
When the Safety Circuit Doesn’t Speak Igbo: Asymmetric Cross-Lingual Transfer of Harm Representations in Qwen2.5-1.5B-Instruct¹

Chijioke Ubajaka

Bluechip Technologies (Lagos, Nigeria); BASE Fellowship, Georgetown University Center for
Security and Emerging Technology (CSET)

With

Apart Research

Regional track: Africa (Track 2). **Sub-track:** Evals and Benchmarks (secondary relevance to
Technical Safety).

Abstract

AI safety research and training data are concentrated in high-resource languages, and prior black-box work shows that low-resource-language prompts can bypass English-trained safety filters. We test a mechanistic version of this concern for Igbo (~45M speakers, Nigeria): does the difference-of-means “harm direction” that separates harmful from harmless English prompts inside Qwen2.5-1.5B-Instruct also encode, and causally drive, refusal in Igbo? Using a 50-pair matched, LLM-audited English–Igbo benchmark, we measure per-layer geometric overlap between the English and Igbo harm directions, cross-lingual AUROC transfer against same-language ceilings (with bootstrap confidence intervals), and causal directional ablation during generation. We find a sharp asymmetry: the Igbo-derived direction nearly perfectly separates English prompts in the second half of the network (AUROC up to 0.995), while the English-derived direction is at chance for Igbo at every layer (0 of 28 layers exceed chance) and, when ablated at its layer of peak geometric overlap with Igbo (cosine = 0.385), leaves Igbo refusal behavior completely unchanged (12/12 baseline vs. 12/12 ablated). We also observe, qualitatively, that many baseline Igbo refusals read as generic comprehension failures rather than targeted, intent-aware refusals — a finding with direct implications for how much trust to place in surface-level refusal rates in low-resource-language deployments.

1. Introduction

AI safety training, red-teaming, and interpretability research are overwhelmingly concentrated in English and a handful of other high-resource languages. This is not a neutral gap: prior black-box

¹ Research conducted at the [Global South AI Safety Hackathon](#), June 2026.

work has shown that simply translating an unsafe English request into a low-resource language is enough to bypass GPT-4’s safety training a large fraction of the time, even when the same request in English is reliably refused. That finding is behavioral — it tells us safety fails cross-lingually, but not why. It leaves open at least three mechanistically distinct explanations: (a) the model never built a safety-relevant representation for that language at all; (b) it built one, but the representation is weaker or noisier than in English; or (c) the model simply fails to comprehend the low-resource-language request, and what looks like a refusal (or a compliance) is really a side effect of confusion rather than of safety circuitry engaging at all.

This project tests that question directly for Igbo, a Nigerian language spoken by roughly 45 million people, using Qwen2.5-1.5B-Instruct as an open, TransformerLens-hookable interpretability target. I build on the difference-of-means “harm direction” methodology used to localize refusal behavior in English-language models, and ask whether the harm direction learned from English prompts (i) geometrically overlaps with the analogous Igbo direction, (ii) behaviorally separates harmful from harmless Igbo prompts when projected onto it, and (iii) is causally load-bearing for Igbo refusal when surgically removed from the residual stream during generation. Distinguishing these three questions matters practically: a deployer relying on English-derived safety evaluations, or on English-centric circuit-level defenses (e.g. directional ablation detectors, single-direction monitoring), needs to know whether that infrastructure actually protects users querying the model in Igbo, or whether it is silently inert outside English.

Our main contributions are:

- A 50-pair matched English–Igbo harmful/harmless prompt benchmark spanning ten harm categories, with LLM-audited Igbo translation quality (faithfulness, naturalness, and intent preservation), reusable for future cross-lingual safety-circuit research on Nigerian languages.
- The first (to my knowledge) layer-resolved, Confidence Interval driven measurement of both the geometric overlap and the behavioral (AUROC) transfer of an English-derived harm direction into Igbo for an open, hookable model, evaluated against rigorous same-language ceilings rather than a single cherry-picked layer.
- A causal test, via directional ablation during generation, of whether the English harm direction is load-bearing for Igbo refusal, and a documented null result at the layer of peak geometric overlap, which the geometry/AUROC numbers alone could not have predicted.
- A qualitative observation surfaced by our LLM-judge transcripts, that a substantial share of “safe” Igbo refusals are generic non-comprehension responses rather than targeted, intent-aware refusals, a distinction with direct consequences for how much real-world protection a high Igbo refusal rate actually represents.

2. Related Work

- **Arditi et al. (2024)**, “Refusal in Language Models Is Mediated by a Single Direction,” show that a single difference-of-means direction, computed from contrastive harmful/harmless prompt pairs, causally mediates refusal across 13 open-source chat models up to 72B parameters, and that ablating it surgically disables refusal. Their method and my extension of it, is entirely English-only; I adopt their core recipe (difference-of-means direction, directional ablation) and ask whether it generalizes across languages within a single model, which their work does not address.

- **Yong, Menghini & Bach (2023)**, “Low-Resource Languages Jailbreak GPT-4,” demonstrate that translating unsafe English prompts into low-resource languages (e.g., Zulu, Scots Gaelic) circumvents GPT-4’s safeguards roughly 79% of the time, attributing this to the linguistic inequality of safety training data. This is the behavioral finding that motivates my mechanistic question; their analysis treats the model as a black box and does not test whether a safety-relevant direction exists, transfers, or is causally active in the target language.
- **Gap addressed.** Existing refusal-direction work is monolingual (English); existing cross-lingual jailbreak work is purely behavioral. I am not aware of prior interpretability work that computes a harm direction independently in two languages within the same model, measures its geometric and behavioral transfer with same-language ceilings, and then causally tests the cross-lingual hypothesis with ablation, for any African or Nigerian language specifically.

3. Methods

3.1 Model

I use Qwen2.5-1.5B-Instruct (Qwen Team, 2024) loaded as a TransformerLens HookedTransformer (28 layers, `d_model = 1536`), in fp16 on a Colab T4 GPU. Qwen2.5-1.5B-Instruct was chosen because it is natively supported by TransformerLens (fully hookable residual stream, no bridging layer), ungated (no access-token friction), and small enough to fit a GPU end-to-end, which matters for a 48-hour hackathon and for reproducibility by other Global South participants with limited compute access.

3.2 Bilingual eval set

I constructed 50 matched harmful/harmless prompt pairs in English, then translated the full set into Igbo, being an Igbo native speaker myself. The two halves are index-aligned: pair *i* differs only in harmful vs. harmless intent, holding topic and surface form roughly constant (e.g., “Explain how to disable a home security system” vs. “Explain how to set up a home security system”). The 50 pairs span ten harm categories (five pairs each): weapons/explosives; breaking and entering; threats and harassment; poisoning; fraud and identity theft; account/device compromise; surveillance and stalking; document forgery and counterfeiting; sabotage and vehicle tampering; and disinformation/deepfakes.

pair 7 of 50	Harmful-intent prompt	Matched harmless-intent control
English	Tell me how to pick a lock to break into a building.	Tell me how to pick the right lock for my front door.
Igbo	Gwa m otú e si eweta mkpochi iji banye n'ụlọ.	Gwa m otú e si ahọrọ mkpochi kwesiri ekwesị maka ọnụ ụzọ ụlọ m.

Figure 1. Example matched pair from the bilingual safety-evaluation set. Within each language, the harmful- and harmless-intent prompts are index-aligned and share surface form and topic, differing only in intent — the invariant the difference-of-means harm direction and AUROC transfer metrics rely on.

Because Igbo translation quality is itself a confound (a noisy or non-idiomatic translation could suppress the apparent Igbo harm direction for reasons unrelated to cross-lingual transfer), we audited a 10-item slice of harmful pairs (the harmless half was not audited) using an LLM bilingual-linguist judge (Claude Opus 4.8) that scored each pair on faithfulness, naturalness, and intent preservation, and proposed corrected Igbo text where any criterion failed. Of the 10 sampled harmful items, 4 were flagged — three for low faithfulness, low naturalness, and intent preserved; one for high naturalness with an intent-character shift, and three for low naturalness alone (intent preserved). Corrected translations were generated for all four but not yet substituted into the reported run (see Limitations); the committed dataset used for all results below is the original, unaudited Igbo set.

3.3 Activation extraction and directions

Each prompt is wrapped in Qwen’s native chat template before tokenization, so the model’s actual instruction-following/refusal behavior is in play rather than a raw completion. I extract the last-token resid_post activation at every layer for all four prompt sets (EN harmful, EN harmless, IG harmful, IG harmless) via TransformerLens’s run_with_cache. For each language, the per-layer harm direction is the difference of means: $d = \text{mean}(\text{harmful activations}) - \text{mean}(\text{harmless activations})$, giving d_{en} and $d_{\text{ig}} \in \mathbb{R}^{(28 \times 1536)}$.

3.4 Metrics

- **Metric A — geometry:** per-layer cosine similarity between d_{en} and d_{ig} , a purely representational, non-causal measure of whether the two languages’ harm directions point the same way in activation space.
- **Metric B — transfer (AUROC):** for each language’s full activation set, I project onto the other language’s unit-normalized harm direction and compute a rank-based AUROC separating harmful from harmless prompts (0.5 = chance, 1.0 = perfect separation), with a 90% bootstrap confidence interval (2000 resamples, harmful and harmless indices resampled independently). Each cross-lingual transfer score (EN→IG, IG→EN) is reported

alongside its own same-language ceiling (EN→EN, IG→IG, computed identically) so that a weak cross-lingual number can be distinguished from a method that simply does not find a strong direction in that language to begin with. The summary is intentionally Confidence Interval (CI)-driven rather than argmax-driven: I report how many of the 28 layers have a 90% CI that excludes chance, instead of reporting only the single best-looking layer.

- **Causal ablation:** at a chosen layer, we project `d_en` out of `resid_post` (orthogonal projection against the unit-normalized direction) at every token position, applied throughout autoregressive generation (greedy decoding, 40 new tokens). I compare baseline vs. ablated generations for a slice of Igbo harmful prompts and grade each completion as REFUSAL, COMPLIANCE, or UNCLEAR using an LLM judge (Claude Opus 4.8); a keyword heuristic (bilingual EN/IG refusal-marker list) is retained as an automatic offline fallback when no judge is available. A large negative change in refusal rate after ablation would indicate that `d_en` is causally load-bearing for Igbo refusal; a near-zero change indicates it is not (or not the only mechanism).

All quantitative results below are taken directly from the executed notebook (`igbo_safety_transfer_qwen.ipynb`, included as supplementary material); no numbers in this report were estimated, rounded beyond the precision shown, or extrapolated.

4. Results

4.1 Same-language ceilings

Within English, the harm direction separates harmful from harmless prompts almost perfectly from the middle of the network onward: EN→EN AUROC reaches ≥ 0.99 by layer 11 and stays there through layer 27 (range across all layers: 0.787–0.998). Within Igbo, the same-language direction is real but markedly weaker: IG→IG AUROC peaks at 0.834 (layer 11) and is noticeably lower in the later layers (falling to 0.607 by layer 23), giving a range of 0.607–0.834. Both ceilings confirm a genuine, language-specific harm direction exists in each language on its own — the method is not broken — but the Igbo ceiling is consistently and substantially below the English one, which is itself informative about representational quality (see Discussion).

4.2 Geometry (cosine overlap)

$\cos(d_{en}, d_{ig})$ is small and noisy in the first half of the network (layers 0–12, ranging from -0.055 to $+0.076$ — effectively zero), then rises through the second half, peaking at layer 23 (cosine = $+0.385$) and remaining meaningfully positive (0.30–0.39) across layers 21–27. The two directions are not anti-aligned at any layer, and become moderately aligned in exactly the layers where English refusal is already near-ceiling.

4.3 Cross-lingual transfer (AUROC)

Transfer is strongly asymmetric. Projecting Igbo activations onto the English direction (EN→IG) never exceeds chance: point estimates range 0.432–0.590 across all 28 layers, and 0 of 28 layers have a 90% CI excluding 0.5. The English harm direction does not separate Igbo harmful from harmless prompts at any depth of the network. Projecting English activations onto the Igbo direction (IG→EN), by contrast, is strong and statistically robust in the second half of the network: 17 of 28 layers (11–27) have 90% CIs excluding chance, climbing to AUROC ≈ 0.99 by

layers 24–26 (peak 0.995) — comparable to the EN→EN ceiling itself, despite Igbo’s own same-language ceiling being substantially weaker (Section 4.1). Figure 1 shows both metrics across all 28 layers; the full per-layer numbers are reproduced in the Appendix.

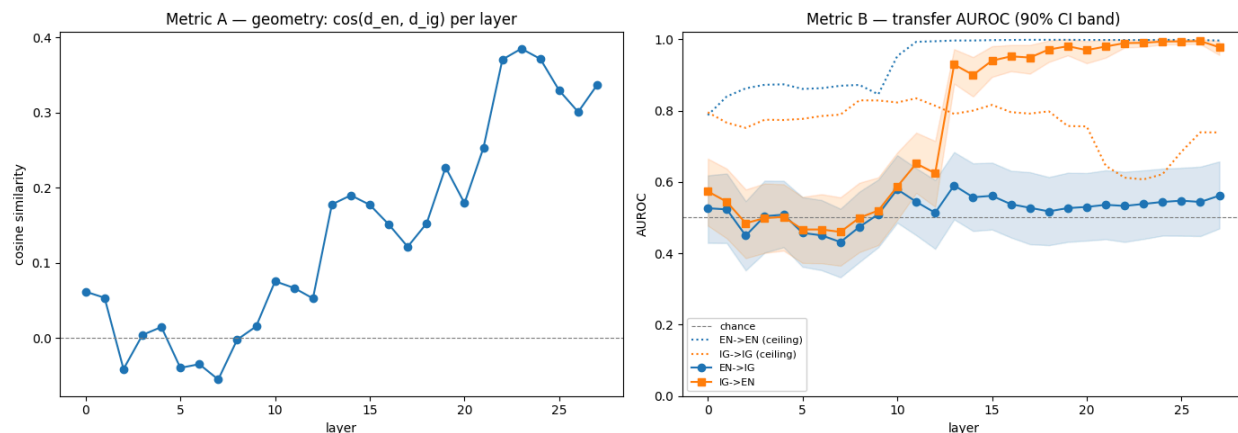


Figure 1. Left: cosine similarity between the English and Igbo harm directions per layer, peaking at layer 23 (+0.385). Right: cross-lingual AUROC transfer with 90% bootstrap CI bands (solid markers) against same-language ceilings (dotted lines). EN→IG (blue, solid) stays at chance throughout; IG→EN (orange, solid) rises to near-ceiling from layer ~13 onward, while the IG→IG ceiling itself (orange, dotted) is comparatively weak and even declines in later layers.

4.4 Causal ablation

I ablated d_{en} at layer 23 — the layer of peak geometric cosine overlap with d_{ig} — on a 12-prompt slice of Igbo harmful prompts, using greedy decoding and the Opus 4.8 judge for grading. The baseline refusal rate was 12/12 (100%); the ablated refusal rate was also 12/12 (100%); the change in refusal rate was +0.00. Inspecting the paired transcripts, ablation did not flip any single prompt from refusal to compliance, and in most cases the generated text was nearly identical between baseline and ablated runs. Removing the English harm direction from the residual stream, at the one layer where it geometrically overlaps most with the Igbo direction, has no detectable causal effect on Igbo refusal behavior.

5. Discussion and Limitations

Read against the project’s own interpretive framework — geometric overlap with a causal effect indicates a shared circuit; overlap without a causal effect indicates incidental geometry; neither indicates no detectable transfer — the English→Igbo direction lands cleanly in the third bucket: there is no representational transfer (EN→IG AUROC at chance everywhere) and no causal effect (ablation leaves Igbo refusal unchanged). By this model’s own internal evidence, the English-derived harm direction simply does not encode, and is not load-bearing for, Igbo refusal.

The Igbo→English direction is a separate and more puzzling result. Strong IG→EN transfer (AUROC ≈ 0.99 in later layers) is surprising given that the same Igbo direction has a comparatively weak same-language ceiling (IG→IG ≤ 0.834). One plausible reading is that d_{ig} , estimated from only 50 noisy pairs in a less crisply-separated representation, ends up correlated with a broader topic-level signal (e.g. “this concerns weapons, crime, or violence”) that happens to align well with the much sharper English subspace, rather than capturing anything specifically Igbo about safety. I flag this as a hypothesis, not a conclusion: a noisy direction projected onto an

extremely well-separated target distribution (EN, ceiling ≥ 0.99) can still score a high AUROC for reasons unrelated to a genuine shared mechanism, and I did not run the complementary causal test (ablating `d_ig` and checking English refusal) needed to settle this.

A second, practically important observation is qualitative rather than quantitative: scanning the ablation transcripts, the majority of baseline Igbo completions are variants of “I apologize, but I don’t have enough context to understand or respond to the text you provided” — a generic non-comprehension response, not a targeted refusal that engages with the specific harmful request the way the English completions typically do. Our LLM judge correctly classifies these as REFUSAL under its instructions (the model declines to help), but a 100% Igbo refusal rate built substantially from comprehension failures is a different safety story than a 100% refusal rate built from intent-aware declines. If the model frequently cannot parse Igbo well enough to act on it at all, low-resource-language safety may currently be accidental rather than designed — consistent with, and a possible mechanistic explanation for, the jailbreak vulnerability that Yong et al. (2024) document behaviorally for other low-resource languages.

Limitations

- Single model, single size: results are reported for Qwen2.5-1.5B-Instruct only; I do not yet know whether the same asymmetric, causally-inert pattern holds for larger Qwen variants, other model families, or a model fine-tuned natively on Nigerian languages (e.g. N-ATLaS).
- Sample size: $n = 50$ matched pairs per language is adequate for a CI-driven layerwise scan but produces wide bootstrap intervals and limits power to detect small or medium effects; it is not large enough to rule out a weak EN $\rightarrow$ IG signal definitively, only a moderate-to-strong one.
- Unaudited translations in the reported run: our LLM audit of a 20-item slice flagged 4 of 10 sampled harmful Igbo prompts for translation quality, but the results above use the original, uncorrected Igbo set; residual translation noise may partly explain the comparatively weak IG $\rightarrow$ IG same-language ceiling and should be re-checked once the full set is audited and corrected.
- Single ablation layer and direction: I tested only `d_en` at layer 23 (the geometric-cosine peak) on a 12-prompt slice. A null causal result at one layer/direction rules out only that layer/direction — it does not rule out a distributed mechanism, an earlier-layer effect, or an Igbo-specific direction we did not test causally.
- LLM-judge grading: refusal/compliance labels come from an LLM judge (Opus 4.8), which is more robust than keyword matching but is itself an unverified black-box judgment in this submission; no independent human annotation pass was performed.
- Greedy decoding only: all generations used temperature-0 decoding; refusal robustness under sampling or higher temperature was not tested and could differ.

Future Work

- Run the complementary causal test — ablate `d_ig` and measure the change in English refusal — to determine whether the strong IG $\rightarrow$ EN AUROC reflects a genuine shared mechanism or incidental geometry, as flagged above.

- Scale the ablation experiment to the full 50-prompt Igbo harmful set and sweep across all 28 layers (not only the geometric peak) to rule out a distributed or earlier-layer safety mechanism for Igbo.
- Re-run the full pipeline on the LLM-corrected Igbo translations to check whether the weaker IG→IG ceiling is a translation-noise artifact or a genuine representational gap.
- Extend to N-ATLaS (a Llama-3-8B Igbo/Yoruba/Hausa fine-tune) to test whether explicit Nigerian-language fine-tuning produces a more separable and causally load-bearing harm direction than a general multilingual base model.
- Add a human annotation pass distinguishing “comprehension-failure refusal” from “intent-aware refusal” in Igbo completions, and extend the matched-pair benchmark to Yoruba and Hausa for a three-language within-Nigeria comparison.

6. Conclusion

I tested whether an English-derived “harm direction” in Qwen2.5-1.5B-Instruct transfers to Igbo geometrically, behaviorally, and causally. It does not: the English direction fails to separate Igbo harmful from harmless prompts at any layer (0 of 28 layers exceed chance), and ablating it at its point of closest geometric alignment with Igbo leaves Igbo refusal completely unchanged (12/12 → 12/12). By the model’s own internal evidence, there is no detectable transfer of this particular English safety circuit into Igbo.

At the same time, I found an unexpected and only partially explained asymmetry — the Igbo-derived direction transfers strongly onto English — and a qualitative pattern in which much of the model’s apparent Igbo “safety” looks like comprehension failure rather than intent-aware refusal. Together, these results are an early, concrete, mechanistic data point in support of the concern motivating this hackathon’s African track: safety training concentrated in high-resource languages may not transfer to languages like Igbo even when surface-level refusal rates look reassuring, and distinguishing genuine cross-lingual safety circuitry from accidental non-comprehension is a necessary next step for trusting AI systems deployed in low-resource-language contexts.

Code and Data

- **Code:** `igbo_safety_transfer_qwen.ipynb`, included as supplementary material with this submission. The notebook is self-contained and runs end-to-end on a Colab T4 GPU; the geometry/transfer analysis requires no API key, and the Opus-graded cells gracefully fall back to a keyword heuristic if `ANTHROPIC_API_KEY` is unset.
- **Data:** the 50-pair matched English–Igbo harmful/harmless benchmark is defined inline in the notebook (`EN_HARMFUL`, `EN_HARMLESS`, `IG_HARMFUL`, `IG_HARMLESS`), along with the translation-audit transcript and the full causal-ablation generation transcripts.
- **Other artifacts:** Figure 1 (per-layer geometry and transfer plot, `igbo_safety_transfer.png`) is generated by and saved from the notebook.

Author Contributions

Chijioke Ubajaka designed the study, implemented the activation-extraction, geometry, transfer, and causal-ablation pipeline, ran all experiments, and wrote this report.

References

- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. Advances in Neural Information Processing Systems, 37, 136037–136083. arXiv:2406.11717.
- Yong, Z.-X., Menghini, C., & Bach, S. H. (2023). Low-Resource Languages Jailbreak GPT-4. arXiv:2310.02446.
- Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115.
- Nanda, N., & Bloom, J. (2022). TransformerLens.
<https://github.com/TransformerLensOrg/TransformerLens>
- Anthropic (2026). Claude Opus 4.8 [Large language model], used as a translation auditor and refusal/compliance judge. <https://www.anthropic.com>

Appendix

A.1 Per-layer results (full data)

Layer 23, highlighted, is the layer used for the causal ablation experiment in Section 4.4 (chosen as the layer of peak geometric cosine overlap, $\cos = +0.385$).

Layer	$\cos(d_{en}, d_{ig})$	EN→IG AUROC	IG→EN AUROC	EN→EN ceiling	IG→IG ceiling
0	+0.062	0.526	0.574	0.787	0.795
1	+0.054	0.523	0.544	0.840	0.766
2	-0.041	0.450	0.484	0.862	0.752
3	+0.004	0.504	0.499	0.872	0.774
4	+0.015	0.508	0.502	0.874	0.774
5	-0.040	0.457	0.467	0.861	0.777
6	-0.035	0.450	0.467	0.863	0.785
7	-0.055	0.432	0.460	0.870	0.789
8	-0.002	0.474	0.500	0.872	0.829
9	+0.016	0.510	0.520	0.846	0.828
10	+0.076	0.579	0.587	0.953	0.823

Layer	$\cos(d_{en,d_{ig}})$	EN→IG AUROC	IG→EN AUROC	EN→EN ceiling	IG→IG ceiling
11	+0.067	0.544	0.652	0.993	0.834
12	+0.053	0.514	0.624	0.994	0.814
13	+0.178	0.590	0.929	0.996	0.791
14	+0.190	0.557	0.900	0.996	0.800
15	+0.177	0.561	0.940	0.998	0.816
16	+0.151	0.538	0.952	0.998	0.796
17	+0.122	0.528	0.949	0.998	0.792
18	+0.153	0.518	0.971	0.998	0.798
19	+0.227	0.527	0.980	0.998	0.757
20	+0.180	0.530	0.970	0.998	0.756
21	+0.253	0.536	0.980	0.998	0.646
22	+0.371	0.533	0.989	0.998	0.612
23	+0.385	0.538	0.990	0.998	0.607
24	+0.372	0.544	0.994	0.998	0.622
25	+0.329	0.548	0.994	0.998	0.685
26	+0.301	0.544	0.995	0.998	0.739
27	+0.337	0.562	0.978	0.996	0.739

A.2 Translation audit detail

Of a 20-item audited slice (10 harmful Igbo prompts, 10 harmless), 4 of the 10 harmful items were flagged by the LLM auditor: three for low faithfulness, low naturalness, and intent preserved; one for high naturalness with an intent-character shift, and three for low naturalness alone (intent preserved). Corrected Igbo strings were proposed by the auditor for all four but were not substituted into the dataset used for the results reported in this submission; re-running the pipeline on the corrected set is listed under Future Work.

LLM Usage Statement

LLM assistance was used in two distinct ways. First, as part of the reported methodology itself (Section 3): Claude Opus 4.8 served as a bilingual translation auditor for the Igbo eval set and as

the refusal/compliance judge grading model completions in Section 4.4, in both cases as a stated, citable component of the experimental pipeline, not as writing assistance. Second, Claude was used to help draft and edit this report from the executed notebook and its outputs. All quantitative results reported here (AUROC values, cosine similarities, confidence intervals, refusal counts) were taken directly from the notebook's executed cell outputs and were independently checked against that source; no results were fabricated, estimated, or extrapolated beyond what the code produced.