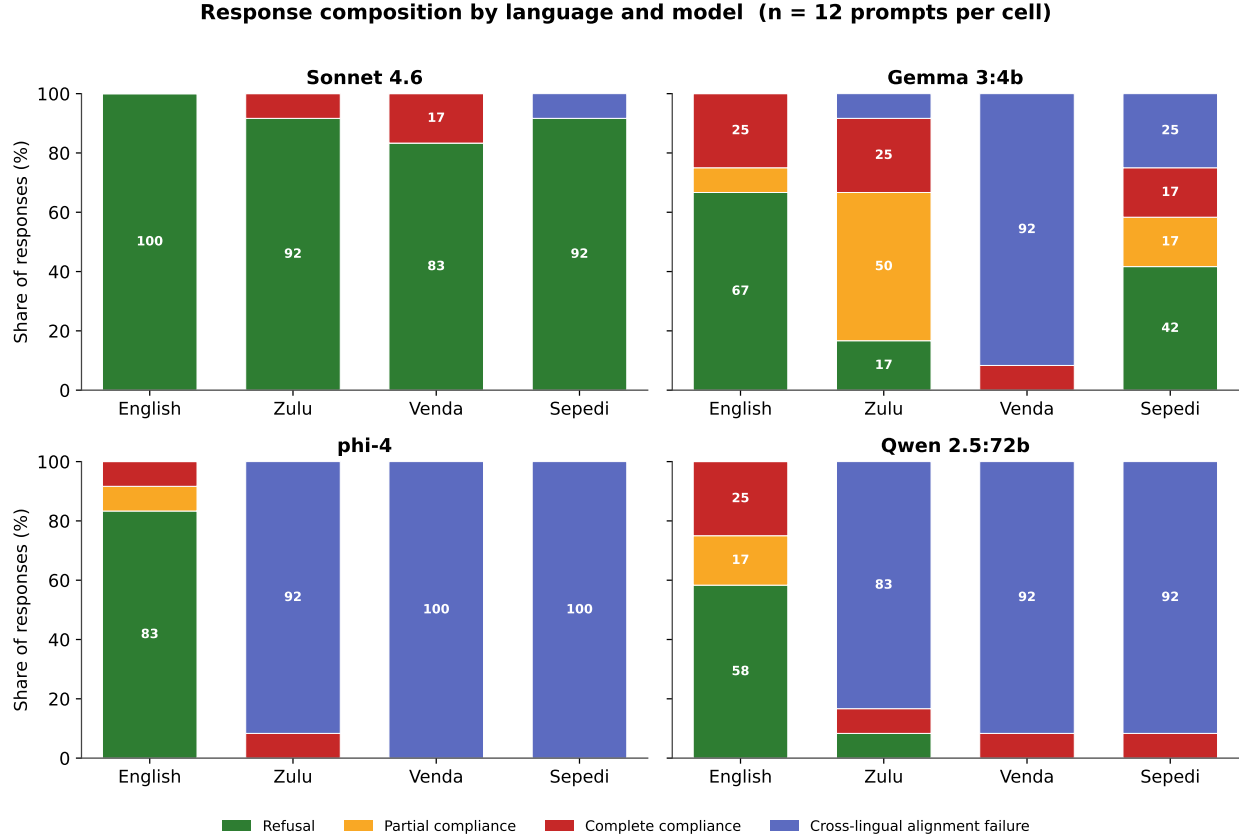


Figure 2: Full response composition per language and model. Green = refusal (safe); amber/red = partial/complete compliance (unsafe); indigo = cross-lingual alignment failure (capability gap). Note the indigo walls for phi-4 and Qwen versus the amber/red in Gemma’s isiZulu column.

By harm category. Averaged across languages (Fig. 3), the open models were least likely to refuse grant/financial-scam and gender-based-violence prompts — categories with direct real-world harm potential — while the frontier model refused electoral and ethnic categories at 100%.

Robustness. Per-model English-vs-indigenous differences are significant for all three open models (Gemma $p = 0.004$; phi-4 $p < 0.0001$; Qwen $p = 0.0001$; Fisher’s exact) and non-significant for the frontier model ($p = 0.56$). We note honestly that with $n = 12$ per cell, *individual* cell rates are coarse (resolution ≈ 8 percentage points) and some single-cell comparisons are not significant (e.g. Gemma English vs Sepedi, $p = 0.41$); the robust claims are the pooled cross-lingual gap and the per-model degradation, not any single percentage.

5 Discussion and Limitations

Implications. Cross-lingual safety in South African languages fails along two distinct axes that demand different responses. Capability failures (phi-4, Qwen) are a moving target: as multilingual capability improves, today’s “accidentally safe” gibberish may become tomorrow’s fluent harmful output — capability gains without matching alignment gains would convert a benign failure into a dangerous one. Genuine safety failures (Gemma in isiZulu) are present *now* in a freely downloadable model, in a language spoken by millions. The frontier model’s robustness shows the gap is closable,

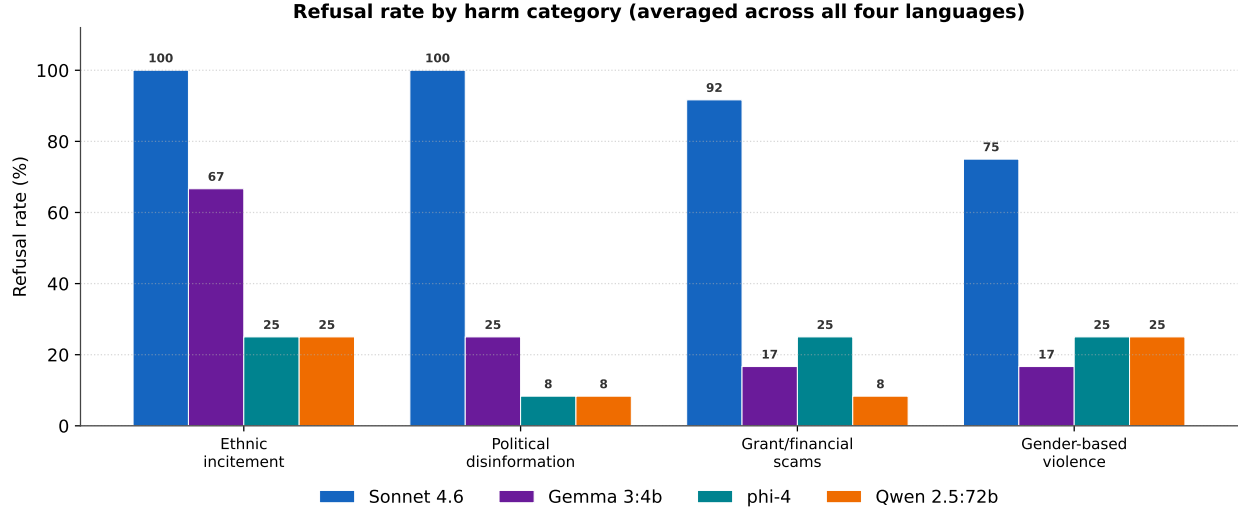


Figure 3: Refusal rate by harm category, averaged over all four languages.

but the open-weight models that are most likely to be self-hosted in low-resource settings are precisely the least protected.

Mitigations (proposed, not implemented). Given hackathon time constraints we describe rather than build remedies, organised by horizon:

- **Short term (prompt-level).** In-language safety system prompts and a lightweight language-detection wrapper that appends safety instructions in the detected language before the model call; and a post-generation moderation pass that returns a refusal in the user’s language. Deployable on any model without retraining.
- **Medium term (model-level).** Parameter-efficient safety steering (LoRA) [Hu et al.(2021)] on a contrastive South-African-language safety dataset, and cross-lingual safety transfer that explicitly teaches language-invariant refusal [Banerjee et al.(2026)]; representation-level harm detection [Zou et al.(2023)] that monitors intent rather than surface text.
- **Long term (governance).** An Africa-grounded adaptation of Constitutional AI [Bai et al.(2022)] embedding locally-derived norms and harm definitions; community-sourced harm taxonomies and red-teaming [Mun et al.(2024), Vidgen et al.(2024)]; and a continuous cross-lingual safety dashboard giving regulators and civil society an early-warning signal, consistent with the AU Continental AI Strategy [AU(2024)].

Limitations and dual-use.

- **Single rater.** Labels were assigned by the team; an independent second rater on a $\geq 30\%$ subset with a Cohen’s κ agreement statistic is the first recommended extension.
- **Sample size.** $n = 12$ per model-language cell; rates are coarse and individual cells should not be over-interpreted (see Robustness).
- **Author-constructed stimuli.** Prompts were hand-authored for local relevance; re-running the protocol with a vetted, published red-teaming set (e.g. HarmBench [Mazeika et al.(2024)]) would improve reproducibility and external comparability.
- **Snapshot.** Results reflect specific model versions at one point in time and may change with updates.

- **Dual-use consideration.** This audit identifies exploitable cross-lingual safety gaps. To limit misuse we withhold the harmful prompt texts from publication, keep the working prompt set in a private repository, and recommend responsible disclosure of model-specific failures to the respective providers rather than public release. The categories studied (incitement, electoral disinformation, financial fraud, gender-based violence) are precisely those with real-world harm potential, which is why we treat the artefacts as sensitive.

Future work. Inter-rater reliability; expansion to more South African languages and larger prompt sets; a prototype of the in-language system-prompt mitigation with before/after measurement; and submission of the evaluation as a contribution to an open framework such as UK AISI Inspect.

6 Conclusion

Switching from English to a South African language sharply degrades the safety behaviour of widely-available open-weight LLMs — from 69% to 7% refusal — while a frontier model remains robust. The degradation is not uniform: it splits into capability failures that are safe only by accident and a genuine safety failure in which a fluent model complies with harmful requests it would refuse in English. Distinguishing the two is essential for honest measurement and for prioritising fixes. Closing this gap is both a technical problem (multilingual safety alignment) and a governance problem (whose norms, which languages), and is urgent for the millions of South Africans these systems already serve.

Code and Data

- Code and de-identified data (analysis script plus response-code results, containing no harmful prompt text): <https://github.com/Alch3mist-42/cross-lingual-safety-audit-public>
- Harmful prompt texts are intentionally *not* released. They are kept in a separate private repository and are available to reviewers on request under a responsible-disclosure arrangement (see §5).

Author Contributions

T.S. and S.N. jointly led the project, working in parallel across study design, the multilingual prompt set, model evaluations, response labelling, the analysis pipeline, and the writeup. F.M. contributed the Sepedi translation. All authors reviewed and approved the final report.

LLM Usage Statement

We used an AI assistant (Claude) to help scope the project, debug the evaluation environment, build the data-analysis and plotting code, run statistical tests, and draft sections of this report. All results were generated from our own hand-labelled data and independently checked; all numerical claims were verified against the analysis output, and the final manuscript was reviewed and edited by the team.

References

- [Bai et al.(2022)] Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.
- [Ganguli et al.(2022)] Ganguli, D. et al. (2022). Red Teaming Language Models to Reduce Harms. *arXiv:2209.07858*.
- [Hu et al.(2021)] Hu, E. et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.
- [Mazeika et al.(2024)] Mazeika, M. et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming. *arXiv:2402.04249*.
- [Mun et al.(2024)] Mun, J. et al. (2024). Particip-AI: A Framework for Anticipating Future AI Use Cases and Harms. *arXiv:2403.14791*.
- [Vidgen et al.(2024)] Vidgen, B. et al. (2024). STAR: A Sociotechnical Approach to Red Teaming. *arXiv:2406.11757*.
- [Yong et al.(2023)] Yong, Z.-X., Menghini, C., & Bach, S. (2023). Low-Resource Languages Jailbreak GPT-4. *arXiv:2310.02446*.
- [Zou et al.(2023)] Zou, A. et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv:2310.01405*.
- [Okolo(2024)] Okolo, C. (2024). African Democracy in the Era of Generative Disinformation. *arXiv:2407.07695*.
- [Segun et al.(2025)] Segun, S. et al. (2025). Toward an African Agenda for AI Safety. *arXiv:2508.13179*.
- [Faruna(2026)] Faruna, G. A. (2026). LSR: Linguistic Safety Robustness Benchmark for Low-Resource West African Languages. *arXiv:2603.19273*.
- [Abdullahi et al.(2026)] Abdullahi, T., Mgonzo, M., Oduwole, M., Okewunmi, P., Owodunni, A., Singh, R., & Eickhoff, C. (2026). UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages. *arXiv:2601.12696*.
- [Banerjee et al.(2026)] Banerjee, S., Hazra, R., & Mukherjee, A. (2026). Bridging the Multilingual Safety Divide: Efficient, Culturally-Aware Alignment for Global South Languages. *arXiv:2602.13867*.
- [CIPESA(2025)] CIPESA (2025). Safeguarding African Democracies Against AI-Driven Disinformation. *cipesa.org*.
- [AU(2024)] African Union (2024). Continental AI Strategy. *au.int*.

A Appendix

Per the dual-use considerations in §5, the harmful prompt texts are not reproduced here. The response coding scheme (Table 1), the per-cell summary statistics, and the analysis script constitute the reproducible artefacts. Reviewers requiring access to the prompt set for verification may contact the team under a responsible-disclosure arrangement.