Life of Research · Published 5 November 2026 · DOI 10.5281/zenodo.23965964

A Missing Slice Is a Missing Measurement: What Language-Level Audits of Safety Datasets Establish, Why the Multi-Turn Residual and Cross-Lingual Transfer Break the Step From Data Gaps to Model Behaviour, and the Contrasts That Would Price Native Data

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Model providers cite multilingual safety benchmarks as evidence that their models are safe for speakers of many languages, and a 2026 audit shows that such claims often dissolve when a single language is inspected: across 25 language slices of 20 safety datasets in Hausa, Swahili and French, it finds overstated provenance, unreported agreement, gated or unreleased data, recycled text, and no native coverage of self-harm or sexual content in either African language. The audit then reads the persistence of multi-turn jailbreaks in low-resource languages, after single-turn translation attacks were largely closed, as the outcome its data gaps would predict. In this analysis the author reviews and synthesises published results to show that the first finding is well supported and the second step is not. A thin or translated slice is first a defect of measurement: translated attack prompts that lose their meaning lower the attack success a benchmark records, so a poor slice tends to make a model look safer, and judges, comprehension failures and deflections distort low-resource scores in ways a slice count does not show. The multi-turn residual is not shown to be larger in Kiswahili than in English in the study the audit relies on: English multi-turn harmful-response rates there run from 52.7 to 83.6 percent, against 41.8 to 70.9 percent for Kiswahili, though the English condition replays conversations that had already succeeded in English and is favoured by construction, and the 2023 Swahili figure the audit cites (83.5 percent) was obtained with an English jailbreak template that already succeeded 72.1 percent of the time in English. Where per-language behaviour gaps do exist, the clearest predictors reported are properties of the model rather than of any safety dataset: how well the model aligns the language with English, which tracked how far English-only preference tuning reduced toxicity in one study, and, on the evidence of a training study that gave every language the same translated safety data, how much of the language pretraining supplied. A 61-configuration study finds 22 configurations least safe in English. The paper consolidates 30 published results in a table and a figure, separates four accounts of a per-language safety gap, gives a reading procedure for a per-language safety claim, and specifies the native-versus-translated, matched-data and difference-in-differences contrasts that would establish what native safety data buys.

Keywords: multilingual safety; low-resource languages; safety datasets; jailbreak evaluation; cross-lingual transfer; dataset documentation; evaluation validity; African languages

1. Introduction

Language models now serve users in hundreds of languages, and the evidence that they are safe for those users is usually a benchmark. Safety benchmarks have multiplied - a systematic review counted 144 open safety datasets and found non-English data notably scarce among them (Rottger et al., 2024) - and a family of multilingual resources now carries most of the weight of non-English safety claims: XSafety in ten languages (Wang et al., 2023a), MultiJail in nine plus English (Deng et al., 2023), RTP-LX in 28 (de Wynter et al., 2024), the Aya red-teaming set (Aakanksha et al., 2024), PolyGuardMix with 1.91 million samples in 17 languages (Kumar et al., 2025), and a growing set of regional benchmarks for African, South Asian and Southeast Asian languages (Abdullahi et al., 2026; Akinode et al., 2026; Pattnayak and Chowdhuri, 2026; Banerjee et al., 2026; Tasawong et al., 2025). A survey of nearly 300 safety papers from 2020 to 2024 found the field English-centric even relative to high-resource non-English languages (Yong et al., 2025).

A 2026 audit by Onuoha, Sunu and Sikiru asks what these collections contain for one language at a time (Onuoha et al., 2026). It examines 21 resources, which resolve under its counting rules into 20 datasets and 25 language slices: seven for Hausa, ten for Swahili and eight for French. Each slice is scored on a 31-field instrument derived from Datasheets for Datasets (Gebru et al., 2018) and checked against the source paper's own tables and the live release. The findings are specific. Provenance described as native traces, in two cases, to English material that was machine-translated. Six of 25 slices print a slice-level agreement statistic. One Hausa dataset's live release holds 300 items where its paper describes 986; another was announced and never shipped. A Swahili resource of 101,014 items is a merge of three corpora already in the audit. Native coverage of the audit's six harm categories is four categories for French, four for Swahili and two for Hausa, and neither

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

African language has native coverage of self-harm or sexual content. In the one pipeline that processed Hausa and Swahili identically, the Hausa policy material scored 66.37 on the pipeline's own translation-quality metric, under its acceptance threshold of 70, while the Swahili material scored 93.30 (Abdullahi et al., 2026; Onuoha et al., 2026).

This is careful work, and the present paper does not dispute any of it. The audit then takes a further step. It notes that single-turn, translation-based jailbreaks in low-resource African languages had largely stopped working by 2026 while multi-turn conversational jailbreaks still succeeded (Marx and Dunaiski, 2026), observes that native multi-turn safety data for Hausa and Swahili does not exist in its corpus, and argues that the asymmetry is "structurally consistent with where our audit finds training and evaluation data thinnest" (Onuoha et al., 2026). It states this as a correspondence rather than a demonstrated cause, and it names the reverse direction as possible. Its recommendations nonetheless lean on the step: providers should treat native multi-turn data as a separately sourced requirement, and funders should pay for native collection rather than translation pipelines.

In this analysis the author reviews and synthesises published results to show that a missing slice is first a missing measurement, and that the step from a slice-level data gap to a per-language behaviour gap is not supported by the evidence the audit cites and is complicated by evidence it does not. Three bodies of published results bear on the step. First, a translated or thin evaluation slice distorts what a benchmark records, and it distorts it in a predictable direction: attack prompts degraded by translation succeed less often, so the poorer the slice, the safer the model looks (Marx and Dunaiski, 2026; Yong et al., 2023). Second, the multi-turn residual the audit treats as a low-resource-language problem is at least as large in English in the very study that measured it, and multi-turn attacks defeat defences in English that hold against single-turn attacks (Li et al., 2024b; Russinovich et al., 2024). Third, where models do behave worse in a language, the clearest predictors reported are properties of the model's representation of that language - how closely its sentences align with English, whether its prompts reach the refusal direction learned in English - which bound what any language-specific safety dataset can buy (Li et al., 2024a; Wang et al., 2025; Shen et al., 2024; Oppong et al., 2026; Ivanova, 2026).

None of this makes native data unimportant. It changes what native data is for and how its value would be shown. The audit's strongest result - that the categories where local idiom carries the signal have no native items in two languages - is a result about what can currently be measured, and it is decisive for evaluation. Whether the same gap is what makes models less safe for Hausa and Swahili speakers is a separate empirical question that the published literature has not settled, and the contrasts that would settle it are cheap relative to the data collection the audit recommends.

The paper proceeds as follows. Section 2 fixes terms and states four accounts of a per-language safety gap. Section 3 reads the anchor audit at the resolution of its design. Sections 4 to 7 take the four accounts in turn: what a slice gap does to a measurement, what the evidence for native and multilingual safety training shows, what the transfer literature shows, and whether the multi-turn residual is language-specific. Section 8 consolidates the evidence in Table 1 and Figure 1. Section 9 states where the accounts disagree (Table 2). Section 10 gives a reading procedure (Algorithm 1). Sections 11 to 15 state the strongest objections, the studies that would settle the open part, what is not known, and the limitations.

2. Scope and terms

A **slice** is the part of a multilingual dataset that is in one language, following the anchor's usage. A **coverage gap** is a difference between what a collection claims for a language and what its slice for that language contains - in size, provenance, verification, access or harm categories. A **behaviour gap** is a difference between a model's safety behaviour in a language and in English on matched content. An **evaluation-validity gap** is a difference between what a safety score in a language measures and what it is taken to measure. The anchor documents coverage gaps directly. Its Section 7.3 links them to a behaviour gap. This paper argues that the link it can most securely support runs through evaluation validity.

Four accounts of why a model might be less safe in a language are in play in the literature, and they are not exclusive.

Account A1 (coverage): Safety training data in the language is thin, translated or missing in some harm categories or formats, and the model's safety behaviour in that language is worse because of it. Adding native data in the language would close the gap. This is the working assumption the anchor attributes to every dataset in its corpus.

Account A2 (transfer): Most safety behaviour is learned in English or a few high-resource languages and reaches other languages through shared representations. The gap in a language is set mainly by how well the model represents that language and maps it onto those shared representations, which pretraining determines. Language-specific safety data helps in proportion to how well the model can use it.

Account A3 (generic threat model): Part of what looks like a language gap is a gap in the defence that exists in every language - multi-turn escalation, role-play templates, shallow refusal - and would remain if language coverage were perfect.

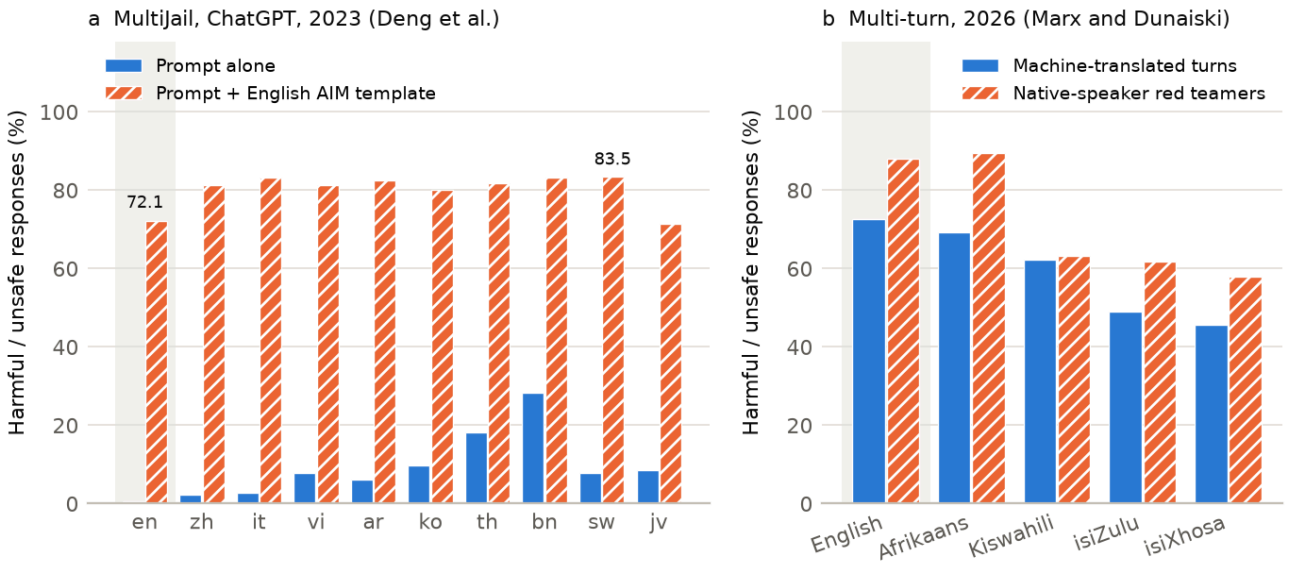


Figure 1. The jailbreak residual is large in English too, and English is at or near the top under multi-turn attack. (a) Unsafe-response rates of ChatGPT (GPT-3.5-turbo-0613) on MultiJail, from Deng et al. (2023), Table 1: the harmful prompt alone (blue) and the same prompt with the English AIM role-play template prepended (orange, hatched); the English prompt-alone value is 0.63 percent. The 83.5 percent Swahili figure cited in the audit is the templated condition, against 72.1 percent for templated English. (b) Mean multi-turn harmful-response rates from the automated-versus-human comparison in Marx and Dunaiski (2026), Table 4, for models tested in 2026: machine-translated turns (blue) and native-speaker red teamers (orange, hatched); English, shaded, is highest under automated attack and Kiswahili rests on one translator. All values are as published in the two cited studies.

Account A4 (measurement): Part of what looks like a language gap, or like its absence, is produced by the instrument: translation that changes a prompt's meaning, a judge that is unreliable in the language, a model that fails to understand and is scored as safe, an English reference that is itself the least safe condition.

The anchor's contribution is strongest under A4, and its recommendations depend on A1. The question this paper asks is how much of the per-language safety picture the published evidence assigns to each.

Two scope notes. The anchor's definition of safety covers content safety and jailbreak safety; agentic safety is a boundary case it sets aside, and so does this paper. And the evidence reviewed here is about models and benchmarks that are public; no paper reviewed reports what safety data any commercial provider used, a limitation the anchor states for itself and that applies equally to every inference here about why a deployed model behaves as it does.

3. The anchor, read at the resolution of its design

The audit belongs to a line of work that treats dataset documentation as a claim to be checked. Data statements (Bender and Friedman, 2018) and datasheets (Gebru et al., 2018) argued that a dataset's fitness for a use cannot be read from its headline description. The closest methodological precedent is an audit of 205 language-specific corpora in five web-crawled multilingual datasets, which found systematic problems concentrated in lower-resource corpora - at least 15 with no usable text, many mislabelled, and a substantial fraction in which fewer than half the sentences were of acceptable quality - and showed that such problems are detectable even by auditors who do not speak the language (Kreutzer et al., 2021). The resource disparities that make such audits necessary were documented across the world's languages by Joshi et al. (2020). The safety audit applies the same discipline to a narrower and higher-stakes kind of data.

The audit's unit is the slice and its method is verification against primary sources: a paper's methodology section and per-language tables, its repository, and the release itself. That method changed the record in four kinds of correction during the audit (a misassigned language; two provenance corrections; release-state corrections, one release smaller than described and one dataset announced but never shipped; and an authorship claim that applied only to English seed material), and the authors present those corrections as evidence that the audit does work that abstracts do not. The instrument and every per-slice judgement are released.

What the audit measures is the state of public resources. What it does not measure is any model. Its link to behaviour is drawn from two external sources: Deng et al. (2023), whom it reports as finding an unsafe-response rate of about 83 percent for Swahili jailbreak attempts against a 2023 model, and Marx and Dunaiski (2026), whom it reports as finding that single-turn translation attacks had been largely closed by 2026 while multi-turn attacks still produced harmful responses 42 to 71 percent

of the time depending on model and language. From these, its Section 7.3 argues that the attack surface with abundant native single-turn data has closed and the attack surface with only synthetic, machine-translated multi-turn data has not.

The audit is explicit about the limits of that argument. It describes the link as a structural correspondence, says it cannot see providers' training data, and notes that providers may have under-invested in multi-turn African-language data because the vulnerability had not been shown, rather than the vulnerability persisting because the data was thin. It also states, in the same section, that the multi-turn attacks "remain effective at rates comparable to English" (Onuoha et al., 2026). That clause is the point at which the present analysis begins. If a vulnerability is as large in English as in Kiswahili, the absence of Kiswahili multi-turn data cannot be what produces it, whatever else such data might do.

Three further features of the audit's design bear on the argument. Its tiering is contested by its own sources: by the CommonCrawl-share rule used in Deng et al. (2023) and Marx and Dunaiski (2026), Swahili is low-resource, while the audit treats it as mid-resource on the strength of curated-resource counts. Its six harm categories are adapted from a trust-and-safety taxonomy and mapped onto the datasets' own labels, which follow "at least ten mutually non-equivalent schemes" (Onuoha et al., 2026), by documented judgement, which the authors say they did not validate against native-speaker reading. And its severity grades are ordinal for five of six gap types, assigned by one annotator per slice. None of these weakens the coverage findings, which rest on counts and verifiable states. They do mean that the gradients the audit reports are gradients in a documented judgement, and that the one metric independent of that judgement is the within-pipeline translation score.

4. Account four first: what a slice gap does to a measurement

The most direct consequence of a thin, translated or unverified slice is that safety numbers computed on it mean less. The literature now documents five distinct ways in which that happens, and they do not all push in the same direction.

Translation that loses the attack lowers recorded attack success. Marx and Dunaiski (2026) machine-translated MultiJail prompts and multi-turn human jailbreak conversations into Afrikaans, Kiswahili, isiZulu and isiXhosa, filtered them hard on back-translation quality (a prompt had to exceed the mean plus one standard deviation on BLEU, METEOR and BERTScore, so that only 47 MultiJail and 55 multi-turn translations survived), and still found that isiZulu and isiXhosa produced the lowest harmful-response rates together with the most unrelated or nonsensical output. When native speakers refined the translations and adapted the conversation, harmful-response rates rose by 12.3 points for isiXhosa and 12.7 points for isiZulu (Table 4 of their paper; Figure 1b here). Their reported correlation between translation quality and jailbreak success is 0.87 to 0.92 depending on the metric. The direction matters: the worse the translation, the lower the measured vulnerability. Yong et al. (2023) show the same mechanism in the 2023 setting, where GPT-4 responses to Hmong and Guarani prompts were classed as unclear - neither a refusal nor on-topic engagement - 66.5 and 65.8 percent of the time. Shen et al. (2024) call this the relevance curse: models give more irrelevant responses to malicious prompts in lower-resource languages, which a harmful-or-safe score counts as safe.

The anchor's Section 6.2 draws on the same Marx and Dunaiski result to argue that a measured translation deficit predicts that safety evaluation in a language "will be less reliable" than its coverage claim implies. This paper agrees and adds the sign, for translation that destroys meaning. Machine translation as such does not lower recorded attack success: in Deng et al. (2023), "machine translation even yields a slightly higher rate of unsafe content, 11.15% on average, compared to human translation, which is 10.19%". But where translation loses the meaning of the attack, as for isiZulu and isiXhosa in Marx and Dunaiski (2026) and the six severe MultiJail mistranslations described below, the deficit is not noise around the true rate; it is a bias toward a lower rate. A Hausa policy slice under its own quality threshold, if used to test a model, would tend to report the model as safer in Hausa than it is.

Translation quality matters most in the tail. The size of the effect is not uniform. A study of 61 configurations from five closed-model families on MultiJail and XSafety, which fits a multi-group item-response model to separate a model's language-agnostic safety from prompt difficulty, a language-wide offset and a prompt-specific cross-lingual gap, finds that translation quality explains about one percent of the variance in the prompt-specific gap (Zhang et al., 2026a). It also finds that the outliers are mistranslations: among prompts with the largest gaps and the lowest native-speaker quality ratings it identifies six severe errors, such as a request to hot-wire a car rendered as a request to warm one. And it finds that 11 percent of the 945 prompt pairs rated by native speakers in Chinese, Thai and Bengali had poor or notable translation problems, in a dataset whose translations were produced and checked by native speakers (Deng et al., 2023). The two results are compatible. With careful human translation, quality is a minor term; with machine translation into a language the translator handles poorly, as for isiZulu and isiXhosa in Marx and Dunaiski (2026) or Hausa policy text in UbuntuGuard (Abdullahi et al., 2026), it can dominate.

Judges are least reliable where the slice is thinnest. Most safety benchmarks score responses with a model judge, often after machine-translating the response back

into English (Deng et al., 2023; Shen et al., 2024); Yong et al. (2023) used human annotation of back-translated responses instead. A study of five judge models over 25 languages finds average agreement across languages near a Fleiss kappa of 0.3, worst in low-resource languages, and not improved by multilingual training or scale (Fu and Liu, 2025). A review of 650 papers using model judges found 33 that address multilingual or low-resource settings, and in them a tendency to overtrust the judge and to rely on a single judge model (Dogruoz et al., 2026). Judges also prefer machine-translated text to human-authored text, most strongly in low-resource languages (Zhang et al., 2026c). The safety benchmarks for African languages report the same pattern directly: judge-human agreement falls in lower-resource languages and less supported scripts in TukaBench (Akinode et al., 2026), and the Kiswahili agentic evaluation the anchor discusses reports Kiswahili among the languages where judge and human assessments diverged most (Onuoha et al., 2026). One study of four African languages declined to report judge-scored attack success at all for this reason and measured the probability of emitting a refusal instead (Oppong et al., 2026).

Non-engagement is scored as safety. A model that does not understand a request does not comply with it. XL-SafetyBench, with 5,500 country-grounded items in ten country-language pairs annotated by two native speakers each, separates principled refusal from non-engagement and finds that among local models a low attack success rate goes with a high rate of neutral non-engagement ($r = -0.81$) (Choi et al., 2026). TukaBench adds a deflection category for the same reason (Akinode et al., 2026). The 61-configuration study checked this for the one family where it looked plausible and found that correcting for incomprehension changed jailbreak success by under one percent (Zhang et al., 2026a), so the effect is model-dependent: large for small or local models in low-resource languages, small for frontier models in languages they handle.

The benchmark items are not all harmful, and not all local. In the same study, a model judge rated 45 of MultiJail's 315 prompts as benign or mostly harmless (Zhang et al., 2026a), and native speakers judged 12 of 30 high-gap prompts with perfect translations to contain concepts foreign to the target language or a framing that lost its adversarial force in that culture - references to United States racial slurs or to institutions such as the FBI. Global MMLU makes the general point for capability benchmarks: 28 percent of MMLU's questions require culturally specific knowledge, and model rankings change depending on whether the full set or only the culturally sensitive subset is scored (Singh et al., 2024). A translated harm item can test a harm that does not exist in the target setting while missing the one that does. RTP-LX was transcreated partly to address this, and models scoring it agree poorly with human judges on holistic toxicity and struggle with subtle, context-dependent harm (de Wynter et al., 2024). Guards inherit the problem: a test suite of seven multilingual toxicity datasets across more than ten languages found existing guardrails ineffective on multilingual toxic input and not robust to jailbreak prompts (Yang et al., 2024).

Taken together, these results make the anchor's audit more important than its own framing suggests. A collection-level claim of coverage in a language is not just under-documented when its slice is thin; the safety number computed on that slice may be biased toward safety, scored by a judge that does not read the language well, and inflated by responses that are safe because they are unintelligible. The audit's demand that coverage claims be verified at slice level is a demand about evaluation validity, and on that the evidence is consistent.

The categories with no native coverage are where this matters most. Self-harm and sexual content are the categories the anchor identifies as carried by idiom and indirection, and they are the categories with no native items in Hausa or Swahili. Field evidence from Amharic shows what fails when such content is unmoderated: speakers searching YouTube for benign, popular queries reported unplanned exposure to policy-violating sexual content, and content creators appeared to exploit weak language technology to target migrant workers and other vulnerable speakers (Nigatu and Raji, 2024). Translated items cannot test whether a model or a guard recognises those forms, so their absence is an absence of measurement in exactly the categories where measurement would be hardest to replace.

5. Account one: what native and multilingual safety training buys

The anchor's recommendations assume that language-specific safety data improves safety in that language. Published training experiments support a weaker form of that assumption, and they say little about the native-versus-translated distinction the anchor's provenance findings turn on.

Adding data in the language helps, usually. Deng et al. (2023) generated multilingual safety data automatically and fine-tuned ChatGPT on it; the unsafe rate fell from 10.19 to 3.95 percent for plain multilingual prompts and from 80.92 to 60.00 percent with an English jailbreak template, with a measured trade-off against usefulness. RefusEU, a refusal-alignment set in 12 European languages built by translating Polish seeds with DeepL and generating items in each language, shows in controlled DPO experiments that English-only alignment does not secure safety in other languages even for the same harm categories, while multilingual training improves safety without loss on Global MMLU (Krasnodebska et al., 2026). The Aya alignment study builds synthetic red-teaming preference data in six

languages by having a multilingual model rephrase seed prompts that native speakers wrote in each language, uses NLLB to translate its general-purpose preference data, and reports that all its training regimes and data mixtures reduce harmful generations relative to the base model on a translated control set, and that its preferred regime does so on the native speakers' unaltered prompts, where it reports the global and local reductions (Aakanksha et al., 2024; NLLB Team et al., 2022). Simple prompting methods that evoke the model's safety knowledge reduced unsafe responses to non-English queries from 19.1 to 9.7 percent for ChatGPT on XSafety (Wang et al., 2023a).

But the gain is smallest where the gap is largest. Shen et al. (2024) fine-tuned LLaMa2-7B on HH-RLHF translated into high- and low-resource languages. The base model was about equally harmful in both groups (77.4 and 80.4 percent). Supervised fine-tuning on the translated data cut the harmful rate to 57.4 percent in high-resource languages but only to 70.6 percent in low-resource ones; a multilingual reward model trained on the translated preferences was 63.3 percent accurate in high-resource languages and 49.4 percent in low-resource ones, which is chance. The authors conclude that the bottleneck lies in pretraining. The language-specific data was present in the same quantity for every language; what differed was the model's ability to learn from it.

And more data in the language can make things worse. Distilling the refusals of a stronger teacher into three open models, using about 28,000 multilingual jailbreak prompts from XSafety, increased jailbreak success on MultiJail for every student, by up to 16.6 points; removing ambiguous boundary refusals from the training data mitigated or reversed the decline (Zhang et al., 2025). The lesson for the anchor's argument is that the content and form of the safety data matter at least as much as its language.

Local harms are partly, not wholly, transferable. The Aya study separates global harms, recognised as harmful across contexts, from local harms that need cultural or historical knowledge. Its preferred training regime reduced global harms by 64.2 percent and local harms by 59.7 percent relative to the base model; training on local examples alone reduced global harms by 77.8 percent, more than any other mixture; and in French, local harm rose while overall harm fell (Aakanksha et al., 2024). The authors offer three explanations for the small global-local difference, one of which is that the model judge is poorly calibrated on local examples. This is the most direct evidence that culturally grounded data has value beyond its language, and it is mixed: local data transfers to global harm well, global data transfers to local harm somewhat less well, and the measurement of local harm is itself in doubt.

Culturally grounded evaluation shows larger gaps than translated evaluation. Benchmarks built from culturally grounded or adapted items, not all of them native-authored, consistently find models weaker than English-centric or translated benchmarks suggest. TukaBench, covering seven African languages with four prompt constructions, finds that African-language prompts reduce refusal relative to English and that culturally adapted prompts produce the least refusal of all (Akinode et al., 2026); the anchor records its Hausa items as machine-translated with native post-editing. UbuntuGuard, whose adversarial queries were written by 155 domain experts as English seed material and whose Hausa and Swahili slices the anchor records as synthetic and machine-translated, concludes that English-centric benchmarks overestimate multilingual safety and that cross-lingual transfer covers only part of the gap (Abdullahi et al., 2026). SEA-SafeguardBench, human-verified across eight Southeast Asian languages, finds guards and models underperform relative to English (Tasawong et al., 2025); IndicSafe finds 12.8 percent cross-language agreement on translated variants of 6,000 culturally grounded prompts across 12 Indic languages (Pattnayak and Chowdhuri, 2026); SurakshaEval finds the weakest performance in native scripts (Banerjee et al., 2026). LSR, with 14 culturally grounded probes per language, reports refusal near 90 percent in English and 35 to 55 percent across Yoruba, Hausa, Igbo and Igala for one model (Faruna, 2026).

These are evaluation results. They show that culturally grounded items find vulnerabilities that translated items miss, which is A4 again - the slice determines what is measured. They do not show that training on such items would close the gaps they find, because none of them trains a model. The one study in this set that compares literal translations of an English safety set (Do-Not-Answer; Wang et al., 2023b) with culturally localised versions of the same harmful prompts, inside the model, found them nearly identical in representation (cosine 0.95 to 0.996) and nearly identical in the probability of emitting a refusal (1.05 and 1.15 percent across four African languages and three 7B-8B models) (Oppong et al., 2026). At that floor, localisation did not change behaviour because almost nothing triggered refusal.

What the training evidence does not contain is the comparison the anchor's provenance findings need: a model trained on native safety data in a language, compared with the same model trained on an equal quantity of translated safety data in that language, evaluated on native items. Every training study reviewed here uses translated or synthetic data, and all but one evaluate mainly on translated benchmarks. The exception is the Aya study (Aakanksha et al., 2024), whose synthetic safety data is rephrased from native-speaker seed prompts and which evaluates partly on prompts written by native speakers, including its global and local results, and partly on a translated control set; it

is the nearest published approach to a test of native provenance in training, but it is not a matched comparison. No study compares native with translated training data of equal size. The claim that native collection is required, rather than merely better for evaluation, is therefore untested.

6. Account two: transfer, and what sets it

The strongest counterweight to A1 is evidence that safety behaviour learned in one language reaches others without data in them, and that how far it reaches is a property of the model.

Some safety behaviour transfers almost completely. Preference tuning against toxicity with English data alone reduced the probability that mGPT-1.3B produced a toxic continuation from 46.8 to 3.9 percent across 17 languages, with similar results for BLOOM, Llama3 and Aya-23 (Li et al., 2024a). Reward models trained on preferences in one language align models in others, and a different-language reward model is sometimes better than a same-language one (Wu et al., 2024). Refusal in chat models is mediated by a single direction in activation space across 13 open models (Arditi et al., 2024), and that direction is shared across languages: a refusal vector extracted from English disables refusal in other safety-aligned languages when ablated, and vectors from German, Chinese or Thai do the same for the rest (Wang et al., 2025). Methods built on this - steering with English refusal directions (Stein et al., 2026), a consistency loss on multilingual prompt variants that needs no responses in the target language (Bu et al., 2026), self-distillation from the model's own high-resource behaviour using only multilingual queries (Qin et al., 2026), and fine-tuning a small set of neurons shared between high- and low-resource languages (Zhang et al., 2026b) - all report gains in lower-resource safety, the first three without response-level supervision in the target language. For Gemma 7B, English steering raised refusal of harmful prompts by 11 points on average across eight languages and by 17 in Bengali, at the cost of 13 points more refusal of pseudo-harmful prompts (Stein et al., 2026).

How far it transfers is predicted by representation alignment. In the detoxification study, the accuracy with which a sentence in a language can be matched to its English translation from the model's own representations correlated at $r = 0.73$ with how much English preference tuning reduced toxicity in that language, and the reduction was smaller for lower-resource languages (Li et al., 2024a). The refusal direction still acts on Yoruba when ablated, raising Gemma-2-9B's Yoruba compliance "from around 58% to over 90%", but the addition result is weaker there. In Yoruba, which the models handle poorly, compliance with harmful requests was 82.9 percent for Llama3.1-8B and 74.0 percent for Qwen2.5-7B (the source's Table 1; its prose gives 74.9), and adding the English refusal vector to Gemma-2-9B reduced Yoruba compliance only from 58 to 28 percent while reducing most languages to zero (Wang et al., 2025). Extending the test to Belarusian, Bashkir and Tajik on Qwen2.5-14B-Instruct finds compliance of 24.7, 87.9 and 92.0 percent respectively against 4.9 in English, and finds that directions extracted from the strongly misaligned languages fail to move refusal in any language, including each other (Ivanova, 2026). Belarusian, a lower-resource language closely related to Russian, is only partially misaligned (24.7 percent) and its refusal direction transfers like English's, which suggests that the outcome tracks the model's grip on the language rather than the language's own safety data.

For the anchor's languages, transfer is weak at the representation level. In Twi, Hausa, Amharic and Swahili, harmful prompts retain less than 10 percent of the English refusal signal for nearly every pairing of language and 7B-8B model; Hausa transfers worst, with 0.97 percent retention in Qwen2.5 and negative retention in two other models; the one exception is Swahili in Llama, at about 27 percent (Oppong et al., 2026). A probe-based study finds that apparent transfer of the harmfulness representation depends on the choice of negatives: with harmless prompts from an unrelated distribution, English-trained probes transfer near-perfectly (AUROC above 0.98), but with benign prompts that resemble harmful ones the mean AUROC drop for Qwen2.5-7B-Instruct rises from 0.003 in English to 0.276 in low-resource languages, and the drop persists after controlling for translation quality (Balani and Panda, 2026). And English is, on a logit-lens reading, close to the model's working concept space; Llama-2 decodes the correct next token in middle layers with higher probability for its English form than for the input language (Wendler et al., 2024). A language that does not map well onto that space does not inherit what was learned there.

Pretraining is the lever the transfer literature points to. Continued pretraining on multilingual translation data that includes low-resource languages slightly reduced the low-resource harmful rate after fine-tuning (from 70.6 to 68.2 percent in Shen et al., 2024). A study of matched English-Hausa clinical questions finds that deployable models of 4 to 9 billion parameters fall from a mean correctness of 1.57 in English to -0.03 in Hausa on a scale where -1 is actively harmful, while a frontier model moves only from 2.00 to 1.75 and gives no harmful answer in either language (Gabriel et al., 2026). The authors' reading is that the deficit belongs to the deployable tier rather than to the language or the clinical material. Capability benchmarks for African languages show the same split: on IrokoBench, the best open model reached 63 percent of GPT-4o's performance, and translating the test into English helped close the gap for larger English-centric models (Adelani et al., 2024). The trade-off between positive transfer and capacity

dilution in multilingual pretraining has been documented since XLM-R (Conneau et al., 2019).

The implication for the anchor is not that native safety data is useless. It is that, under A2, the marginal value of language-specific safety data depends on a property of the model - its representation of the language - that the data does not supply, and that a large part of a language's safety gap may close, or fail to close, for reasons that have nothing to do with its safety slice. Shen et al. (2024) is the cleanest demonstration: equal data, unequal gain. The anchor's A1 and this A2 make different predictions, and the difference is measurable; Section 9 states it.

7. Account three: is the multi-turn residual language-specific?

The anchor's link from data gaps to behaviour rests on one asymmetry: single-turn translation attacks closed, multi-turn attacks open. For that asymmetry to be evidence of a language-specific data gap, the multi-turn vulnerability has to be larger in the low-resource language than in English. The published numbers show the reverse or parity.

In the source study, English is at or near the top under multi-turn attack. Marx and Dunaiski (2026) report English multi-turn harmful-response rates from 52.7 percent (Claude 3.5 Haiku) to 83.6 percent (GPT-4o-mini), Afrikaans from 60.0 to 78.2 percent, and Kiswahili from 41.8 to 70.9 percent. In the study's comparison of automated and human red-teaming (its Table 4; the human sessions used GPT-4o, Gemini-2-flash and Grok), the mean harmful-response rate under automated attack is 72.73 percent in English and 62.34 percent in Kiswahili; under human red-teaming it is 88.05 percent in English and 63.33 percent in Kiswahili, though the Kiswahili figure rests on a single translator (Figure 1b). In their single-turn experiment, English shows the lowest harmful-response rate for every model, and once unrelated responses are grouped with safe ones, language does not significantly predict harmful output ($p = 0.089$). The pattern the anchor reads as a closed single-turn surface and an open multi-turn surface is, in this study, the same for English as for the African languages: single-turn attacks fail everywhere and multi-turn attacks succeed everywhere, with English at or near the top. Under human red-teaming in Table 4 only Afrikaans, at 89.39 percent, edges above English. The English condition is favoured by construction: it replays conversations from the MHJ dataset, which Marx and Dunaiski describe as containing "successful multi-turn jailbreak conversations collected through human red-teaming trials" (Marx and Dunaiski, 2026; Li et al., 2024b), while the other languages receive machine translations of the same conversations.

Multi-turn vulnerability is a property of current defences in English. The conversations Marx and Dunaiski translated come from a dataset of 537 multi-turn human jailbreaks (2,912 prompts) collected against English-language defences, which exceed 70 percent attack success on HarmBench against defences that report single-digit success under automated single-turn attack (Li et al., 2024b; Mazeika et al., 2024). Crescendo, which escalates gradually from a benign opening, succeeds across ChatGPT, Gemini, Llama and Anthropic models in English (Russinovich et al., 2024). One account of why points to alignment that adjusts a model's output distribution mainly over its first few tokens, leaving later tokens less constrained (Qi et al., 2024) and, on this paper's inference, later turns of a conversation as well; an older one distinguishes competing objectives from safety training that fails to generalise to domains where the model is capable (Wei et al., 2023). Neither account is about language, and the residual both predict exists where native data is most abundant.

The 2023 baseline was mostly the jailbreak template. The anchor cites an unsafe rate of about 83 percent for Swahili in Deng et al. (2023). That figure, 83.49 percent for ChatGPT, is the intentional scenario, in which an English role-play jailbreak (AIM) is prepended to the translated harmful prompt; the same template prepended to the English prompt produced 72.06 percent (Figure 1a). Without the template, the Swahili unsafe rate for the same model was 7.94 percent, lower than Thai (18.10) and Bengali (28.25). The decline the anchor attributes to providers closing a translation attack surface between 2023 and 2026 is a decline from a figure dominated by an English attack template.

Safety is largely one capacity, and English is not always its best language. Factor analysis of 1.9 million responses from 61 configurations finds safety on MultiJail essentially unidimensional - models that refuse one harm category refuse others - and 22 of the 61 configurations are least safe in English, including every Grok and DeepSeek configuration and two of five Claude configurations; for GPT models the least safe languages are often Swahili or Javanese (Zhang et al., 2026a). In a factorial study of Hausa and English with three frontier models, Claude 4.5 Opus was safer in Hausa (45.0 percent safe) than in English (36.7 percent), while past-tense framing bypassed defences (15.6 percent safe) and future-tense framing triggered conservative refusals (57.2 percent safe) (Said and Sani, 2025). Cross-language differences are real for some families and some prompts - in the same 61-configuration study, 34 percent of prompts are harder in Swahili than in English net of the language offset, against 44 percent for Javanese and 17 to 27 percent for the other seven languages (Zhang et al., 2026a) - but they are not a uniform low-resource penalty, and in the one study reviewed that measures both formats they are not larger under multi-turn than single-turn attack.

Language mixing is the attack that is language-specific. The attacks that do exploit language identity are mixtures: code-switching red-teaming elicited 46.7 percent more attacks than standard English attacks (Yoo et al., 2024), blending tokens across languages raised bypass rates to 67.23 percent on GPT-3.5 and 40.34 percent on GPT-4o (Song et al., 2024), and sandwiching a harmful question among benign ones in other languages succeeded against several frontier models (Upadhayay and Behzadan, 2024). A multilingual benchmark spanning 18 languages and four resource tiers finds a sharp change in safety between its second and third tiers and code-switching attacks that stay effective down to the lowest tier (Ifebi et al., 2026). Here the anchor's corpus is ahead of its high-resource comparison: code-switching is modelled in four of its African-language slices and none of its French ones. If native data has a distinctive role in defending against attacks, the evidence points to mixed-language and transliterated input rather than to multi-turn escalation.

The conclusion is narrow. The multi-turn residual is real, it affects low-resource-language users, and native multi-turn data would be needed to measure it properly in their languages. But its existence is not evidence that language-specific data gaps cause it, because it is not shown to be larger in Kiswahili than in English, even with an English baseline favoured by construction. The quantity that would bear on A1 is a difference in differences: the multi-turn gap between a language and English, minus the single-turn gap between them. On the published numbers that quantity is not positive for Kiswahili.

8. Consolidated evidence

Table 1 collects the results the argument rests on, one row per result, with its source. Values are those the source reports for the setting named; rows use different metrics, models and years and should not be compared across rows. The last column names the account each row bears on.

Three things stand out. The rows bearing on A1 are training studies on translated or synthetic data, evaluated mainly on translated benchmarks; the exception is the Aya row, whose synthetic data derives from native seed prompts and whose evaluation is partly on native-speaker prompts, including its global and local results, and partly on translated prompts, but none compares native with translated training data of equal size, so none isolates native provenance. The rows bearing on A3 show the multi-turn and template residual at full strength in English. And the rows bearing on A2 locate the largest per-language deficits in languages the model handles poorly, with a measurable predictor.

Figure 1 sets the two studies the audit relies on beside the English baseline each reports. In both, the condition with the most native data is not the condition with the least vulnerability.

9. Predictions that separate the accounts

The four accounts are not exclusive, and the evidence suggests that each holds for some languages, models and harm categories. What makes them useful is that they disagree about measurements that can be made. Table 2 states five.

The first column distinguishes A1 and A4 from A2 and A3 only partly, because both A1 and A4 predict that native evaluation finds more vulnerability. That is why native evaluation alone, however valuable, cannot establish that native training is needed. The second column is the decisive one for the anchor's recommendation: if a model trained on native safety data in a language improves, on native evaluation in that language, more than the same model trained on an equal quantity of translated data, A1's provenance claim is supported. If the two improve equally, and both improve less in languages the model represents poorly, A2 carries the gap. The third column is decisive for the anchor's Section 7.3: A1 predicts that the multi-turn gap between a low-resource language and English exceeds the single-turn gap; A3 predicts it does not. The fourth column is the cheapest: if a measure of how well a model aligns a language with English, such as bilingual sentence retrieval accuracy (Li et al., 2024a), predicts the per-language safety gap across languages better than any count of the language's safety data, A2 is the better first account.

10. A procedure for reading a per-language safety claim

Before a per-language safety number is cited as evidence that a model is safe in a language, or a per-language gap is cited as evidence that the language lacks safety data, it should pass the steps in Algorithm 1. The first four need only the benchmark and the model; the last two need training runs.

A per-language gap that survives steps 2 and 3, in a language with good alignment to English, is a candidate for A1, and step 5 tests it. A gap in a language with poor alignment is a candidate for A2, and the remedy may lie in pretraining or in representation-level methods rather than in a safety slice. A gap that is as large in English is A3. A number that fails step 2 is A4 and should not be cited either way.

11. The strongest case against this reading

"**The anchor called it a correspondence, not a cause; this paper attacks a claim it did not make.**" The anchor is careful, and Section 3 says so. But its contribution list describes the multi-turn persistence as evidence that the gaps are a live safety problem rather than a documentation problem, and two of its recommendations - treating native multi-turn data as a

Table 1. Published results on per-language safety coverage, transfer and measurement, with the accounts each informs (A1 coverage, A2 transfer, A3 generic threat, A4 measurement).

Source	Setting	Reported value	Bears on
Onuoha et al. 2026	25 slices, HA/SW/FR	Native harm categories 2 / 4 / 4 of 6	A1, A4
Onuoha et al. 2026	Agreement reporting	6 of 25 slices print slice-level IAA	A4
Abdullahi et al. 2026	One pipeline, HA vs SW	Policy MT 66.37 vs 93.30 (threshold 70)	A4
Deng et al. 2023	ChatGPT, prompt alone	EN 0.63, SW 7.94, BN 28.25 % unsafe	A2, A4
Deng et al. 2023	ChatGPT, English AIM template	EN 72.06, SW 83.49 % unsafe	A3
Deng et al. 2023	Self-Defense fine-tuning	80.92 to 60.00 % with template	A1
Yong et al. 2023	GPT-4, AdvBench	LRL combined 79.04 % bypass; HMN unclear 66.5 %	A2, A4
Marx and Dunaiski 2026	Multi-turn, tested 2026	EN 52.7-83.6, SW 41.8-70.9 % harmful	A3
Marx and Dunaiski 2026	Human vs MT turns, ZU/XH	+12.7 / +12.3 points with natives	A4
Marx and Dunaiski 2026	Single-turn, language effect	Not significant, $p = 0.089$	A3
Li et al. 2024b	Multi-turn human, English	Over 70 % ASR vs single-digit automated	A3
Zhang et al. 2026a	61 configurations, MultiJail	22 least safe in English	A3, A4
Zhang et al. 2026a	Translation quality	About 1 % of prompt-gap variance	A4
Zhang et al. 2026a	Native rating, 945 pairs	11 % with translation problems	A4
Said and Sani 2025	Claude 4.5 Opus, HA vs EN	45.0 vs 36.7 % safe	A3
Shen et al. 2024	LLaMa2-7B, translated SFT	Harmful 80.4 to 70.6 (low), 77.4 to 57.4 (high)	A1, A2
Shen et al. 2024	Translated reward model	49.4 % accurate, low-resource	A2, A4
Li et al. 2024a	English-only DPO, mGPT	Toxic 46.8 to 3.9 % over 17 languages	A2
Li et al. 2024a	Transfer predictor	Bilingual retrieval $r = 0.73$	A2
Wang et al. 2025	Yoruba compliance	82.9 (Llama3.1-8B), 74.0 (Qwen2.5-7B) %	A2
Ivanova 2026	Qwen2.5-14B compliance	EN 4.9, BE 24.7, BA 87.9, TG 92.0 %	A2
Oppong et al. 2026	4 African languages, 7-8B	Under 10 % English refusal signal kept	A2
Oppong et al. 2026	Literal vs localised prompts	Refusal 1.05 vs 1.15 %	A1, A2
Balani and Panda 2026	Hard-negative probes	AUROC drop 0.003 EN, 0.276 low tier	A2, A4
Gabriel et al. 2026	Clinical, EN vs HA	Deployable 1.57 to -0.03; frontier 2.00 to 1.75	A2
Aakanksha et al. 2024	DPO(SFT), 6 languages	Global -64.2 %, local -59.7 % harm	A1
Krasnodebska et al. 2026	DPO, 12 EU languages	English-only insufficient; multilingual helps	A1
Zhang et al. 2025	Refusal distillation	JSR up by as much as 16.6 points	A1
Stein et al. 2026	English steering, Gemma 7B	Refusal +11 points; pseudo-harmful +13	A2
Fu and Liu 2025	Judges, 25 languages	Fleiss kappa about 0.3	A4

Table 2. Predictions of each account for five measurements; "no change" means the per-language gap is unchanged.

Account	Native vs translated eval	Matched native vs MT training	Language minus English, multi-turn	Transfer predictor	English-only intervention
A1 Coverage	Native finds more	Native closes more	Larger than single-turn gap	Weak	Leaves gap
A2 Transfer	No change if model fails both	Similar gain, bounded by model	Same as single-turn gap	Strong	Closes gap in aligned languages
A3 Generic	No change	No change in residual	Near zero or negative	n/a	Residual stays in all languages
A4 Measurement	Native finds more, or less if MT fails	n/a	Depends on translation	n/a	n/a

separately sourced requirement, and funding native collection rather than translation pipelines - follow only if the correspondence is causal. The point here is not that the anchor overclaims its audit. It is that the correspondence it offers is with a residual that the source study shows in English, which removes the correspondence rather than merely leaving its direction open.

"Kiswahili is not Hausa, and Marx and Dunaiski did not test Hausa." This is correct, and it limits the multi-turn argument to the languages tested. For Hausa, the evidence available is single-turn: LSR's 14 probes for one model (Faruna, 2026), the factorial frontier study in which one model was safer in Hausa than in English (Said and Sani, 2025), the representation-level evidence that Hausa transfers worst of four African languages (Oppong et al., 2026), and the clinical study in which a frontier model answered Hausa questions competently (Gabriel et al., 2026). Taken together, these place the Hausa gap mainly with the model tier and its representation of Hausa, which is A2. No published study measures multi-turn safety in Hausa, so the anchor's correspondence for Hausa is neither supported nor refuted; it is unmeasured, and the step-3 contrast in Algorithm 1 is the way to measure it.

Algorithm 1. Reading a per-language safety claim for language L

```
Input: model M; benchmark slice S_L and its English source S_EN; harm categories C
step 1 – Characterise the slice.
   Record provenance (native, transcreated, human or machine translated, synthetic),
   size, release state and slice-level agreement for S_L; list categories in C
   with no native item. A category with no native item is unmeasured, not safe.
step 2 – Check that the instrument can see L.
   Report comprehension or deflection rate on S_L beside the harmful rate.
   Validate the judge on a native-labelled sample of M's L responses; report kappa.
   If translation quality is below the slice's own threshold, mark the L number as
   a likely underestimate of vulnerability.
step 3 – Compare like with like.
   Report the English rate on the same items and the same attack format.
   For each format f (single-turn, multi-turn, template, mixed-language):
      gap_f = rate_L,f − rate_EN,f
   Language-specific residual for format f: gap_f − gap_single
   A residual that is present in English is not evidence about L's data.
step 4 – Estimate transfer capacity.
   Measure M's alignment of L with English (e.g. bilingual retrieval accuracy,
   cosine of L and EN refusal directions). Low alignment predicts that
   English-derived safety will not reach L, whatever L's safety data.
step 5 – Price native data (training run).
   Train M on native S_L-train vs equal-size translated S_L-train; evaluate on
   native held-out items in L and on English. Report the difference.
step 6 – Price transfer (training run).
   Apply an English-only or language-agnostic intervention; report the change in L.
Output: for L, whether the reported number is valid (steps 1–2), whether any gap is
   language-specific (step 3), and whether native data, transfer or the generic defence
   is the lever (steps 4–6).
```

"**Lower Kiswahili rates under multi-turn attack reflect translation loss, so the true Kiswahili vulnerability may exceed English.**" That is possible, and Section 4 is the reason to take it seriously. But it cuts against the anchor's inference, not for it. If the Kiswahili rate is depressed by poor translation, the measured multi-turn rate in Kiswahili is not evidence of an open low-resource attack surface; it is evidence of an instrument that cannot see the language well. The human red-teaming condition, which removes most translation loss, gives 63.33 percent for Kiswahili against 88.05 percent for English, with the caveat that it rests on one Kiswahili translator.

"**Transfer methods are measured on translated benchmarks, so their success is an artefact of the same slices the audit criticises.**" This objection is right and applies symmetrically. The evidence for A2 and the evidence for A1 are both measured mostly on translated items, and Section 5's observation that culturally grounded items find larger gaps applies to transfer methods as much as to anything else. That is why Table 2 makes native evaluation a prerequisite for every decisive contrast rather than a test of any one account. It does not rescue A1, because, apart from one study that trains on synthetic data derived from native seed prompts and evaluates partly on native-speaker prompts, including its global and local results, and partly on translated prompts (Aakanksha et al., 2024), the training evidence for A1 is no more native than the evidence for A2, and that study does not compare native with translated training data.

"**Even if native data does not close behaviour gaps, it is the only way to measure them, so the recommendations stand.**" This paper agrees with the conclusion and disputes only the reason. Native items in the categories the audit finds empty - self-harm and sexual content in Hausa and Swahili, and native multi-turn conversations - are justified as measurement, because without them a provider cannot tell whether its model is safe for those users. Framing them as training data that will close a gap commits funders to an outcome that the evidence has not shown and that A2 predicts may not follow for languages the model represents poorly.

"**Unidimensional safety and English reversals come from closed commercial models and one benchmark.**" The 61-configuration study does rest on MultiJail and XSafety and on five closed families, and finds unidimensionality strong on MultiJail but borderline on XSafety (Zhang et al., 2026a). Its English reversal concentrates in particular families. The claim made here from it is narrow: an English baseline is not a safe floor against which every other language is measured, and per-language differences are family- and prompt-specific rather than a uniform penalty by resource level.

12. Studies that would settle the open part

Study 1: native versus translated evaluation on the same model. For one model and one language, build native items and translated items for the same harm categories and attack formats, have the same native annotators judge responses to both, and report harmful, refused and deflected rates for each. This isolates A4's provenance effect on measurement. TukaBench's four constructions (Akinode et al., 2026) and LoDNA's literal and localised pairs (Oppong et al., 2026) are close to this design and could be extended with human judgement

and multi-turn items.

Study 2: matched native versus translated safety training. Fine-tune one open model on native safety data in a language and, separately, on an equal quantity of translated data in the same language, for the same categories, and evaluate on held-out native items and on English. Repeat for a language the model represents well and one it represents poorly. A1 predicts a native advantage in both; A2 predicts similar gains bounded by the model's representation of the language. This is the comparison the anchor's provenance findings need and no published study reports.

Study 3: the multi-turn difference in differences. For each of several languages and models, run single-turn and multi-turn attacks with native red teamers in the language and in English on matched objectives, and report the multi-turn gap between the language and English minus the single-turn gap. A positive value is evidence for a language-specific multi-turn vulnerability; a value near zero assigns the residual to A3. The Marx and Dunaiski (2026) protocol, with more than one translator per language and with Hausa added, would serve.

Study 4: what predicts the per-language gap. Across 20 or more languages and several open models, regress the per-language safety gap, measured with native judgement, on the model's bilingual retrieval accuracy for the language, on the cosine between its refusal direction in the language and in English, on tokenizer fertility, and on an audited count of the language's safety data. If representation measures explain the gap and data counts add little, A2 is the first account; if data counts add substantially once representation is held fixed, A1 gains support.

Study 5: native evaluation for the empty categories. Build native self-harm and sexual-content items in Hausa and Swahili with native annotators and slice-level agreement, and evaluate deployed models and guards on them. This study is justified under every account, since it measures a category that is currently unmeasured, and it would show whether the absence of native items hides a behaviour gap or merely an evaluation gap.

13. What is not known

It is not known whether training on native safety data in a language improves safety in that language more than training on an equal quantity of translated data. It is not known whether multi-turn attacks are more effective in Hausa than in English, because no study has measured multi-turn safety in Hausa. It is not known how much of the per-language safety gap, in any language, is explained by representation alignment rather than by safety-data coverage, because no study has measured both for the same models. It is not known what safety data any commercial provider has used in any language. It is not known how models behave on native self-harm or sexual-content items in Hausa or Swahili, because such items do not exist in the published record. And it is not known how far the judges used to score low-resource safety agree with native annotators for the harm categories the anchor audits, beyond the general finding that agreement falls in lower-resource languages.

14. Limitations

The paper relies on published results and inherits their limits. The multi-turn evidence comes from one study with small filtered samples (47 single-turn and 55 multi-turn items after translation filtering) and a single Kiswahili translator, and that study's abstract and conclusion attribute some range endpoints to different models, and its appendix Table 6 gives per-model human red-teaming rates whose means differ from its Table 4 (about 79.5 rather than 88.05 percent for English, about 78.8 rather than 61.74 for isiZulu), so model-level attributions here are avoided where the source is inconsistent and the human red-teaming figures used here are those of Table 4. The 61-configuration study uses closed models, two benchmarks and model judges, though it reports native-speaker checks. The representation-level studies use 7B to 14B open models, and frontier models may transfer differently; the clinical study suggests they do. The training studies reviewed use translated or synthetic data, so the paper's claim that native provenance is untested for training is a claim about the absence of a study, which a single new paper could change. The four accounts are a framing chosen for this analysis, and other decompositions are possible; the item-response decomposition of Zhang et al. (2026a) is one, and maps imperfectly onto them. Of the paper's results, the reader should weigh most heavily the sign of the bias from meaning-destroying translation and the English baselines in Figure 1, since both are read directly from the sources the anchor cites; the transfer account less, because its strongest evidence comes from small open models; and the predictions in Table 2 least, since none has been tested.

15. Conclusion

A slice-level audit of multilingual safety datasets shows that claims of coverage in Hausa, Swahili and French rest on less than they appear to: translated material described as native, unmeasured agreement, unreleased data, recycled text, and no native items at all for self-harm and sexual content in the two African languages. That finding is about what can be measured, and the published evidence on measurement makes it stronger than its authors state: translations that lose an attack's meaning bias attack benchmarks toward reporting safety, judges are least reliable where slices are thinnest, and non-engagement counts as refusal. The further step, from those gaps to the persistence of multi-turn jailbreaks in low-resource languages, does not hold on the evidence cited for it. In the study the audit relies on,

English multi-turn rates, on conversations that had already succeeded in English, are the highest on average of the five languages tested under automated attack and, in the study's Table 4, within two points of the highest under human red-teaming; the 2023 Swahili baseline is the effect of an English jailbreak template that succeeded nearly as often in English; and a larger study finds more than a third of model configurations least safe in English. Where models are less safe in a language, the clearest predictors are the model's alignment of that language with English and the share of the language in pretraining, which set how far safety learned in English reaches. Native safety data is therefore justified first as measurement, and its value as training data is a testable proposition: a matched comparison of native and translated training, evaluated natively, and a multi-turn difference in differences against English, would establish what it buys.

References

Aakanksha, Ahmadian, A., Ermis, B., Goldfarb-Tarrant, S., Kreutzer, J., Fadaee, M., et al. (2024). The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm. arXiv:2406.18682. doi:10.48550/arXiv.2406.18682.

Abdullahi, T., Mgonzo, M., Oduwole, M., Okewunmi, P., Owodunni, A., Singh, R., et al. (2026). UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages. arXiv:2601.12696. doi:10.48550/arXiv.2601.12696.

Adelani, D. I., Ojo, J., Azime, I. A., Zhuang, J. Y., Alabi, J. O., He, X., et al. (2024). IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. arXiv:2406.03368. doi:10.48550/arXiv.2406.03368.

Akinode, V., Li, S., Hamidouche, W., Zamir, W., Becker-Reshef, I., & Adelani, D. I. (2026). TukaBench: A Culturally Grounded Jailbreak Benchmark for African Languages. arXiv:2606.01322. doi:10.48550/arXiv.2606.01322.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717.

Balani, P., & Panda, S. (2026). Hard Negatives Reveal What Easy Negatives Hide: Cross-Lingual Harmfulness Representations Degrade with Resource Tier Under Hard Negatives. arXiv:2609.27758. doi:10.48550/arXiv.2609.27758.

Banerjee, D., Kavitha, K. R., Borah, A., Han, X., Wang, Y., Krishnamurthy, P., et al. (2026). SurakshaEval: An Indic Safety Benchmark for Multilingual LLMs. arXiv:2608.07862. doi:10.48550/arXiv.2608.07862.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6, 587-604. doi:10.1162/tacl_a_00041.

Bu, Y., Liu, X., Ren, Z., Yang, Y., & Dai, J. (2026). Align Once, Benefit Multilingually: Enforcing Multilingual Consistency for LLM Safety Alignment. arXiv:2602.16660. doi:10.48550/arXiv.2602.16660.

Choi, D., Kim, E., Noh, J., Seo, S., Kim, E., Oh, M., et al. (2026). XL-SafetyBench: A Country-Grounded Cross-Cultural Benchmark for LLM Safety and Cultural Sensitivity. arXiv:2605.05662. doi:10.48550/arXiv.2605.05662.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., et al. (2019). Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116. doi:10.48550/arXiv.1911.02116.

de Wynter, A., Watts, I., Wongsangaroonsri, T., Zhang, M., Farra, N., Altintoprak, N. E., et al. (2024). RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios? Proceedings of the AAAI Conference on Artificial Intelligence, 39(27). doi:10.1609/aaai.v39i27.35011. arXiv:2404.14397. doi:10.48550/arXiv.2404.14397.

Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2023). Multilingual Jailbreak Challenges in Large Language Models. ICLR 2024. arXiv:2310.06474. doi:10.48550/arXiv.2310.06474.

Dogruoz, A. S., Liao, X., Blaschke, V., Prange, J., Li, S., & Adelani, D. I. (2026). Challenges and Recommendations for LLM-as-a-Judge in Multilingual Settings and for Low-Resource Languages. arXiv:2607.02235. doi:10.48550/arXiv.2607.02235.

Faruna, G. A. (2026). LSR: Linguistic Safety Robustness Benchmark for Low-Resource West African Languages. arXiv:2603.19273. doi:10.48550/arXiv.2603.19273.

Fu, X., & Liu, W. (2025). How Reliable is Multilingual LLM-as-a-Judge? arXiv:2505.12201. doi:10.48550/arXiv.2505.12201.

Gabriel, A. O., Olawuyi, D., Ajayi, T., & Aderemi, T. (2026). Safety That Does Not Transfer: Cross-Lingual Clinical Correctness Drift in Deployable Medical Language Models. arXiv:2607.17270. doi:10.48550/arXiv.2607.17270.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume, H., et al. (2018). Datasheets for Datasets. arXiv:1803.09010. doi:10.48550/arXiv.1803.09010.

Ifebi, C., Kong, B., & Mehrotra, A. (2026). Minionese: Comprehensive Benchmark and Mechanistic Study of Multilingual LLM Safety. arXiv:2607.10112. doi:10.48550/arXiv.2607.10112.

Ivanova, A. (2026). Is Refusal Direction Universal across Low-Resourced Languages? arXiv:2610.28897. doi:10.48550/arXiv.2610.28897.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. arXiv:2004.09095. doi:10.48550/arXiv.2004.09095.

Krasnodebska, A., Kusa, W., & Lipani, A. (2026). Multilingual Refusal Alignment for Safer Large Language Models. arXiv:2606.07535. doi:10.48550/arXiv.2606.07535.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., et al. (2021). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Transactions of the Association for Computational Linguistics, 10, 50-72. doi:10.1162/tacl_a_00447. arXiv:2103.12028. doi:10.48550/arXiv.2103.12028.

Kumar, P., Jain, D., Yerukola, A., Jiang, L., Beniwal, H., Hartvigsen, T., et al. (2025). PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages. arXiv:2504.04377. doi:10.48550/arXiv.2504.04377.

Li, N., Han, Z., Steneker, I., Primack, W., Goodside, R., Zhang, H., et al. (2024b). LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet. arXiv:2408.15221. doi:10.48550/arXiv.2408.15221.

Li, X., Yong, Z., & Bach, S. H. (2024a). Preference Tuning For Toxicity Mitigation Generalizes Across Languages. arXiv:2406.16235. doi:10.48550/arXiv.2406.16235.

Marx, D., & Dunaiski, M. (2026). Multilingual jailbreaking of LLMs using low-resource languages. arXiv:2605.18239. doi:10.48550/arXiv.2605.18239.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv:2402.04249. doi:10.48550/arXiv.2402.04249.

Nigatu, H. H., & Raji, I. D. (2024). "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. arXiv:2405.16656. doi:10.48550/arXiv.2405.16656.

NLLB Team, Costa-jussa, M. R., Cross, J., Celebi, O., Elbayad, M., Heafield, K., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672. doi:10.48550/arXiv.2207.04672.

Onuoha, C. P., Sunu, B. E., & Sikiru, R. (2026). Language-Specific Gaps in AI Safety Training Datasets. arXiv:2608.13695. doi:10.48550/arXiv.2608.13695.

Oppong, A., Sahil, P. S., Belay, T. D., Mukhtar, M. I., Abdu, E. A., Abdullahi, T., et al. (2026). The Illusion of Cross-Lingual Safety in Low-Resource Languages. arXiv:2608.11146. doi:10.48550/arXiv.2608.11146.

Pattnayak, P., & Chowdhuri, S. (2026). IndicSafe: A Benchmark for Evaluating Multilingual LLM Safety in South Asia. arXiv:2603.17915. doi:10.48550/arXiv.2603.17915.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., et al. (2024). Safety Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv:2406.05946. doi:10.48550/arXiv.2406.05946.

Qin, R., Wang, Q., Liu, D., Li, Q., Wei, Z., & Shen, W. (2026). Multilingual Safety Alignment via Self-Distillation. arXiv:2605.02971. doi:10.48550/arXiv.2605.02971.

Rottger, P., Pernisi, F., Vidgen, B., & Hovy, D. (2024). SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety. arXiv:2404.05399. doi:10.48550/arXiv.2404.05399.

Russinovich, M., Salem, A., & Eldan, R. (2024). Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. arXiv:2404.01833. doi:10.48550/arXiv.2404.01833.

Said, M. A., & Sani, M. S. (2025). Safe in the Future, Dangerous in the Past: Dissecting Temporal and Linguistic Vulnerabilities in LLMs. arXiv:2512.24556. doi:10.48550/arXiv.2512.24556.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., et al. (2024). The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts. arXiv:2401.13136. doi:10.48550/arXiv.2401.13136.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., et al. (2024). Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. arXiv:2412.03304. doi:10.48550/arXiv.2412.03304.

Song, J., Huang, Y., Zhou, Z., & Ma, L. (2024). Multilingual Blending: LLM Safety Alignment Evaluation with Language Mixture. arXiv:2407.07342. doi:10.48550/arXiv.2407.07342.

Stein, E. V., Meier, D., Ruas, T., Wahle, J. P., & Gipp, B. (2026). BabelSteering: Multilingual Safety Alignment via English Steering Vectors. arXiv:2608.16577. doi:10.48550/arXiv.2608.16577.

Tasawong, P., Ngui, J. G., Aji, A. F., Cohn, T., & Limkonchotiwat, P. (2025). SEA-SafeguardBench: Evaluating AI Safety in SEA Languages and Cultures. arXiv:2512.05501. doi:10.48550/arXiv.2512.05501.

Upadhayay, B., & Behzadan, V. (2024). Sandwich attack: Multi-language Mixture Adaptive Attack on LLMs. Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024). doi:10.18653/v1/2024.trustnlp-1.18. arXiv:2404.07242. doi:10.48550/arXiv.2404.07242.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J., Jiao, W., et al. (2023a). All Languages Matter: On the Multilingual Safety of Large Language Models. arXiv:2310.00905. doi:10.48550/arXiv.2310.00905.

Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2023b). Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. arXiv:2308.13387. doi:10.48550/arXiv.2308.13387.

Wang, X., Wang, M., Liu, Y., Schutze, H., & Plank, B. (2025). Refusal Direction is Universal Across Safety-Aligned Languages. arXiv:2505.17306. doi:10.48550/arXiv.2505.17306.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? arXiv:2307.02483. doi:10.48550/arXiv.2307.02483.

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do Llamas Work in English? On the Latent Language of Multilingual Transformers. arXiv:2402.10588. doi:10.48550/arXiv.2402.10588.

Wu, Z., Balashankar, A., Kim, Y., Eisenstein, J., & Beirami, A. (2024). Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment. arXiv:2404.12318. doi:10.48550/arXiv.2404.12318.

Yang, Y., Dan, S., Roth, D., & Lee, I. (2024). Benchmarking LLM Guardrails in Handling Multilingual Toxicity. arXiv:2410.22153. doi:10.48550/arXiv.2410.22153.

Yong, Z., Ermis, B., Fadaee, M., Bach, S. H., & Kreutzer, J. (2025). The State of Multilingual LLM Safety Research: From Measuring the Language Gap to Mitigating It. arXiv:2505.24119. doi:10.48550/arXiv.2505.24119.

Yong, Z., Menghini, C., & Bach, S. H. (2023). Low-Resource Languages Jailbreak GPT-4. arXiv:2310.02446. doi:10.48550/arXiv.2310.02446.

Yoo, H., Yang, Y., & Lee, H. (2024). Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding. arXiv:2406.15481. doi:10.48550/arXiv.2406.15481.

Zhang, H., Chen, K., Bai, X., Pan, Y., Xiang, Y., Wang, J., et al. (2026c). Mitigating Translationese Bias in Multilingual LLM-as-a-Judge via Disentangled Information Bottleneck. arXiv:2603.10351. doi:10.48550/arXiv.2603.10351.

Zhang, M., Liu, D., Zhang, K., Franco, J., & Liu, H. (2025). Response-Based Knowledge Distillation for Multilingual Jailbreak Prevention Unwittingly Compromises Safety. arXiv:2602.11157. doi:10.48550/arXiv.2602.11157.

Zhang, M., Patel, A., Truong, S. T., & Koyejo, S. (2026a). Why Do Safety Guardrails Degrade Across Languages? arXiv:2605.17173. doi:10.48550/arXiv.2605.17173.

Zhang, X., Xie, C., Zhu, L., Yang, Y., Zhao, W., Cheng, Z., et al. (2026b). Who Transfers Safety? Identifying and Targeting Cross-Lingual Shared Safety Neurons. arXiv:2602.01283. doi:10.48550/arXiv.2602.01283.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23965964