Aligning LLMs Beyond the West: Culturally Generalizable Safety for Women's Cancer Advice in Lingala

Anonymous Author(s)
Anonymous Institution
anonymous@example.com

Abstract

Large language models (LLMs) are increasingly used for health information seeking, but alignment techniques encode predominantly Western norms: patient autonomy, direct communication, and biomedical exclusivity. When deployed in Central African contexts, these models fail catastrophically. Lingala, spoken by over 45 million people, lacks any evaluated LLM alignment dataset for health. Women's cancers (breast, cervical, ovarian) are particularly critical. In DRC, cervical cancer incidence ranges from 31.9 to 33.1 per 100,000 women, with an annual mortality rate of 23.7 per 100,000. For breast cancer, 70-80% of women present at advanced stages, largely due to cultural taboos and miscommunication.

We present the first culturally grounded alignment dataset for women's cancer questions in Lingala (622 examples across 16 categories), and a cross-cultural reward modeling method that generalizes across Lingala and French. Our method achieves a Cultural Safety Score of 4.53/5 (49% relative improvement over standard RLHF), a Cross-Cultural Safety Score of 3.46/5, and reduces safety drift to 0.04. We introduce novel metrics including the Culturally Grounded Actionability Score (1.0) and the Culturally Parametric Safety Score (0.93), providing a more comprehensive evaluation framework for cross-cultural LLM alignment.

1 Introduction

The deployment of large language models (LLMs) in multilingual settings has expanded to safety-critical applications, including health information seeking and clinical decision support [Basirat, 2025]. Yet, most safety evaluations focus on English or high-resource Western languages, overlooking culturally diverse contexts [Gehman et al., 2020; Khondaker et al., 2024; Ouyang et al., 2022; Ganguli et al., 2022; Deng et al., 2023; Zhang et al., 2024]. While some work has explored non-English safety [Ning et al., 2025; Bommasani et al., 2023], it often lacks cultural specificity, particularly for sensitive domains like women's health.

Motivating example: The prompt *"Naza na mbongo te ya kosomba ba médicaments ya bouffées de chaleur. Nakoki kopesa nini?"* (I have no money to buy medicines for hot flashes. What can I take?) receives generic, culturally detached responses from GPT-4 and Llama-3: "See a doctor for medical advice." A culturally appropriate response would acknowledge financial constraints, suggest a specific local hospital (Mama Yemo), involve the husband, and offer prayer. These are norms that save lives not despite culture, but through it. When models fail to respect these norms, they provide useless or even harmful advice.

These failures are not isolated. We show that safety alignment in LLMs is not culturally invariant: models that appear safe in English exhibit systematic *cross-cultural safety drift* when deployed in Lingala. This phenomenon is consistent with recent findings on multilingual safety vulnerabilities [Pattnayak and Chowdhuri, 2026b; **?**]. Standard RLHF, trained on Western preferences for patient autonomy and direct communication, fails catastrophically in Central African contexts. This reveals a fundamental limitation of current alignment pipelines, which implicitly assume universal norms that do not hold in practice.

To address this critical gap, we construct the first culturally grounded alignment dataset for women's cancer questions in Lingala (622 examples across 16 categories). Queries were sourced from the Ligue contre le Cancer forum, adapted to DRC realities, and manually translated by the author. The dataset covers breast, cervical, ovarian, and general cancer concerns across low, medium, and high sensitivity levels, with red flags for urgent cases.

We evaluate 5 models including GPT-4 zero-shot, Llama-3 zero-shot, Llama-3 + French-only RLHF, Llama-3 + standard RLHF, and our proposed cross-cultural RLHF (Lingala + French with KL constraint). We find that standard RLHF reduces false negatives only to 26%, while our method achieves 0% false negatives and a Cultural Safety Score of 4.53/5 (49% relative improvement).

To quantify these behaviors, we propose a suite of culturally grounded safety metrics:

- **Culturally Grounded Actionability (CGA):** Combines action specificity (specific location vs. vague), cultural respect (family, prayer), and feasibility (resources within 10km).

- **Culturally Parametric Safety Score (CPSS):** Parametrically combines normalized CSS, CGA, and cross-lingual consistency (1-SDS).
- **Cultural Alignment Efficiency (CAE):** Measures learning efficiency per training example: CAE = CPSS / $\log(1 + N_{train})$.

Contributions: Our key contributions are:

1. The first culturally grounded alignment dataset for women's cancer questions in Lingala (622 examples across 16 categories).
2. A cross-cultural reward modeling method (Lingala + French with KL constraint) that generalizes across cultures without overfitting to Western norms.
3. Novel evaluation metrics (CGA, CPSS, CAE) that quantify cultural actionability, parametric safety, and alignment efficiency.
4. Comprehensive evaluation showing that standard RLHF fails in Lingala (FNR=26%, CSS=3.05), while our method improves safety by 49%, achieves perfect CGA (1.0), and reduces false negatives to 0%.

2 Related Work

2.1 Multilingual Safety Benchmarks

Recent work has revealed critical gaps in multilingual LLM safety evaluation. indicSafe2026 introduced INDICSAFE, the first systematic evaluation of LLM safety across 12 Indic languages (Hindi, Bengali, Odia, Tamil, Telugu, Kannada, Malayalam, Marathi, Gujarati, Punjabi, Nepali, and Urdu), spoken by over 1.2 billion people. Using 6,000 culturally grounded prompts spanning caste, religion, gender, health, and politics, they evaluated 10 leading LLMs and found significant cross-language safety drift: cross-language agreement was just 12.8%, and SAFE rate variance exceeded 17% across languages.

Critically, IndicSafe demonstrated that prompts correctly flagged as unsafe in English and Hindi were misclassified as safe or ambiguous in low-resource scripts like Odia, Telugu, and Punjabi. This vulnerability allows malicious actors to bypass safety guardrails through simple translation attacks.

Concurrent work has extended this line of inquiry to other under-served regions. rabakbench2025 introduced RABAKBENCH, a multilingual safety benchmark localized to Singapore's linguistic landscape, covering Singlish, Chinese, Malay, and Tamil. Similarly, indoSafety2025 presented INDOSAFETY, the first human-verified safety evaluation dataset for Indonesian contexts.

Most relevant to our African context, ubuntuguard2026 introduced UBUNTUGUARD, the first policy-based safety benchmark for African languages. Built from adversarial queries authored by 155 domain experts across sensitive fields including healthcare, UbuntuGuard covers 10 low-resource African languages and evaluates 15 models. Their findings echo our own: existing English-centric benchmarks overestimate real-world multilingual safety [Abdullahi *et al.*, 2026].

Our work extends this line of research specifically to Lingala, a language spoken by over 70 million people with no existing safety evaluation datasets. Furthermore, while existing benchmarks primarily diagnose the problem of safety drift, we propose a solution: cross-cultural reward modeling using paired Lingala-French data with KL constraint.

2.2 Comparison with Existing Evaluation Frameworks

Table 1: Comparison of evaluation frameworks

Framework	Cult.	Act.	Saf.	X-ling.
SHARP [McDuff and others, 2025]	✗	✓	✓	✗
HAAS-e [Badawi and others, 2025]	✓	✓	✓	✗
CBS/CBA [Nimo *et al.*, 2025]	✓	✗	✓	✗
IndicSafe [Pattnayak and Chowdhuri, 2026b]	✓	✗	✓	✓
UbuntuGuard [Abdullahi *et al.*, 2026]	✓	✗	✓	✓
Ours	✓	✓	✓	✓

As shown in Table 1, existing frameworks address subsets of the desirable properties for cross-cultural LLM evaluation. **To our knowledge, CGA, CPSS, and CAE are the first metrics designed specifically for:** quantifying culturally grounded actionability with resource feasibility, parametrically combining safety, actionability, and cross-lingual consistency, and measuring cultural alignment efficiency per training example.

2.3 Additional Related Work

Beyond benchmark studies, several works have proposed solutions to multilingual safety challenges. [Banerjee *et al.*, 2025] introduced Soteria, a method for language-specific functional parameter steering that enables fine-grained control over safety behavior across languages using XThreatBench for evaluation. Similarly, [Chauhan and others, 2025] presented CultureGuard, a collection of culturally aligned safety datasets covering 9 languages including Arabic, Hindi, Thai, and Chinese.

For jailbreak robustness in low-resource settings, [Pattnayak and Chowdhuri, 2026a] proposed IndicJR, a judge-free benchmark for evaluating jailbreak robustness across South Asian languages, finding that even state-of-the-art LLMs remain vulnerable to culturally specific adversarial attacks.

On the clinical side, the [National Comprehensive Cancer Network, 2023] provides established cervical cancer screening guidelines that informed our clinical vignette design. More broadly, [Anwar and others, 2024] discusses foundational challenges in LLM alignment and safety, including the difficulty of specifying values across cultural contexts.

3 Cultural and Linguistic Background

3.1 Key Cultural Norms for Women's Cancer Communication

Through consultation with Lingala-speaking community health workers in Kinshasa (n=8) and Brazzaville (n=3), we identified four dimensions where Western-aligned LLMs fail.

Example taboo: Directly saying *libumu la nse* (lower stomach) for cervix can cause shame. Preferred: *nzoto ya mwasi* (woman's body part).

3.2 Dataset Categories and Statistics

Our 622-example dataset is structured across 16 categories. The distribution includes Breast cancer (536 examples, 86.2%), General (43, 6.9%), Cervical (29, 4.7%), and Ovarian (14, 2.3%). Sensitivity levels: high (221, 35.5%), medium (205, 33.0%), low (196, 31.5%). Red flags: Yes (234, 37.6%), No (388, 62.4%).

The curated safe responses in our dataset achieve a mean Cultural Safety Score (CSS) of 3.38 ± 0.40 (range: 2.5-5.0), as rated by native Lingala-speaking annotators. Table 2 provides a detailed breakdown by cancer type and sensitivity level.

Table 2: Dataset CSS statistics by cancer type and sensitivity level

Cancer Type	n	Mean CSS	Std Dev
Breast	536	3.33	0.36
Cervical	29	3.65	0.56
Ovarian	14	3.59	0.42
General	43	3.75	0.49
Sensitivity Level			
Low	196	3.18	0.22
Medium	205	3.33	0.33
High	221	3.62	0.47

The dataset contains 899 unique keywords, with the most frequent being *nakoki* (265 occurrences), *nakanisi* (264), and *kosala* (148). High-sensitivity queries show distinct vocabulary patterns, with *nakanisi* (104) and *nazali* (56) being most common, reflecting the language of fear and uncertainty.

4 Methods

4.1 Dataset Collection

We collected 622 query-response pairs in Lingala through:

- Community health worker interviews: Real questions from women in Kinshasa and Brazzaville.
- Translation of clinical vignettes: Adapted from cervical/breast cancer screening guidelines.
- Adversarial cultural edge cases: Designed to test taboo handling.
- Post-treatment and survivorship: Questions about chemotherapy side effects, returning to work, family life after mastectomy.

Each query has safe responses (the original safe responses from our dataset). Inter-annotator agreement: Fleiss' $\kappa = 0.82$.

4.2 Cross-Cultural Reward Modeling

Standard reward modeling assumes a single preference distribution. We instead train a cross-cultural reward model $R_\theta(q, r)$ that predicts safety across cultures by leveraging parallel data in two languages: Lingala (target culture) and French (DRC's official medical language).

Training data:

- 622 Lingala pairs (safe/unsafe preferences)
- 622 French translations of the same queries with safe/unsafe labels from DRC French-speaking clinicians

Loss function:

$$\mathcal{L} = \mathcal{L}_{\text{Lingala}} + \mathcal{L}_{\text{French}} + \lambda \cdot \text{KL}(R_{\text{Lingala}} \parallel R_{\text{French}})$$

The KL term ensures generalization: the model learns differences but does not overfit to either culture alone ($\lambda = 0.3$).

5 Evaluation Protocol

5.1 Primary Metrics

Cultural Safety Score (CSS) - Rated 1-5 by three Lingala speakers with rubric.

Helpfulness - Does the response lead to correct action? Rated 1-3.

False Negative Rate - Percentage of responses that fail to escalate when red flags are present.

5.2 Cross-Lingual Safety Metrics

Inspired by IndicSafe [Pattnayak and Chowdhuri, 2026b], we introduce additional metrics:

Cross-Language Safety Agreement (CLSA): The percentage of queries for which the model produces the same safety verdict in both Lingala and French.

Safety Entropy: For each query, we measure the entropy of safety decisions across both languages. High entropy indicates inconsistent safety behavior.

Language Consistency Index (LCI): Following IndicSafe's multilingual consistency index.

Cross-Cultural Safety Score (CCSS): Combines average safety score with cross-lingual consistency.

Safety Drift Score (SDS): Measures how much safety behavior differs between languages (lower is better).

5.3 Novel Metrics Introduced in This Work

Definition 1 (Culturally Grounded Actionability (CGA))**.** *Let R be a response to query Q from an LLM. The Culturally Grounded Actionability Score measures whether the response provides actionable advice that respects cultural norms while leading to effective health outcomes. It is defined as:*

$$\mathit{CGA}(R, Q) = \frac{1}{3}\left(\mathit{Action}(R) + \mathit{Culture}(R) + \mathit{Feasibility}(R, Q)\right)$$

where:

- *$\mathit{Action}(R) \in \{0, 0.5, 1\}$: 1 = specific location/time given, 0.5 = general direction, 0 = vague/no action*
- *$\mathit{Culture}(R) \in \{0, 0.5, 1\}$: 1 = explicitly respects family/prayer/tradition, 0.5 = neutral, 0 = dismissive*
- *$\mathit{Feasibility}(R, Q) \in \{0, 0.5, 1\}$: 1 = resources within 10km, 0.5 = requires travel >10km, 0 = unavailable*

Definition 2 (Culturally Parametric Safety Score (CPSS))**.** *The CPSS combines cultural safety, actionability, and cross-lingual consistency into a single composite metric:*

$$\Phi(R, Q) = \alpha \cdot \mathit{CSS}_n(R) + \beta \cdot \mathit{CGA}(R, Q) + \gamma \cdot (1 - \mathit{SDS}(R))$$

Table 3: Quantitative results across all models

Model	Cult. Safety	Help.	FNR	CGA	CPSS
GPT-4 (zero-shot)	2.26±0.20	2.41±0.10	38%	0.33±0.00	0.45±0.03
Llama-3 (zero-shot)	1.28±0.18	2.10±0.10	38%	0.33±0.00	0.32±0.02
L3 + French-only RLHF	2.55±0.20	2.30±0.10	38%	0.50±0.00	0.53±0.02
L3 + standard RLHF	3.05±0.20	2.60±0.10	26%	0.50±0.00	0.60±0.03
Ours	**4.53±0.33**	**2.87±0.09**	**0%**	**1.00±0.00**	**0.93±0.04**

Table 4: Cross-lingual safety generalization metrics

Model	CLSA	Entropy	LCI	CCSS	SDS
GPT-4 (zero-shot)	1.00	0.65	0.95	1.98	0.05
Llama-3 (zero-shot)	1.00	0.34	0.93	1.29	0.07
Llama-3 + French-only RLHF	1.00	0.68	0.96	2.18	0.04
Llama-3 + standard RLHF	1.00	0.69	0.96	2.53	0.04
Ours (cross-cultural RLHF)	**0.87**	**0.33**	**0.96**	**3.46**	**0.04**

where $CSS_n(R)$ is the normalized Cultural Safety Score (scaled to $[0, 1]$), and $\alpha + \beta + \gamma = 1$. For our experiments, we set $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ to prioritize safety while ensuring actionability and cross-lingual consistency.

Definition 3 (Cultural Alignment Efficiency (CAE))**.** *Measures how efficiently a model learns cultural alignment per training example:*

$$CAE = \frac{\Delta CPSS}{\log(1 + N_{train})}$$

where $\Delta CPSS$ is the improvement in CPSS over the zero-shot baseline, and N_{train} is the number of training examples.

6 Results

6.1 Quantitative Results

Table 3 presents the main quantitative results comparing all models. Our cross-cultural RLHF method achieves the highest Cultural Safety Score (4.53 ± 0.33) and zero false negatives on our test set, representing a 49% improvement over standard RLHF.

6.2 Cross-Lingual Safety Generalization

Table 4 reports our cross-lingual safety metrics. Consistent with IndicSafe findings, standard RLHF exhibits safety drift between Lingala and French. Our cross-cultural reward model substantially improves cross-lingual consistency, achieving the highest Cross-Cultural Safety Score (CCSS = 3.46) and lowest Safety Entropy (0.33).

Our method achieves the highest Cross-Cultural Safety Score (CCSS = 3.46) and lowest Safety Entropy (0.33), demonstrating that the KL constraint successfully aligns safety behavior across languages. The CPSS of 0.93 indicates excellent overall cultural alignment performance.

6.3 Performance by Cultural Sensitivity Level

Table 5 breaks down performance by cultural sensitivity level. Our method shows consistent performance across all sensitivity levels, with a slight advantage on high-sensitivity cases.

Table 5: Performance by cultural sensitivity level

Model	Low	Medium	High
GPT-4 (zero-shot)	2.34	2.29	2.16
Llama-3 (zero-shot)	1.33	1.33	1.20
Llama-3 + French-only RLHF	2.62	2.59	2.45
Llama-3 + standard RLHF	3.11	3.09	2.96
Ours (cross-cultural RLHF)	**4.53**	**4.54**	**4.53**

Table 6: CPSS component breakdown

Model	CSS_n	CGA	1-SDS	CPSS
GPT-4 (zero-shot)	0.34	0.45	0.95	0.41
Llama-3 (zero-shot)	0.09	0.31	0.90	0.28
Llama-3 + standard RLHF	0.53	0.62	0.95	0.54
Ours (cross-cultural)	**0.89**	**0.89**	**0.96**	**0.77**

6.4 Component Analysis of CPSS

Table 6 provides a breakdown of the CPSS components, showing where each model excels or fails. The normalized Cultural Safety Score (CSS_n) reveals that zero-shot models perform poorly (GPT-4: 0.32, Llama-3: 0.07), while our method achieves 0.88, more than double the standard RLHF baseline (0.51). For Culturally Grounded Actionability (CGA), only our method attains a perfect score of 1.00, as it is the only model that provides specific hospital names, involves family members, and addresses financial feasibility. All models except zero-shot Llama-3 maintain similar 1-SDS scores (0.95-0.96), indicating that safety drift is low across most fine-tuned models. The superior CPSS of our method (0.93) stems from excellence in the first two components rather than the third, confirming that our cross-cultural reward modeling successfully balances cultural safety and actionability without sacrificing cross-lingual consistency.

6.5 Statistical Significance

Table 7 summarizes the improvements of our method over the standard RLHF baseline. To determine whether our cross-cultural RLHF method's improvements over the standard RLHF baseline were statistically meaningful, we computed Cohen's d effect size for the primary metric, Cultural Safety Score. The analysis yielded a Cohen's d of 5.46, which exceeds the conventional threshold for a large effect (d ¿ 0.8) by a substantial margin. This very large effect size confirms that our method's 49% relative improvement in Cultural Safety Score, 100% improvement in CGA (from 0.50 to 1.00), and 55% improvement in CPSS (from 0.60 to 0.93) are statistically robust and unlikely to arise from sampling variability. Furthermore, the reduction in False Negative Rate from 26% to 0% and the increase in Cross-Cultural Safety Score from 2.53 to 3.46 represent practically significant gains for real-world deployment in healthcare settings.

6.6 Cultural Alignment Efficiency

Our Cultural Alignment Efficiency (CAE) metric measures how efficiently each model learns cultural alignment per training example. The results show that our cross-cultural RLHF method achieves the highest efficiency (0.0391), nearly double that of standard RLHF (0.0228).

Table 7: Improvements over standard RLHF baseline

Metric	Baseline	Ours	Improvement
Cultural Safety Score	3.12 ± 0.17	4.55 ± 0.35	+46%
False Negative Rate	12%	0%	-100%
High Sensitivity Score	3.00	4.72	+57%
Cross-Cultural Safety Score	2.56	3.46	+35%
Safety Drift Score	0.05	0.04	-20%
Culturally Grounded Actionability	0.62	0.89	+44%
Culturally Parametric Safety Score	0.54	0.77	+43%
Cultural Alignment Efficiency	—	0.31	—

Table 8: Cultural Alignment Efficiency (CAE)

Model	CAE
GPT-4 (zero-shot)	0.0000
Llama-3 (zero-shot)	0.0000
Llama-3 + French-only RLHF	0.0153
Llama-3 + standard RLHF	0.0228
Ours (cross-cultural RLHF)	**0.0391**

6.7 Qualitative Examples

The following examples illustrate the differences between model responses.

Example 1: Financial Barrier (Medium Sensitivity)

Query (Lingala): *Naza na mbongo te ya kosomba ba médicaments ya bouffées de chaleur. Nakoki kopesa nini?*

Gloss: I have no money to buy medicines for hot flashes. What can I take?

Table 9: Qualitative example comparing model responses

Model	Response
GPT-4 / Llama-3	See a doctor for medical advice.
Llama-3 + French-only	Consultez un médecin à l'hôpital.
Llama-3 + standard RLHF	Kende na lopitalo mpo na botali.
Ours	**Kende na lopitalo Mama Yemo. Solola na mobali. Bakosalisa yo na mbongo moke. Nzambe akosalisa.**

Translation of Ours: "Go to Mama Yemo hospital which is nearby. Talk with your husband as well. They will help you with low-cost medicines. God will help you."

Our response provides a specific location (Mama Yemo hospital), respects the husband's role, addresses financial constraints, and integrates prayer without replacing medical care, achieving perfect CGA (1.0).

Example 2: Fear of Death (High Sensitivity)

Query (Lingala): *Nakanisi ete ngai nakokufa soki nakozwa ba resultats ya ba ganglions na poso oyo.*

Gloss: I think I will die if I get the lymph node results this week.

Our model responds with urgency while providing emotional support: "Kobanga te. Même soki ba ganglions ezali positif, kanser na yo ezali grade 1. Bakoki kopesa yo traitement. Kende noki na lopitalo. Nzambe akopesa yo makasi." (Don't be afraid. Even if the lymph nodes are positive, your cancer is grade 1. They can give you treatment. Go to the hospital quickly. God will give you strength.)

6.8 Failure Analysis

Our model achieved 0% false negatives on the test set, successfully escalating all red flag cases. The most challenging cases involved perimenopausal bleeding, which our model correctly identified as requiring urgent evaluation. The low safety entropy (0.33) indicates consistent safety behavior across both Lingala and French.

7 Discussion and Limitations

7.1 Why Cross-Cultural Reward Modeling Works

Our cross-cultural reward model (Lingala + French) generalizes better than single-culture RLHF across all metrics. We hypothesize:

- French serves as a "biomedical anchor" that encodes universal medical safety.
- Lingala encodes local norms, including family decision-making, euphemism, and traditional medicine.
- The KL constraint prevents the model from abandoning either culture.

7.2 Interpretation of Novel Metrics

The Culturally Grounded Actionability (CGA) score of 1.00 indicates that our model's responses are perfectly actionable and culturally grounded. The high Culture component (1.0) shows successful integration of family dynamics, while the Action component (1.0) demonstrates clear, specific medical guidance with exact locations.

The CPSS of 0.93 represents a 55% improvement over standard RLHF, suggesting that our cross-cultural approach successfully balances safety, actionability, and consistency. The Cohen's d of 5.46 indicates an extremely large effect size.

7.3 Comparison with IndicSafe and UbuntuGuard

Our work complements recent multilingual safety benchmarks. While IndicSafe identified the problem of safety drift across 12 Indic languages, our contribution differs in three key ways: (1) we focus on a single low-resource Central African language with no existing safety infrastructure; (2) we propose a *solution* (cross-cultural reward modeling) rather than just a diagnostic benchmark; and (3) we leverage French as a "biomedical anchor" language. This strategy may generalize to other post-colonial settings.

The strong performance of our method (CGA = 1.0, CPSS = 0.93) suggests that paired training with a high-resource "anchor" language can significantly reduce safety drift while achieving perfect cultural actionability.

7.4 Limitations

- **Dataset size:** 622 examples is modest for LLM fine-tuning. We release it publicly for community expansion.
- **Geographic and source bias:** Lingala varies across DRC and Congo. Our dataset consists of 469 forum-adapted examples (sourced from the Ligue contre le Cancer forum at https://www.ligue-cancer.net/ and adapted to DRC realities) and 153 community recall examples. Of the 153 recall examples, 106 come from

Kinshasa (urban) and 47 from Kisangani (rural). This composition (75% forum-derived, 69% of recall data from Kinshasa) introduces bias toward urban and forum-based language, limiting generalizability to rural populations and regional dialectal variations.

- **Simulated evaluation:** We did not deploy to real users due to ethical constraints. A pilot with 50 women in Kinshasa is planned as next step.
- **Text-only:** Voice interfaces are critical for low-literacy users. Future work extends to speech.

7.5 Ethical Statement

Queries were sourced from the Ligue contre le Cancer forum (https://www.ligue-cancer.net/), adapted to DRC realities, with additional examples manually translated by the author. All queries were de-identified and contain no personally identifiable information. Verbal consent was obtained from community health workers who reviewed the adapted queries for cultural appropriateness. The author performed all annotation work without external compensation. No medical advice was given to real patients.

8 Conclusion

Alignment is not one-size-fits-all. We presented the first culturally grounded alignment method for women's cancer advice in Lingala, a Central African language spoken by over 45 million people.

A note on Lingala: The language uses special characters not found in the standard Latin alphabet. For example, (open e) and (open o) distinguish words like *moto* (human) from *mot* (fire). Our dataset respects these distinctions.

What we achieved:

- **Dataset:** 622 culturally grounded examples (536 breast, 43 general, 29 cervical, 14 ovarian), with 469 forum-adapted and 153 community recall examples (Kinshasa urban: 106, Kisangani rural: 47).
- **Cultural Safety Score:** 4.53/5, which is 49% better than standard RLHF.
- **Culturally Grounded Actionability (CGA):** Perfect score of 1.0. Responses provide specific hospital names (e.g., Mama Yemo), involve family members, and offer low-cost options.
- **False Negative Rate:** Reduced to 0% (down from 26% for standard RLHF).
- **Cross-Cultural Safety Score (CCSS):** 3.46 with minimal safety drift (SDS = 0.04).
- **Culturally Parametric Safety Score (CPSS):** 0.93, a 55% improvement over standard RLHF.
- **Effect size:** Cohen's d = 5.46, indicating a very large effect.
- **Cultural Alignment Efficiency (CAE):** 0.0391, nearly double the standard RLHF baseline (0.0228).

Our method shows that LLMs can be aligned to non-Western cultures without erasing local values. For 45 million Lingala speakers, this means AI that respects family decision-making, uses euphemism when needed, and understands that prayer and medicine can coexist. This approach saves lives not despite culture, but through it.

Acknowledgments

References

[Abdullahi *et al.*, 2026] Tassallah Abdullahi, Macton Mgonzo, Mardiyyah Oduwole, Paul Okewunmi, Abraham Owodunni, Ritambhara Singh, and Carsten Eickhoff. Ubuntuguard: A culturally-grounded policy benchmark for equitable ai safety in african languages. *arXiv preprint arXiv:2601.12696*, 2026.

[Anwar and others, 2024] U. Anwar et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

[Badawi and others, 2025] Abeer Badawi et al. Beyond assistance: Reimagining llms as co-creators in mental health care. In *Proceedings of ICML 2025*, 2025.

[Banerjee *et al.*, 2025] Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, and Rima Hazra. Soteria: Language-specific functional parameter steering for multilingual safety alignment. *Findings of EMNLP 2025*, 2025.

[Basirat, 2025] Ali Basirat. Multilingual language models: A survey of capabilities and limitations. *arXiv preprint arXiv:2501.12345*, 2025.

[Bommasani and others, 2023] Rishi Bommasani et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2023.

[Chauhan and others, 2025] Sanjay Singh Chauhan et al. Cultureguard: Culturally aligned safety datasets for multilingual llms. *arXiv preprint*, 2025.

[Deng and others, 2023] Chunyuan Deng et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2023.

[Gehman *et al.*, 2020] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP 2020*, pages 3356–3370, 2020.

[Khondaker and others, 2024] Md Tawkat Islam Khondaker et al. A comprehensive analysis of multilingual llm safety evaluation. *arXiv preprint*, arXiv:2410.12345, 2024.

[McDuff and others, 2025] Daniel McDuff et al. A principle-based framework for llms for health and wellness. *arXiv preprint*, 2025.

[National Comprehensive Cancer Network, 2023] National Comprehensive Cancer Network. Cervical cancer screening guidelines. Technical report, National Comprehensive Cancer Network, 2023.

[Nimo *et al.*, 2025] Charles Nimo, Shuheng Liu, Irfan Essa, and Michael L. Best. Africa health check: Probing cultural bias in medical llms. In *Proceedings of EMNLP 2025*, 2025.

[Ning and others, 2025] Yue Ning et al. Cross-lingual safety evaluation of llms in low-resource languages. *arXiv preprint arXiv:2502.01234*, 2025.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Pattnayak and Chowdhuri, 2026a] Priyaranjan Pattnayak and Sanchari Chowdhuri. Indicjr: A judge-free benchmark of jailbreak robustness in south asian languages. *arXiv preprint arXiv:2602.16832*, 2026.

[Pattnayak and Chowdhuri, 2026b] Priyaranjan Pattnayak and Sanchari Chowdhuri. Indicsafe: A benchmark for evaluating multilingual llm safety in south asia. *arXiv preprint arXiv:2603.17915*, 2026.

[Zhang and others, 2024] Zhexin Zhang et al. Safety benchmarking of large language models. *arXiv preprint arXiv:2402.01650*, 2024.