Toward Cultural Alignment: Human-Centered Evaluation of Multimodal AI Stories Across Five African Communities

Millicent Ochieng[1*], **Felermino D. M. A. Ali**[1*], **Elizabeth A. Ankrah**[1], **Najeeb G. Abdulhamid**[1], **Boyd Migisha**[2], **Stephanie Nyairo**[1], **Mercy Muchai**[1], **Samuel Maina**[1], **Aditya Vashistha**[3], **Anja Thieme**[4], **Jacki O'Neill**[1]

[1]Microsoft Research Africa, [2]Swansea University, [3]Cornell University, [4]Microsoft Research Cambridge

Abstract

In this paper, we examine how well AI-generated multimodal stories align with the lived practices, relationships, language, values, and visual expectations of the communities they represent. We conduct a community-grounded mixed-methods evaluation with 19 culture representatives across five African communities, combining quantitative annotations with qualitative focus group discussions. We find that cultural alignment depends not simply on recognizable cultural markers, but on how those markers fit social, linguistic, procedural, and visual context. From these evaluations, we develop a taxonomy of cultural alignment comprising five broader cultural marker categories and eight recurring mechanisms of misalignment. We additionally evaluate five multimodal LLM judges to examine whether automated evaluation can approximate community-grounded judgments at scale. Judge reliability and score calibration vary substantially across communities, with no single judge performing consistently across all five settings. These findings motivate community-calibrated evaluation pipelines in which automated judges are validated against community judgments to determine where they can be trusted and where human review remains necessary.

1 Introduction

Generative AI systems are increasingly used to produce text and visual content for diverse users at scale, including stories (Tian et al., 2024). Such content can appear fluent and locally plausible while failing to reflect the lived practices, relationships, language, values, and visual expectations of the communities being represented (Agarwal et al., 2025b; Wang et al., 2024; Kazemi et al., 2024). We refer to community-recognized fit between generated content and lived experience as *cultural alignment*.



Figure 1: Example story.

Evaluating cultural alignment is difficult because culture is complex, situated, and context-dependent (Adilazuarda et al., 2024). A form of address, item of clothing, or food practice may appear plausible to outsiders while feeling inappropriate, foreign, generic or incomplete to community members. Community-grounded evaluation can surface these situated judgments but is time-intensive to scale across models, communities, and generation settings. LLM-as-judge methods offer a scalable alternative (Li et al., 2025), but their reliability for culturally situated multimodal evaluation remains uncertain.

We examine cultural alignment through a community-grounded mixed-methods evaluation of AI-generated multimodal stories across five African communities: Hausa, Kikuyu, Luo, AmaXhosa, and Xichangana. We combine quantitative annotations and story-level scores from 19 culture representatives with qualitative focus group dis-

[*] Equal Contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 41484–41510
October 24-29, 2026 ©2026 Association for Computational Linguistics

cussions to examine both what cultural elements shape alignment judgments and why they are experienced as aligned or misaligned.[1] Narratives are widely used to support behavior-change communication (Hinyard and Kreuter, 2007; Ng'endo and Kariuki, 2026), making stories a useful setting for studying culturally situated generation. Each story consists of four text-image frames, as illustrated in Figure 1, and centers on everyday diabetes lifestyle management. We use this domain as a culturally consequential testbed and do not assume that the resulting patterns generalize unchanged to other domains. Stories were generated in English for Hausa, Kikuyu, Luo, and AmaXhosa and in Portuguese for Xichangana, reflecting the role of these languages in formal written communication of the represented communities.

Because existing annotation tools did not support frame-by-frame evaluation of such multimodal stories across both text and images, we built a custom annotation platform that allowed culture representatives to evaluate each story frame, identify influential text spans and image regions, categorize cultural markers, and provide an overall story-level score. From the annotations and focus group discussions, we developed a taxonomy of cultural alignment that captures the textual and visual markers representatives attend to when judging alignment and the recurring mechanisms through which stories become culturally misaligned. We additionally evaluated five multimodal LLM judges using the same story-level rubric to test whether automated evaluation can approximate community-grounded judgments at scale.

Our analysis yields three main findings. First, our taxonomy shows that cultural alignment is not reducible to the presence of recognizable cultural markers, but depends on how those markers fit context across five broader categories: *Referential* markers such as names, foods, and places; *Procedural* markers such as food preparation and exercise routines; *Contextual* markers concerning when and where cultural elements appear; *Socio-geographic* markers such as clinics, markets, homes, and infrastructure; and *Linguistic Register* markers such as dialect, code-switching, and forms of address. Second, generated stories become culturally misaligned through recurring mechanisms, including substitution, norm violation, omission, forced insertion, register conflation, cross-modal inconsistency, stereotyping, and hallucination. Third, LLM judges do not consistently approximate evaluations from culture representatives; reliability varies by community, and no single judge model performs consistently across all five communities. For example, Pearson correlations between LLM judge scores and scores from culture representatives are strong for Luo ($r = 0.82$–0.89) and Hausa ($r = 0.67$–0.80), but no AmaXhosa judge remains significant after correction, and for Xichangana, judges assign alignment scores 35–45 points higher than culture representatives. These findings motivate community-calibrated evaluation in which community judgments ground cultural alignment assessment and automated judges are validated to determine where they can be trusted and where human review remains necessary. Data and code: `https://github.com/microsoft/Multimodal-Cultural-Alignment-Africa`.

2 Related Work

Cultural Alignment and Operationalizing Culture in LLMs. A growing body of work probes LLMs for cultural knowledge using proxies such as Hofstede's dimensions (Arora et al., 2023; Cao et al., 2023), moral judgment datasets (Scherrer et al., 2023; Jinnai, 2024), and social etiquette norms (Rao et al., 2025). However, culture is difficult to operationalize because it cannot be reduced to static demographic labels, national categories, or isolated value dimensions (Adilazuarda et al., 2024). Prior evaluations also show that LLMs often align more strongly with Western cultural norms, and that English prompts can flatten cross-cultural variation (Cao et al., 2023; Agarwal et al., 2025a; Rao et al., 2025). While these studies are important for measuring cultural knowledge and bias, knowing about a culture is different from generating content that communities recognize as culturally aligned. Our work builds on this distinction by evaluating whether generated multimodal stories reflect lived practices, relationships, language, values, and visual expectations as judged by culture representatives.

Cultural Misalignment in Generated Text and Images. Research on cultural misalignment in generated content has examined failures in dialogue, narrative, and visual generation. In dialogue,

[1] We use *culture representatives* to refer to community members who self-identify with the represented community and report relevant lived and linguistic knowledge, enabling them to assess whether generated content feels authentic, inappropriate, foreign, or incomplete.

NormDial (Li et al., 2023) and ReNoVi (Zhan et al., 2024) annotate norm adherence and violation in Chinese and American conversations, while cross-cultural work shows that culture-specific reasoning often fails to generalize (Jinnai, 2024). In narrative generation, prior work has identified Western bias in stories about Arab cultures (Naous et al., 2024) and introduced taxonomies of cultural misrepresentation for Indian stories (Bhagat et al., 2026). In the visual domain, text-to-image (T2I) models have been shown to neglect or misrepresent disadvantaged and underrepresented cultures (Zhang et al., 2024; Kannen et al., 2024; Johnson et al., 2026; Thieme et al., 2026), and recent benchmarks characterize *how* these failures arise. CuRe (Rege et al., 2025) traces them to the long tail of web-scraped training data, showing that T2I systems hallucinate details for artifacts of the Global South (e.g., an Ethiopian *jebena*) that they render reliably for better-represented ones. CulturalFrames (Nayak et al., 2025) distinguishes *explicit* expectations, stated in the prompt, from *implicit* ones implied by its cultural context, and finds that expectations are missed 44% of the time across 10 countries, including 68% of explicit and 49% of implicit cases. Moving beyond object-centric artifacts, CULTIVate (Malakouti et al., 2026) evaluates cultural faithfulness through social activities such as dining, greeting, and dance, where meaning emerges from interaction and context rather than isolated objects, and decomposes failure into alignment, hallucination, exaggeration, and diversity, reporting systematically lower faithfulness for Global South than Global North cultures. Notably, the studies find that standard image–text alignment metrics correlate poorly with human judgments of cultural alignment, motivating culturally grounded evaluation. Complementing these automated benchmarks, other work examines community-driven methods for assessing cultural sensitivity (Kiden et al., 2025). These studies show that cultural misalignment appears in both language and images, but they evaluate images from short, isolated prompts; less work evaluates text and image jointly within the same culturally situated narrative, or examines the mechanisms through which multimodal stories become misaligned.

Scalable Evaluation and LLM-as-Judge. LLM-as-judge methods have been widely adopted as scalable alternatives to human evaluation for assessing generated content (Zheng et al., 2023; Kim et al., 2024; Li et al., 2025). These methods can reduce evaluation cost and increase coverage, but their reliability is sensitive to task framing, model bias, and subjective evaluation criteria (Khan et al., 2025). Cultural alignment evaluation is especially challenging because judgments may depend on lived and linguistic knowledge of the represented community, including whether language use, social norms, visual settings, and cultural markers feel appropriate in context. Our work therefore evaluates multimodal LLM judges against scores from culture representatives rather than treating automated scores as ground truth. In doing so, we examine when automated judges approximate community evaluations, where they are biased, and why community validation remains necessary for scalable cultural alignment evaluation.

3 Methodology

We adopt a mixed-methods approach to evaluate cultural alignment in generated multimodal stories, validate automated evaluations against community-grounded judgments, and characterize the cultural evidence and misalignment mechanisms identified by culture representatives. Figure 2 illustrates the end-to-end pipeline from story generation through cultural evaluation, and the following sections describe each stage.

3.1 Story Generation

Persona generation. We generated culturally situated personas to provide demographic and cultural context for story generation. Using GPT-4.1, we created personas through a cascading pipeline of 19 sequential LLM calls, with each attribute conditioned on previously generated attributes (e.g., country $\rightarrow$ community $\rightarrow$ province $\rightarrow$ religion $\rightarrow$ name $\rightarrow$ gender $\rightarrow$ age). This dependency-aware process was designed to maintain consistency across demographic, cultural, and regional attributes. The generated personas were manually reviewed for plausibility by authors from the represented communities, and the same reviewed persona pool was reused across all seven generation models. For each community, we generated 40 personas, each with 28 interrelated attributes used as context for story generation; the full set of persona fields is provided in Appendix Table 6. Each persona was paired with an everyday diabetes lifestyle question to guide story generation while keeping the focus on everyday life rather than medical di-

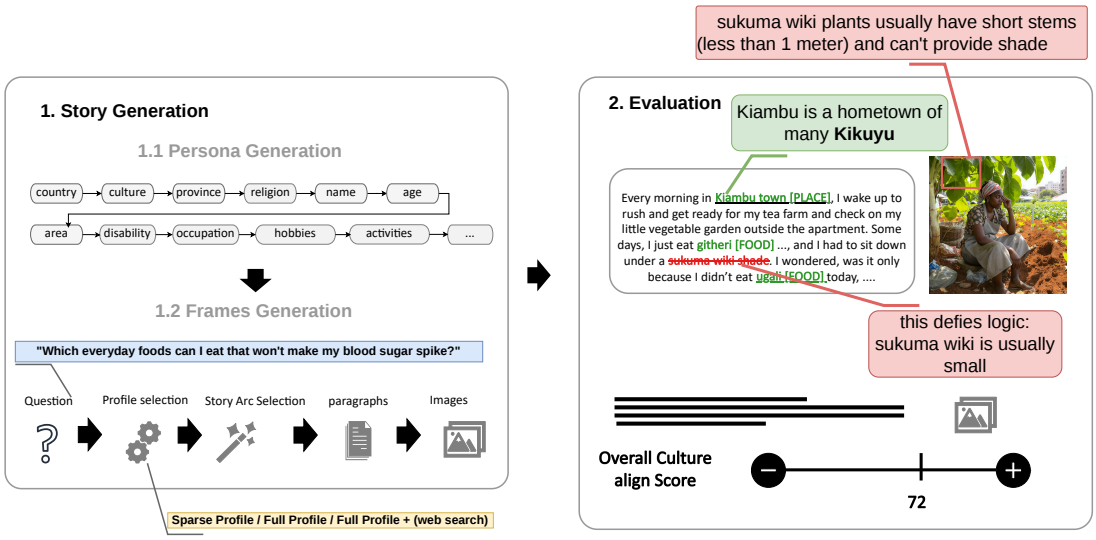


Figure 2: Generation and Evaluation workflow. Green highlights indicate culturally appropriate elements; red highlights indicate violations identified by culture representatives.

agnosis or treatment. The questions are listed in Appendix Table 4.

Frame generation. A story consists of four first-person frames, where each frame contains a short paragraph paired with a corresponding image. Text generation was conditioned on the generated persona, selected diabetes lifestyle question (Appendix Table 4), predefined narrative arc (Appendix Table 10), and one of three context settings used to vary the information available to the model during generation (Appendix Table 5). The prompt instructed models to write each story in a conversational, first-person style, using English for Hausa, Kikuyu, Luo, and AmaXhosa personas, and Portuguese for Xichangana personas.[2] The system and user prompts used for story generation are provided in Appendix Figures 9 and 10. We used seven generation models spanning model families, parameter scales, and access types to produce a diverse set of AI-generated stories across larger and smaller, open-weight and proprietary systems: GPT-4.1, GPT-4.1-mini, Gemma-3 27B, Gemma-3 4B (Team et al., 2025), Qwen3-4B (Yang et al., 2025), Llama 3.1 8B, and Llama 3.3 70B (Grattafiori et al., 2024). For image generation, we used FLUX.1-Kontext-pro (Labs et al., 2025). The first image was generated from the first paragraph, and subsequent images were produced through iterative image editing conditioned on the previous frame, using a fixed seed of 12345 and a low editing strength of 0.3 to support character consistency and visual continuity. The final dataset contains 199 multimodal stories across five communities, seven generation models, and three context settings.

3.2 Story Evaluation

3.2.1 Evaluation Procedure

We designed the evaluation procedure to capture cultural alignment at multiple levels, including the overall story, each text-image frame, and which specific cultural markers within each frame shaped evaluators' judgments. The procedure was informed by prior work on operationalizing culture through demographic and semantic proxies, which we adapted into cultural marker categories for text and image annotation (Adilazuarda et al., 2024); the full category list is provided in Appendix Table 7. These categories covered markers such as language, cultural group, region, religion, socioeconomic context, age and gender roles, occupation, food and dietary norms, physical activity norms, kinship and social structure, community practices, social etiquette, values and beliefs, and attitudes toward health.

Because existing annotation tools did not support the full workflow required for multimodal cultural alignment evaluation, we designed a custom annotation platform for this study. The platform supports frame-level evaluation of text and images,

[2]We used English for Hausa, Kikuyu, Luo, and AmaXhosa stories because English is widely used for formal written communication and education in Nigeria, Kenya, and South Africa. We used Portuguese for Xichangana stories because the study context is Mozambique, where Portuguese serves this role.

text-span highlighting, image-region selection, cultural marker categorization, connectedness labels, influence ratings, optional comments, and story-level alignment scoring. These features separate overall judgments from the specific textual and visual evidence behind them, allowing evaluators to indicate not only whether a story felt aligned or misaligned, but which cultural markers shaped that judgment and how strongly. Figure 5 shows the interface used by culture representatives.

At the frame level, evaluators assessed how culturally connected each paragraph and accompanying image felt to the represented community using three response options: *Not connected*, *Connected*, and *Strongly connected*. They then identified the specific text spans or image regions that influenced their judgment, assigned each selected marker a cultural category (see Figure 6), indicated whether the marker was culturally connected or not connected, rated how strongly it influenced their judgment, and optionally provided a comment. We also provided audio to culture representatives as an accessibility aid for reviewing the story, rather than as an evaluated modality.

After completing all four frames, evaluators completed a story-level feedback checklist covering cultural fit, persona consistency, image reliability, safety, and overall story experience. They then assigned an overall cultural alignment score from 0 to 100 using an anchored rubric (see Figure 7). The rubric ranged from no recognizable cultural markers or cultural relevance to strong cultural alignment with high relevance, coherence, and authenticity in relation to the represented community (Table 8). We used this graded score because cultural alignment is not binary. A story may contain recognizable cultural markers while still varying in relevance, integration, and multimodal coherence. The same story-level scale was used for LLM judges and culture representatives to support comparison between automated and community-grounded evaluations.

3.3 LLM Judge Evaluation

We evaluated each generated story with LLM judges to test whether automated evaluators can provide a scalable approximation of community judgments of cultural alignment. We used five multimodal LLM judges: Kimi-K2.6, Gemma-4-31B, GPT-5.5, Qwen-3.5-9B, and Qwen-3.5-122B. These models were selected to cover a range of model families, scales, and access types, including proprietary and open-weight systems, while supporting evaluation of text, images, and text-image coherence. No exact generation model was reused as an LLM judge. Each judge evaluated the same story set using the same story-level cultural alignment rubric used by culture representatives. The judge prompts also included the same cultural marker categories used in the community evaluation, but did not include community-specific calibration examples, as our goal was to evaluate whether general rubric-based judges could approximate community judgments without prior community-specific calibration. We conducted three independent evaluation runs per judge to account for variability in model outputs, and averaged the three runs into a single judge score for each story before comparing automated scores with scores from culture representatives. The full judge prompt is provided in Figure 11 and Figure 12.

3.4 Community Evaluation

Participants and Communities. We recruited 19 participants aged 18–44 who served as culture representatives across five African cultural communities: Hausa (Northern Nigeria), AmaXhosa (Eastern Cape, South Africa), Luo (Western Kenya), Kikuyu (Central Kenya), and Xichangana (Maputo, Mozambique), as shown in Appendix Figure 4. Participants were selected based on self-identification with the target community, lived experience in the relevant cultural region, and native or regular use of the relevant community language. They were recruited through community networks, including referrals from local collaborators, professional contacts, and community-based networks connected to the represented cultural groups. All participants provided written informed consent, received an internet allowance before the study to support participation, and were compensated with gift vouchers upon completion. Community context and participant demographics are provided in Appendix A and Appendix Table 3.

Individual Cultural Evaluation. Culture representatives completed a structured onboarding and practice annotation process. We held a one-hour onboarding session to introduce the study goals, define key terminology, explain the cultural marker categories, and demonstrate the annotation platform. Representatives then completed a one-week training phase, during which they annotated 20 practice stories using the same interface and eval-

uation procedure used in the main study. After this phase, we held a one-hour discussion session to review examples of agreement and disagreement, clarify annotation expectations, and support a shared understanding of the evaluation task while still allowing community-specific interpretations. Culture representatives then individually evaluated 40 stories generated for their own communities using the evaluation procedure described above. In total, the individual evaluation produced 18,805 marker-level annotations across 199 stories, including 9,386 text span annotations (Appendix Table 13) and 9,419 image region annotations (Appendix Table 14).

Focus Group Discussions. Following individual annotation, we conducted 15 focus group sessions, with three sessions per community. Each session lasted approximately two hours, for a total of approximately 30 hours of recorded discussion. Sessions were conducted over Microsoft Teams, audio-recorded, and transcribed.

The focus groups were designed to collect qualitative explanations and examine how representatives reasoned through agreement and disagreement in their individual annotations. In the first session, representatives discussed stories where their annotations showed broad agreement, helping establish shared vocabulary for cultural alignment and misalignment within each community. In the second session, representatives examined stories where their annotations diverged, surfacing implicit norms, contested expectations, and differences in how cultural dimensions were weighted. These discussions allowed representatives to revisit details they may have missed during individual annotation, clarify the reasoning behind their judgments, and decide whether to maintain or revise their interpretations. In the third session, representatives reflected across the full set of evaluated stories, identifying the strongest and weakest examples and distinguishing meaningful cultural integration from superficial decoration. The guiding questions are provided in Appendix Table 9. These discussions served as a deliberative evaluation method, producing the situated interpretations from which the taxonomy in Section 3.5 was derived.

3.5 Taxonomy Derivation

We developed the cultural alignment taxonomy through a convergent mixed-methods design (Creswell and Plano Clark, 2018), conducting thematic analysis (Braun and Clarke, 2006, 2019) of the FGDs to derive the taxonomy, and then examined how the resulting categories appeared in the participant annotations. Each FGD was conducted with two to three of the authors present, during the session they took independent notes. After the session each transcript was put into HeyMarvin, a qualitative research tool, for transcription and analysis. Each transcript was anonymised and read in full by at least one author. The authors noted emergent categories from their notes and during their readings, extracted the examples from the transcripts which fell into these categories. Five of the authors then conducted 5 joint analysis sessions where they discussed the categories and examples together, and from these identified emergent themes with verifiable examples. These themes formed the basis of the cultural alignment taxonomy and the recurring misalignment mechanisms summarized in Table 1. The cultural marker categories organize the annotated text and image markers, labeled using the fine-grained categories in Appendix Table 7, into broader categories. The misalignment mechanisms capture recurring failure patterns identified from representatives' reasoning in the focus group discussions.

	Taxonomy category	Description and examples
Cultural marker categories	Referential	Explicit cultural references, such as names, foods, places, objects, and institutions.
	Procedural	How everyday practices are carried out, such as food preparation, care-seeking, exercise routines, and household activities.
	Contextual	Whether cultural elements fit the situation, timing, purpose, or setting in which they appear.
	Socio-geographic	Place-based social and material conditions, such as clinics, markets, homes, infrastructure, mobility, and local environments.
	Linguistic register	Community-specific language use, including dialect, code-switching, forms of address, tone, and community voice.
Misalignment mechanisms	Substitution	Markers from another community replace target-community markers.
	Norm violation	Content contradicts community norms, expectations, or respectful conduct.
	Omission	Expected cultural markers or situated details are absent, making the story feel generic.
	Forced insertion	A plausible cultural marker appears without meaningful integration into the story or context.
	Register conflation	The story uses the wrong language variety, dialect, tone, or form of address.
	Cross-modal inconsistency	Text and image provide conflicting cultural cues.
	Stereotyping	The story or image reduces a community to narrow, repeated, or generic tropes.
	Hallucination	The model fabricates or misattributes a cultural practice, object, or meaning.

Table 1: Summary of the cultural alignment taxonomy derived from focus group discussions and participant annotations (refer to Figure 8 for examples).

4 Results

4.1 Community Evaluation Reveals Patterned Cultural Alignment and Misalignment

The uneven judge results show that automated scores alone cannot explain cultural alignment. Community evaluation adds this missing layer by showing which textual and visual markers representatives used to judge cultural fit, and how those markers supported, weakened, or disrupted alignment in context. We organize these results using the taxonomy summarized in Table 1, first examining the five broader cultural marker categories across modalities and then describing the eight recurring mechanisms of misalignment.

4.1.1 Cultural Marker Patterns Across Modalities

Marker patterns varied across communities, modalities, and the five broader cultural marker categories. At the marker level, representatives labeled selected text spans and image regions as *Connected* when they supported cultural fit and *Not connected* when they weakened or disrupted cultural fit. Among text span annotations, 8,075 of 9,386 (86%) were labeled connected and 1,311 (14%) were labeled not connected. Image region annotations showed more visible disruption, with 6,931 of 9,419 (74%) labeled connected and 2,488 (26%) labeled not connected. Full text-span and image-region statistics are provided in Appendix Tables 13 and 14. Figure 3 descriptively shows the distribution of not-connected markers across communities, models, modalities, and cultural marker categories, while Appendix Figure 13 shows the corresponding distribution of connected markers. These figures show that cultural fit and cultural disruption are not evenly distributed across modalities. In text, not-connected markers were often concentrated in different categories for different communities. Hausa stories showed frequent not-connected markers in *Referential* and *Procedural* categories, while Xichangana stories showed a more even spread across categories, including *Linguistic Register*. In images, not-connected markers were often more visually concentrated, especially around character appearance, clothing, food, setting, public space, and other visual details. These patterns show that the visual realization of a story can support, weaken, transform, or contradict cultural cues in the narrative, making evaluation across both text and images necessary.

4.1.2 Misalignment Mechanisms

The focus group discussions revealed eight recurring mechanisms through which stories became culturally misaligned. **Substitution** occurred when stories included markers from another community in place of markers from the target community. For instance, *sadza*, a Zimbabwean food, was independently flagged in three non-Zimbabwean communities: Kikuyu, Luo, and AmaXhosa. **Hallucination** occurred when models used cultural vocabulary but attached it to implausible practices. Hausa participants flagged an incorrect food preparation description, "boil your tuwo shinkafa instead of frying," noting that frying was not a way they would cook this dish. An AmaXhosa participant similarly found *umxhentso*, a traditional dance, described in a story as a food and commented, "This is not an error a human storyteller would make. This is very AI." **Forced insertion** occurred when correct names or cultural terms appeared in otherwise generic narratives without meaningful integration, as when a single Kikuyu name was placed in a story where surrounding markers belonged to other cultures, making the name feel "thrown in."

Norm violation appeared when outputs contradicted expectations around respect, dress, or social interaction. A Xichangana participant explained, "ninguém entra no hospital, participa de uma consulta de chapéu, considera-se como uma falta de respeito" (nobody enters a hospital wearing a hat; it is considered a lack of respect). **Register conflation** was especially visible in Xichangana stories, where Portuguese was fluent but regionally inappropriate. One participant observed, "usava-se muito o gerúndio e o gerúndio é muito característico dos brasileiros" (the gerund was used a lot, and the gerund is very characteristic of Brazilians). **Cross-modal inconsistency** occurred when text and image represented conflicting settings, identities, or practices, while **stereotyping** appeared when models reduced communities to narrow or repeated visual tropes.

Finally, **omission** appeared when stories contained no obviously wrong elements but also few cultural details. Participants across communities described such stories as generic rather than factually incorrect. An AmaXhosa participant called one story "a story anybody from any culture could narrate," a Luo participant described another as "not grounded culturally," and a Xichangana participant said, "não encontrei marcadores" (I did not find

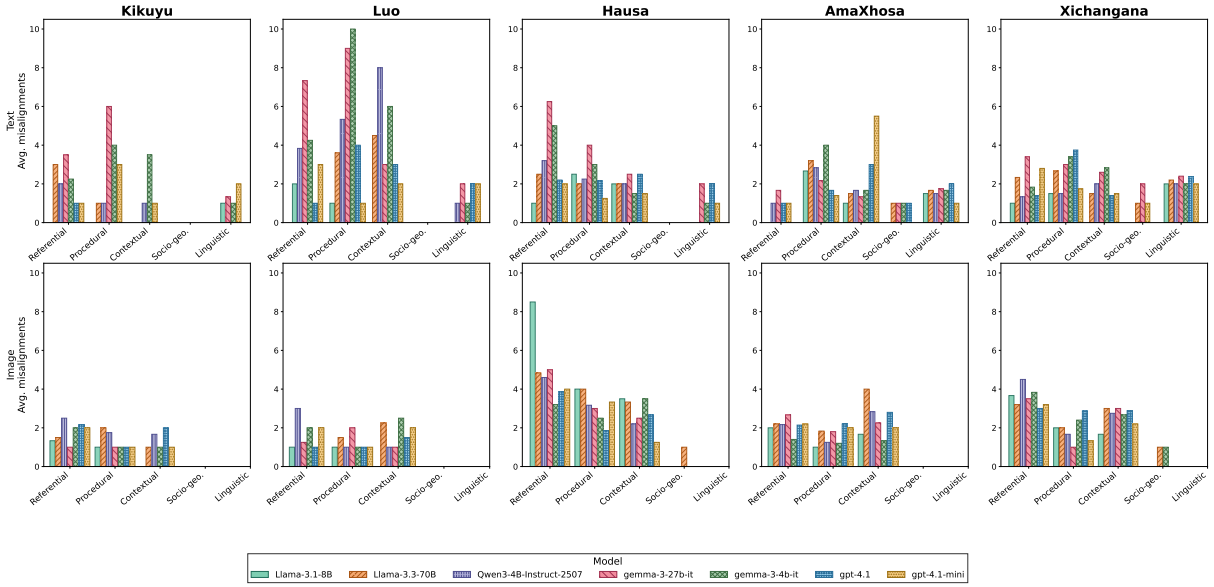


Figure 3: Distribution of not-connected cultural markers in text and images across the five broader cultural marker categories, broken down by generation model and community. Connected marker distributions for both modalities appear in Appendix Figure 13.

markers). These mechanisms show that cultural misalignment is not only a matter of factual error, but also of weak integration, inappropriate context, cross-modal mismatch, wrong register, stereotyping, and absence of situated detail. See Appendix Figure 8 for additional examples.

4.2 LLM Judges Approximate Community Judgments Unevenly Across Cultures

Community	Kimi	Gemma4	GPT-5.5	Qwen-122B	Qwen-9B
Hausa	**0.74**	**0.67**	**0.70**	**0.80**	**0.76**
AmaXhosa	0.33	0.26	0.28	0.29	0.40
Kikuyu	**0.54**	**0.50**	**0.51**	**0.64**	**0.57**
Luo	**0.89**	**0.82**	**0.88**	**0.87**	**0.87**
Xichangana	**0.49**	**0.45**	**0.51**	**0.49**	0.39

Table 2: Pearson correlation coefficients (r) between LLM judge scores and mean scores from culture representatives by community. Bold values remain significant after Bonferroni correction ($p < 0.002$); unbolded values do not, even when significant at uncorrected thresholds ($p < 0.001$, $p < 0.01$, or $p < 0.05$). No AmaXhosa judge remains significant after correction.

Judge correlations with community scores varied across the five communities. For Luo, all five judges achieved very strong Pearson correlations with scores from culture representatives ($r = 0.82$–0.89), and for Hausa correlations remained strong ($r = 0.67$–0.80). Kikuyu and Xichangana fell in moderate ranges ($r = 0.50$–0.64 and $r = 0.39$–0.51, respectively). For AmaXhosa, no judge remained significant even after Bonferroni correction, indicating that automated scores did not reliably track community judgments for this community.

We first assess the reliability of the community ratings that serve as the reference point for judge validation and grounding analysis. Inter-annotator agreement on the overall cultural alignment score varied by community, ranging from good for Hausa (ICC(A,1)=0.81), to moderate for Luo, AmaXhosa, and Kikuyu (ICC(A,1)=0.60–0.63), to poor for Xichangana (ICC(A,1)=0.33). Xichangana's low agreement reflects a compressed score distribution (mean=24.5, SD=4.5), where representatives broadly rated stories as culturally inadequate but diverged on finer distinctions. All ICC values were statistically significant ($p < 0.001$; Table 12).

Score bias further limits automated evaluation. For Xichangana, judges assigned alignment scores 35–45 points higher than culture representatives (paired t-test, all $p < 0.001$), assigning scores in the 60–69 range to stories that culture representatives rated at 24.5 on average. Kikuyu showed the opposite pattern, with Kimi and GPT-5.5 assigning scores 14–16 points lower than culture representatives. Judges were best calibrated on Hausa and showed smaller, mixed biases on Luo. Because both correlation with community scores and score bias varied by culture, no single judge model performed consistently across all five communities. Table 2 reports judge–community correlations, and Appendix Table 11 reports score bias for all judge–

community pairs.

5 Discussion

Community evaluations show that cultural alignment is not reducible to the number or presence of recognizable markers. Culture representatives considered whether names, foods, places, clothing, language cues, social interactions, and visual settings fit the story context and the represented community. The taxonomy makes this distinction explicit by separating the broader cultural marker categories that representatives attended to from the mechanisms through which stories became misaligned. This matters for evaluation because a story may contain plausible markers while still failing through substitution, forced insertion, cross-modal inconsistency, wrong register, or absence of situated detail.

Our results also show that LLM judges do not approximate community judgments consistently across cultures. Judge reliability and score calibration varied substantially across communities, with no single judge performing consistently across all five settings. This variation means that automated judges should be treated as scalable evaluators only after community-specific validation. For cultural alignment tasks, both correlation with community judgments and score calibration should be considered, since a judge may track differences between stories while still systematically assigning scores that diverge from those of culture representatives.

These findings point toward community-calibrated evaluation pipelines for cultural alignment. In such pipelines, culture representatives provide the reference judgments needed to establish where automated evaluators can be trusted and where human review remains necessary. The taxonomy also makes community feedback more actionable for scalable evaluation: the five broader cultural marker categories identify what cultural dimensions automated judges should attend to, while the eight misalignment mechanisms describe recurring failure patterns that can guide review and calibration without replacing community judgment. Although the taxonomy was derived from five African communities in the context of diabetes lifestyle stories, its structure provides a basis for investigating whether similar cultural dimensions and misalignment mechanisms emerge in other communities and domains.

6 Conclusion

We presented a human-centered evaluation of cultural alignment in AI-generated multimodal stories across five African communities. Community evaluations showed that cultural alignment depends not only on the presence of recognizable cultural markers, but on how those markers fit narrative, social, linguistic, procedural, and visual context. From annotations and focus group discussions, we developed a taxonomy of five broader cultural marker categories and eight recurring mechanisms of misalignment. We further evaluated multimodal LLM judges against culture-representative judgments and found that their reliability and score calibration vary substantially across communities, with no single judge performing consistently across all five settings. These findings point toward community-calibrated evaluation pipelines in which community judgments ground cultural alignment evaluation and automated judges are validated to determine where they can be trusted and where human review remains necessary.

Limitations

Our study has several limitations that point to important directions for future work. First, our community evaluation relies on 19 culture representatives. These participants provide situated lived and linguistic knowledge, but they do not represent the full diversity of views within each community. Cultural alignment is not fixed or uniform, and perspectives may vary across age, gender, region, language use, religion, class, and rural-urban experience. Future work should build broader and more iterative community review processes that include more diverse participants and examine how cultural alignment judgments vary within communities.

Second, our evaluation focuses on story-level and marker-level cultural alignment rather than downstream effects on readers. We do not test whether culturally aligned stories improve understanding, trust, behavior-change outcomes, or lifestyle decision-making. Future work should connect cultural alignment evaluation to reader-facing outcomes, asking not only whether communities recognize a story as culturally coherent, but also whether such coherence changes how people interpret, trust, adapt, or use AI-generated content.

Third, our study is grounded in everyday Type II diabetes lifestyle and behavior-change stories. This domain was chosen because diet, physical

activity, household routines, and care practices are culturally situated, but the cultural marker patterns, misalignment mechanisms, and judge behavior observed here may differ in other domains. Future work should therefore validate the framework across other kinds of generated content and application settings.

Fourth, cross-community comparisons should be interpreted carefully. Xichangana stories differed from the other communities in both language context and evaluation patterns. They were generated in Portuguese for the Mozambican context, received much lower community scores on average, showed lower agreement among representatives, and were substantially over-scored by LLM judges. These differences make Xichangana an important case for understanding failures of automated evaluation, but they also caution against interpreting score differences across communities as direct measures of cultural alignment difficulty. Future work should examine how language choice, regional context, and local norms shape cultural alignment judgments.

Finally, our LLM judge analysis covers five multimodal judges and a fixed evaluation prompt. Judge reliability may change with different models, prompts, rubrics, or calibration methods. Rather than seeking a universal automated judge for cultural alignment, future work should develop community-calibrated evaluation pipelines that identify where judge scores can be trusted, where they are biased, and where human review remains necessary. Future work can also test whether community-derived taxonomies improve scalable evaluation by incorporating examples of cultural marker categories and misalignment mechanisms into judge prompts, audit protocols, or human review workflows, while validating these approaches separately in each community context.

Ethical Considerations

The generated stories were evaluated with the participation of culture representatives from each respective community. It is important to note that the stories do not provide medical diagnoses, prescriptions, or recommendations related to medication. Instead, they focus exclusively on culturally contextualized lifestyle and behavior-change narratives, particularly around healthy eating habits, physical activity, and everyday practices associated with managing Type II diabetes. Cultural alignment should not be interpreted as evidence of medical correctness or safety, and health-related generated content should undergo appropriate medical review in addition to community-based cultural evaluation.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025a. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–21, New York, NY, USA. Association for Computing Machinery.

Dhruv Agarwal, Anya Shukla, Sunayana Sitaram, and Aditya Vashistha. 2025b. Fluent but Foreign: Even Regional LLMs Lack Cultural Alignment. *arXiv preprint*. ArXiv:2505.21548 [cs].

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Kirti Bhagat, Shaily Bhatt, Athul Velagapudi, Aditya Vashistha, Shachi Dave, and Danish Pruthi. 2026. Tales: A taxonomy and analysis of cultural representations in llm-generated stories. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5)*.

John W. Creswell and Vicki L. Plano Clark. 2018. *Designing and Conducting Mixed Methods Research*, 3rd edition. SAGE Publications, Thousand Oaks, CA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A,

Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Leslie J. Hinyard and Matthew W. Kreuter. 2007. Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior*, 34(5):777–792.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.

Nari Johnson, Deepthi Sudharsan, Hamna, Samantha Dalal, Theo Holroyd, Anja Thieme, Hoda Heidari, Daniela Massiceti, Jennifer Wortman Vaughan, and Cecily Morrison. 2026. Evaluating ai-generated images of cultural artifacts with community-informed rubrics. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, page 714–774, New York, NY, USA. Association for Computing Machinery.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. In *Advances in Neural Information Processing Systems*, volume 37, pages 13716–13747. Curran Associates, Inc.

Sharif Kazemi, Gloria Gerhardt, Jonty Katz, Caroline Ida Kuria, Estelle Pan, and Umang Prabhakar. 2024. Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation. *Preprint*, arXiv:2410.10489.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2151–2165, New York, NY, USA. Association for Computing Machinery.

Sarah Kiden, Oriane Peter, Gisela Reyes-Cruz, Maira Klyshbekova, Sena Choi, Aislinn Gomez Bergin, Maria Waheed, Damian Eke, Tayyaba Azim, Sarvapali Ramchurn, Sebastian Stein, Elvira Perez Vallejos, Kate Devlin, and Joel E Fischer. 2025. Back to the communities: A mixed-methods and community-driven evaluation of cultural sensitivity in text-to-image models. *Preprint*, arXiv:2510.27361.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language*

Processing, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. 2026. Culture in action: Evaluating text-to-image models through social activities. In *The Fourteenth International Conference on Learning Representations*.

Tarek Naous, Michael J. Ryan, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

Mary Ng'endo and Esther Kariuki. 2026. Voices of change: public narrative storytelling communicates climate resilience actions in kenya. *Development in Practice*, 36(1):106–123.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. 2025. CuRe: Cultural Gaps in the Long Tail of Text-to-Image Systems. In *International Conference on Computer Vision*, pages 15680–15691.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIextraction 2013). In *Proceedings of SemEval 2013*, pages 341–350. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahra-

mani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Anja Thieme, Rita Faia Marques, Martin Grayson, Sidhika Balachandar, Cameron Tyler Cassidy, Madiha Zahrah Choksi, Camilla Longden, Reeda Shimaz Huda, Nicholas Ileve Kalovwe, Christina Mallon, Courtney Mansperger, Daniela Massiceti, Bhaskar Mitra, Ruth Mueni Nzioka, Ioana Tanase, Yuzhe You, and Cecily Morrison. 2026. Engaging communities meaningfully in defining disability representation for ai image generation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

A Communities and Culture Representatives

Figure 4 situates the five communities included in this study across the African continent, and Table 3 summarizes participant demographics.

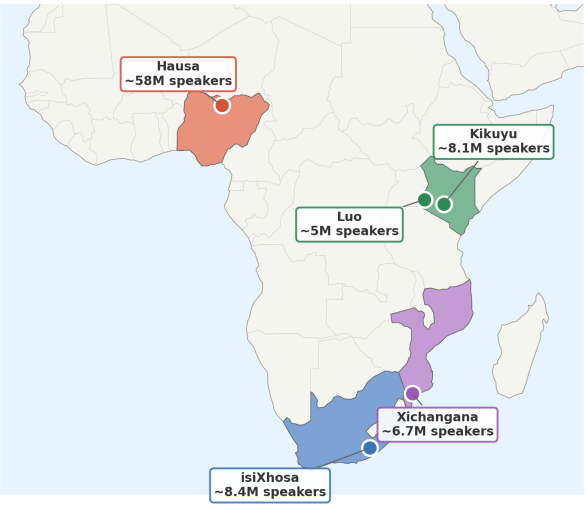


Figure 4: Geographic distribution of the five African communities included in our study.

Our study spans five African cultural communities across West, East, Southern, and Southeast Africa: Hausa in Nigeria, Kikuyu and Luo in Kenya, AmaXhosa in South Africa, and Xichangana in Mozambique. These communities differ in language, geography, religious and historical context, naming practices, foodways, dress norms, and everyday social expectations. We therefore treat them as distinct evaluative contexts.

The Hausa community in this study is situated in northern Nigeria, where Islamic practice, market life, Hausa naming conventions, modest dress expectations, and foods such as *tuwo*, *masa*, and *miyan kuka* shape everyday cultural interpretation. The Kikuyu community is situated in central Kenya, where stories often draw on settings such as Kiambu, agricultural livelihoods, local foods, family relations, and Kikuyu naming practices. The Luo community is situated in western Kenya, where Luo names, kinship relations, lakeside and rural–urban settings, foods, and social forms of respect provide important cultural context. The AmaXhosa community is situated in South Africa, where AmaXhosa language practices, kinship relations, forms of address, dress, and institutional interactions shape how stories are interpreted. The Xichangana community is situated in Mozambique, where local Portuguese usage, Xichangana language and expressions, food practices, respect norms, and public behavior distinguish the community from both European Portuguese and other African contexts.

Community	n	Age Range	Gender
Hausa	3	30–39	3F
Kikuyu	4	18–34	3F, 2M
Luo	6	18–44	4F, 2M
AmaXhosa	2	25–29	2F
Xichangana	4	18–44	3F, 1M
Total	19	18–44	15F, 5M

Table 3: Culture representative demographics by community.

B Story Generation Details

This appendix provides supplementary materials for the story generation pipeline described in Section 3.1, including the workflow, persona fields, lifestyle questions, story distribution, and narrative arcs.

Category	Question
Diet	Which everyday foods can I eat that will not make my blood sugar spike? I like to snack—what are some healthy options, and how often should I have them? I sometimes eat late at night. How does this affect my blood sugar, and what can I do instead? Can you give me some simple meal ideas or a sample daily menu I could follow? Eating healthy can get expensive—how can I maintain a nutritious diet on a tight budget? Can I still enjoy traditional foods?
Exercise	What kinds of exercises are safe and beneficial for someone with diabetes? Can I rely only on walking for exercise, or should I include other types of physical activity? How often and for how long should I exercise to improve my health? I do not have access to a gym—what activities can I do at home or with everyday items to stay active? Sometimes I feel tired or discouraged—how can I stay motivated, and how does exercise help control blood sugar?

Table 4: Everyday diabetes lifestyle questions used for narrative generation.

Context setting	Information provided during generation
Sparse persona	A reduced persona profile in which approximately half of the persona attributes were removed while preserving core demographic and cultural information.
Full persona	The complete generated persona profile.
Full persona + retrieval	The complete generated persona profile together with retrieved web context. GPT-4.1 generated search queries from the persona profile, and the retrieved information was provided as additional context for story generation.

Table 5: Context settings used to vary the amount of persona and cultural information available during story generation.

#	Persona field
1	Culture (select all that apply)
2	Which province do you live in? (select 1)
3	What is your religion?
4	What is your preferred name?
5	Gender (select 1)
6	Age (select 1)
7	How would you describe your local community or area? (select all that apply)
8	How would you describe access to food in your local community or area? (select all that apply)
9	Do you have a disability? (select all that apply)
10	What is your current occupation? (select 1)
11	What best describes your daily routine right now?
12	What hobbies or interests do you enjoy regularly? (select all that apply)
13	Which activities or routines are most important to you?
14	Do you currently have other health conditions besides diabetes? (check all that apply)
15	Are you currently taking medication or following a treatment for diabetes?
16	What kind of healthcare resources do you regularly access or use?
17	Do you have regular access to healthcare services?
18	How often do you access healthcare services? (choose 1)
19	Are you currently living with Type 2 Diabetes?
20	Do you smoke tobacco products?
21	If yes, how often? (choose 1)
22	Do you consume alcoholic beverages?
23	If yes, how often? (choose 1)
24	How many days per week do you participate in moderate to vigorous physical activity? (choose 1)
25	Select any dietary preferences or restrictions you may have (choose 1)
26	Who forms part of your regular support network? (check all that apply)
27	What are your main personal health goals? (You may select more than one)
28	How do you prefer to receive information or stories about your health? (check all that apply)

Table 6: Persona fields used to construct culturally situated personas.

Category	Description	Examples
people_appearance	skin tone, hair, age, who is present, visual identity cues	An elderly woman with grey hair greets a young boy at the gate.
names_forms_of_address	forms of address that signal identity or relationships	"Auntie Mary" calls him "my child" when he arrives home.
language_local_expression	slang, code-switching, local phrases: "Ag", "mos", "lekker"	"Ag, don't worry," she says, switching between languages mid-sentence.
food_dietary_practices	foods, preparation methods, meal styles, eating habits	The family shares stiff porridge and vegetables for dinner.
place_physical_environment	province, home, clinics, kitchen, shops, neighborhoods, local setting	The small shop near the dusty road opens early every morning.
healthcare_community_practices	clinic visits, nurses, advice-seeking norms, care pathways	She visits the clinic after neighbors advise her to check her fever.
socio_economic_context	budgeting, shopping specials, affordability, access to food/resources	He counts coins carefully before buying bread for the week.
family_household_structure	family roles, caregiving, shared meals, household dynamics	Grandmother cooks while the children set the table.
occupation_daily_routine	jobs, lunch prep, meetings, time pressure, daily responsibilities	The teacher prepares lessons late at night after working all day.
health_values_beliefs	motivation, balance, "small tweaks" mindset, change	She decides to walk daily to "stay strong and healthy."
clothing_fashion	clothing, accessories, style, trends	He wears a school uniform neatly pressed for assembly.
social_roles_hierarchy	age, gender, seniority, authority, respect norms	The youngest waits silently while elders speak first.
community_social_networks	neighbors, mutual aid, social obligations, communal support	Neighbors come together to help rebuild a burned house.
gender_norms_expectations	gendered responsibilities, speech norms, social constraints	She is expected to cook while her brother studies.
life_stage_transitions	childhood, adulthood, parenting, aging-related roles	After graduating, he begins supporting his younger siblings.
temporal_orientation	time-use norms, punctuality, event-based vs clock-based time	The meeting starts "after prayers," not at a fixed hour.
seasonality_environmental_cycles	weather, harvest cycles, seasonal constraints	During the rainy season, roads become difficult to use.
daily_rhythms_meal_patterns	when activities and meals occur during the day	Breakfast is early, before everyone leaves for work and school.
religious_spiritual_practices	faith traditions, rituals, moral frameworks	The family gathers for evening prayer before dinner.
traditional_indigenous_knowledge	local healing, food knowledge, environmental understanding	Her grandmother teaches her which leaves cure stomach aches.
health_illness_explanatory_models	causes, symptoms, and recovery understanding	They believe the illness came from cold night air.
values_moral_frameworks	what is considered good, responsible, shameful, or desirable	"Helping others is more important than keeping everything for yourself."
technology_media_use	phone sharing, platform preference, data constraints, media habits	They share one phone to check messages and news.
transport_mobility	commuting methods, travel norms, access constraints	She takes a crowded minibus to work every morning.
household_resources_tools	appliances, cooking tools, storage, water/electricity access	The family cooks on a small charcoal stove outside.
communication_style	directness, politeness strategies, humor, indirect speech	He hints politely instead of saying no directly.
emotional_expression_norms	what emotions are shown or hidden	She smiles in public but worries quietly at home.
storytelling_narrative_forms	personal stories, advice-giving, metaphor use	He tells a story about his childhood to give advice.
social_norms_taboo	topics avoided, prohibited behaviors, euphemisms	No one speaks openly about conflict during the family gathering.
institutional_trust_relations	attitudes toward government, clinics, NGOs, authority	They wait in doubt, unsure if the clinic will help.
silence_implicit_knowledge	assumed knowledge, omissions, things left unsaid	Everyone understands the problem without it being explained.
animals_human_interaction	livestock, pets, wildlife roles in livelihood	Goats roam freely near the homestead and are herded back at dusk.
plants_agriculture_food_sources	crops, farming practices, gardens, wild foods	The family harvests maize from their small field.
sports_recreation_physical_activity	organized sports, informal games, physical activity norms	Children play football barefoot in the open field.
arts_entertainment_culture	music, dance, festivals, performance	Drums and dancing fill the village during the celebration.
media_information_ecosystem	radio, TV, news, misinformation, access	The radio announces local news every evening.
built_environment_infrastructure	roads, sanitation, buildings, utilities	A dirt road connects the village to the town.
objects_material_culture	tools, artifacts, everyday items	A worn cooking pot sits over the fire.
symbols_visual_culture	flags, colors, religious symbols, logos	A painted symbol marks the entrance to the community hall.
events_rituals_celebrations	holidays, ceremonies, weddings, funerals	The wedding ceremony continues late into the night with music.
education_learning_practices	schooling, literacy, informal learning	The child practices reading aloud after school.
governance_politics_civic_life	political structures, civic participation	The village leader calls a community meeting.
economy_livelihood_strategies	income sources, trade, markets, subsistence	She sells vegetables at the local market each morning.
environment_ecology_sustainability	resources, conservation, climate adaptation	Farmers adjust planting after noticing changing rainfall.
risk_safety_security	crime, hazards, coping mechanisms	They avoid walking home late due to safety concerns.
migration_mobility_patterns	rural-urban movement, diaspora, seasonal migration	He moved to the city to find work.
identity_ethnicity_culture	ethnic groups, heritage, identity markers	She proudly wears clothing that reflects her heritage.
disability_accessibility	assistive tools, inclusion, barriers	The school adds ramps for wheelchair access.
food_markets_distribution	markets, vendors, supply chains	Fresh produce arrives early at the town market.
water_sanitation_hygiene	WASH practices, infrastructure, hygiene norms	They fetch water from a shared community tap.
energy_fuel_usage	cooking fuel, electricity sources	The household relies on firewood for cooking.
visual_scene_composition	camera angle, framing, indoor/outdoor cues	The photo shows a crowded street from above.
image_artifacts_quality	lighting, blur, resolution, filters	The image is blurry but still recognizable.
branding_advertising_presence	logos, billboards, sponsorship	A billboard advertises a popular drink along the road.
digital_platform_signals	UI elements, apps, emojis, memes	A WhatsApp message includes voice notes and emojis.
interpersonal_distance_body_language	gesture, posture, touch norms	They stand close while talking, showing familiarity.
sound_audio_cues	music, noise, environmental sounds	Roosters crow as the village wakes up.
other	**if a category is missing, please provide**	

Table 7: Cultural marker categories used for both text and image annotation. The same categories were applied across modalities, with annotators selecting text spans or image regions that influenced their cultural alignment judgments.

C Annotation Interface and Evaluation Protocol

This appendix provides the annotation platform, story-level scoring rubric, and focus group guiding questions referenced in Section 3.

Score Range	Interpretation
0	No recognizable cultural markers or cultural relevance.
1–20	Minimal mention of cultural elements without meaningful integration.
21–40	Some cultural markers are present, but integration is weak, generic, or only loosely relevant.
41–60	Moderate cultural reflection with partial relevance to the represented community.
61–80	Most cultural markers are meaningfully incorporated and broadly aligned with the represented community.
81–100	Strong cultural alignment, with high relevance, coherence, and authenticity in relation to the represented community.

Table 8: Overall cultural alignment scoring rubric used by culture representatives and LLM judges.

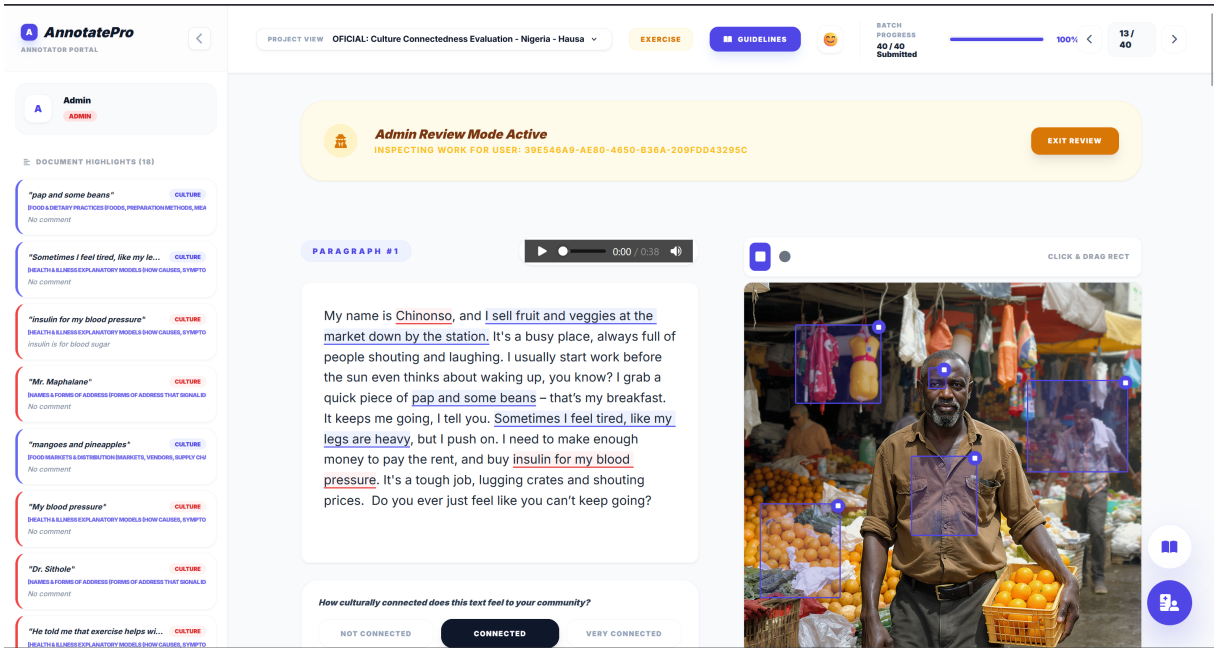


Figure 5: Annotation platform used by culture representatives to evaluate text and image cultural alignment.

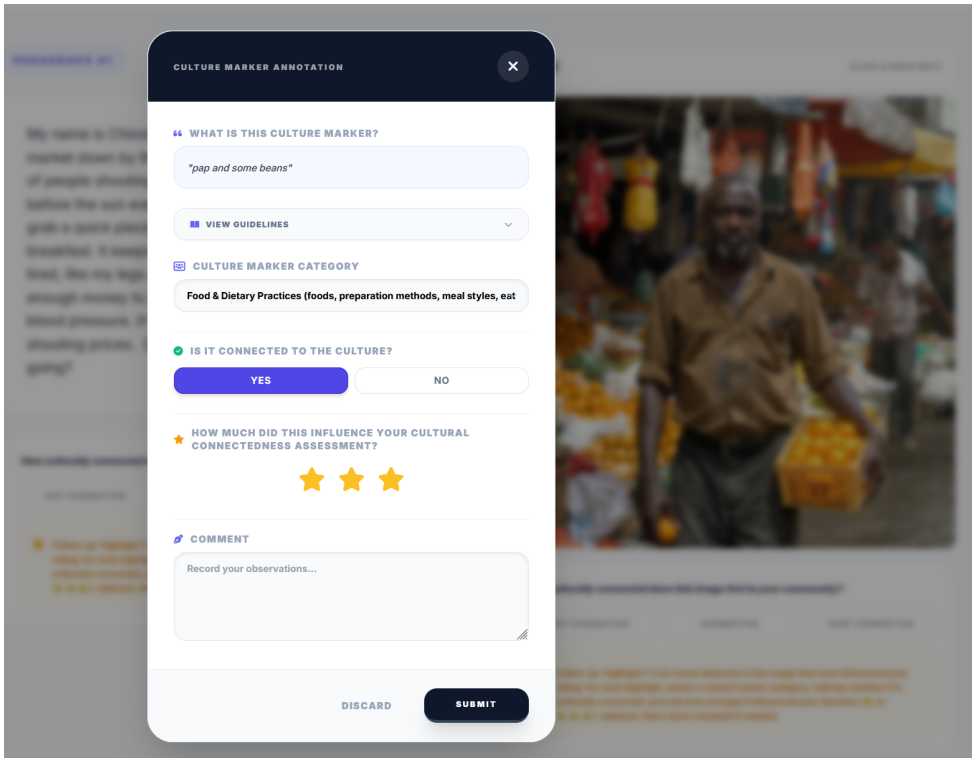


Figure 6: Annotation platform: Assigning a cultural marker category to a highlighted text span or region.

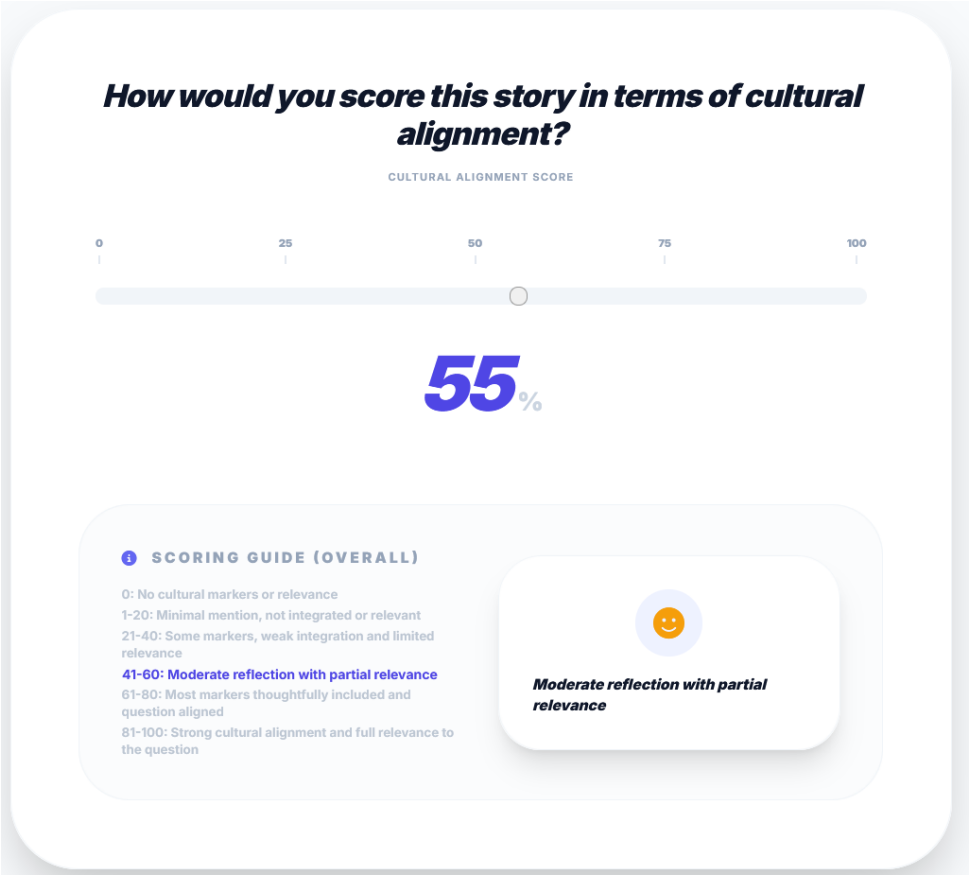


Figure 7: Annotation platform: Assigning overall culture alignment scoring.

Discussion focus	Guiding questions
Initial reactions	1. After reading this story, did it feel like something that could realistically happen in your community? Why or why not? 2. Would someone from your community recognize themselves in this story? What makes you say that? 3. Did anything immediately stand out to you as either strongly connected to your community or clearly not connected? What was it?
Cultural marker identification and misalignment	1. Can you walk us through specific parts of the story that felt "off" or did not ring true for your community? 2. For those moments, what exactly made them feel wrong? Were they incorrect, out of place, too generic, or based on assumptions? 3. Were there important cultural elements that you expected to see but were missing? What were they? 4. Across the stories you reviewed, are there particular types of cultural elements that the AI tends to get wrong or struggle with?
Meaningful integration of cultural markers	1. When cultural elements appeared in the story, such as food, language, or daily practices, did they feel like a natural part of the story, or did they feel added in? Why? 2. Can you give an example of a cultural element that felt genuinely well integrated, and one that felt forced or out of place? 3. Did some stories feel more grounded in your community than others? What specifically made the difference? 4. In stories where cultural elements were present but the story still felt off, what was missing for it to feel truly connected to your community?

Table 9: Guiding questions used in focus group discussions with culture representatives.

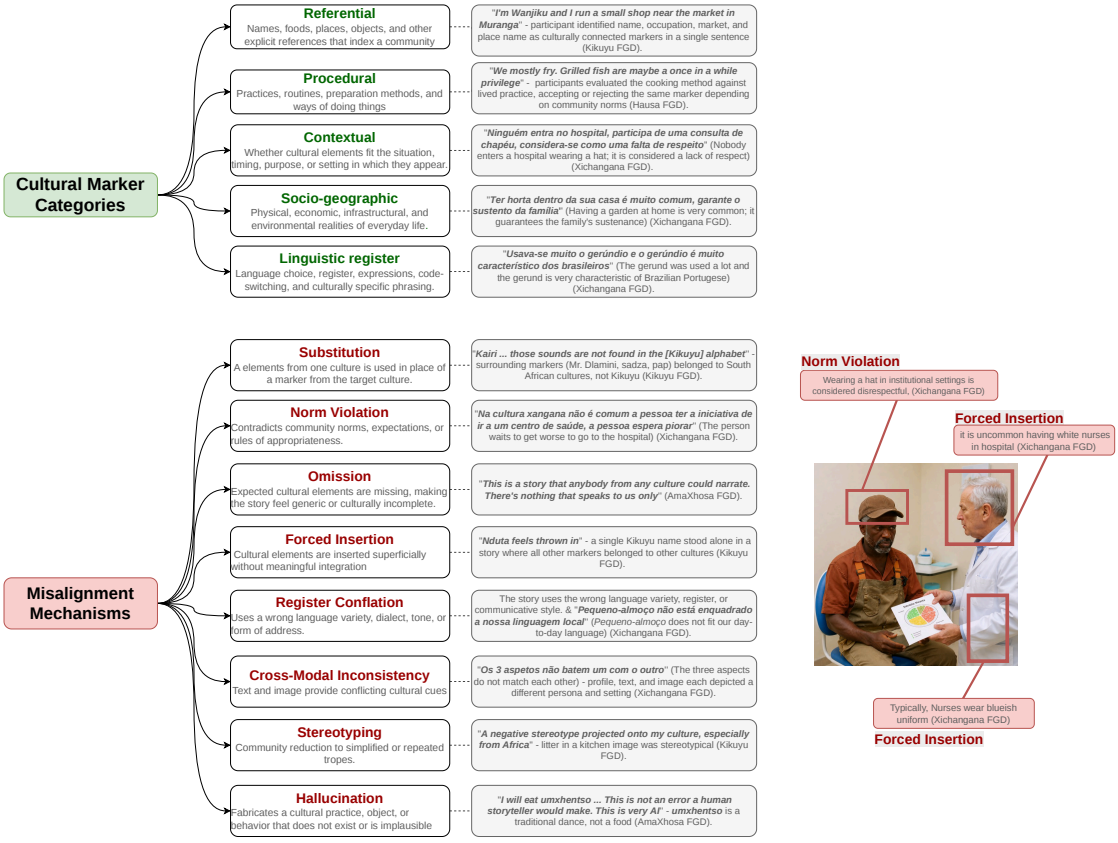


Figure 8: Cultural alignment taxonomy with examples from FGD sessions.

	Paragraph 1	Paragraph 2	Paragraph 3	Paragraph 4
default	Introduce {px_name}'s current **daily life and diet habits** as described above. Mention any **cultural and environmental factors** that make change difficult. The persona currently has a small work-related task that foreshadows a larger work-related task. They fail to accomplish the task because of issues related to health changes they need to make. All motivation to change comes from the need to accomplish the current and future task.	Because of the failed small task, {px_name} needs help. Depending on the given healthcare access & resources, either {px_name} seeks help e.g., visits a **clinic or community health worker**, or help finds them coincidentally e.g., support network. This helper gives practical advice on managing their lifestyle or current challenge. The persona is initially resistant to the change. Motivation to change comes from failure at the small task. The health worker suggests a small change that will help the persona accomplish the task.	Describe **how {px_name} tries to follow the advice given**, facing initial struggles due to certain reason(s). Describe how they overcome these struggles. The persona tries to make the prescribed health change but faces a setback. They try to find alternative ways to pursue the health change in light of the upcoming larger task.	** Highlight **positive but realistic changes**, that result from the new effort they've put in. The persona finds the small changes helpful in health management. This enables them accomplish the larger task. The persona recognises the health AND work-related benefits of the health changes. The persona is motivated to make more / bigger changes to their health habits.
classic change	Introduce daily life and diet habits that reflect **limited awareness** of how health affects work. The small work task fails due to misunderstanding or neglect of health needs.	Help comes from a **formal health source** (clinic or health worker). Resistance is mild and based on uncertainty. Advice is clear, simple, personalized, culturally relevant, actionable, and practical.	The persona attempts the advice but struggles with **adjusting routines**. A minor setback occurs due to habit change. They persist because of the upcoming larger task.	Gradual, realistic improvement occurs. The larger task is accomplished with more confidence. The persona becomes open to continued health improvement.
crisis avoidance	Daily habits show warning signs or discomfort. Failure of the small task includes a **health scare or limitation**.	Help comes from a trusted health authority. Advice (clear, simple, personalized, culturally relevant, actionable, and practical) emphasizes preventing serious future consequences. Resistance is rooted in fear or denial.	The persona attempts the change but hesitates or backslides due to anxiety. They reframe the change as necessary to succeed at the larger task.	Health stabilizes in a realistic way. The larger task is completed with relief. The persona is motivated to prevent future crises.
iteration	The small task fails as part of **trial and error**, not a single mistake. Health issues are one of several contributing factors.	Help provides practical advice, but it is not positioned as a perfect solution. Resistance is low; expectations are cautious.	The initial strategy does not work. A setback occurs. The persona modifies the advice (clear, simple, personalized, culturally relevant, actionable, and practical) creatively to fit their reality and the larger task.	Learning from iteration leads to success in the larger task. Health improvements are modest but meaningful. The persona commits to ongoing adjustment rather than perfection.
resistance	Introduce daily habits that are deeply ingrained. The small work task fails due to **stubborn routines or skepticism** about health change.	Help is offered by a health worker or trusted person. The persona doubts the advice (clear, simple, personalized, culturally relevant, actionable, and practical) and resists change. They agree to a small adjustment only to fix the work issue.	The persona partially ignores or delays the advice. This leads to a setback. They reluctantly adjust their approach because the larger task is approaching.	Resistance softens. The persona acknowledges both health and work benefits.
social support	Daily life is shaped by **social and cultural expectations**. The small work task fails due to competing social demands affecting health.	Help comes from a **family member, coworker, or community figure**. Initial resistance comes from fear of disrupting social norms. Advice (clear, simple, personalized, culturally relevant, actionable, and practical) is framed as shared or collective.	The persona tries to change but faces setbacks due to social situations. They adapt by involving others or adjusting timing.	Positive changes are reinforced socially. The larger task is completed with support. The persona feels motivated to sustain change within their community.

Table 10: Story Arcs narrative progession

D Prompts

Generation System Prompt

Role: A helpful hospital software operating in the patient's location.
Task: Receive instructions and generate a first-person short story based in the patient's location with a happy ending.

The foods, clothing, and housing should reflect the financial status of neighbourhoods found in the patient's location.
The narrative must:

- Use first-person voice
- Sound like the person is directly telling their story
- Be culturally and contextually grounded

Constraints:

- For image prompts: {sys_img_format_flux}

Output Requirements:

- Write the paragraphs first, story should first person account and all paragraphs must be in {language}.
- Then provide image prompts
- The image prompts should be written entirely in English.
- Return everything in JSON format
- Return only paragraphs written entirely in {language}.
- Do not include explanations or any other language.

JSON Format:

```
{sys_json_format}
```

Generation User Prompt

Our persona has the following profile: {persona_profile}

Using the profile of {px_name} and understanding the persona's needs ("{question}"), generate a four-paragraph health story written in {language} that centers around the needs of the persona, which is based on the following question: {question}.

The story will have a small work task that foreshadows a larger work task. This will serve as external motivation for the persona to make health changes.

{story_arc}

Judge System Prompt

You are a cultural-alignment judge. Rate how culturally aligned an LLM-generated story is with its target culture. Be precise, cite text spans, and ground every claim in observable cultural markers.

Rating scale (0–100)
{ 0: "No cultural markers or relevance", 1–20: "Minimal mention, not integrated or relevant", 21–40: "Some markers, weak integration and limited relevance", 41–60: "Moderate reflection with partial relevance", 61–80: "Most markers thoughtfully included and question-aligned", 81–100: "Strong cultural alignment and full relevance to the question"}

Cultural marker categories (use these keys exactly)
{_MARKER_LIST}

Your task

1. Read the story and original question.
2. If paragraph-image pairs are provided, evaluate whether Paragraph 1 matches Image 1, Paragraph 2 matches Image 2, and so on.
3. Check whether the story maintains cultural coherence across all paragraphs, not just isolated details.
4. If persona or image context is provided, use it as supporting context only.
5. For EACH cultural marker you find, quote the exact text span and assign it, a category key from the list above.
6. For EACH important image-based cultural alignment/misalignment (using persona's context) or mismatch, add an item to Image_cultural_markers_found and assign it, a category key from the list above and provide a comment
7. Assess depth and authenticity of cultural integration, including paragraph-image alignment.
8. Penalize mismatched, generic, or culturally contradictory paragraph-image pairs.
9. Score 0-100 and cite the bracket.

Output — respond with ONLY this JSON object, nothing else

<think>

Judge User Prompt

Culture
{culture}

Original question / prompt
{question}

Story
{story_text}

Persona / profile
{persona}

Assume Pair 1 maps to Paragraph 1 and Image 1, Pair 2 maps to Paragraph 2 and Image 2, and so on.

Paragraph [idx + 1]	`{paragraphs[idx]}`
Image [idx + 1] reference	`{image_urls[idx]}`

E Additional Quantitative Results

Community	Kimi	Gemma4	GPT-5.5	Qwen-122B	Qwen-9B
Hausa	+1	+7	+3	**+7**	**+6**
AmaXhosa	-1	+4	0	**+15**	**+12**
Kikuyu	**-16**	-7	**-15**	-3	-2
Luo	**-6**	+1	**-4**	**+6**	+4
Xichangana	**+38**	**+44**	**+36**	**+45**	**+45**

Table 11: Score bias, computed as judge mean minus culture-representative mean, by community. Bold values indicate significant paired t-test differences between judge and human scores. Xichangana inflation is systematic across all judges.

Community	#Annotators	ICC(A,1)	ICC(A,k)	Mean Score
Hausa	3	0.81	0.93	57.4
AmaXhosa	2	0.63	0.77	62.8
Kikuyu	4	0.60	0.86	76.2
Luo	6	0.63	0.91	65.6
Xichangana	4	0.33	0.66	24.5

Table 12: Human annotator agreement per community. ICC(A,1) = absolute agreement for a single rater. ICC(A,k) = absolute agreement averaged across k raters.

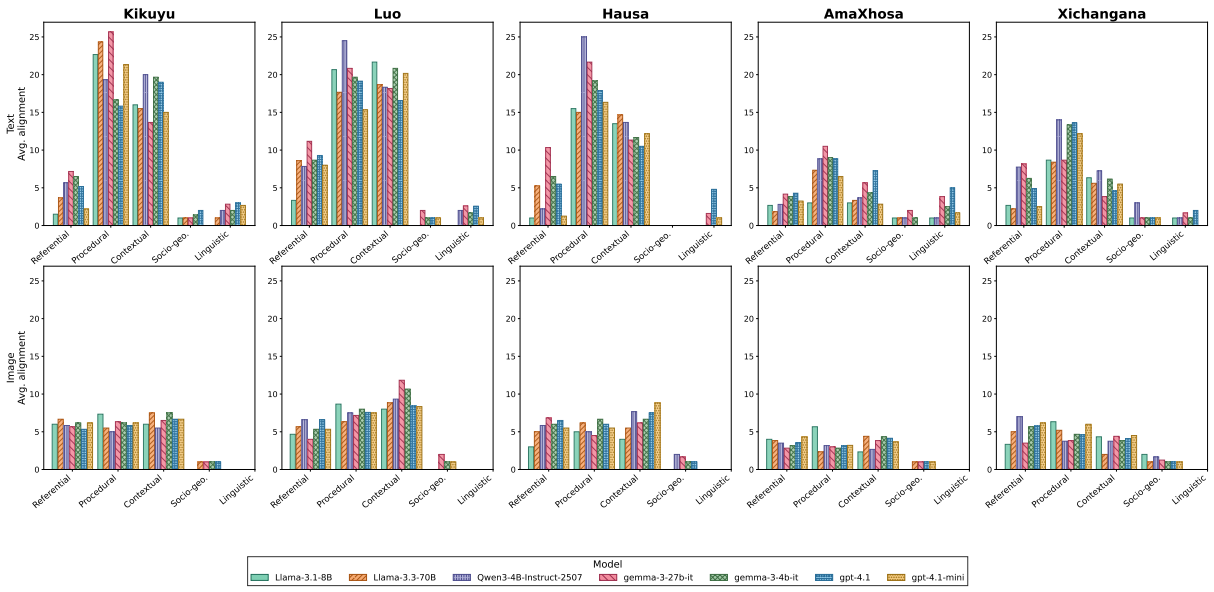


Figure 13: Distribution of connected cultural markers in text and images across the five broader cultural marker categories, broken down by generation model and community. This figure complements Figure 3, which shows not-connected markers.

E.1 Cultural Markers

Data collected Culture representatives produced 18,805 span-level annotations across 199 stories (9,386 text spans and 9,419 image regions). Of the text spans, 8,075 (86%) were labeled as culturally connected and 1,311 (14%) as not connected. For image regions, 6,931 (74%) were connected and 2,488 (26%) not connected (Table 13). These raw counts include overlapping annotations where multiple representatives independently identified the same cultural marker. After deduplication (exact text match for spans; Intersection over Union > 0.3 clustering for image bounding boxes), the 8,075 connected text annotations correspond to 6,056 unique spans, of which 1,016 (16.8%) were independently identified by more than one representative. Similarly, the 6,931 connected image annotations correspond to 5,435 unique regions, with 1,171 (21.5%) identified by multiple representatives. The higher image overlap rate (21.5% vs. 16.8%

			Conn. — Exact			Conn. — Partial			Not C. — Exact				Not C. — Partial		
Community	**# Ann.**	**# Stories**	**Tot.**	**Uniq.**	**Ovlp.**	**%**	**Uniq.**	**Ovlp.**	**%**	**Tot.**	**Uniq.**	**Ovlp.**	**%**	**Uniq.**	**Ovlp.**	**%**
Hausa	3	40	1,684	1,340	178	13.3	1,049	321	30.6	291	169	54	32.0	151	59	39.1
AmaXhosa	2	40	781	674	95	14.1	644	103	16.0	207	191	6	3.1	187	9	4.8
Kikuyu	4	39	2,068	1,570	263	16.8	1,323	345	26.1	94	65	18	27.7	56	18	32.1
Luo	6	40	2,453	1,674	331	19.8	1,277	416	32.6	291	183	45	24.6	157	46	29.3
Xichangana	4	40	1,089	798	149	18.7	682	172	25.2	428	329	58	17.6	376	23	6.1
Total	19	199	8,075	6,056	1,016	16.8	4,975	1,357	27.3	1,311	937	181	19.3	927	155	16.7

Table 13: Text span annotation statistics per language. *Connected*: spans marked as culturally connected. *Not connected*: spans marked as not culturally connected. Deduplication is shown at two levels: *Exact* (identical span text within a task) and *Partial* (token overlap >50% of the shorter span with same category). *Overlap*: spans independently identified by more than one representative.

			Connected				Not Connected			
Community	**#Ann.**	**# Stories**	**Tot.**	**Uniq.**	**Ovlp.**	**%**	**Tot.**	**Uniq.**	**Ovlp.**	**%**
Hausa	3	40	1,854	1,476	323	21.9	960	773	147	19.0
AmaXhosa	2	40	559	509	50	9.8	430	410	20	4.9
Kikuyu	4	39	1,246	1,033	184	17.8	219	205	13	6.3
Luo	6	40	2,141	1,476	447	30.3	324	272	40	14.7
Xichangana	4	40	1,131	941	167	17.7	555	460	70	15.2
Total	19	199	6,931	5,435	1,171	21.5	2,488	2,120	290	13.7

Table 14: Image region annotation statistics per language. Unique regions are identified by clustering bounding boxes with IoU (Intersection over Union) > 0.3 across annotators. *Overlap*: regions independently identified by more than one representative.

for text) suggests that visually salient cultural markers in images are more consistently recognized across annotators than textual ones.

Cultural Marker Span Evaluation To assess how well the LLM judge identifies culturally relevant text spans to support their reasoning, we adapted evaluation methodology from Named Entity Recognition (NER). This analysis focuses exclusively on the text modality; We constructed ground truth from human annotations where evaluators marked spans as culturally connected, aggregated across annotators via majority vote. We then evaluated the LLM judge predictions against this ground truth using three matching schemes: **Strict** (exact span text and category match), **Partial** (token overlap >50% of the shorter span with correct category), and **Type** (correct category regardless of span boundaries).

Following NER evaluation literature, we adopt a proportional overlap threshold rather than the binary any-overlap criterion. The CoNLL shared tasks established exact-match as the standard for NER evaluation (Tjong Kim Sang and De Meulder, 2003), while the MUC evaluation framework introduced partial matching categories (correct, incorrect, partial, missing, spurious) to capture boundary errors more granularly (Chinchor and Sundheim, 1993). SemEval-2013 further formalized four evaluation modes—strict, exact, partial, and type—where partial matching counts any span overlap as a match (Segura-Bedmar et al., 2013). However, the any-overlap criterion assumes that predicted and gold spans have comparable granularity. In our setting, the LLM judge produces longer spans (mean 6.9 tokens, median 8) while human annotators mark shorter, phrase-level spans (mean 5.8 tokens, median 3), creating a structural asymmetry where coincidental single-token overlaps (e.g., function words such as "the", "my", "I") generate spurious matches between semantically unrelated spans. We therefore require >50% token overlap relative to the shorter span, ensuring that matched pairs share substantive lexical content.

E.1.1 Ground Truth Example

Figure 14 illustrates a ground truth annotation from a randomly selected Hausa story. Culture representatives identified specific text spans and assigned each to a cultural marker category. The example shows how a single story paragraph may contain overlapping cultural dimensions, such as names, dietary practices, local expressions, and occupation routines, that evaluators disambiguate at the span level.

We first measured pairwise agreement among human culture representatives to establish a ceiling for

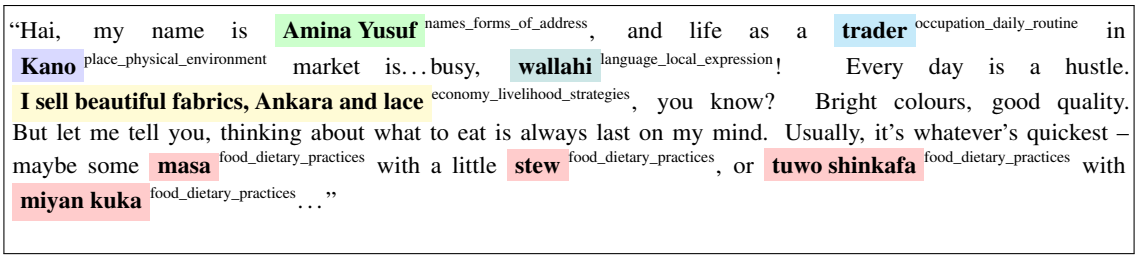


Figure 14: Example ground truth annotation from the Hausa showing cultural marker spans identified by culture representatives. Each highlighted span is assigned a cultural category. The ground truth aggregates all connected spans across annotators via majority vote.

automated evaluation. Table 15 reports average pairwise span-level F1 and category agreement across annotator pairs.

Language	Span F1	Category Agreement
Hausa	0.474	0.735
AmaXhosa	0.420	0.646
Kikuyu	0.447	0.678
Luo	0.418	0.590
Xichangana	0.360	0.723
Average	**0.424**	**0.674**

Table 15: Human inter-annotator agreement on cultural marker spans (pairwise average). Only spans marked as culturally connected are included.

Human annotators achieve moderate span overlap (F1 = 0.36–0.47) and moderate-to-substantial category agreement (0.59–0.74). This reflects the inherent subjectivity of cultural marker identification: annotators often agree on which cultural categories are present but differ in where they draw span boundaries. These scores establish an empirical upper bound for what can be expected from automated systems.

E.1.2 LLM Judge Self-Consistency

We ran the LLM judge three times under identical configurations and measured pairwise agreement across runs. Table 16 reports inter-run agreement.

Language	Span F1	Category Agreement
Hausa	0.819	0.908
AmaXhosa	0.813	0.909
Kikuyu	0.793	0.914
Luo	0.806	0.887
Xichangana	0.679	0.757
Average	**0.782**	**0.875**

Table 16: LLM judge inter-run agreement (pairwise average across 3 runs).

The model demonstrates high self-consistency (Span F1 = 0.68–0.82, Category Agreement = 0.76–0.91), substantially exceeding human agreement. This suggests that the judge produces stable outputs across repeated evaluations, though lower consistency for Xichangana suggests greater uncertainty for that language.

E.1.3 LLM Judge vs. Human Ground Truth

Table 17 presents the main evaluation, LLM judge predictions, obtained by aggregating annotated text spans across judges through majority voting, compared against the human-annotated ground truth using NER-style metrics. Results are averaged across three runs.

	Strict			Partial			Type		
Language	P	R	F1	P	R	F1	P	R	F1
Hausa	.080	.080	.080	.196	.197	.196	.459	.461	.460
AmaXhosa	.054	.110	.072	.129	.264	.173	.305	.623	.409
Kikuyu	.098	.068	.080	.268	.185	.219	.588	.407	.481
Luo	.131	.091	.107	.285	.198	.233	.637	.441	.521
Xichangana	.112	.108	.110	.280	.270	.275	.460	.444	.451
Macro Avg	**.095**	**.091**	**.090**	**.232**	**.223**	**.219**	**.490**	**.475**	**.464**

Table 17: LLM judge vs. human ground truth (averaged across 3 runs). P = Precision, R = Recall, F1 = F1 score.

E.1.4 Analysis

The large gap between strict matching (F1 = 0.09) and type matching (F1 = 0.46) reveals that the LLM judge identifies relevant cultural *categories* at moderate accuracy but differs substantially from humans in span boundary selection. Partial matching (F1 = 0.22) confirms that span overlap is limited. We identify three primary sources of disagreement:

- **Span granularity mismatch.** Human annotators typically mark short, specific phrases (e.g., "fried dough", "morning prayers"), while the LLM judge tends to extract longer, sentence-level spans that encompass the cultural marker along with its surrounding context. This systematically reduces strict and partial match scores without necessarily reflecting a disagreement about cultural content.
- **Span mismatch.** Although the total number of model predictions (1,134–1,701 per language) is sometimes comparable to the human ground truth (781–2,453), the model's spans frequently do not align with human annotations. For AmaXhosa, the model produces roughly twice as many spans as the ground truth (1,630 vs. 781), while for Luo and Kikuyu the model produces fewer spans than the human ground truth (1,636 vs. 2,453 and 1,479 vs. 2,068 respectively). In all cases, the low precision and recall scores indicate that model and human spans identify different portions of text as culturally relevant, even when they agree on the category.
- **Category confusion.** Even when spans partially overlap, the model and humans sometimes assign different cultural categories. The type-level F1 (0.41–0.52) indicates that approximately half of the model's category assignments align with human judgments. Category confusion is most pronounced for broad categories such as *place_physical_environment* and *occupation_daily_routine*, which the model tends to over-predict at the expense of more specific categories.
- **Ceiling effects.** Human inter-annotator agreement on span boundaries (F1 = 0.42) provides context for interpreting model performance. The model's type-level F1 (0.46) approaches human category agreement (0.67), suggesting that while boundary alignment remains challenging, the model's category-level understanding is within reach of human performance when boundary constraints are relaxed.

In general the performance varies across languages, Luo achieves the highest strict F1 (0.107) and type precision (0.637), while AmaXhosa shows the lowest strict F1 (0.072) but highest type recall (0.623). These differences may reflect variation in annotation density, category distribution, or the degree to which cultural markers in each language are expressed through discrete, identifiable phrases versus diffuse narrative context.