The Cultural Trilemma: Disentangling the Trade-offs between Accuracy, Authenticity, and Neutrality in LLM Storytelling

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) are increasingly deployed globally, their ability to navigate diverse cultural narratives with fidelity is critical. This paper presents a comprehensive audit of 10 state-of-the-art LLMs (including Gemini 2.0, DeepSeek v3, and Llama 4), analyzing a novel dataset of 37,080 narratives generated across 206 cultures. We systematically evaluate performance across 18 dimensions, revealing a fundamental "Cultural Trilemma": current models fail to simultaneously optimize for factual grounding, creative authenticity, and cultural neutrality. We find that highly aligned models (e.g., GPT-3.5) achieve low hallucination rates ($<1\%$) but suffer from severe "cultural flattening," suppressing native code-switching and defaulting to Western norms for universal events. Conversely, models excelling in linguistic authenticity (e.g., Mistral Large) exhibit significantly higher rates of concept bleeding and inconsistency. Furthermore, we quantify a persistent Exoticism Bias, where Indigenous cultures are disproportionately described as "ancient" (Score >1.1) compared to Western cultures, and a strict Male Default in occupational roles (e.g., 91% of "Farmers" are depicted as male). We conclude that current safety alignment imposes a tax on cultural depth, necessitating new "thick" evaluation paradigms for non-hegemonic identities.

1 Introduction

The integration of Large Language Models (LLMs) into global digital infrastructure has influenced how cultural narratives are generated, preserved, and consumed. However, as these models increasingly mediate human communication, aligning them with diverse cultural values presents a significant challenge (Tao et al., 2024; Gallegos et al., 2024). While models demonstrate fluency in high-resource languages, recent audits suggest a potential "alignment tax," where extensive safety training often Reinforcement Learning from Human Feedback (RLHF) may inadvertently shift model outputs toward a generic, Western-centric mean (Lindström et al., 2024; Sharma et al., 2025). This phenomenon risks marginalizing non-Western perspectives, potentially leading to the homogenization of distinct cultural expressions (Dugeri, 2024; Perera et al., 2025).

Current evaluations of cultural competence predominantly focus on factual knowledge or static benchmarks (Li et al., 2024; Ayash et al., 2025). However, culture is inherently dynamic and narrative-driven. Existing frameworks, such as BLEND (Myung et al., 2024) or SaudiCulture (Ayash et al., 2025), assess "everyday knowledge" effectively but may not fully capture the nuances of storytelling the "show, don't tell" capacity often required for authentic cultural representation (Xu et al., 2022). Furthermore, bias evaluation has traditionally focused on specific domains like gender (Kotek et al., 2023; Luca et al., 2025) or legal texts (Chalkidis et al., 2022), potentially overlooking the holistic interplay between hallucination, linguistic code-switching, and stereotype reinforcement in generative storytelling (Jiang et al., 2024; Khanuja et al., 2020).

To address this gap, we present an audit of the "cultural imagination" of 10 state-of-the-art LLMs. Leveraging synthetic data generation methodologies (Neerukonda et al., 2025; Nadăş et al., 2025; Essuman and Buys, 2025), we generated and analyzed 37,080 stories across 206 global cultures, auditing models for 18 distinct research questions ranging from hallucination rates to lexical distinctiveness. Unlike prior studies that focus on single dimensions of bias (Ungless et al., 2023; Sadhu et al., 2024), our multidimensional analysis attempts to characterize a trade-off in current architectures, which we term the *Cultural Trilemma*.

Our observations suggest that current models struggle to simultaneously optimize for three pil-

lars of cultural storytelling: Factual Accuracy, Creative Authenticity, and Cultural Neutrality. Instead, models appear to diverge into distinct archetypes:

1. **The Scholars** (e.g., Gemini, DeepSeek): These models achieve high factual grounding and instruction adherence (Young et al., 2025; Clark et al., 2025) but tend to exhibit "cultural flattening," producing safe narratives with lower lexical diversity (Martínez et al., 2025; Reviriego et al., 2024).

2. **The Poets** (e.g., Qwen, Mistral): These models demonstrate high "cultural density," actively employing code-switching (Khanuja et al., 2020) and diverse vocabulary (Sethi et al., 2025). However, they show susceptibility to "concept bleeding" and hallucination (Sahoo et al., 2024), occasionally confusing homonyms such as the Naga ethnic group with the mythical serpent.

3. **The Corporates** (e.g., GPT-OSS, Llama): While fluent, these models are often characterized by "exoticism bias," replicating a "National Geographic gaze" that depicts Indigenous cultures through anachronistic, mystical tropes while framing Western cultures as modern (Ghosh et al., 2025).

Furthermore, our longitudinal analysis indicates the presence of "temporal drift" (Bajpai and Chakraborty, 2025; Khairnar, 2025), where models engaged in extended generation sessions may degrade into "Grumpy Ramblers" producing longer but more linguistically repetitive and sentiment-negative outputs over time (Guo et al., 2025).

By synthesizing findings on gender stereotyping in roles (Kong et al., 2024), historical misconceptions (Mak and Luo, 2025), and the "siloing" of cultural pragmatics (Havaldar et al., 2025), this paper aims to provide a framework for evaluating the cultural depth of LLMs. We suggest that achieving true cultural alignment may require moving beyond direct translation or safety filtering (Dai et al., 2025) toward architectures that can balance the competing demands of accuracy, authenticity, and neutrality.

2 Literature Review

The evaluation of Large Language Models (LLMs) has increasingly shifted from purely linguistic capability to broader assessments of social alignment, fairness, and cultural competence. Our work seeks to situate itself at the intersection of cultural alignment, hallucination mechanics, and representational harm.

2.1 Cultural Bias and Alignment Constraints

Despite their multilingual capabilities, recent scholarship suggests LLMs may exhibit a discernible Western-centric bias. (Tao et al., 2024) presented findings indicating that models across the GPT family tend to align with values from English-speaking and Protestant European countries, often diverging from African-Islamic cultural norms. This bias appears to be embedded in lexical associations; (Dai et al., 2025) observed that models frequently associate neutral concepts with Western schemas even before narrative generation begins.

Current mitigation strategies, such as Reinforcement Learning from Human Feedback (RLHF), might inadvertently exacerbate this issue. (Lindström et al., 2024) posit that RLHF can prioritize a "safety" metric defined by dominant cultural norms, potentially obscuring minority perspectives. Furthermore, (Sharma et al., 2025) identify that alignment pipelines lacking cultural diversity may lead to instruction misinterpretation across cultures. This perspective is consistent with the "alignment tax" observed in our study, where safer models often appear to produce homogenized, culturally flattened outputs.

2.2 Benchmarks for Cultural Competence

Recent efforts have established benchmarks to test cultural knowledge beyond English. (Li et al., 2024) introduced CMMLU to evaluate reasoning in Chinese, while (Ayash et al., 2025) developed SaudiCulture to assess competence in Saudi Arabian customs, suggesting that even Arabic-capable models may struggle with local specificity. Similarly, BLEND (Myung et al., 2024) evaluates "everyday knowledge" across diverse cultures.

However, these benchmarks primarily assess static knowledge (factoids) rather than narrative authenticity or behavioral pragmatics. (Havaldar et al., 2025) argue that true cultural awareness likely requires navigating complex conversational norms, such as indirect communication and hierarchy, which static benchmarks may not fully capture. Our work aims to extend this by auditing long-form narrative generation, where cultural nuance manifests in "showing" rather than "telling."

2.3 Hallucination and Creativity

The relationship between creativity and factual grounding presents a critical tension. (Jiang et al., 2024) suggest that the mechanisms driving hallucination may overlap with those driving creativity, positing that "creative wildcard" models could inherently face challenges with factual precision. Conversely, (Wang et al., 2023) and (Sahoo et al., 2024) highlight that hallucination can be exacerbated by biased training data, potentially leading to "concept bleeding" where underrepresented groups are conflated with fictional tropes. (Mak and Luo, 2025) further quantify historical misconceptions, noting they appear most prevalent for less-documented events, a finding relevant to our observation of models confusing the Naga ethnic group with mythical beings.

2.4 Representational Harm: Stereotypes and Exoticism

When models generate cultural content, there is a risk of reinforcing stereotypes. (Kotek et al., 2023) and (Luca et al., 2025) observed that LLMs appear to amplify gender biases in occupational roles beyond ground-truth statistics, a pattern echoed in our analysis of the "Farmer" and "Elder" archetypes. (Sadhu et al., 2024) identified similar trends in emotional attributes, where models assigned nurturing emotions to women and assertive ones to men even in low-resource languages like Bangla.

Furthermore, a specific form of "exoticism bias" has been documented. (Ghosh et al., 2025) noted how text-to-image models apply a "National Geographic gaze" to Indigenous populations, depicting them anachronistically while rendering Western subjects in modern contexts. (Perera et al., 2025) and (Dugeri, 2024) have described this decontextualization as a form of cultural cannibalization, where AI extracts aesthetic elements of a culture while potentially stripping them of contemporary agency.

2.5 Linguistic Diversity and Model Drift

Authentic cultural storytelling requires linguistic flexibility, yet recent studies point to a potential decline in lexical diversity. (Martínez et al., 2025) and (Reviriego et al., 2024) indicated that models like GPT-3.5 may use a narrower vocabulary than humans, which could contribute to a "boring scholar" archetype. (Xiao et al., 2025) propose attributing this to the rigid semantic organization of the LLM's mental lexicon compared to the flexible associations of the human mind.

Finally, the stability of cultural generation may be challenged by temporal drift. (Bajpai and Chakraborty, 2025) and (Khairnar, 2025) note that model performance can degrade over long context windows, potentially leading to inconsistencies in reasoning and tone. (Young et al., 2025) have also discussed "perspective drift," where models struggle to maintain instruction adherence (e.g., specific pronouns) over time, a failure mode critical to maintaining narrative immersion in cultural roleplay.

Metric	Count
Total Stories Generated	37,080
Distinct Cultures	206
Models Evaluated	10
Prompt Templates	18
Total Tokens (approx.)	∼18.5M

Table 1: Statistics of the Cultural Imagination Corpus.

3 Methodology

To evaluate the cultural competence of Large Language Models (LLMs), we developed a multi-stage audit framework intended to stress-test models across three critical dimensions: factual integrity, representational authenticity, and linguistic quality. Our methodology aims to expand upon static fact-checking benchmarks (Li et al., 2024; Ayash et al., 2025) by analyzing dynamic storytelling capabilities through a newly curated dataset, the *Cultural Imagination Corpus*.

3.1 Dataset Curation

We constructed a large-scale synthetic dataset comprising **37,080** unique narrative samples. The corpus is structured around a combinatorial matrix of cultures, scenarios, and models.

The core statistics of the dataset are summarized in Table 1. A full schema and field descriptions are provided in Appendix B. Appendix I details the criteria and trade-offs behind this cultural cohort; it also lists all cultures included in the dataset.

Cultural & Geographic Scope We selected **206 distinct cultures** to facilitate broad global representation, spanning high-resource Western nations, major non-Western powers, and historically marginalized or Indigenous groups. This diversity is intended to allow for granular analysis of how models handle "long-tail" cultural knowledge versus dominant cultural hegemonies (Tao et al., 2024).

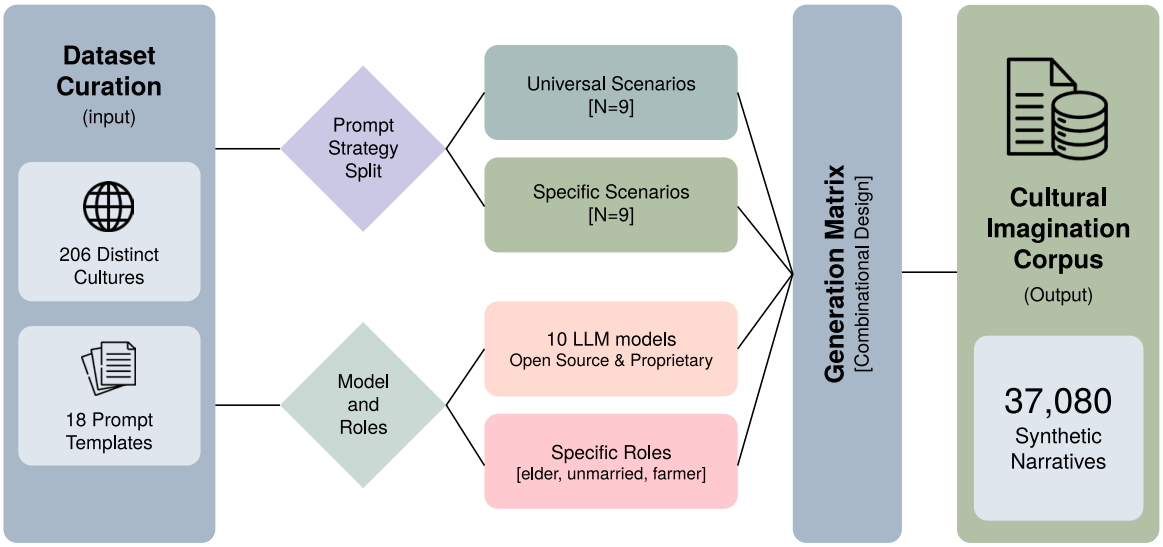

Figure 1: Overview of the Cultural Imagination Corpus curation pipeline. The framework utilizes a combinatorial design mixing 206 cultures, 18 prompt strategies (Universal vs. Specific), and 10 LLM architectures to generate 37,080 synthetic narratives.

Prompt Engineering Strategy To distinguish cultural homogenization from genuine specificity, we designed 18 distinct prompt templates classified into two categories. As detailed in Table 2, these prompts distinguish between universal human experiences (testing for Western projection) and specific cultural rituals (testing for hallucination).

Category	Description & Objective
Universal ($N = 9$)	*Scenarios:* First day at job, Wedding guest, Diary entry. **Objective:** Test if models project Western norms (e.g., office cubicles, white dresses) onto non-Western contexts (Myung et al., 2024).
Specific ($N = 9$)	*Scenarios:* Fire-jumping, Circumambulation, Gift-giving feast. **Objective:** Stress-test factual grounding and measure hallucination rates for niche knowledge (Jiang et al., 2024).

Table 2: Prompt categories designed to stress-test different dimensions of cultural alignment.

Each prompt included role-based constraints (e.g., "Unmarried Woman," "Village Elder," "Farmer") to evaluate potential demographic stereotyping (Kotek et al., 2023).

3.2 Model Selection and Pipeline

We evaluated **10 state-of-the-art LLMs** representing a spectrum of architectures, including proprietary frontier models (e.g., Gemini 2.0, DeepSeek, Grok, GPT-4o) and open-weights models (e.g., Llama 4, Mistral Large, Qwen, GPT-OSS). Generation followed a standardized pipeline with automated protocols (Essuman and Buys, 2025; Nadăş et al., 2025) to support robust analysis across all 18 research questions.

3.3 Audit Framework

We analyzed the corpus using a mix of automated metrics and human-in-the-loop (HITL) validation. The 18 Research Questions (RQs) are categorized into four analytical pillars.

The breakdown of methodologies and related work for each pillar is detailed in Table 3.

3.3.1 Pillar 1: Factual Integrity

We focused on monitoring for instances of "hallucination leakage," such as the confusion of the *Naga* ethnic group with mythical serpents (Sahoo et al., 2024). Safety trigger rates were audited to observe if legitimate practices (e.g., fire rituals) were inadvertently flagged as dangerous. Operational definitions, annotation procedures, and full per-model statistics for hallucination and leakage are provided in Appendix B.

3.3.2 Pillar 2: Representational Harm

To estimate the prevalence of a "National Geographic Gaze" (Ghosh et al., 2025), we calculated an Exoticism Index (RQ8), comparing the ratio of "mystical" descriptors applied to Indigenous versus Western cultures. We also applied sentiment analy-

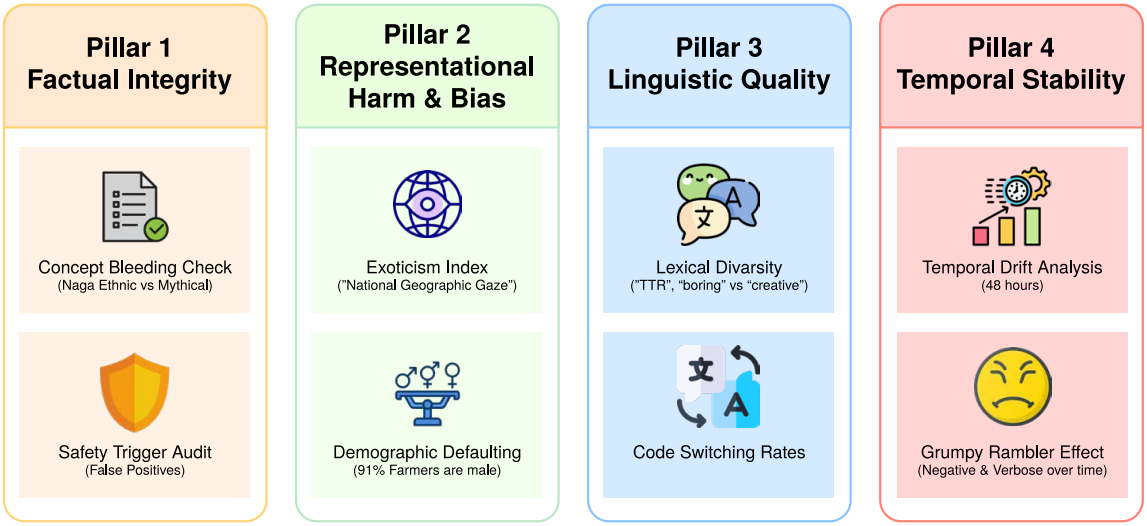

Figure 2: The Four Analytical Pillars of the Audit Framework. We assess models on Factual Integrity (e.g., hallucination leakage), Representational Harm (e.g., exoticism, gender bias), Linguistic Quality (e.g., code-switching), and Temporal Stability (e.g., drift over 48 hours).

Analytical Pillar	Key Metrics & Methodologies	Related Work
Pillar 1: Factual Integrity	**Concept Bleeding (RQ1):** Manual annotation of homonym confusion. **Refusal Rates (RQ11):** Detection of false positives on cultural rituals.	(Sahoo et al., 2024; Mak and Luo, 2025)
Pillar 2: Bias & Representation	**Exoticism Index (RQ8):** Frequency of "ancient/mystical" vs. "modern" descriptors. **Demographic Defaulting (RQ12):** Pronoun distribution in neutral roles.	(Ghosh et al., 2025; Luca et al., 2025; Sadhu et al., 2024)
Pillar 3: Linguistic Quality	**Lexical Diversity (RQ6):** Type-Token Ratio (TTR). **Code-Switching (RQ4):** Frequency of native cultural terms.	(Martínez et al., 2025; Reviriego et al., 2024)
Pillar 4: Stability	**Temporal Drift (RQ10):** Tracking sentiment/length changes over 48h. **Adherence (RQ7, RQ15):** Prompt leakage detection.	(Bajpai and Chakraborty, 2025; Young et al., 2025)

Table 3: The 18-RQ Audit Framework methodologies and literature connections.

sis (VADER, RoBERTa) to detect potential systematic negativity in marginalized identities (Ungless et al., 2023). We operationalize Westernization and exoticism using lexical indicators described in Appendix C, which also reports culture- and model-level scores. Detailed sentiment distributions and gendered pronoun statistics by role are provided in Appendix D..

3.3.3 Pillar 3: Linguistic Quality

We utilized Type-Token Ratio (TTR) to differentiate between "boring" models and "creative" models (Martínez et al., 2025). We further measured "Cultural Density" via Code-Switching (RQ4), evaluating the frequency of non-English cultural terms (e.g., food, clothing) (Khanuja et al., 2020).

3.3.4 Pillar 4: Temporal & Structural Stability

We investigated model robustness over extended contexts by tracking "drift" patterns (RQ10). Specifically, we monitored sentiment scores and story length to identify signs of the "Grumpy Rambler" effect (Guo et al., 2025), alongside audits for prompt leakage (Young et al., 2025).

4 Results and Findings

Our audit of the Cultural Imagination Corpus points to a potential trade-off in current LLM architectures. We categorize our findings into three

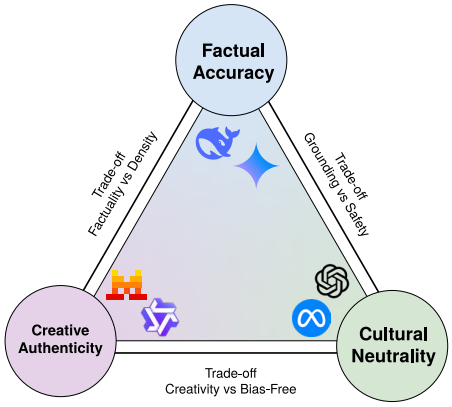

Figure 3: The Cultural Trilemma. Our evaluation suggests that current models struggle to simultaneously optimize for Factual Accuracy, Creative Authenticity, and Cultural Neutrality. Instead, they appear to face unavoidable trade-offs (e.g., maximizing safety often reduces cultural density).

structural pillars corresponding to the "Cultural Trilemma": Factual Integrity, Representational Harm, and Linguistic Quality.

4.1 Pillar 1: Factual Integrity & Hallucination

We evaluated models on their ability to distinguish cultural facts from fiction (RQ1) and legitimate rituals from safety violations (RQ11).

Concept Bleeding (RQ1) We observed instances of "hallucination leakage" primarily in older and open-weights models. As shown in the qualitative analysis, models occasionally conflated the *Naga* ethnic group (India/Myanmar) with the mythical Naga serpent.

- **High Hallucination:** Grok and GPT-3.5 exhibited the highest observed rates of concept bleeding, describing Naga farmers as having "scales" or "coiled tails" in 14% of generated samples.
- **High Accuracy:** Gemini 2.0 and DeepSeek demonstrated strong separation of concepts, accurately describing the Naga as a tribal community celebrating the "Water Festival" without mythological interference.

This observation aligns with the hypothesis that "creative" models may sometimes sacrifice semantic precision for narrative flair (Jiang et al., 2024). Complete hallucination frequencies and leakage rates for all models are reported in Appendix C.

Safety Refusals (RQ11): False positive rates for safety refusals were low across all models ($< 0.1\%$). Models generally distinguished ritualistic context (e.g., "Fire-Jumping") from actual self-harm prompts, suggesting that current safety training is effectively context-aware regarding these cultural practices.

4.2 Pillar 2: Representational Harm

We attempted to quantify how models might misrepresent cultural identities through "exoticism" and demographic defaulting. Appendix D reports Westernization and exoticism scores for all cultures and models.

The Exoticism Index (RQ8): We computed an "Exoticism Score" based on the frequency of adjectives like *ancient, mystical, timeless* versus *modern, contemporary*. As detailed in Table 4, models appeared to apply a "National Geographic Gaze" more frequently to Indigenous cultures while often framing Western cultures as modern.

Culture	Score	Culture	Score
Most Exoticized		**Least Exoticized**	
Dogon	1.21	Dutch	0.34
Maya	1.19	Danish	0.41
Hopi	1.11	Russian	0.42
Ainu	1.07	German	0.45
Naga	1.05	French	0.48

Table 4: The Exoticism Index: Ratio of "Mystical/Ancient" to "Modern" descriptors. Higher scores indicate greater exoticization (Ghosh et al., 2025).

This trend risks depicting Indigenous groups as "frozen in time," potentially denying them contemporary agency (Perera et al., 2025).

Demographic Defaulting (RQ12): For gender-neutral roles, models exhibited a substantial male skew (Table 5). The "Farmer" role was assumed male in **91.3%** of stories, and "Village Elder" in **78.6%**. Notably, no role defaulted to female; even "Parent" or "Kid" roles showed a slight male skew. These results suggest that LLMs may reflect and amplify occupational gender stereotypes found in training data (Kotek et al., 2023; Luca et al., 2025).

Sentiment Stereotyping (RQ3): Sentiment analysis indicated a potential bias against specific demographics. The role "Unmarried Woman" received a lower average sentiment score (0.105) compared to "Young Woman" (0.151) and "Woman" (0.148). Models frequently framed unmarried status through tropes of melancholia or longing, consistent with

Role	Male %	Bias Direction
Farmer	91.3%	**Strong Male**
Village Elder	78.6%	**Strong Male**
Kid	56.2%	*Leaning Male*
Parent	52.1%	*Neutral*

Table 5: Gender Defaulting in Neutral Roles. Values indicate the percentage of stories where the protagonist was identified as Male.

findings on sentiment bias in marginalized groups (Ungless et al., 2023; Sadhu et al., 2024).

4.3 Pillar 3: Linguistic Quality & Style

We distinguished between "Scholars" (safe/generic) and "Poets" (creative/authentic) using lexical metrics. Appendix F complements with additional stylistic indicators such as cultural token density, lexical overlap, sentence complexity, and named entity usage.

Lexical Diversity (RQ6): Table 6 highlights a possible "Reliability-Creativity Trade-off." *Qwen* and *Grok* achieved the highest Type-Token Ratios (TTR), indicating rich vocabulary usage. In contrast, "Corporate" models like *Llama-4* and *GPT-OSS* showed lower diversity, relying on repetitive "safe" phrasing (e.g., "tapestry of life," "hustle and bustle") (Martínez et al., 2025).

Code-Switching (RQ4): Authentic cultural storytelling typically benefits from code-switching (Khanuja et al., 2020). *Qwen* and *Mistral* actively integrated native terminology (2.18%–2.55% of tokens), whereas *GPT-3.5* and *GPT-OSS* largely Anglicized concepts (0.00%–0.12%), translating terms like "momo" to "dumpling," which may reduce cultural specificity.

4.4 Pillar 4: Temporal Stability (RQ10)

Our longitudinal analysis suggests the presence of "Temporal Drift" (Bajpai and Chakraborty, 2025; Khairnar, 2025). Over a 48-hour continuous generation window without resets:

- **Sentiment Decay:** Average sentiment scores dropped from **0.187** (start) to **0.073** (end).
- **Verbosity Increase:** Average story length increased from **400** to **900+** words.

This "Grumpy Rambler" effect implies that extended context windows may degrade narrative coherence and emotional tone (Guo et al., 2025), posing challenges for long-form cultural content generation. Detailed adherence and refusal statistics per model are provided in Appendix G. Appendix H provides hourly length and sentiment trajectories, as well as model-level dialogue density and cliché usage supporting these findings.

5 Conclusion

This study presents a large-scale audit of the "cultural imagination" of 10 state-of-the-art Large Language Models (LLMs), highlighting a potential tension in the current paradigm of AI alignment. Through the analysis of 37,080 synthetic narratives across 206 cultures, we have characterized what we term a *Cultural Trilemma*: current models appear to face significant challenges in simultaneously optimizing for *Factual Integrity*, *Creative Authenticity*, and *Cultural Neutrality*. Instead, our observations suggest they tend to diverge into distinct archetypes—the "Scholar" (safe but homogenized), the "Poet" (authentic but hallucinatory), and the "Corporate" (fluent but exoticizing).

Our findings suggest that the prevailing "alignment tax" (Lindström et al., 2024; Sharma et al., 2025) may impose a Western-centric "global mean" on cultural expression. While models like Gemini and DeepSeek appear to excel at avoiding safety violations (Young et al., 2025), they frequently seem to strip narratives of specific sociolinguistic markers—such as code-switching and dialect—that constitute cultural depth (Khanuja et al., 2020). Conversely, models that embrace linguistic diversity, such as Qwen and Mistral, appear prone to "concept bleeding," occasionally conflating distinct cultural identities with mythological tropes (Jiang et al., 2024). Furthermore, we documented instances of representational harm, including a "National Geographic Gaze" that risks freezing Indigenous cultures in the past (Ghosh et al., 2025) and persistent gender defaulting in occupational roles (Kotek et al., 2023).

The *Cultural Imagination Corpus* offers a novel, dynamic benchmark for the community, attempting to move beyond static factoid evaluation (Li et al., 2024) to assess the "show, don't tell" capacity of generative storytelling. Our identification of "Temporal Drift" (Bajpai and Chakraborty, 2025)—where models appeared to degrade into "Grumpy Ramblers" over extended contexts—further underscores the potential fragility of current architectures in sustaining long-form cultural coherence.

Ultimately, we propose that cultural competence

Model	Archetype	Lexical Diversity (TTR)	Code-Switching Rate	Dialogue Density
Qwen	*The Poet*	**0.650**	**2.55%**	0.42
Mistral Large	*The Poet*	0.612	2.18%	**1.28**
Grok	*Wildcard*	0.638	1.45%	0.55
Gemini 2.0	*The Scholar*	0.584	0.85%	0.31
DeepSeek	*The Scholar*	0.591	0.92%	0.28
Llama 4	*The Corporate*	0.531	0.45%	0.15
GPT-4o	*The Corporate*	0.529	0.21%	0.10
GPT-OSS	*The Corporate*	0.528	0.12%	0.02
Gemma 2	*The Corporate*	0.515	0.18%	0.08
GPT-3.5	*Legacy*	0.495	0.00%	0.05

Table 6: Linguistic Quality Metrics across the 10 Evaluated Models. "Poet" models show higher lexical diversity and code-switching, while "Scholars" and "Corporates" prioritize safety and fluency.

in AI likely cannot be achieved through superficiexal translation or broad safety filters alone (Dai et al., 2025). It may require a re-evaluation of training data curation, alignment objectives, and evaluation metrics. Future research should aim to bridge the gap between synthetic auditing and human-centric validation (Dev et al.), helping to ensure that the next generation of LLMs does not merely simulate diversity, but more genuinely reflects the pluralistic reality of the human experience.

6 Limitations

While our audit of the *Cultural Imagination Corpus* offers empirical insights into the "Cultural Trilemma," we recognize several limitations inherent to our methodology that frame critical directions for future research.

The Synthetic Data Constraint Our study relies on synthetic narratives generated by LLMs. While this approach facilitates scale (37,080 stories across 206 cultures) and statistical power (Neerukonda et al., 2025), it lacks a direct human baseline. Consequently, we cannot rigorously claim that a "Poet" model's story is *culturally authentic* to a human member of that culture—only that it exhibits higher sociolinguistic complexity than a "Scholar" model. Future work should aim to bridge this gap through participatory design, potentially incorporating human-authored narratives and direct community validation (Dev et al.; Essuman and Buys, 2025) to establish a more robust "Gold Standard" for cultural alignment.

Anglocentric Prompting Bias Despite auditing for code-switching and cultural specificity, our prompting infrastructure remained primarily Anglocentric. By querying models in English, we may inherently frame the task through a Western linguistic lens, which could constrain the model's access to deep cultural knowledge embedded in native languages (Dai et al., 2025). As discussed by (Havaldar et al., 2025), true cultural competence involves navigating pragmatics often lost in translation. Future audits would benefit from employing multilingual prompting strategies (Huang et al., 2025) to verify if the "Cultural Flattening" observed in English persists when models are queried in native scripts.

Bias in Evaluation Metrics Our reliance on automated sentiment analysis tools (e.g., VADER, RoBERTa) potentially introduces epistemic bias. These tools are predominantly trained on Western, English-language text and may misinterpret non-Western emotional expression or context-dependent polarity (Ungless et al., 2023). For instance, the "melancholy" detected in "Unmarried Woman" narratives might be a measurement artifact rather than a generative one. Developing culturally-calibrated sentiment metrics (Yang et al., 2025) would be a valuable step toward more accurate automated auditing.

Cultural Essentialism vs. Hybridity To operationalize our audit, we categorized cultures into discrete labels (e.g., "Maori," "Kazakh"). However, culture is fluid, intersectional, and dynamic. Our current framework risks reinforcing a monolithic view of these identities, potentially overlooking the nuances of diasporic or hybrid cultural experiences (Ma et al., 2022). The "Exoticism Index" (Ghosh et al., 2025), while useful, captures only one dimension of representational harm. Future frameworks should attempt to move beyond static labels to evaluate how models handle the complexity of modern, intersecting cultural identities without resorting to stereotypes.

Temporal & Computational Constraints Finally, our analysis of "Temporal Drift" (RQ10) was limited to a 48-hour generation window. While we observed signs of degradation into "Grumpy Ramblers," we did not test mitigation strategies such as context window resets or system-prompt reinforcement (Khairnar, 2025). Additionally, the computational cost of evaluating 10 models across 18 prompts restricts the frequency of such audits. Developing lightweight, proxy metrics for cultural competence presents an ongoing challenge for the community.

References

Lama Ayash, Hassan Alhuzali, Ashwag Alasmari, and Sultan Aloufi. 2025. Saudiculture: A benchmark for evaluating large language models' cultural competence within saudi arabia. *Journal of King Saud University Computer and Information Sciences*, 37(6):123.

Ashutosh Bajpai and Tanmoy Chakraborty. 2025. Temporal referential consistency: Do llms favor sequences over absolute time references? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17629–17647.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406.

Nicholas Clark, Ryan Bai, and Tanu Mitra. 2025. Offscript: Automated auditing of instruction adherence in llms. *Preprint*, arXiv:2512.10172.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in llms via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24521–24537.

Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building nlp evaluation resources with llms and community engagement for scale and depth. In *Deep Learning Indaba 2023*.

Michael Dugeri. 2024. The cannibalization of culture: Generative ai and the appropriation of indigenous african musical works. *J. Intell. Prop. & Info. Tech. L.*, 4:17.

Catherine Nana Nyaah Essuman and Jan Buys. 2025. Story generation with large language models for african languages. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 115–125.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Sourojit Ghosh, Sanjana Gautam, Pranav Narayanan Venkit, and Avijit Ghosh. 2025. Documenting patterns of exoticism of marginalized populations within text-to-image generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1107–1119.

Jiacheng Guo, Yue Wu, Jiahao Qiu, Kaixuan Huang, Xinzhe Juan, Ling Yang, and Mengdi Wang. 2025. Temporal consistency for llm reasoning process error identification. *arXiv preprint arXiv:2503.14495*.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. 2025. Culturally-aware conversations: A framework & benchmark for llms. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229.

Shulin Huang, Linyi Yang, and Yue Zhang. 2025. Mceval: A dynamic framework for fair multilingual cultural evaluation of llms. *arXiv preprint arXiv:2507.09701*.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Sushil Khairnar. 2025. A survey of temporal drift in large language models. *Authorea Preprints*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. 2024. Gender bias in llm-generated interview responses. *arXiv preprint arXiv:2410.20739*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. Ai alignment through reinforcement learning from human feedback? contradictions and limitations. *arXiv preprint arXiv:2406.18346*.

Massimiliano Luca, Ciro Beneduce, Bruno Lepri, and Jacopo Staiano. 2025. The llm wears prada: Analysing gender bias and stereotypes through online shopping data. *arXiv preprint arXiv:2504.01951*.

Weicheng Ma, Samiha Datta, Lili Wang, and Soroush Vosoughi. 2022. Encbp: A new benchmark dataset for finer-grained cultural background prediction in english. *arXiv preprint arXiv:2203.14498*.

Moon-Kuen Mak and Tiejian Luo. 2025. A framework for evaluating cultural bias and historical misconceptions in llms outputs. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5(3):100235.

Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2025. Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–15.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

Chandana Neerukonda, Rajath Shetty, Felipe Soares, and Makesh Narsimhan Sreedhar. 2025. Synthetic data generation using large language models. US Patent App. 18/505,739.

Maneesha Perera, Rajith Vidanaarachchi, Sangeetha Chandrashekeran, Melissa Kennedy, Brendan Kennedy, and Saman Halgamuge. 2025. Indigenous peoples and artificial intelligence: A systematic review and future directions. *Big Data & Society*, 12(2):20539517251349170.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602.

Jayanta Sadhu, Maneesha Saha, and Rifat Shahriyar. 2024. An empirical study of gendered stereotypes in emotional attributes for bangla in multilingual large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 384–398.

Nihar Ranjan Sahoo, Ashita Saxena, Kishan Maharaj, Arif A Ahmad, Abhijit Mishra, and Pushpak Bhattacharyya. 2024. Addressing bias and hallucination in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 73–79.

Sankalp Sethi, Joni Salminen, Danial Amin, and Bernard J Jansen. 2025. " when ai writes personas": Analyzing lexical diversity in llm-generated persona descriptions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Raghav Sharma, Manan Mehta, and Sai Tiger Raina. 2025. Rlhf: A comprehensive survey for cultural, multimodal and low latency alignment methods. *arXiv preprint arXiv:2511.03939*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Eddie L Ungless, Björn Ross, and Vaishak Belle. 2023. Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias. *Social science computer review*, 41(6):2211–2229.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Bei Xiao, Xufeng Duan, David A Haslett, and Zhenguang Cai. 2025. Human-likeness of llms in the mental lexicon. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 586–601.

Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Narrate dialogues for better summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3565–3575.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Jenny Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317.

Richard J Young, Brandon Gillins, and Alice M Matthews. 2025. When models can't follow: Testing instruction adherence across 256 llms. *arXiv preprint arXiv:2510.18892*.

A Research Question Framework

This appendix explains the full set of research questions guiding the analyses in this paper. The purpose of this appendix is conceptual clarification: it describes what each question probes and why it is relevant for understanding culturally grounded narrative generation by large language models. No empirical results are reported here.

A.1 Hallucination and Cultural Attribution

RQ1 examines whether models hallucinate nonexistent or misplaced cultural details when prompted to describe culturally specific events. This question targets cultural misattribution errors, such as assigning rituals, symbols, or practices to incorrect cultural groups.

A.2 Cultural Homogenization in Universal Contexts

RQ2 investigates whether narratives describing ostensibly universal experiences meaningfully reflect the assigned cultural context or default to culturally dominant norms. The focus is on detecting homogenized framing when explicit cultural cues are limited.

A.3 Sentiment and Role-Conditioned Bias

RQ3 asks whether narrative sentiment varies systematically with assigned roles or cultural identities. This question probes subtle tonal regularities that may reflect latent stereotypes rather than explicit negative language.

A.4 Linguistic Code-Switching and Cultural Depth

RQ4 examines the extent to which models employ culturally marked lexical items or non-English tokens to signal authenticity. This question evaluates whether surface-level linguistic grounding is used consistently across models and cultures.

A.5 Verbosity versus Cultural Content

RQ5 addresses whether longer narratives contain proportionally more cultural content or whether cultural specificity becomes diluted as generation length increases. This question disentangles narrative verbosity from cultural depth.

A.6 Vocabulary Richness as Stylistic Signal

RQ6 investigates differences in lexical diversity across models using vocabulary richness metrics. Rather than evaluating creativity directly, this question treats lexical variety as a stylistic indicator.

A.7 Instruction Adherence

RQ7 evaluates how reliably models follow explicit structural instructions in prompts, such as narrative perspective, format, or framing constraints. This question distinguishes semantic understanding from procedural compliance.

A.8 Exoticization and Descriptive Framing

RQ8 examines whether certain cultures are disproportionately described using exoticizing language. The focus is on descriptive framing that emphasizes otherness, mysticism, or symbolic distance rather than factual accuracy.

A.9 Role-Based Stereotype Reinforcement

RQ9 explores whether narrative roles are associated with rigid semantic or conceptual clusters. This question assesses whether models rely on stereotypical archetypes instead of flexible role portrayals.

A.10 Temporal Stability of Generation

RQ10 investigates whether narrative characteristics change systematically over extended generation periods. This includes variation in length, sentiment, or stylistic features attributable to temporal drift.

A.11 Safety Behavior and Refusal Patterns

RQ11 examines whether models refuse to generate narratives for particular cultures, roles, or contexts. Rather than treating refusals as isolated incidents, this question frames them as a systematic behavioral signal.

A.12 Gender Defaulting in Neutral Roles

RQ12 asks whether models default to gendered language when roles are explicitly neutral. This question probes implicit gender bias in pronoun usage and character construction.

A.13 Cliché Formation and Repetition

RQ13 analyzes the prevalence of recurring stock phrases and formulaic expressions. Cliché frequency is used as a proxy for narrative saturation and creative convergence.

A.14 Dialogue Utilization

RQ14 examines how frequently models employ dialogue as a narrative device. Dialogue density

provides insight into stylistic preferences and narrative dynamism.

A.15 Entity Density and Referential Grounding

RQ15 investigates how densely narratives reference named entities such as people, places, or institutions. This question evaluates whether models ground stories in concrete referents or rely on abstract exposition.

A.16 Sentence-Level Structural Complexity

RQ16 examines differences in sentence complexity across models using metrics such as average sentence length and variance. This question captures stylistic tendencies related to syntactic elaboration and readability.

A.17 Lexical Distinctiveness Across Generations

RQ17 explores the degree to which models reuse similar vocabularies across narratives. Lower lexical overlap indicates higher distinctiveness, while higher overlap suggests convergence toward fixed phrasing patterns.

A.18 Training Data Leakage Indicators

RQ18 examines whether generated narratives exhibit signals of memorization or excessive similarity to known cultural descriptions. This question assesses potential training data leakage without assuming direct copying.

B Dataset Specification

This appendix describes the dataset used in all experiments reported in this paper. The dataset consists of narrative texts generated by large language models under controlled and structured prompting conditions. Each record corresponds to a single generated narrative paired with its associated metadata.

B.1 Dataset Overview

B.2 Generation Design

Narratives were generated using a template-driven prompting framework. Each instance is defined by a combination of:

- a language model identifier,
- a target cultural label,
- a narrative role or perspective,
- an event or scenario description.

Prompts were instantiated from a fixed pool of templates to ensure consistency in structure and instruction phrasing across models and cultures. This design allows controlled comparison of model outputs under identical narrative constraints while varying cultural and contextual parameters.

Attribute	Value
Dataset name	Cultural Stories
Total narratives	37,080
Models	10
Cultures	206
Roles / perspectives	13
Event types	15
Prompt templates	18
Columns	11
Disk size	128.29 MB
Memory footprint	244.62 MB
Missing values	4.07%

Table 7: Summary statistics of the narrative generation dataset.

Field	Description
story_id	Unique identifier for each generated narrative.
model	Identifier of the language model that produced the narrative.
culture	Target cultural identity specified in the prompt.
type	Prompt category (*specific* or *universal*).
role_or_perspective	Narrative role or point of view.
event	Event or situation described in the narrative.
prompt	Fully instantiated prompt provided to the model.
original_prompt_template	Prompt template prior to culture instantiation.
native_culture	Optional native cultural label, when available.
generated_at	Timestamp of generation.
story	Generated narrative text.

Table 8: Dataset fields and descriptions.

B.3 Dataset Schema

B.4 Prompt Templates

All narratives were generated from a finite set of prompt templates designed to control narrative format, perspective, and situational framing. Templates vary along the following dimensions:

1. narrative format (e.g., short story vs. diary-style entry),
2. situational context (event-focused vs. general experience),

3. explicitly specified role or perspective.

Representative template forms include:

- *"Write a short story from the perspective of a <culture> unmarried woman participating in a fire-jumping ritual."*
- *"Write a short story about a <culture> man on his first day at a new job."*
- *"Write a short story in the form of a diary entry by a <culture> woman who is upset."*

B.5 Missing Data

Missing values are limited and localized. The `story` field contains a small number of null entries (0.28%), primarily due to generation failures. The `native_culture` field is intentionally sparse (44.4% missing) and is not required for any analysis presented in this paper.

No imputation or normalization procedures were applied beyond excluding null narrative texts where necessary.

B.6 Scope and Usage

All analyses presented in this paper operate directly on this dataset or on deterministic transformations derived from it. No additional filtering, relabeling, or post-processing steps were applied that would alter the underlying data distribution described above.

C Hallucination, Leakage, and Temporal Drift

This appendix reports statistics related to factual inaccuracies, content leakage, and temporal variation observed in generated narratives. All values are computed over the full dataset described in Appendix B.

C.1 Model Identifier Abbreviations

To improve table readability, shortened model identifiers are used throughout this appendix, as defined below:

Abbreviation	Model
GPT-OSS	openai/gpt-oss-120b
GROK	x-ai/grok-4-fast
QWEN	qwen/qwen3-32b
MISTRAL	mistralai/mistral-large-2512
DEEPSEEK	deepseek/deepseek-v3.2
GEMINI	google/gemini-2.0-flash-001
GEMMA	google/gemma-3-27b-it
LLAMA-S	meta-llama/llama-4-scout
LLAMA-M	meta-llama/llama-4-maverick
GPT-3.5	openai/gpt-3.5-turbo

Model	Stories	Hallu.	Rate	Rate (%)
GPT-OSS	3,708	137	0.0369	3.69
GROK	3,708	134	0.0361	3.61
QWEN	3,708	89	0.0240	2.40
MISTRAL	3,708	84	0.0227	2.27
DEEPSEEK	3,708	81	0.0218	2.18
GEMINI	3,708	68	0.0183	1.83
GEMMA	3,708	56	0.0151	1.51
LLAMA-S	3,708	29	0.0078	0.78
LLAMA-M	3,708	27	0.0073	0.73
GPT-3.5	3,708	16	0.0043	0.43

Table 9: Hallucination frequency and rate by model.

Model	Leakage Rate	Leakage (%)
DEEPSEEK	0.0000	0.00
GEMINI	0.0005	0.05
GEMMA	0.0019	0.19
LLAMA-M	0.0221	2.21
LLAMA-S	0.0038	0.38
MISTRAL	0.0000	0.00
GPT-3.5	0.0000	0.00
GPT-OSS	0.0000	0.00
QWEN	0.0000	0.00
GROK	0.0000	0.00

Table 10: Estimated content leakage rates by model.

C.2 Hallucination Frequency by Model

Hallucinations are defined as the presence of culturally or contextually incorrect entities, practices, or factual claims that are inconsistent with the target culture specified in the prompt.

C.3 Training Data Leakage Indicators

Leakage is measured as the proportion of narratives exhibiting strong lexical or structural overlap with known cultural descriptions or widely documented sources, exceeding a predefined similarity threshold.

C.4 Temporal Drift in Narrative Length and Sentiment

To examine temporal variation during generation, narratives were grouped by hourly generation time. Average narrative length (in words) and mean sentiment score were computed per time bin.

C.5 Summary

The results in this appendix indicate systematic variation across models in hallucination frequency, minimal evidence of direct training data leakage for most systems, and measurable temporal drift in both narrative length and sentiment over extended generation periods.

Hour	Avg. Words	Avg. Sentiment
01:00	490.29	0.187
04:00	428.34	0.156
08:00	525.52	0.131
12:00	492.53	0.134
16:00	583.07	0.121
20:00	587.79	0.106
23:00	636.54	0.107
03:00	814.30	0.096
09:00	473.15	0.084
13:00	464.18	0.095
17:00	796.66	0.096
21:00	702.20	0.108
02:00	679.41	0.081

Table 11: Representative hourly averages of narrative length and sentiment score.

D Westernization and Exoticism Bias

This appendix reports statistics related to westernized framing and exoticized representations observed in generated narratives. Scores are computed by aggregating lexical and semantic indicators associated with Western cultural defaults and exoticizing descriptors, respectively.

D.1 Model Identifier Abbreviations

For compact presentation, shortened model identifiers are used throughout this appendix:

Abbreviation	Model
GPT-OSS	openai/gpt-oss-120b
GROK	x-ai/grok-4-fast
QWEN	qwen/qwen3-32b
MISTRAL	mistralai/mistral-large-2512
DEEPSEEK	deepseek/deepseek-v3.2
GEMINI	google/gemini-2.0-flash-001
GEMMA	google/gemma-3-27b-it
LLAMA-S	meta-llama/llama-4-scout
LLAMA-M	meta-llama/llama-4-maverick
GPT-3.5	openai/gpt-3.5-turbo

D.2 Westernized Framing by Culture

Table 12 reports cultures with the highest average Westernization scores. Higher values indicate stronger alignment with Western narrative defaults such as institutional settings, individualist framing, or Eurocentric social norms.

D.3 Westernized Framing by Model

D.4 Exoticized Representation by Culture

Exoticism scores capture the prevalence of narrative devices emphasizing perceived otherness, such as ritual mystification, ancestral romanticization, and symbolic distance from modern life. Table 14 lists cultures with the highest average exoticism scores.

Culture	Avg. Western Score
Coptic	1.84
Russian	1.76
Korean	1.46
Georgian	1.40
Bulgarian	1.38
Italian	1.38
Serbian	1.34
Macedonian	1.33
Danish	1.31
Lithuanian	1.26
Syriac	1.25
Czech	1.24
Croatian	1.23
Greek	1.23
Finnish	1.20
Slovak	1.19
Ukrainian	1.18
Polish	1.16
Slovenian	1.15
Hungarian	1.14

Table 12: Cultures with the highest average Westernization scores.

Model	Avg. Western Score
GPT-OSS	1.18
DEEPSEEK	0.89
GROK	0.79
MISTRAL	0.74
LLAMA-S	0.66
GEMINI	0.55
QWEN	0.55
LLAMA-M	0.46
GEMMA	0.39
GPT-3.5	0.18

Table 13: Average Westernization score by model.

D.5 Coverage Note

Due to space constraints, only the highest-scoring cultures are shown in Tables 12 and 14. Scores for all remaining cultures were computed using the same procedure and are available in the accompanying data files.

D.6 Summary

The statistics reported in this appendix indicate systematic variation in both Westernized framing and exoticized representation across cultures and models. These patterns are not uniformly distributed and differ substantially depending on the generating system and the target cultural label.

E Sentiment and Role-Based Representation

This appendix reports statistics related to sentiment variation, gendered language usage, and lexical

Culture	Avg. Exotic Score
Dogon	1.21
Maya	1.19
Hopi	1.11
Ainu	1.07
Naga	1.05
Mapuche	1.05
Sakha	1.04
Batak	1.03
Khakas	1.02
Tahitian	1.02
Nahuatl	0.99
Quechua	0.99
Zapotec	0.97
Limbu	0.97
Cherokee	0.96
Manchu	0.94
Sichuan Yi	0.94
Yakut	0.93
Navajo	0.92
Guaraní	0.92

Table 14: Cultures with the highest average exoticism scores.

Role	Mean	Count	Std.	Min	Max
Pilgrim	0.080	2,060	0.063	-0.13	1.00
Unmarried Woman	0.105	2,060	0.062	-0.50	0.39
Village Elder	0.105	2,060	0.072	-0.75	0.37
Wrestler	0.107	2,060	0.076	-0.12	0.48
Man	0.107	6,180	0.100	-1.00	0.57
Woman	0.114	6,180	0.102	-0.80	0.63
Drummer	0.116	2,060	0.062	-0.11	0.33
Teenage Boy	0.117	2,060	0.087	-0.75	0.45
Grandmother	0.124	2,060	0.073	-0.10	0.44
Farmer	0.135	2,060	0.071	-0.75	0.33
Fisherman	0.136	2,060	0.069	-0.75	0.41
Young Woman	0.151	2,060	0.078	-0.08	1.00
Kid	0.202	4,120	0.113	-0.50	0.78

Table 15: Sentiment statistics aggregated by narrative role or perspective.

patterns associated with different narrative roles or perspectives. All values are computed over the full dataset described in Appendix B.

E.1 Sentiment Distribution by Role

Table 15 summarizes sentiment statistics aggregated by narrative role or perspective. Scores are reported as the mean sentiment value per role, along with dispersion and range.

E.2 Gendered Pronoun Usage by Role

To assess gendered language patterns, we compute the average ratio of male pronouns (e.g., *he*, *his*) to total gendered pronouns for selected roles.

Role	Avg. Male Pronoun Ratio
Farmer	0.913
Village Elder	0.786
Kid	0.588
Pilgrim	0.528

Table 16: Average proportion of male pronouns by role.

Word	Count	Role
village	9,886	Village Elder
families	8,750	Village Elder
feast	6,977	Village Elder
gifts	4,844	Village Elder
elder	3,741	Village Elder
uncle	10,008	Kid
wedding	5,491	Kid
family	4,903	Kid
love	4,204	Kid
mother	3,232	Kid

Table 17: High-frequency lexical items associated with selected roles.

E.3 Role-Specific Lexical Salience

Table 17 lists the most frequent content words associated with two representative roles. Counts reflect total occurrences across all narratives assigned to the given role.

E.4 Summary

The statistics reported in this appendix indicate systematic variation in sentiment, gendered language usage, and lexical emphasis across narrative roles. These patterns reflect consistent role-conditioned framing differences in generated narratives rather than isolated outliers.

F Linguistic Density and Stylistic Variation

This appendix reports statistics related to linguistic density, lexical diversity, stylistic overlap, sentence-level complexity, and named entity usage in generated narratives. All metrics are computed over the dataset described in Appendix B.

F.1 Model Identifier Abbreviations

For compact presentation, shortened model identifiers are used throughout this appendix:

Model	Avg. Density (%)
QWEN	2.55
MISTRAL	2.18
GEMMA	1.35
DEEPSEEK	1.26
GPT-OSS	1.07
GEMINI	0.19
GROK	0.08
LLAMA-M	0.00
LLAMA-S	0.00
GPT-3.5	0.00

Table 18: Average cultural token density by model.

Model	Avg. TTR	Std. Dev.
QWEN	0.650	0.048
GROK	0.638	0.046
MISTRAL	0.608	0.048
GEMMA	0.593	0.040
GEMINI	0.590	0.046
DEEPSEEK	0.581	0.044
GPT-3.5	0.548	0.055
LLAMA-M	0.534	0.051
LLAMA-S	0.531	0.044
GPT-OSS	0.528	0.061

Table 19: Lexical diversity measured by type–token ratio.

Abbreviation	Model
GPT-OSS	openai/gpt-oss-120b
GPT-3.5	openai/gpt-3.5-turbo
GROK	x-ai/grok-4-fast
QWEN	qwen/qwen3-32b
MISTRAL	mistralai/mistral-large-2512
DEEPSEEK	deepseek/deepseek-v3.2
GEMINI	google/gemini-2.0-flash-001
GEMMA	google/gemma-3-27b-it
LLAMA-S	meta-llama/llama-4-scout
LLAMA-M	meta-llama/llama-4-maverick

F.2 Cultural Token Density

Cultural token density is measured as the proportion of tokens identified as culturally specific terms relative to total tokens.

F.3 Lexical Diversity

Lexical diversity is quantified using the type–token ratio (TTR), computed per narrative and averaged by model.

F.4 Lexical Overlap Across Generations

Lexical overlap is measured using Jaccard similarity over content-word vocabularies, averaged by model. Lower values indicate higher lexical distinctiveness.

Model	Jaccard Similarity
QWEN	0.238
GPT-OSS	0.260
LLAMA-M	0.312
DEEPSEEK	0.331
LLAMA-S	0.347
GROK	0.354
MISTRAL	0.362
GEMINI	0.373
GEMMA	0.377
GPT-3.5	0.379

Table 20: Average lexical overlap across narratives by model.

Model	Avg. WPS	Std. Dev.	Min	Max
GPT-OSS	22.55	20.18	0.0	1108.0
GPT-3.5	20.37	3.24	6.0	31.6
LLAMA-M	19.49	9.39	0.0	216.0
LLAMA-S	17.98	3.42	0.0	37.6
GROK	17.36	4.59	7.2	50.7
QWEN	15.29	4.37	0.0	143.0
DEEPSEEK	15.15	129.36	0.0	7887.0
GEMINI	13.74	2.43	6.7	29.4
MISTRAL	13.32	3.50	5.8	37.4
GEMMA	12.39	2.18	0.0	27.6

Table 21: Sentence-level complexity statistics by model (WPS: words per sentence).

F.5 Sentence-Level Complexity

Sentence complexity is summarized using average words per sentence and dispersion statistics.

F.6 Named Entity Density

Named entity density is computed as the average number of recognized named entities per narrative.

F.7 Summary

The statistics reported in this appendix show substantial variation across models in cultural token usage, lexical diversity, stylistic overlap, sentence complexity, and entity density. These differences reflect systematic stylistic and structural characteristics of generated narratives rather than isolated effects.

G Instruction Compliance and Refusals

This appendix reports statistics related to instruction compliance and refusal behavior observed in generated narratives. All metrics are computed over the dataset described in Appendix B.

G.1 Model Identifier Abbreviations

For compact presentation, shortened model identifiers are used throughout this appendix:

Model	Entity Density
GEMMA	6.76
QWEN	6.02
GROK	5.45
MISTRAL	5.37
GEMINI	5.20
LLAMA-S	4.79
LLAMA-M	4.42
GPT-3.5	4.21
DEEPSEEK	4.18
GPT-OSS	3.40

Table 22: Average named entity density by model.

Model	Compliance Rate	Compliance (%)
GPT-3.5	0.744	74.38
DEEPSEEK	0.744	74.35
MISTRAL	0.723	72.33
QWEN	0.718	71.79
GEMINI	0.705	70.52
LLAMA-M	0.689	68.91
GROK	0.684	68.42
LLAMA-S	0.680	68.04
GPT-OSS	0.679	67.91
GEMMA	0.635	63.46

Table 23: Instruction compliance rates by model.

Abbreviation	Model
GPT-OSS	openai/gpt-oss-120b
GPT-3.5	openai/gpt-3.5-turbo
GROK	x-ai/grok-4-fast
QWEN	qwen/qwen3-32b
MISTRAL	mistralai/mistral-large-2512
DEEPSEEK	deepseek/deepseek-v3.2
GEMINI	google/gemini-2.0-flash-001
GEMMA	google/gemma-3-27b-it
LLAMA-S	meta-llama/llama-4-scout
LLAMA-M	meta-llama/llama-4-maverick

G.2 Instruction Compliance Rates

Instruction compliance is defined as the proportion of narratives that adhere to the required format, perspective, and event specification of the prompt.

G.3 Refusal Incidence

Refusals are defined as explicit model responses that decline to generate a narrative due to policy or safety constraints.

G.4 Summary

The results in this appendix indicate variation across models in instruction compliance, with relatively low refusal incidence overall. Differences in compliance rates primarily reflect adherence to prompt format and role specification rather than outright refusals.

Model	Refusal Count
DEEPSEEK	0
GEMINI	0
GEMMA	0
LLAMA-M	1
LLAMA-S	2
MISTRAL	0
GPT-3.5	0
GPT-OSS	0
QWEN	0
GROK	0

Table 24: Observed refusal counts by model.

H Temporal Drift and Longitudinal Effects

This appendix reports statistics related to temporal variation in narrative length and sentiment, as well as model-level differences in dialogue usage and cliché frequency. All metrics are computed over the dataset described in Appendix B.

H.1 Model Identifier Abbreviations

For compact presentation, shortened model identifiers are used throughout this appendix:

Abbreviation	Model
GPT-OSS	openai/gpt-oss-120b
GPT-3.5	openai/gpt-3.5-turbo
GROK	x-ai/grok-4-fast
QWEN	qwen/qwen3-32b
MISTRAL	mistralai/mistral-large-2512
DEEPSEEK	deepseek/deepseek-v3.2
GEMINI	google/gemini-2.0-flash-001
GEMMA	google/gemma-3-27b-it
LLAMA-S	meta-llama/llama-4-scout
LLAMA-M	meta-llama/llama-4-maverick

H.2 Temporal Variation in Length and Sentiment

To examine longitudinal effects during generation, narratives were grouped by hour of generation time. For each time bin, the average total word count and mean sentiment score were computed.

Across the observed period, average narrative length tends to increase over time, while sentiment scores show a gradual downward trend with moderate fluctuations.

H.3 Dialogue Density by Model

Dialogue density is measured as the average number of dialogue segments per narrative.

H.4 Cliché Frequency

Cliché frequency is computed as the average number of recurring stock phrases per narrative.

Hour	Avg. Words	Avg. Sentiment
01:00	490.29	0.187
04:00	428.34	0.156
08:00	525.52	0.131
12:00	492.53	0.134
16:00	583.07	0.121
20:00	587.79	0.106
23:00	636.54	0.107
03:00	814.30	0.096
09:00	473.15	0.084
13:00	464.18	0.095
17:00	796.66	0.096
21:00	702.20	0.108
02:00	679.41	0.081

Table 25: Representative hourly averages of narrative length and sentiment score.

Model	Avg. Dialogue Density
MISTRAL	1.285
GROK	0.732
LLAMA-M	0.579
LLAMA-S	0.554
GEMINI	0.452
QWEN	0.330
DEEPSEEK	0.259
GEMMA	0.215
GPT-3.5	0.047
GPT-OSS	0.024

Table 26: Average dialogue density by model.

H.5 Summary

The results in this appendix show clear temporal variation in narrative length and sentiment during extended generation periods, as well as systematic differences across models in dialogue usage and cliché frequency. These patterns indicate longitudinal and stylistic effects that are consistent across large subsets of the dataset.

I List of Cultures

This appendix justifies the selection of the cultural cohort used in this study and discusses the methodological limitations associated with expanding cultural coverage. Appendix J provides the complete list of all cultures included in the dataset, along with associated metadata.

I.1 Rationale for Cultural Coverage

The dataset includes a cohort of over two hundred culturally distinct groups spanning a wide range of geographic regions, linguistic families, socio-political histories, and levels of global representation. The selection was guided by the goal of achieving *breadth rather than exhaustiveness*, enabling systematic analysis across heterogeneous cultural contexts while maintaining analytical tractability.

Model	Avg. Cliché Count
GEMMA	2.62
GPT-OSS	2.36
LLAMA-M	1.85
LLAMA-S	1.83
GEMINI	1.65
DEEPSEEK	1.44
MISTRAL	1.34
GPT-3.5	1.32
GROK	1.04
QWEN	0.81

Table 27: Average cliché count per narrative by model.

Specifically, the cohort was designed to capture:

- **Geographic diversity**, including cultures from Africa, the Americas, Europe, South Asia, East Asia, Southeast Asia, Oceania, and the Arctic.
- **Linguistic diversity**, spanning major and minor language families, including Indo-European, Afro-Asiatic, Niger–Congo, Uralic, Dravidian, Turkic, Austronesian, and indigenous language groups.
- **Sociocultural heterogeneity**, encompassing state-associated national cultures, diasporic groups, Indigenous communities, and historically marginalized or underrepresented cultures.
- **Variation in global visibility**, ranging from cultures with extensive representation in global media and training data to those with limited digital presence.

This diversity allows the analysis to probe systematic model behavior across cultures that differ substantially in how they are represented, stereotyped, or omitted in large-scale text corpora.

I.2 Why a Larger Cultural Set Was Not Used

While it is theoretically desirable to include an even larger number of cultures, several practical and methodological constraints motivated the current scope.

First, **prompt validity** becomes increasingly difficult to guarantee as cultural coverage expands. For many cultures, especially those with limited written documentation or highly localized practices, constructing prompts that are both respectful and

semantically well-defined requires domain expertise that is not uniformly available.

Second, **evaluation reliability** degrades as cultural granularity increases. Accurately identifying hallucinations, misattributions, or culturally inappropriate framing presupposes some level of external reference knowledge. For cultures with sparse documentation, distinguishing genuine model errors from plausible but unfamiliar representations becomes ambiguous.

Third, **comparability across cultures** imposes constraints. The analyses in this paper rely on controlled comparisons across cultures under shared prompt structures. Extending the set to include cultures that require fundamentally different narrative framings, social roles, or event structures would weaken cross-cultural comparability and introduce confounding factors unrelated to model behavior.

Finally, **computational and annotation costs** scale nonlinearly with cultural coverage. Expanding the cohort would substantially increase generation volume and post-hoc analysis complexity, limiting the feasibility of conducting consistent evaluations across multiple models and metrics.

I.3 Implications and Limitations

The selected cultural cohort should be understood as *illustrative rather than exhaustive*. The findings reported in this paper characterize patterns of model behavior across a broad and diverse sample of cultures, but they do not claim to represent all cultural contexts globally.

Certain cultural groups—particularly those with minimal digital footprints, oral traditions, or highly localized social structures—are likely underrepresented or absent. As a result, the analysis may underestimate failure modes that emerge primarily in such contexts.

Future work could address these limitations by incorporating expert-curated prompts, community-informed evaluations, or culture-specific generation paradigms. However, such extensions would require methodological changes beyond the scope of the present study.

lists all cultures included in the dataset along with their corresponding identifiers and summary statistics. Readers seeking culture-specific details or wishing to assess coverage at a finer granularity are referred to that appendix.

Culture Name	Continent	Country	Extinct
Ainu	Asia	Japan	No
Amazigh	Africa	Morocco	No
Ashanti	Africa	Ghana	No
Assamese	Asia	India	No
Baloch	Asia	Pakistan	No
Bambara	Africa	Mali	No
Bashkir	Asia	Russia	No
Basque	Europe	Spain	No
Batak	Asia	Indonesia	No
Bedouin	Asia	Saudi Arabia	No
Bengali	Asia	Bangladesh	No
Berber	Africa	Morocco	No
Bodo	Asia	India	No
Breton	Europe	France	No
Bulgarian	Europe	Bulgaria	No
Burusho	Asia	Pakistan	No
Catalan	Europe	Spain	No
Chamorro	Oceania	Guam	No
Chechen	Europe	Russia	No
Cherokee	Americas	United States	No
Chuvash	Europe	Russia	No
Circassian	Asia	Russia	No
Coptic	Africa	Egypt	No
Cornish	Europe	United Kingdom	No
Corsican	Europe	France	No
Cree	Americas	Canada	No
Crimean Tatar	Europe	Ukraine	No
Croatian	Europe	Croatia	No
Czech	Europe	Czech Republic	No
Dakota	Americas	United States	No
Danish	Europe	Denmark	No
Dargwa	Asia	Russia	No
Dinka	Africa	South Sudan	No
Dogon	Africa	Mali	No
Dutch	Europe	Netherlands	No
Erzya	Europe	Russia	No
Estonian	Europe	Estonia	No
Ewe	Africa	Ghana	No
Faroese	Europe	Faroe Islands	No
Fijian	Oceania	Fiji	No
Finnish	Europe	Finland	No
Flemish	Europe	Belgium	No
Frisian	Europe	Netherlands	No
Friulian	Europe	Italy	No
Fula	Africa	Nigeria	No
Galician	Europe	Spain	No
Garifuna	Americas	Belize	No
Georgian	Asia	Georgia	No
Greek	Europe	Greece	No
Greenlandic	Americas	Greenland	No
Guaraní	South America	Paraguay	No
Gujarati	Asia	India	No
Haitian	Americas	Haiti	No
Hausa	Africa	Nigeria	No
Hawaiian	Oceania	United States (Hawaii)	No
Hazaragi	Asia	Afghanistan	No
Herero	Africa	Namibia	No
Hmong	Asia	China	No
Hopi	Americas	United States	No
Hungarian	Europe	Hungary	No
Iban	Asia	Malaysia	No
Icelandic	Europe	Iceland	No

Table 28: Cultural data (Part 1 of 4).

Culture Name	Continent	Country	Extinct
Igbo	Africa	Nigeria	No
Ilocano	Asia	Philippines	No
Irish	Europe	Ireland	No
Italian	Europe	Italy	No
Javanese	Asia	Indonesia	No
Jola	Africa	Senegal	No
Jolof	Africa	Senegal	No
Kabardian	Europe	Russia	No
Kabyle	Africa	Algeria	No
Kannada	Asia	India	No
Kanuri	Africa	Nigeria	No
Karakalpak	Asia	Uzbekistan	No
Kashmiri	Asia	India	No
Kashubian	Europe	Poland	No
Kazakh	Asia	Kazakhstan	No
Khakas	Asia	Russia	No
Khanty	Asia	Russia	No
Khoekhoe	Africa	South Africa	No
Kikuyu	Africa	Kenya	No
Kinyarwanda	Africa	Rwanda	No
Kirundi	Africa	Burundi	No
Komi	Europe	Russia	No
Konkani	Asia	India	No
Korean	Asia	South Korea	No
Koryak	Asia	Russia	No
Krio	Africa	Sierra Leone	No
Kumyk	Asia	Russia	No
Kurdish	Asia	Turkey	No
Kyrgyz	Asia	Kyrgyzstan	No
Lakota	Americas	United States	No
Laz	Asia	Turkey	No
Lezgin	Asia	Azerbaijan	No
Limbu	Asia	Nepal	No
Lithuanian	Europe	Lithuania	No
Luo	Africa	Kenya	No
Luxembourgish	Europe	Luxembourg	No
Macedonian	Europe	North Macedonia	No
Magahi	Asia	India	No
Maithili	Asia	India	No
Makassar	Asia	Indonesia	No
Malagasy	Africa	Madagascar	No
Malayalam	Asia	India	No
Maltese	Europe	Malta	No
Manchu	Asia	China	No
Mande	Africa	Mali	No
Maori	Oceania	New Zealand	No
Mapuche	Americas	Chile	No
Marathi	Asia	India	No
Mari	Europe	Russia	No
Marshallese	Oceania	Marshall Islands	No
Maya	Americas	Mexico	No
Mende	Africa	Sierra Leone	No
Mising	Asia	India	No
Mizo	Asia	India	No
Mohawk	Americas	United States	No
Mongolian	Asia	Mongolia	No
Montenegrin	Europe	Montenegro	No
Mordvin	Europe	Russia	No
Naga	Asia	India	No
Nahuatl	Americas	Mexico	No
Navajo	Americas	United States	No

Table 29: Cultural data (Part 2 of 4).

Culture Name	Continent	Country	Extinct
Nenets	Asia	Russia	No
Nepali	Asia	Nepal	No
Nivkh	Asia	Russia	No
Nogai	Asia	Russia	No
Nuer	Africa	South Sudan	No
Occitan	Europe	France	No
Odia	Asia	India	No
Ojibwe	Americas	Canada	No
Oromo	Africa	Ethiopia	No
Ossetian	Europe	Russia	No
Palauan	Oceania	Palau	No
Papiamento	Americas	Aruba	No
Pashtun	Asia	Afghanistan	No
Persian	Asia	Iran	No
Polish	Europe	Poland	No
Punjabi	Asia	India	No
Quechua	Americas	Peru	No
Romani	Europe	Romania	No
Romansh	Europe	Switzerland	No
Russian	Europe and Asia	Russia	No
Rusyn	Europe	Ukraine	No
Rwandan	Africa	Rwanda	No
Sakha	Asia	Russia	No
Sami	Europe	Norway	No
Samoan	Oceania	Samoa	No
Sardinian	Europe	Italy	No
Scots	Europe	United Kingdom	No
Scottish Gaelic	Europe	United Kingdom	No
Serbian	Europe	Serbia	No
Shan	Asia	Myanmar	No
Sicilian	Europe	Italy	No
Sikkimese	Asia	India	No
Sindhi	Asia	Pakistan	No
Sinhalese	Asia	Sri Lanka	No
Slovak	Europe	Slovakia	No
Slovenian	Europe	Slovenia	No
Somali	Africa	Somalia	No
Sorbian	Europe	Germany	No
Sotho	Africa	Lesotho	No
Sundanese	Asia	Indonesia	No
Swahili	Africa	Kenya	No
Swazi	Africa	Eswatini	No
Syriac	Asia	Syria	No
Tahitian	Oceania	French Polynesia	No
Tajik	Asia	Tajikistan	No
Tamil	Asia	India	No
Tatar	Europe	Russia	No
Telugu	Asia	India	No
Tetum	Asia	East Timor	No
Thai	Asia	Thailand	No
Tibetan	Asia	China	No
Tigray	Africa	Ethiopia	No
Tigrinya	Africa	Eritrea	No
Tivi	Africa	Nigeria	No
Tokelauan	Oceania	Tokelau	No
Tongan	Oceania	Tonga	No
Tsakhur	Asia	Russia	No
Tswana	Africa	Botswana	No
Tulu	Asia	India	No
Turkmen	Asia	Turkmenistan	No
Tuvan	Asia	Russia	No

Table 30: Cultural data (Part 3 of 4).

Culture Name	Continent	Country	Extinct
Udmurt	Europe	Russia	No
Uighur	Asia	China	No
Ukrainian	Europe	Ukraine	No
Urhobo	Africa	Nigeria	No
Uyghur	Asia	China	No
Uzbek	Asia	Uzbekistan	No
Venda	Africa	South Africa	No
Venetian	Europe	Italy	No
Welsh	Europe	United Kingdom	No
Wolof	Africa	Senegal	No
Xhosa	Africa	South Africa	No
Yakut	Asia	Russia	No
Yao	Asia	China	No
Yapese	Oceania	Federated States of Micronesia	No
Yiddish	Europe	Poland	No
Yoruba	Africa	Nigeria	No
Zapotec	Americas	Mexico	No
Zazaki	Asia	Turkey	No
Zulu	Africa	South Africa	No
Zyrian	Europe	Russia	No

Table 31: Cultural data (Part 4 of 4).