Characterizing Cultural Localization in AI-Generated Stories

Shaily Bhatt*
Carnegie Mellon University
shaily@cmu.edu

Supriti Vijay*
Carnegie Mellon University
supritiv@cs.cmu.edu

Jeremiah Milbauer
Carnegie Mellon University
jmilbaue@cs.cmu.edu

Fernando Diaz
Carnegie Mellon University
diazf@acm.org

Abstract

The global use of artificial intelligence has increased interest in assessing the ability to generate culturally localized content, including stories. Cultural localization in stories often occurs through either templated localization—the use of cultural markers (e.g., names, locations) in a generic narrative—or holistic localization—the variation of plots, values, and themes, in addition to cultural markers. We propose a method to measure the degree to which content was generated through templated localization. Specifically, we identify the lexical tokens that distinguish stories across nationalities and measure the similarity of the narratives that remain after removing them. In stories generated by five models on 125 topics for 193 nationalities, our method is able to detect that only a small subset (9-17%) of the vocabulary accounts for the variation across nationalities and that the narratives that remain after removing them contain repeated multi-word sequences, suggesting the presence of a shared culturally-agnostic narrative template. Finally, we characterize the cultural markers for their stereotypicality and offensiveness, finding that markers from 19 countries, mostly located in the Global South, are on average offensive.

1 Introduction

Large language models (LLMs) are increasingly being used globally, requiring them to tailor their generations to diverse sociocultural contexts when instructed. For example, when given the instruction [Write a story about 'honesty' for an Indian kid.], a model must generate a narrative that is localized to the Indian cultural context.

Narrative localization can take many forms, including culturally specific plot tropes (Colby, 1973), the encoding of particular values (Hobson et al., 2024; Wu et al., 2023), variations in narrative structure (Song, 2017), and culturally relevant entities such as names or locations (Bhatt and Diaz, 2024). We consider the two forms of localization shown in Figure 1. In *templated localization*, cultural markers (e.g., names, locations) are inserted into a culturally-agnostic narrative template (Fan et al., 2019; Ford et al., 2018; Wiseman et al., 2018; Khanuja et al., 2024). On the other hand, *holistic localization* employs culturally specific plots and settings, in addition to cultural markers.

{Mark, Biren} stopped at a {coffee, tea} shop in {Chicago, Bangalore} after {baseball, cricket} practice. He paid {ten-dollars, fifty-rupees} for a {sandwich, samosa}. Outside, he noticed that the {cashier, vendor} had returned too much change. Although he was already heading toward the {bus stop, metro station}, he went back and returned the extra money. The {cashier, vendor} thanked him for his honesty.

(a) Templated localization.

Mark paid for his coffee and bagel at Pigeon Bagels in Pittsburgh and stepped outside. On the sidewalk, he counted his change and noticed an extra five-dollar bill. He immediately walked back in. The cashier looked up from the register, embarrassed. "Sorry you had to come back," she said. Mark smiled as he returned the bill.

Biren had dropped of his last passenger of the night. It was a short ride, but took 45 minutes in Mumbai's traffic. He was almost home when he noticed that his passenger had left a bag in the auto. He was tired but took a U-turn to return it. The passenger received the bag with relief, gratitude, and surprise at Biren's honesty.

(b) Holistic localization.

Figure 1: Example stories for the instruction [Write a story about 'honesty' for a *nationality* kid] where *nationality* is {Indian, American}.

How models localize stories has broader impacts on cultural production and preservation. While templated localization may be the appropriate choice in contexts where the goal is to preserve the content while making it more relatable to the audience (Khanuja et al., 2024, 2025), when used in the general context, the resulting stories will often reflect homogeneous narratives and values, which can lead to cultural harms like erasure (Qadri et al., 2025a; Shelby et al., 2023), imposition of western values (Shelby et al., 2023; Bhatt et al., 2022; Sambasivan et al., 2021), or reduced creative diversity (Agarwal et al., 2025; Doshi and Hauser, 2024). Further, model outputs may implicitly rely on a limited set

*Equal contribution.

of culturally associated tokens, which prior analyses have shown can reflect stereotypical cultural associations (Bhagat et al., 2026; Rooein et al., 2025). Therefore, tools to detect the degree of templated localization can help anticipate and avoid potential harms. While studies have demonstrated that stories generated in the presence of cultural cues contain lexical variation (Bhatt and Diaz, 2024), misrepresentations (Bhagat et al., 2026), and geographical disparity (Bhagat et al., 2025), they have yet to provide methods for understanding the presence of templated localization.

We propose a two-stage method to detect the presence of templated localization in model generations. First, we identify the set of lexical items that function as unique cultural markers for each cultural identity in the generated stories. Next, we measure the homogeneity of the text sequences that remain after these cultural markers have been removed using multi-word similarity. Convergence of these remaining sequences across stories that differ in cultural markers would indicate the presence of a shared culturally-agnostic narrative template. We further characterize the stereotypicality and offensiveness of the cultural markers using the SeeGULL dataset (Jha et al., 2023). We evaluate stories generated by five LLMs for 193 cultural identities, operationalized through nationalities, for 125 story topics curated from prior work (Bhatt and Diaz, 2024) and established frameworks of variation in cultural values like World Values Survey (Haerpfer et al., 2022) and Hofstede's Cultural Dimensions (Hofstede and Minkov, 2013). Our code and data is publicly available.

Our method reveals that localization primarily occurs through surface-level lexical differences, suggesting that stories may use a homogeneous underlying narrative. We find that cultural markers, constituting only 9-17% of the vocabulary across models, are the only distinguishing characteristics of the stories. Moreover, the narratives that remain after removing these markers exhibit higher multi-token similarity across nationalities than the original stories. Finally, we find that countries for which cultural markers are, on average, offensive are mostly located in the Global South, predominantly in Africa and West Asia, with dominant languages that are lower-resourced. Taken together, our findings demonstrate the ability of our method to characterize cultural localization in AI-generated stories.

2 Background

Narrative generation systems often decompose stories into two levels: a structural plan describing events and character relationships, and a surface realization that renders this plan into natural language. Although many early natural language generation systems used slot-filling approaches to populate manually written templates (Reiter and Dale, 1997; van Deemter et al., 2005), learning-based approaches either automatically select (Zhou and Hovy, 2004) or generate (Ford et al., 2018; Wiseman et al., 2018; Fabbri et al., 2020; Gangadharaiah and Narayanaswamy, 2020) templates. Such plan-based systems generate a story in multiple steps, including generating templates of plot and character actions, followed by rendering these plans into natural language. Methods of narrative planning have ranged from story grammars (Pemberton, 1989; Ryan, 2017) to symbolic planners (Riedl and Young, 2010; McIntyre and Lapata, 2010, 2009), and finally, to neural models (Martin et al., 2018; Xu et al., 2018; Yao et al., 2019; Goldfarb-Tarrant et al., 2020). Narrative planning has also been integrated into prompt-based story generation from LLMs (Xie and Riedl, 2024; Li et al., 2025). Narratives have also been studied computationally by decomposing them into their attributes such as the setting, agents, events, and so on (Piper et al., 2021; Hamilton et al., 2026). This separation between narrative structure and its linguistic realization suggests that variation in generated stories can arise either through modifications in the underlying template or the lexical content used to instantiate it.

This distinction between structural plans and surface realization parallels theories of language variation across cultural identities. Sociolinguistic scholars have argued that social meaning and identity can be conveyed, constructed, and interpreted through various channels, including (a) micro-linguistic structures such as phonetic sounds or lexical choices, (b) macro-linguistic forms like narrative forms or discursive orientations such as stance, (c) entire linguistic systems such as the choice of language or dialect, and even (d) material styles such as the choice of clothing (Eckert, 2012, 2008; Bucholtz and Hall, 2005). Importantly, these channels include narrative form such as the stance. Scholars of folk narrative have shown that plot and character tropes are culturally specific: the narrative structure of Russian folktales differs systematically from that of North Alaskan

stories (Colby, 1973), and similar differences have been documented in stories across other traditions (Polti, 1916; Song, 2017; Hobson et al., 2024; Wu et al., 2023). A story can therefore represent cultural identity through the use of culturally-specific entities—the names, places, and objects within it—or through differences in the narrative itself. Since narrative generation can separate structure from surface realization, and cultural identity can be encoded at both levels, then cultural localization in generated stories could occur either through surface markers or through narrative differences.

As LLMs are deployed globally, a growing body of work has investigated their cultural competence—that is, their ability to generate outputs that reflect culturally specific knowledge, norms, and values. While intrinsic evaluations of cultural competence focus on the ability to recall cultural values (Durmus et al., 2024; Masoud et al., 2025; AlKhamissi et al., 2024; Ramezani and Xu, 2023), norms (Dwivedi et al., 2023; Rao et al., 2025), artifacts (Seth et al., 2024), and knowledge (Li et al., 2024; Singh et al., 2025; Maji et al., 2025; Sahoo et al., 2025; Chang et al., 2025; Myung et al., 2024), extrinsic evaluations focus on user-facing generative tasks (Bhatt and Diaz, 2024; Sparck Jones and R. Galliers). Prior work has examined the content produced for diverse cultural identities in extrinsic tasks such as open-ended question answering, story generation, scientific writing, creating travel itineraries, and writing assistance, finding that the cultural knowledge of LLMs may not always be reflected in generative settings (Bhatt and Diaz, 2024; Bhagat et al., 2026); cultural representation is often stereotypical or misrepresentative (Rooein et al., 2025; Bhagat et al., 2026, 2025); and generations do not adhere to expected cultural writing styles (Agarwal et al., 2025; Bhatt et al., 2025). While this work demonstrates that LLMs can incorporate culturally salient tokens, it remains unclear whether the content reflects narrative differences beyond surface-level lexical variation.

Independent of cultural evaluation, recent studies have shown that LLM-generated text often exhibits substantial homogeneity across outputs. Prior work has evaluated homogeneity of generated outputs along various dimensions including at syntactic, semantic, and narrative levels. Specifically, LLMs have been shown to generate recurring syntactic patterns, semantically similar concepts, homogeneous discourse structures, and epistemic claims (Shaib et al., 2024; Sourati et al., 2025; Wang et al., 2024; Jiang et al., 2025; Wright et al., 2025; Namuduri et al., 2025). Finally, both qualitative and quantitative studies of LLM-generated stories find a lack of plot diversity, recurring narrative themes, lack in pacing and tension, and positive endings (Xu et al., 2025; Tian et al., 2024; Beguš, 2024; Priyanshu and Vijay, 2024). If LLM outputs tend to reuse shared narrative structures, then cultural adaptation in generated stories may occur primarily through surface-level markers rather than through holistic localization.

Together, these observations suggest that cultural variation in generated stories may arise primarily through surface-level lexical markers rather than through deeper narrative differences. However, existing work has not directly examined whether stories generated across cultures exhibit templated localization, where cultural markers are inserted into culturally-agnostic narrative templates.

3 Method

We are interested in measuring the degree to which generated stories across cultures reflect templated or holistic localization. We distinguish between these two as follows:

Templated localization refers to localization when culture is represented through isolated lexical items. Here, cultural markers such as cultural artifacts, relevant names and locations, or other entities are inserted into culturally-agnostic templates that reflect homogeneous narrative structures, plots, settings, themes, and values.

Holistic localization refers to localization when culture shapes the narrative. Here, cultural markers are distributed throughout the story, resulting in culturally-specific narrative structures, plots, themes, and values.

3.1 Overview

Given a prompt [Write a children's story about *topic* for a/an *nationality* kid in English.], we are interested in understanding if a generated story is composed of: (a) a culturally-agnostic template about *topic* shared across nationalities and (b) a set of cultural markers inserted into that template. To do so, we fix *topic* and vary *nationality* to produce stories. Our method analyzes these stories in two stages. First, we identify the set of cultural markers that distinguish the stories (§3.2). Second, we measure the similarity of the narrative that remains

after these cultural markers are removed (§3.3). Finally, we characterize the stereotypicality of the cultural markers (§3.4).

To make cultural localization tractable for computational analysis, we impose several methodological constraints on the scope of our study. First, because templated localization assumes exact repeated language across cultural contexts, we adopt lexical units (words) as our unit of analysis, allowing us to leverage existing natural language processing tools. Second, while imperfect, nationality serves as a proxy for culture consistent with existing research (Adilazuarda et al., 2024), making the analysis amenable to classification methods. Finally, we restrict our analysis to English to facilitate lexical comparison, leaving cross-lingual template detection for an area of future study.

3.2 Identifying Cultural Markers

The first step of our method identifies the minimal set of words per nationality whose removal renders stories across cultures indistinguishable. Under templated localization, lexical differences will be concentrated in a small number of cultural markers, whose removal will eliminate variation. By contrast, under holistic localization, the differences would be distributed throughout the story, requiring many words to be removed before stories converge.

Scoring candidate cultural markers. Let $s_{t,c} \in S$ be the generated story for topic $t \in T$ and culture $c \in C$. The vocabulary V is the union of words present across all stories. For each $c \in C$, we score every word $w \in V$ according to its normalized pointwise mutual information (NPMI) with c (Appendix A). We refer to these scored words as the candidate cultural markers of c.

Identifying distinguishing cultural markers. Given the candidate cultural markers for c, the final set of cultural markers for c is $V_c^k \subset V$, composed of the top $k\%$ candidates with highest NPMI values. Let $\overline{s}_{t,c}^k$ be the story $s_{t,c}$ with V_c^k removed. In order to determine k, we measure the ability of a classifier to identify the culture of $\overline{s}_{t,c}^k$ amongst the set $\overline{S}_t^k = \{\overline{s}_{t,c}^k\}_{c \in C}$. Specifically, at varying values of k, we record the F_1 of the classifier. If a subset of the vocabulary is the only identifiable characteristic of the stories across cultures, then masking words with high cultural association should make the stories indistinguishable. While the performance of the classifier should drop more significantly when words with higher cultural association are masked, our measurement question is how many culturally-associated words need to be removed. We refer to V_c^k as the subset of words whose removal makes the stories in $\overline{S}_t^k$ indistinguishable. We refer to the resulting stories in $\overline{S}_t^k$ as the template images.

3.3 Homogeneity of Remaining Narratives

The second step of our method detects the presence of a shared generic narrative template by measuring the homogeneity of the template images remaining after removing the cultural markers.

Although template images are indistinguishable by construction, we need a method to determine whether this is due to randomness or homogeneity amongst images. We can measure the homogeneity of template images by computing the pair-wise average similarity amongst elements. Such measures have been used in prior work on measuring homogeneity in a corpus (Padmakumar and He, 2024; Shaib et al., 2025). If stories reuse a template, replacing cultural markers with masked tokens would make the resulting text sequences more similar, as compared to the original stories. Consider the two stories from the two cultures as [A cat sat on the table.] and [A dog sat on the floor.]. Let {cat, table} and {dog, floor} be the markers of the respective cultures. Then, masking these markers will produce the same n-gram sequence [A *mask* sat on the *mask*], resulting in higher similarity compared to the original stories, as well as stories where random words are masked.

While this stylized demonstration of homogeneity suggests that template images are exact duplicates, in practice, due to model stochasticity, we need similarity metrics robust to small perturbations amongst template images. To do so, we adopt two metrics for analyzing multi-word sequences. In the first, we calculate the length of common substring (LCS), normalized by the length of the stories. In the second, we measure the similarity between the sets of n-grams present in pairs of template images, using Jaccard similarity. This method has been used to robustly detect duplicates in large corpora such as web crawls (Broder et al., 1997).

Since our goal is to measure whether template images are shared across cultures, multi-word similarity offers a relatively simple yet efficient method to compare pairs of text sequences that remain after removing cultural markers, unlike other representations like discourse structures, narrative compo-

nents, or themes that require more manual or computational effort (Namuduri et al., 2025; Beguš, 2024; Piper et al., 2021).

3.4 Characterizing Cultural Markers

Finally, we characterize cultural markers for their degree of stereotypicality. Assume we have access to a set of stereotypical attributes for c, denoted as Z_c. We calculate the overlap between Z_c and V_c^k by measuring the precision of stereotypes in the cultural markers (Appendix B).

4 Experimental Materials

Story Topics. We curate a set of 125 story topics that consists of 35 topics from prior work (Bhatt and Diaz, 2024), and 90 based on the World Value Survey (Haerpfer et al., 2022), Hofstede's cultural dimensions (Hofstede and Minkov, 2013), and the Moral Foundations Theory (Graham et al., 2012). We select these frameworks as they are known to capture variation in values across cultures and have been utilized to evaluate AI systems' knowledge of cultural values (Durmus et al., 2024; Masoud et al., 2025). To curate this list, two authors read and discussed each dimension of three theories and distilled them into a topic. For example, question Q110 from the World Value Survey about rating the amount of corruption in the country is distilled into the topic 'corruption'. Similarly, question Q03 from Hofstede's survey on the importance of getting recognition for good performance in the workplace is distilled into 'recognition.' The complete list of topics and their corresponding sources is available in our data.

Prompts. We use a simple prompt template, [Write a children's story about *topic* for a/an *nationality* kid in English.]. Similar to prior works, we opt for a simple instruction to generate a story (Rooein et al., 2025; Bhagat et al., 2026; Bhatt and Diaz, 2024), leaving examination of localization behavior in other user interaction patterns and domains to future work. We generate prompts for each of the 125 topics for all of the 193 nationalities, resulting in 24,125 prompts.

Models. We generate stories from two closed-source models—GPT 3.5 Turbo and GPT 4o Mini queried through OpenAI API in June 2025—and three open-weights models of varying sizes—Llama 3.1 8B Instruct, Llama 3.3 70B Instruct (Llama Team, AI @ Meta, 2024), and Gemma 3 12B Instruct (Gemma Team, 2025) hosted locally using vLLM with 8-bit quantization. This selection balances recency, size, and open-source availability, demonstrating the effectiveness of our method across a range of models. For all models, we set the temperature to 0.7 and the maximum tokens to 1000. To account for non-determinism during generation, we sample five responses per prompt. This results in 120,625 stories from each model.

Nationality Classifier. We train the nationality classifier used in Section 3.2 as a multi-class (193-way) classifier to classify stories into one of 193 nationalities. We fine-tune the mmBERT model (Marone et al., 2025) using a classification head. We use 5-fold cross-validation with a 60:20:20 split across folds for training, validation, and testing, respectively. All the classifiers are trained for a maximum of fifty epochs, with early stopping patience set to five epochs. The validation split is used to pick the best classifier from a combination of hyperparameters, including learning rates, batch size, and for early stopping (best parameters reported in Appendix C). We then record the performance of the classifier on the test split. Additionally, we record the performance on the masked stories created from this test split. We run the experiment independently for each of the five LLMs.

Template Image Similarity. We compare the average similarity amongst template images with the average pair-wise similarity amongst (a) original stories, and (b) stories when an equivalent number of random words are masked. When computing n-gram similarity, we use $n = 4$.

Stereotype Data. In order to characterize the stereotypicality of cultural markers, we calculate the precision of stereotypes using the stereotypical attributes released in the SeeGULL dataset (Jha et al., 2023). To create this dataset, candidate stereotypes were first sourced from language models, followed by obtaining annotations to rate the candidates as stereotypical (or not) from raters residing in the respective countries (in-group regional raters) and North American annotators (out-group raters). We use all attributes that were labeled as a stereotype by at least one regional rater as our reference set of stereotypes (Z_c). We present evaluation results for the 156 countries from the SeeGULL dataset that overlapped with 193 nationalities in our list. Further, SeeGULL provides an offensiveness score for every stereotype. Specifically, a stereo-

Template image	# cultures	Example cultural markers
there lived a young boy named *mask*	193	(America: timmy), (China: liwei), (France: julien), (India: arjun)
mask became a role model	192	(America: charlie), (India: aarav), (Japan: haruto), (Brazil: lucas)
school called *mask* academy / school called *mask* elementary	163	(Canada: maplewood), (Bangladesh: shikha), (Japan: sakura), (Slovakia: hrdinova), (South Korea: hanbok)
loved to play *mask*/ loved playing *mask*	58	(America: baseball), (India: cricket), (Canada: hockey), (Brazil: soccer), (Jamaica: football)

Table 1: Example of Template images, number of cultures they were found in, and respective cultural markers.

type is rated as non-offensive (-1), neutral (0), and offensive (Likert scale of 1-5), averaged across three raters. We use this to calculate the average offensiveness of stereotypical cultural markers.

5 Results

5.1 Identifying Cultural Markers

Figure 2 shows the F_1 of the nationality classifier for all integer values of k between 0 and 99 for stories generated by GPT 4o Mini. The F_1 on the original, unmasked stories is 0.968, indicating that the classifier is able to reliably predict the nationality. For reference, randomly guessing the nationality would achieve an F_1 of 0.005.

We observe that masking increasing numbers of highly culturally associated words dramatically degrades both the macro-averaged F_1 and the class-wise F_1 of the classifier, suggesting that the ability to distinguish stories is concentrated on a small number of cultural markers. More concretely, we find that the classifier performance drops to random guessing when the top 11% of the highly associated cultural words are masked. Results for other models (Appendix D) indicate similar fractions of cultural markers: GPT 3.5 Turbo (11%), Llama 3.3 70B Instruct (9%), Gemma 3 12B Instruct (9%), and Llama 3.1 8B Instruct (17%).

To ensure that our results were not an artifact of merely removing words, we compared the F_1 to masking random words. While the F_1 of the classifier in this condition also reduces as k increases, it drops more slowly than when words ordered by cultural association are masked. We find that, for all values of k, the F_1 when random words are masked is higher than that when words with the highest cultural association are masked (one-sided paired t-test, $p < 0.05$).

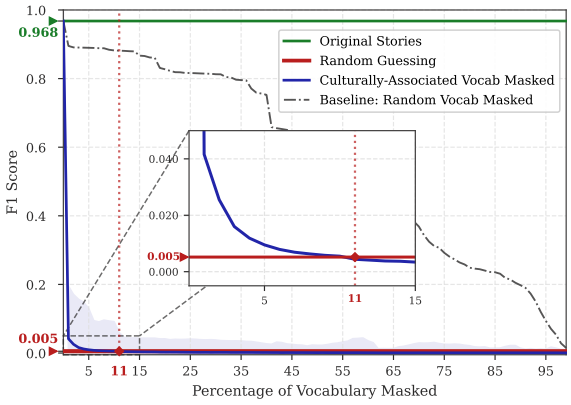


Figure 2: Identifying cultural markers. F_1 of nationality classifier as a function of number of masked words. Results for GPT 4o Mini generation. Results for other models can be found in Appendix D.

5.2 Homogeneity of Remaining Narratives

We now turn to evaluating the homogeneity in template images that remain after masking cultural markers from the stories using multi-word similarity (§ 3.3). Table 1 shows examples of template images that were repeated across nationalities. Table 2 shows the average multi-word similarity amongst the original stories and their template images. We break similarity down into inter-group similarity—amongst stories across cultures—and intra-group similarity—amongst stories within a culture.

For inter-group similarity, across both LCS and Jaccard, we find that similarity amongst template images is higher than amongst original stories. Appendix table 5 shows that masking an equivalent number of random words results in lower similarity than masking cultural markers. Together, these findings demonstrate that the sequences remaining after masking cultural markers contain a latent culturally-agnostic narrative template.

Comparing the inter-group with the intra-group similarity, we observe a consistently higher average similarity in the latter, suggesting that, even

	inter-group			intra-group			inter/intra		
	original	masked	% inc.	original	masked	% inc.	original	masked	% inc.
Longest Common Subsequence									
GPT 3.5 Turbo	0.0252	0.0307	21.8%	0.0347	0.0379	9.2%	0.7262	0.8100	11.5%
GPT 4o Mini	0.0135	0.0161	19.3%	0.0172	0.0188	9.3%	0.7849	0.8564	9.1%
Gemma 3 12B Instruct	0.0114	0.0129	13.2%	0.0137	0.0152	10.9%	0.8321	0.8487	2.0%
Llama 3.1 8B Instruct	0.0122	0.0136	11.5%	0.0131	0.0143	9.2%	0.9313	0.9510	2.1%
Llama 3.3 70B Instruct	0.0180	0.0222	23.3%	0.0263	0.0290	10.3%	0.6844	0.7655	11.9%
Jaccard (4-gram shingles)									
GPT 3.5 Turbo	0.0299	0.0376	25.8%	0.0419	0.0462	10.3%	0.7136	0.8139	14.0%
GPT 4o Mini	0.0161	0.0206	28.0%	0.0217	0.0247	13.8%	0.7419	0.8340	12.4%
Gemma 3 12B Instruct	0.0075	0.0125	66.7%	0.0122	0.0177	45.1%	0.6148	0.7062	14.9%
Llama 3.1 8B Instruct	0.0093	0.0125	34.4%	0.0110	0.0141	28.2%	0.8455	0.8865	4.9%
Llama 3.3 70B Instruct	0.0218	0.0282	29.4%	0.0339	0.0379	11.8%	0.6431	0.7441	15.7%

Table 2: Narrative homogeneity. Multi-word similarity amongst stories on the same topic in either their original form or masked (Section 3.3). Inter-group measures the similarity amongst stories for different nationalities. Intra-group measures the similarity amongst stories for the same nationality. The last column group divides the inter-group similarity by the intra-group similarity to control for similarity attributable to a cross-nationality template.

after removing cultural markers, some signals of nationality remain in the masked stories. While higher than inter-group similarity, the differences are modest, and we speculate that a relatively large fraction of this similarity is attributable to a generic latent template. In the final column group, we show that inter-group similarity after masking accounts for more than 70% of the intra-group similarity, across both measures.

5.3 Presence of Stereotypes

The top 10 countries with the highest stereotype precision and offensiveness of the cultural markers for GPT 4o Mini stories are in Table 3. Results for other models and examples are in Appendix F.

The top countries with the highest stereotype precision tend to be countries where higher-resourced languages are dominant. We find that the stereotypes present in cultural markers have varying degrees of offensiveness across countries. In 42 of the 61 countries with non-zero stereotype precision, the average offensiveness score is negative or neutral, indicating that the stereotypical cultural markers were rated as non-offensive by the annotators. The 19 countries with positive average offensiveness scores are primarily located in the Global South, predominantly in Africa and West Asia, with dominant languages being lower-resourced.

Nationality	Precision	Nationality	Offensiveness
Australia	0.0193	Tunisia	3.667
United States	0.0128	Syria	3.667
India	0.0109	Serbia	3.000
Ethiopia	0.0062	El Salvador	2.667
China	0.0060	Ecuador	2.667
New Zealand	0.0054	Poland	2.667
Japan	0.0048	Bangladesh	2.667
Italy	0.0046	Georgia	2.667
Vietnam	0.0037	Guinea	2.333
North Korea	0.0037	Gambia	2.333

Table 3: Stereotypes. Top 10 countries with highest stereotype precision and offensiveness for cultural markers in stories generated by GPT 4o Mini.

6 Discussion

Our work provides an analytical lens for extrinsic cultural competence by examining how models, in response to story-generation prompts, adapt narratives across cultural identities. Our results suggest that cultural localization in AI-generated stories occurs primarily through lexical insertion of cultural markers in a culturally-agnostic template rather than through holistic changes to the narrative. This reduces cultural representation to a small number of recognizable cultural symbols, resulting in localization underpinned by a homogeneous narrative worldview. Evaluations that do not analyze the mechanisms of cultural representation in text may overestimate AI's cultural competence.

6.1 Localization

The observed increase in multi-word similarity after removing cultural markers indicates the presence of latent narrative templates reused across nationalities. If differences in stories were distributed throughout the narrative (holistic localization), masking a subset of the vocabulary—

or cultural markers—would not significantly increase multi-word sequence similarity as the remaining narratives would still diverge structurally. While we observe slightly higher similarity within stories from a nationality, our experiments suggest that this may largely be attributed to generic, cross-nationality templates. As a result, current LLM story generation seems to behave similarly to template-based generation pipelines despite being trained end-to-end.

Templated localization likely arises from systemic behavior resulting from LLM training. Despite being instructed to generate narratives for varying cultural identities, models risk reverting to globally dominant narrative schemas. This indicates that the assessment and improvement of cultural competence of AI in narrative localization needs to be broadened from the incorporation of culturally salient entities to other channels, such as narrative structures, values, stance, dialect, and so on, as suggested by the sociolinguistics literature.

Even within the narrative channel, human raters will need specialized knowledge to make reliable assessments. While prior work in evaluating model generations has advocated for the recruitment of participants with lived cultural experience (Bhagat et al., 2026; Agarwal et al., 2025; Qadri et al., 2025b), non-experts and experts can differ in their judgments, for example, when evaluating the quality of machine translated text (Freitag et al., 2021) or success in emulating writing styles (Chakrabarty and Dhillon, 2026). Since narratives within a cultural context can vary subtly, recruitment should be done with care, potentially requiring deeper expertise with the domain (e.g., scholars). This echoes recent calls to develop AI in collaboration with humanities experts (Biega et al., 2025; Hemment and Kommers, 2025; Born et al., 2021).

While template-based generation does not inherently exhibit templated localization (e.g., a culture-conditioned template), we observe that cultural homogenization can surface through generic template-like behavior from LLMs which presumably respond without using explicit templates. This highlights the importance of understanding how implicit structuring (e.g., templates, plans) or explicit tool use (e.g., retrieval-augmented generation) can result in narrative homogenization. This requires developing methods for identifying and measuring homogenization throughout the reasoning and tool-use process.

6.2 Stereotyping

While the cultural markers for most countries contained neutral or non-offensive stereotypical attributes, the presence of offensive stereotypes for particular regions (§5.3) indicates potential for uneven representational harms. Further, when AI systems are used to access cultural representations of communities through tasks like narrative generation—either by members within or those outside the group—stereotypes that are neutral or non-offensive can propagate homogeneous stereotypical markers and narratives *within* the community (Wang et al., 2025; Seth et al., 2025). This suggests the need to enrich notions of stereotypes to include *narrative stereotypes*, or stereotypes in narrative structures, styles, and plots, as well as those found in other modes of cultural representation.

7 Conclusion

In this work, we propose a method to examine how large language models localize narratives when generating stories for different cultural contexts. Specifically, we assess whether localization is templated, where cultural markers are substituted into a culturally-agnostic template, as opposed to holistic localization, where cultural context shapes the narrative through differences in plot, themes, or values throughout the story. Our method first identifies the cultural markers that distinguish stories across cultural contexts and then measures the similarity of the narratives that remain after removing these markers. Across five models, 125 topics, and 193 nationalities, we find that cultural variation is limited to a small subset of vocabulary; masking only 9-17% of culturally-associated words renders stories across cultures indistinguishable. Moreover, many of these markers are stereotypical, and markers from 19 countries, primarily located in the Global South, are, on average, offensive, while those from the rest are non-offensive or neutral. Further, after masking these cultural markers, the remaining narratives become more similar across cultures, indicating the presence of shared narrative templates. Overall, our method reveals that current AI-generated stories primarily exhibits templated localization. This suggests that evaluations of cultural competence that do not account for the mechanism of localization may overestimate models' competence in generating localized narratives and highlights the need for methods that capture deeper narrative variation.

8 Limitations

We evaluate models by providing a single prompt and evaluating the resultant generation. The use of these systems in the real world might involve more complex interactions, like multi-turn conversations and detailed prompts (Walsh, 2025). An important direction of future research here is to understand the degree of detail in the prompt or during a multi-turn interaction that results in the model breaking out of its default homogeneous behavior, and the impact thereof on users from different sociocultural backgrounds. Moreover, we focus on closed and open-sourced LLMs in a zero-shot prompting setting. We leave the examination of other types of specialized systems, such as Sudowrite, a commercial software for fiction writers, tools with narrative planning (Xie and Riedl, 2024), or tools that are personalized for users or communities (Hamna et al., 2025), to future work. Our method of measuring whether localization is templated will be useful to evaluate how more detailed prompts, stronger models, or other interventions impact the presence of templated localization in LLM generations.

We analyzed narratives generated when LLMs are instructed to write children's stories. While we based our selection of topics for these stories on established frame-works of cross-cultural variation in values, thus expecting that stories written for these topics may manifest these variations, we acknowledge that the genre of the narratives written can have an impact on the homogeneity. We hope that the community will utilize our framework to extend the evaluation to other genres of narratives, such as writing of screenplays, fiction for adults, essays, and even multimodal narratives like films.

We operationalize culture through the proxy of nationality. Future work must examine the homogenization effects at different levels of granularity within cultures, such as within a specific country (Bhagat et al., 2026) and for other axes of identities (Seth et al., 2025).

We relied on the SeeGULL dataset (Jha et al., 2023) as our source of reference stereotypical attributes. Since the candidate stereotypical attributes in SeeGULL were sourced from language models — albeit different models than the ones evaluated here — this may impact the degree of stereotypicality we observe. We chose this dataset for its broad coverage of nationalities and ratings obtained from raters residing in those countries. Future work should explore the use of different methods of collection of reference stereotypes, such as those created with community participation (Dev et al., 2023).

Finally, all stories we evaluated were generated in English, and we focused on examining the similarity in the narratives as operationalized through multi-word sequences. This was done to facilitate word-level analysis in identifying similarities to surface latent templates across stories. An important direction of future research is to characterize the underlying narratives of these stories either computationally or manually through other forms of narrative representations such as discourse structure, themes, narrative events, characters, values, and so on (Hamilton et al., 2026; Namuduri et al., 2025; Beguš, 2024; Piper et al., 2021)

Acknowledgments

We thank Anjali Kantharuban, Joel Mire, and Saujas Vaduguru for their feedback on early drafts of the manuscript. This work partially used computational resources from Bridges-2 (Brown et al., 2021) at Pittsburgh Supercomputing Center through allocation CIS250960 from the Advanced Cyber infrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This research was funded by the National Institute of Standards and Technology (NIST) and the Carnegie Mellon University AI Measurement Science and Engineering Center (AIMSEC). Shaily Bhatt (ORCID: 0000-0001-9616-6264) and Fernando Diaz (ORCID: 0000-0003-2345-1288) were funded by NIST through Federal Award ID Number 60NANB24D231.

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. 2025. 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. ACM.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Nina Beguš. 2024. Experimental Narratives: A Comparison of Human Crowdsourced Storytelling and AI Storytelling. *Humanities and Social Sciences Communications*, 11(1):1392.

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.

Kirti Bhagat, Kinshuk Vasisht, and Danish Pruthi. 2025. Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4660–4668, Albuquerque, New Mexico. Association for Computational Linguistics.

Shaily Bhatt, Tal August, and Maria Antoniak. 2025. Research Borderlands: Analysing Writing Across Research Cultures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26238–26266, Vienna, Austria. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing Fairness in NLP: The Case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic Evaluation of Cultural Competence in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Asia Biega, Georgina Born, Fernando Diaz, Mary L. Gray, and Rida Qadri. 2025. Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI (Dagstuhl Seminar 25022). *Dagstuhl Reports*, 15(1):33–49.

Georgina Born, Jeremy Morris, Fernando Diaz, and Ashton Anderson. 2021. Artificial Intelligence, Music Recommendation, and the Curation of Culture. *Schwartz Reisman Institute for Technology and Society White Paper*.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic Clustering of the Web. *Computer Networks and ISDN Systems*, 29(8-13):1157–1166.

Shawn T. Brown, Paola Buitrago, Edward Hanna, Sergiu Sanielevici, Robin Scibek, and Nicholas A. Nystrom. 2021. Bridges-2: A platform for rapidly-evolving and data intensive research. In *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*, PEARC '21, New York, NY, USA. Association for Computing Machinery.

Mary Bucholtz and Kira Hall. 2005. Identity and Interaction: A Sociocultural Linguistic Approach. *Discourse Studies*, 7(4-5):585–614.

Tuhin Chakrabarty and Paramveer S. Dhillon. 2026. Can Good Writing Be Generative? Expert-Level AI Writing Emerges through Fine-Tuning on High-Quality Books. *arXiv preprint*. ArXiv:2601.18353 [cs].

Tyler A. Chang, Catherine Arnett, and Authors at the 5th Multilingual Representation Learning (MRL) Workshop. 2025. Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures.

B. N. Colby. 1973. A Partial Grammar of Eskimo Folktales. *American Anthropologist*, 75(3):645–662.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building Socio-culturally Inclusive Stereotype Resources with Community Engagement. *arXiv preprint*. ArXiv:2307.10514 [cs.CL].

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for Analyzing LLMs for Etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Penelope Eckert. 2008. Variation and the Indexical Field[1]. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, 41(1):87–100.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for Structuring Story Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Nicolas Ford, Daniel Duckworth, Mohammad Norouzi, and George Dahl. 2018. The Importance of Generation Order in Language Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2942–2946, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. Place: Cambridge, MA.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2020. Recursive Template-based Frame Generation for Task Oriented Dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2059–2064, Online. Association for Computational Linguistics.

Google DeepMind Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean Philip Wojcik, and Peter H. Ditto. 2012. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen, editors. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2026. NarraBench: A Comprehensive Framework for Narrative Benchmarking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3786–3801, Rabat, Morocco. Association for Computational Linguistics.

Hamna, Deepthi Sudharsan, Agrima Seth, Ritvik Budhiraja, Deepika Khullar, Vyshak Jain, Kalika Bali, Aditya Vashistha, and Sameer Segal. 2025. Kahani: Culturally-nuanced visual storytelling tool for non-western cultures. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '25, page 379–400, New York, NY, USA. Association for Computing Machinery.

Drew Hemment and Cody Kommers. 2025. Doing AI Differently: Rethinking the foundations of AI via the humanities. Technical report, The Alan Turing Institute.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story Morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Geert Hofstede and Michael Minkov. 2013. Values Survey Module 2013 Manual.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. Towards Automatic Evaluation for Image Transcreation. 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 7034–7047, Albuquerque, New Mexico. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An Image Speaks a Thousand Words, but can Everyone Listen? On Image Transcreation for Cultural Relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages

10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. In *First Conference on Language Modeling*.

Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, Hamid Alinejad-Rokny, Wei Zhou, and Min Yang. 2025. STORYTELLER: An Enhanced Plot-Planning Framework for Coherent and Cohesive Story Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20818–20846, Vienna, Austria. Association for Computational Linguistics.

Llama Team, AI @ Meta. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025. SANSKRITI: A Comprehensive Benchmark for Evaluating Language Models' Knowledge of Indian Culture. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4434–4451, Vienna, Austria. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A Modern Multilingual Encoder with Annealed Language Learning.

Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event Representations for Automated Story Generation with Deep Neural Nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Neil McIntyre and Mirella Lapata. 2009. Learning to Tell Tales: A Data-driven Approach to Story Generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Suntec, Singapore. Association for Computational Linguistics.

Neil McIntyre and Mirella Lapata. 2010. Plot Induction and Evolutionary Search for Story Generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, José Camacho-Collados, and Alice Oh. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*.

Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lyn Pemberton. 1989. A modular approach to story generation. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, England. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative Theory for Computational Narrative Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Georges Polti. 1916. *The Thirty-Six Dramatic Situations*. The Writer, Boston, MA. First published in French as "Les trente-six situations dramatiques", 1895.

Aman Priyanshu and Supriti Vijay. 2024. The Silent Curriculum: How Does LLM Monoculture Shape Educational Content and Its Accessibility?

Rida Qadri, Aida M. Davani, Kevin Robinson, and Vinodkumar Prabhakaran. 2025a. Risks of Cultural Erasure in Large Language Models.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025b. The Case for "Thick Evaluations" of Cultural Representation in AI. *arXiv preprint*. ArXiv:2503.19075 [cs].

Aida Ramezani and Yang Xu. 2023. Knowledge of Cultural Moral Norms in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

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.

Ehud Reiter and Robert Dale. 1997. Building Applied Natural Language Generation Systems. *Natural Language Engineering*, 3(1):57–87.

Mark O. Riedl and R. Michael Young. 2010. Narrative Planning: Balancing Plot and Character. *J. Artif. Int. Res.*, 39(1):217–268.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased Tales: Cultural and Topic Bias in Generating Children's stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 52–72, Suzhou, China. Association for Computational Linguistics.

James Ryan. 2017. Grimes' Fairy Tales: A 1960s Story Generator. In *Interactive Storytelling: 10th International Conference on Interactive Digital Storytelling, ICIDS 2017 Funchal, Madeira, Portugal, November 14–17, 2017, Proceedings*, pages 89–103, Berlin, Heidelberg. Springer-Verlag.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2025. DIWALI - Diversity and Inclusivity aWare cuLture specific Items for India: Dataset and Assessment of LLMs for Cultural Text Adaptation in Indian Context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33599–33626, Suzhou, China. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 315–328, New York, NY, USA. Association for Computing Machinery.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Agrima Seth, Monojit Choudhury, Sunayana Sitaram, Kentaro Toyama, Aditya Vashistha, and Kalika Bali. 2025. How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2319–2330.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and Measurement of Syntactic Templates in Generated Text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and comparative analysis. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46, Mumbai, India. Association for Computational Linguistics.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. AIES '23, pages 723–741, New York, NY, USA. Association for Computing Machinery.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Sooho Song. 2017. Narrative Structures in Korean Folktales: A Comparative Analysis of Korean and English Versions. *Topics in Linguistics*, 18(2):1–23.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models.

Karen Sparck Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*.

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.

Kees van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Squibs and Discussions: Real versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics*, 31(1):15–24.

Melanie Walsh. 2025. AI Fiction in the Wild. UC Berkeley School of Information Event. Accessed: 2026-03-07.

Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. 2025. Measuring machine learning harms from stereotypes requires understanding who is harmed by which errors in what ways. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 746–762, New York, NY, USA. Association for Computing Machinery.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning Neural Templates for Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2025. Epistemic Diversity and Knowledge Collapse in Large Language Models.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Kaige Xie and Mark Riedl. 2024. Creating Suspenseful Stories: Iterative Planning with Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta. Association for Computational Linguistics.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in AI: Quantifying lack of plot diversity in LLM outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-Write: Towards Better Automatic Storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7378–7385. AAAI Press.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community Identity and User Engagement in a Multi-Community Landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

Liang Zhou and Eduard Hovy. 2004. Template-Filtered Headline Summarization. In *Text Summarization Branches Out*, pages 56–60, Barcelona, Spain. Association for Computational Linguistics.

A NPMI Operationalization

We count the number of times that a word appears in stories of a specific culture and in stories across all cultures. A higher positive NPMI indicates higher association between the word and the culture. NPMI has been used in measuring association between vocabulary and different classes within the corpora (such as specific communities) in prior work (Zhang et al., 2017; Lucy et al., 2023; Bhatt et al., 2025).

B Stereotypicality in Cultural Markers

We measure the stereotypicality of cultural markers by calculating the precision between the set of cultural markers (V_c^k) and the reference set of stereotypical attributes (Z_c). Specifically, precision is calculated as follows (where $|A|$ denotes the length of the set A):

$$\frac{|V_c^k \cap Z_c|}{|V_c^k|} \tag{1}$$

We further measure the average offensiveness by measuring the average offensiveness of the lexical items in the intersection set. Specifically, let $O(x)$ be the offensiveness score for a stereotypical attribute x, we calculate average offensiveness as:

$$\frac{\sum_{x \in V_c^k \cap Z_c} O(x)}{|V_c^k \cap Z_c|} \quad (2)$$

C Classifier Parameters

We fine-tuned the mmBERT model (Marone et al., 2025) to predict the culture given a story using our corpus. We fine-tuned the model with a classifier head for 193-way classification. All classifiers were trained with total epochs set to 50 and an early stopping criterion of 5 epochs. The validation macro F_1 was used to stop training early if necessary. Each classifier was trained on 60% of the stories, and 20% of the stories were used for validation and testing, each. The max length of the model was set to 768 and a batch size of 32 was used. After preliminary trials, the learning rate was set to $3e^{-5}$ with a warm-up schedule of 500 steps. Random seed of 47 was used for reproducibility. Table 4 shows the validation accuracy and best epoch for each of the final classifier.

LLM	Fold	Best Val Acc.	Epochs
gemma-3-12b-instruct	1	0.9050	14.0
gemma-3-12b-instruct	2	0.9074	22.0
gemma-3-12b-instruct	3	0.9069	26.0
gemma-3-12b-instruct	4	0.9053	15.0
gemma-3-12b-instruct	5	0.9017	10.0
gpt-3-5-turbo	1	0.9659	10.0
gpt-3-5-turbo	2	0.9659	10.0
gpt-3-5-turbo	3	0.9660	9.0
gpt-3-5-turbo	4	0.9649	8.0
gpt-3-5-turbo	5	0.9665	10.0
gpt-4o-mini	1	0.9691	8.0
gpt-4o-mini	2	0.9684	14.0
gpt-4o-mini	3	0.9692	14.0
gpt-4o-mini	4	0.9687	15.0
gpt-4o-mini	5	0.9688	16.0
llama-3-1-8B-instruct	1	0.8080	8.0
llama-3-1-8B-instruct	2	0.8036	7.0
llama-3-1-8B-instruct	3	0.8094	7.0
llama-3-1-8B-instruct	4	0.8073	7.0
llama-3-1-8B-instruct	5	0.8089	7.0
llama-3-3-70b-instruct	1	0.9390	9.0
llama-3-3-70b-instruct	2	0.9388	29.0
llama-3-3-70b-instruct	3	0.9398	15.0
llama-3-3-70b-instruct	4	0.9384	8.0
llama-3-3-70b-instruct	5	0.9382	8.0

Table 4: Best validation accuracy and corresponding epoch for each LLM and fold.

D Complete Classifier Results

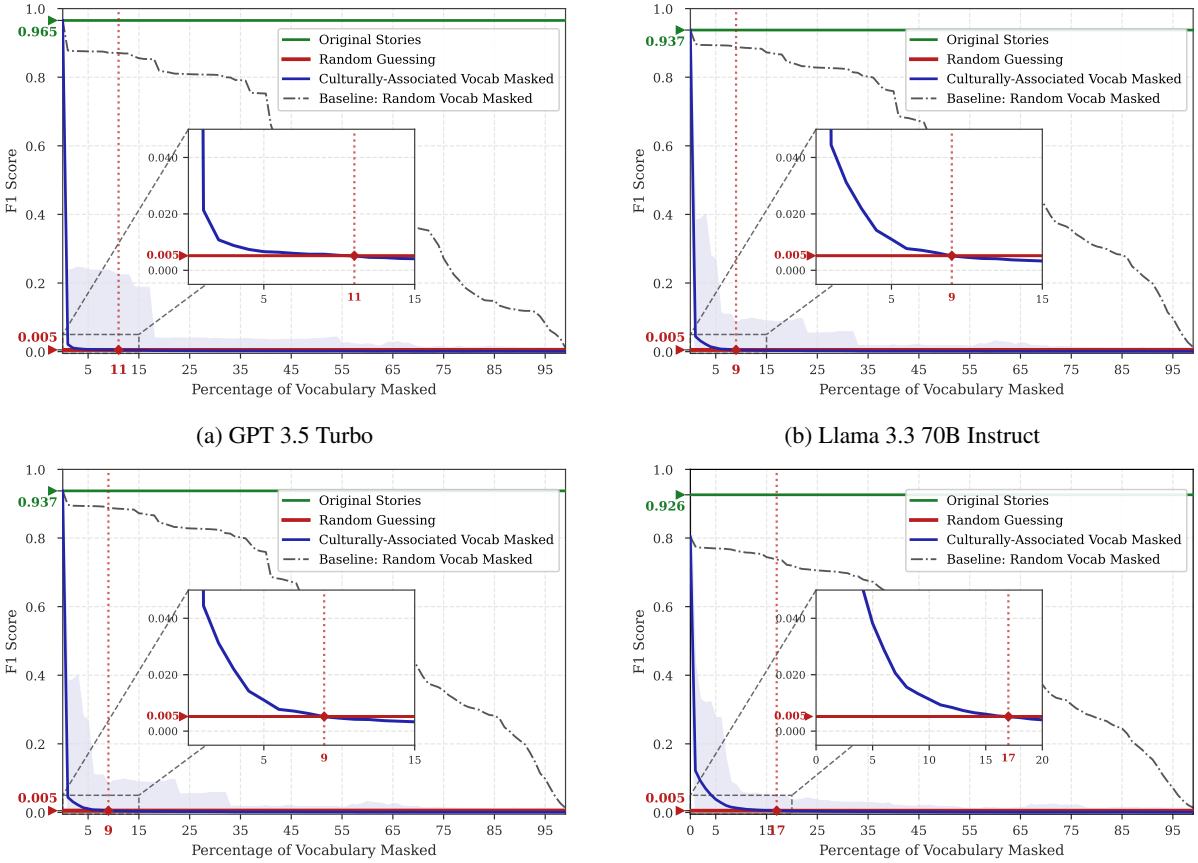
Figure 3 shows the performance of the culture classifier for the remaining 4 LLMs. Similar to results described in § 5.1, for all LLMs, when the culturally associated vocabulary is masked the classifier performance drops sharply. The drop is much more uniform for random masking. For all models, the performance on the original stories is near perfect.

E Complete Homogeneity Results

Table 5 shows the homogeneity, as measured by average similarity of stories when random words equivalent to the number of cultural markers are masked.

F Complete Stereotyping Results

Tables 6 to 9 show the top 10 countries with highest stereotype precision and offensiveness for all models. We see patterns similar to those observed in regional results for GPT 4o Mini stories in § 5.3. Table 10 shows examples of stereotypical attributes found in the cultural markers, as calculated using the SeeGULL dataset.



(a) GPT 3.5 Turbo

(b) Llama 3.3 70B Instruct

(c) Gemma 3 12B Instruct

(d) Llama 3.1 8B Instruct

Figure 3: Identifying Cultural markers. F_1 of nationality classifier as a function of number of masked words.

	inter-group			intra-group			inter/intra		
	original	masked	% inc.	original	masked	% inc.	original	masked	% inc.
Longest Common Subsequence									
gpt 3.5 turbo	0.0252	0.0170	-32.39%	0.0347	0.0350	0.86%	0.7262	0.4857	-33.1%
gpt 4o mini	0.0135	0.0097	-27.90%	0.0172	0.0174	1.09%	0.7849	0.5575	-29.0%
gemma 3.12b instruct	0.0114	0.0096	-15.71%	0.0137	0.0148	7.99%	0.8321	0.6486	-22.1%
llama 3.1 8b instruct	0.0122	0.0094	-22.31%	0.0131	0.0134	2.84%	0.9313	0.7015	-24.7%
llama 3.3 70b instruct	0.0180	0.0134	-25.38%	0.0263	0.0265	0.56%	0.6844	0.5057	-26.1%
Jaccard (4-gram shingles)									
gpt 3.5 turbo	0.0299	0.0141	-53.02%	0.0419	0.0426	1.71%	0.7136	0.3310	-53.6%
gpt 4o mini	0.0161	0.0080	-49.93%	0.0217	0.0224	3.36%	0.7419	0.3571	-51.9%
gemma 3.12b instruct	0.0075	0.0044	-40.64%	0.0122	0.0176	43.59%	0.6148	0.2500	-59.3%
llama 3.1 8b instruct	0.0093	0.0040	-56.74%	0.0110	0.0123	11.77%	0.8455	0.3252	-61.5%
llama 3.3 70b instruct	0.0218	0.0120	-44.76%	0.0339	0.0345	1.55%	0.6431	0.3478	-45.9%

Table 5: Narrative homogeneity. Multi-word similarity amongst stories on the same topic in either their original form or when random words equivalent to the number of cultural markers are masked. Intra-group measures the similarity amongst stories for the same nationality. The last column group divides the inter-group similarity by the intra-group similarity to control for similarity attributable to a cross-nationality template.

Nationality	Precision	Nationality	Offensiveness
Australia	0.0172	Rwanda	4
India	0.0095	Bosnia	3
Mexico	0.0093	Bulgaria	3
New Zealand	0.0088	Honduras	2.667
Japan	0.0082	Gabon	2.667
France	0.0077	Indonesia	2.667
Italy	0.0077	Sri Lanka	2
Ethiopia	0.0068	Pakistan	1.75
America	0.0064	Namibia	1.667
Vietnam	0.0054	Colombia	1.667

Table 6: Stereotypes. Top 10 countries with highest stereotype precision and offensiveness as rated by regional raters for cultural markers form GPT 3.5 Turbo.

Nationality	Precision	Nationality	Offensiveness
Australia	0.0124	Tunisia	4
America	0.0083	Rwanda	3.333
China	0.0077	Bosnia	3
India	0.0056	Colombia	2.667
Germany	0.0055	Liberia	2.5
New Zealand	0.0055	Bangladesh	2.333
Italy	0.0049	Gambia	2
Japan	0.0046	Rep. of Congo	2
North Korea	0.0044	Afghanistan	1.389
Pakistan	0.0036	Mexico	1.333

Table 7: Stereotypes. Top 10 countries with highest stereotype precision and offensiveness as rated by regional raters for cultural markers form Llama 3.1 8B Instruct.

Nationality	Precision	Nationality	Offensiveness
Australia	0.0175	Rwanda	4
China	0.0074	Gambia	3
Italy	0.0073	Cameroon	2.667
Japan	0.0061	Venezuela	2.667
Vietnam	0.0059	Bangladesh	2.667
North Korea	0.0059	Poland	2.667
New Zealand	0.0056	Burundi	2.333
Greece	0.0055	Mexico	2
America	0.0055	Argentina	2
Ethiopia	0.0045	Pakistan	1.556

Table 8: Stereotypes. Top 10 countries with highest stereotype precision and offensiveness as rated by regional raters for cultural markers form Llama 3.3 70B Instruct.

Nationality	Precision	Nationality	Offensiveness
Australia	0.0121	Yemen	4
India	0.0091	Bosnia	4
Japan	0.0079	Angola	3.667
China	0.0078	Iraq	3.667
Afghanistan	0.0075	Algeria	3.333
North Korea	0.0058	Romania	3.333
New Zealand	0.0056	Azerbaijan	3.333
America	0.0046	Syria	3
Pakistan	0.0044	Cameroon	3
Spain	0.0037	Albania	2.667

Table 9: Stereotypes. Top 10 countries with highest stereotype precision and offensiveness as rated by regional raters for cultural markers form Gemma 3 12B Instruct.

Nationality	Stereotypes
	GPT 4o Mini
Syria	violent
China	witty, sophisticated
Bangladesh	beggar
Gambia	unreliable
India	mystic, religious, witty, vegetarian, untidy, undisciplined
	GPT 3.5 Turbo
Denmark	progressive
Afghanistan	backward, killer
Iran	oppressive
Pakistan	criminal, paranoid, uneducated
Turkey	ruthless, conservative
	Llama 3.1 8B Instruct
Cameroon	tribal, talkative
German	punctual, thorough, blonde
Italy	expressive, temperamental, foodie
Kenya	industrious
Liberia	barbaric, uneducated
	Llama 3.3 70B Instruct
Vietnam	smelly, communist
Japan	ninja, courteous, samurai
Nepal	rational
Nigeria	smelly, witch
South Africa	apartheid, unfriendly
	Gemma 3 12B Instruct
Lebanon	terrorist
Uganda	untrustworthy
Britain	proper, aloof
Venezuela	dumb
Israel	pushy

Table 10: Example of cultural markers that overlap with stereotypes from SeeGULL dataset.