

Lost in Localization: Benchmarking Intralingual Cultural Adaptation in Large Language Models

Soumedhik Bharati^{1*} Maharaj Brahma² Maunendra Sankar Desarkar²

¹Mohamed bin Zayed University of Artificial Intelligence, UAE

²Natural Language and Information Processing Lab (NLIP), IIT Hyderabad, India

soumedhik.bharati@mbzuai.ac.ae

cs23resch01004@iith.ac.in, maunendra@cse.iith.ac.in

Abstract

Cultural adaptation, rewriting text for a different cultural audience, requires more than lexical substitution. A model must decide what to change, whether each replacement is locally valid, and whether it preserves the original item’s functional role. We formalize this as *function-preserving cultural transfer* and introduce a dataset, a taxonomy, and an evaluation framework constructed around that principle. CULTURALADAPT-6K identifies 6,000 culturally rich texts in four discourse domains, while the *Cultural Rewriting and Adaptation Failure Taxonomy* (CRAFT) classifies errors into five categories. To mitigate the Anglocentric priors of current models, our proposed framework anchors each culturally significant item in a Wikidata-linked function frame, capturing its contextual role rather than surface denotation. Across 191,968 adaptations by eight instruction-tuned LLMs, we find that the models that substitute the most are not the ones that substitute best. High localization increases failure exposure, while conservative retention increases validity scores, yet both converge to the same overall functional quality. Current generative models can certainly dress text in a target culture’s vocabulary but ultimately cannot reliably preserve the contextual roles the vocabulary is meant to serve.

1 Introduction

“The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached.” – Sapir (1929)

Large language models can reproduce cultural value surveys, classify social norms, and generate culturally marked vocabulary on cue. However, when asked to rewrite text for a different cultural audience, they generally produce text that is stereotyped, structurally broken, or numerically inconsistent. This disconnect is well documented, but

*This work was carried out while the author was a research intern at the NLIP Lab, IIT Hyderabad, India.

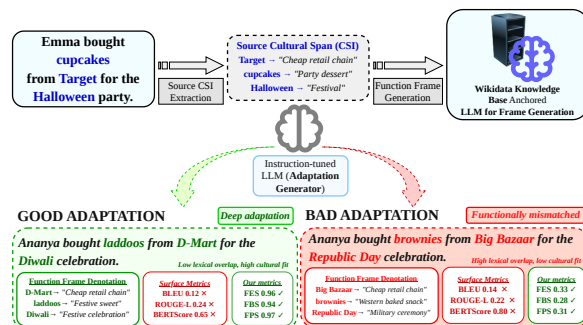


Figure 1: Traditional metrics fail by rewarding semantic overlap over cultural fit. By comparing Wikidata-anchored *function frames* (abbreviated here for brevity) of **culture-specific items** (CSIs), we evaluate whether target-culture replacements preserve contextual roles rather than surface form.

poorly diagnosed (Havaladar et al., 2025). Existing evaluations measure *whether* an output looks culturally appropriate or not. They do not check *where* in the adaptation process things might have gone wrong: Did the model fail to identify what needed to be changed? Did it pick an invalid replacement? Or did it swap the right word but destroy the contextual role that the item was serving? These are failure categories that require a distinct evaluation. We argue that the right unit of evaluation is not the surface form of a cultural reference, but its *contextual function* within the target culture. To address these gaps, we present three core contributions: (1) **CULTURALADAPT-6K**, a culturally rich 6,000-text dataset; (2) **CRAFT**, a five-category taxonomy to isolate adaptation errors; and (3) a **Wikidata-grounded evaluation framework** to measure cross-cultural text adaptation¹.

2 Related Work

The foundational work in cross-cultural NLP establishes that language and culture are entangled but not interchangeable, with variations emerging

¹Code and resources are available at <https://github.com/nlip-lab/cultural-evaluation>.

in common ground, style, topical importance, and values (Hershcovich et al., 2022). Early evaluations of LLMs treated culture as alignment or recall, probing models with value surveys, persona-conditioned questionnaires, Hofstede-style benchmarks, or regional cultural knowledge benchmarks (AlKhamissi et al., 2024; Wang et al., 2024b; Masoud et al., 2025; Sukiennik et al., 2025; Maji et al., 2025). Parallel work on cultural dominance demonstrated that models default to Anglocentric priors under multilingual prompting (Wang et al., 2024a; Naous et al., 2024; Huang et al., 2025). Although these benchmarks diagnose what models know and whose values they privilege, they have not evaluated the quality of intralingual adapted text.

The second strand moves from the recall of cultural knowledge to its application. Culturally-aware Natural Language Inference (NLI) uses cultural variation in interpretation as label disagreement (Huang and Yang, 2023). Evaluation under progressively weaker cues in culture-aware prompting shows that norm knowledge does not transfer to norm application (Rao et al., 2025). Perturbing nationality cues changes the vocabulary, but not the values (Bhatt and Diaz, 2024). Benchmarks on conversational style and story generation further show that models that answer cultural questions do not necessarily produce culturally grounded text (Havaladar et al., 2025; Bhagat et al., 2025).

Sahoo et al. (2025) and Bharati et al. (2026) evaluate LLMs on cultural text adaptation. These studies reveal limitations ranging from selective sub-regional coverage to surface-level cultural representations. Singh et al. (2024) assess correctness, naturalness and stereotypicality, comparing human and LLM-generated dialog adaptation. Across recipe adaptation and machine translation, cultural appropriateness consistently emerges as the dimension automatic metrics fail to capture (Cao et al., 2024; Liao et al., 2025). What remains missing is a framework that treats cultural appropriateness as a primary evaluation dimension, separated from fluency, preservation, and stereotype harms.

3 Data

We introduce CULTURALADAPT-6K, a dataset of 6,000 *English texts curated from material from the United States for evaluating intralingual cultural adaptation*. The items in the dataset are distributed evenly in four domains specifically chosen to span different registers and modes of cultural ground-

ing: educational math reasoning (GSM8K; Cobbe et al. 2021), five-sentence narratives (ROCStories; Mostafazadeh et al. 2016), conversational dialog (DailyDialog; Li et al. 2017), and procedural recipe text (RecipeNLG; Bieñ et al. 2020). The unfiltered corpus has approximately 2.35 million candidates and is filtered in the following stages.

1. Candidate Filtering. Domain-specific structural criteria are applied as follows. RecipeNLG candidates are kept when they exhibit a coherent procedural structure with US-standard measurements and culturally anchored ingredients (50,000 candidates). ROCStories candidates must satisfy the five-sentence structure within a coherent length range (8,960 candidates). GSM8K (8,792) and DailyDialog (13,118) do not require structural filtering. The resulting pool has 80,870 candidates.

2. Cultural Sensitivity Identification. The central challenge is distinguishing adaptation-critical candidates from culturally neutral English text. Each candidate is independently assessed for the presence of *US-specific cultural references* by a Qwen2.5-7B model, which produces a binary judgment with confidence (high: clear cultural grounding, low: ambiguous evidence), and a named entity recognizer (spaCy en_core_web_trf; Honnibal et al. 2020), which extracts mentions of place, affiliation, and organization as grounding signals. A rule-based conflict checker flags disagreements between the two systems. Low-confidence positives from the LLM and positives conflicting with entity evidence receive a second LLM pass with NER evidence injected into the judgment prompt. This system yields 23,947 candidates (acceptance rates: 32% GSM8K, 12% DailyDialog, 17% ROCStories, 36% RecipeNLG), with 97.6% receiving high-confidence labels. A sample of 500 candidates was annotated by two annotators, who provided binary judgments on whether each candidate contained a US-specific cultural reference. The annotations achieved Cohen’s $\kappa = 0.86$ and 92% inter-annotator agreement (IAA).

3. Diversity Sampling. We embed all verified candidates into a 768-dimensional space using a transformer (all-mpnet-base-v2; Reimers and Gurevych 2019) and apply K -means clustering ($K=1,500$ per domain). From each cluster, one item is chosen by a weighted score $S_i = \alpha P_i + (1-\alpha)R_i$, thus balancing centroid proximity (P_i) with cultural richness (R_i). Here, R_i aggregates

Category	Subtype	Severity
C1 Integrity	C1.1 Output Integrity Failure	***
C2 Scope	C2.1 Residual Source Retention	**
	C2.2 Cultural Saturation	**
	C2.3 Over-Globalization	*
C3 Substitution	C3.1 Affordance Violation	***
	C3.2 Scale/Numeric Transfer Error	***
	C3.3 Dependency Propagation Failure	**
	C3.4 Identity-Marker Mismatch	**
	C3.5 Figurative Equivalence Failure	*
C4 Preservation	C4.1 Core Intent Shift	***
	C4.2 Scope/Granularity Distortion	**
	C4.3 Pragmatic/Register Drift	**
C5 Plausibility	C5.1 Internal Cultural Incoherence	**
	C5.2 Stereotypical Portrayal	**
	C5.3 Perspective Inauthenticity	*

Table 1: CRAFT. Categories ordered by pipeline stage. Severity: *Critical* (***), *Substantive* (**), *Surface* (*).

marker density, confidence tier, and NER entity count into a single cultural-richness score, with $\alpha=0.65$ set through pilot ablations. This yields 1,500 final items per domain and a dataset of 6,000 items with strict domain balance and broad semantic coverage. Detailed filtering results, samples, and statistics are reported in the Appendix A.

4 The CRAFT Error Taxonomy

We expect the adaptations to follow certain desired properties. Each cultural element must be handled appropriately: it should be modified if and only if such a modification is necessary and maintains the role and intent of the original element. Additionally, the modified element should remain coherent with the surrounding text. For a structured assessment of these aspects, we create the CULTURAL REWRITING AND ADAPTATION FAILURE TAXONOMY (CRAFT). The appropriateness of an adaptation is evaluated through stages corresponding to the five CRAFT categories (C1 to C5) and their subtypes, indicated in Table 1.

Code	Adapted Excerpt	Diagnosis
C2.1	"Priya bought 3 pounds at Safeway for \$4.99 ... Fourth of July."	Name swapped; unit, store, currency, and holiday remain unadapted, i.e. American.
C3.2	"... mutton at Rs. 4.99/kg for Independence Day."	Labels localized, values unchanged: Rs. 4.99/kg is implausible and therefore arithmetic breaks.
C5.3	"... spiced meat patties for the Republic Day barbecue."	Right referents, outsider framing: the register and event pairing read as foreign-authored rather than local.

Table 2: CRAFT categories on a sample (*Sarah bought 3 pounds of ground beef at Safeway for \$4.99/lb. She used half to make cheeseburgers for the Fourth of July barbecue. How much did the unused beef cost?*). Full example in Table 16.

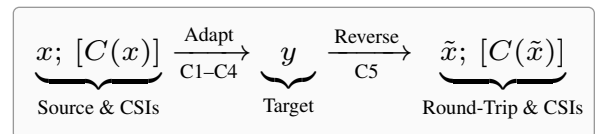
C1 (Integrity) flags degenerate outputs, compromising evaluation in subsequent stages. **C2 (Scope)** evaluates whether the model identified the correct cultural spans and applied the right degree of change. **C3 (Substitution)** evaluates the local validity of each replacement. A replacement may be culturally plausible in isolation but violate real-world affordances, numeric conventions, or figurative norms of the target culture. **C4 (Preservation)** evaluates functional equivalence. Surface-level substitutions can silently shift intent, alter scope, or lose pragmatic force. Such failures emerge only when the replacement is evaluated in context. **C5 (Plausibility)** evaluates holistic authenticity. Here, failures arise from the interaction of choices across an adaptation rather than from an isolated substitution. Severity levels serve as qualitative guidelines for error diagnosis. The subtypes are defined in the Appendix F, with a sample diagnosis in Table 2.

5 Proposed Methodology

For evaluating cultural adaptations, we propose a six-stage evaluation framework grounded in CRAFT. The first stage extracts culturally significant spans from the source text. The subsequent stages correspond to the five CRAFT categories: Output Integrity Check, CSI Scope Check, Substitution Validity Check, Function Preservation Check, and Round-Trip Plausibility Check. Following a root-cause principle, cascading failures receive the first applicable CRAFT label.

Following recent LLM-as-judge work (Liu et al., 2023; Zheng et al., 2023), two judges (Gemini 2.5 Pro and GPT-4o) score adaptations using matched prompts. The scalar scores are averaged, and categorical disagreements are adjudicated². To mitigate circularity bias, we pair open-weight models as adapters with closed-weights frontier models as judges, alongside robust human validation.

For each source x , we evaluate the adaptation y and a reverse-adapted round-trip reconstruction $\tilde{x}$ using C1 to C5. During this process, we pay a closer look at the cultural elements (also called Culture Specific Items or CSIs) in the content. The overall process can be summarized as follows:



²Disagreements between the two judges are resolved by a third independent pass from the higher-confidence judge.

Schema Field	Source Function Frame	Adapted Function Frame
Span & Meta	<i>Fourth of July</i>	<i>Republic Day</i>
Wikidata ID	Q86591 (US Indep. Day)	Q1139536 (Republic Day)
Denotation	US federal holiday (July 4) marking national independence.	Indian public holiday (Jan 26) honoring the national constitution.
Cultural Role	Major national marker of patriotic identity and history.	Major national marker of patriotic identity and history.
Context Function	Social anchor for an informal, outdoor summer celebration.	Social anchor for a formal, patriotic winter observance.
Affordances	{festive, casual, outdoor, BBQ-norm}	{festive, formal, ~~BBQ-norm~~ }
Evaluation Outcome	The denotation and cultural role align perfectly, but the affordance coverage fails : <i>Republic Day</i> is a formal event that breaks the required BBQ context.	

Table 3: Function-frame schema for “*Fourth of July*”. Wikidata grounding injects factual attributes for frame generation

5.1 Source CSI Extraction

Culture-specific items (CSIs) are spans whose meaning or contextual role is tied to a particular culture. Good adaptations should modify appropriate CSIs rather than rewrite culturally neutral text, while adjusting the surrounding context when needed to support the adapted item. As CSIs vary in how strongly culture-bound they are, we assign each CSI a cultural salience score $s_i \in [0, 1]$ to strongly weigh highly culture-dependent spans.

For each source text x , an LLM is used to extract a set of CSIs $C(x) = \{c_i\}_{i=1}^m$. Each c_i is assigned a category, a cultural salience score s_i , and an adaptation-necessity judgment n_i (ADAPT, PERMIT or RETAIN). We also construct a *function frame* for each c_i , encoding its denotation, cultural role, contextual function, and required affordances. Table 3 presents a two sample function frames for the CSIs “Fourth of July” and “Republic Day”, from the adaptation example in Table 2.

To mitigate the Anglocentric prior in models (Wang et al., 2024a; Naous et al., 2024), *frame generation is grounded via Wikidata entity linking*, where a CSI is first resolved to a Wikidata entity. Factual attributes are then injected into the LLM prompt, anchoring the subsequent judgments in verified knowledge over parametric recall.

5.2 Output Integrity Check (C1)

We apply rule-based checks to filter degenerate adaptations before further evaluation. Using boolean indicator functions $\mathbb{I}(\cdot)$, each adapted text y is flagged for non-English output $\mathbb{I}_{NE}(y)$, truncation $\mathbb{I}_{Tr}(y)$, and length anomaly $\mathbb{I}_{Len}(y)$, where the adapted length is set to lie within $\pm 50\%$ of the source length. The integrity score is calculated as:

$S_{Itg}(y) = \mathbb{I}_{NE}(y) + \mathbb{I}_{Tr}(y) + \mathbb{I}_{Len}(y)$. Outputs with $S_{Itg} = 0$ pass, outputs with $S_{Itg} = 1$ are retained as soft failures, and outputs with $S_{Itg} \geq 2$ are discarded from further evaluation.

5.3 CSI Scope Check (C2)

We evaluate the manner of adaptation for each source CSIs with three questions: did the model change CSIs that should change, preserve CSIs that should stay, and avoid adding invalid target-culture material? Each $c_i \in C(x)$ is classified into $n_i \in \{\text{ADAPT}, \text{RETAIN}, \text{PERMIT}\}$, where ADAPT requires replacement, RETAIN requires preservation, and PERMIT allows either choice. Next, each c_i is aligned to its adapted counterpart (if present), and assigned action a_i : *retained* (kept unchanged), *localized* (replaced by a target-culture equivalent), *globalized* (replaced by a culture-neutral item), *transformed* (rewritten to preserve role), *omitted* (removed), or *invalid* (unalignable or unsupported).

We report three salience-weighted scores. **Cultural Foreignness Score (CF)** measures required adaptations that remain source-culture specific (under-adaptation). **Cultural Transfer Score (CT)** measures appropriate adaptation actions for each CSI (ideal case). **Cultural Saturation Score (CS)** measures unsupported target-culture insertions not grounded in any source CSI (over-adaptation).

$$\text{CF}(y) (\downarrow) = \frac{\sum_i s_i \mathbb{1}[n_i = \text{ADAPT} \wedge a_i = \text{retained}]}{\sum_i s_i \mathbb{1}[n_i = \text{ADAPT}] + \epsilon},$$

$$\text{CT}(y) (\uparrow) = \frac{\sum_i s_i T(a_i, n_i)}{\sum_i s_i + \epsilon}, \quad \text{CS}(y) (\downarrow) = \frac{\sum_{u \in U(y)} \omega(u)}{\sum_i s_i + \epsilon}.$$

Here, $T(a_i, n_i)$ is a compatibility score: $T = 1$ for highly compatible actions (ADAPT with *localized* or *transformed*, RETAIN with *retained*, and PERMIT with *retained*, *localized*, or *transformed*). $T = 0.5$ for acceptable neutralization via *globalized* when $n_i \in \{\text{ADAPT}, \text{PERMIT}\}$; and $T = 0$ for *omitted*, *invalid*, or actions changing a RETAIN item. $U(y)$ is the set of unsupported additions, $\omega(u) \in (0, 1]$ is the LLM-assigned severity weight, and ϵ avoids division by zero. We define aggregate $C_2 = \max(0, \text{CT} - \text{CF} - \frac{1}{2}\text{CS})$ rewarding transfer and penalizing underadaptation (residual foreignness) and overadaptation (unsupported saturation).

5.4 Substitution Validity Check (C3)

We evaluate whether the replacements made in y are *locally defensible*. For each c_i with a replacement (neither retained nor omitted), five dimensions are scored by an LLM judge in $[0, 1]$: affordance validity (A_i), scale transfer (S_i), dependency

propagation (D_i), identity-marker fit (I_i), and figurative equivalence (F_i). The score is a weighted sum: $C3(c_i) = \alpha A_i + \beta S_i + \gamma D_i + \delta I_i + \zeta F_i$.³ To ground outputs, S_i is overridden by a deterministic checker if unit/currency conversions are implausible. We compute C_3 as the salience-weighted mean of $C3(c_i)$ across all eligible c_i s in y .

5.5 Function Preservation Check (C4)

We evaluate if each CSI replacement in y serves the *same contextual role* as the original. For each aligned replacement, we compare the source function frame (f_i^{src}) and its adapted target frame (f_i^{tgt}) (Table 3) using three measures. Frame Embedding Similarity (**FES**) and BERTScore-F1 (**FBS**; Zhang et al. 2020) are computed on frame descriptions with denotation, cultural role, contextual function, and affordances. Using sentence encoder E (**all-mpnet-base-v2**), FES measures frame-level semantic proximity; FBS measures token-level functional overlap on the same. Frame Preservation Score $\text{FPS}(c_i)$ is the LLM judges’ mean scalar score in $[0, 1]$ on five areas: denotation match, functional equivalence, affordance coverage, scope preservation, and register fit. Aggregate score C_4 is the mean FPS across adapted items.

5.6 Round-Trip Plausibility Check (C5)

We reverse-adapt valid adaptations y back into the source culture using the same model⁴. We compare the round-trip output $\tilde{x}$ against the original x . At *surface level*, BLEU, ROUGE-L, and BERTScore measure string recovery. At *CSI level*, extracted CSIs $C(\tilde{x})$ are aligned to $C(x)$ and categorized into: recovered source CSIs R , substituted source CSIs S , unreverted target-culture CSIs L , and missing source CSIs M . Corresponding scores $|R|/|C(x)|$, $|S|/|C(x)|$, $|L|/|C(\tilde{x})|$, and $|M|/|C(x)|$ are reported. At the *semantic level*, two LLMs judge the pair $(\tilde{x}, x)$ on meaning, structure, detail, and overall equivalence (S_{Ov}) on $[0, 10]$. As the reverse-adaptation phase is easier given LLMs’ Anglocentric priors (Naous et al., 2024), degraded round-trip fidelity strongly implicates forward-pass coherence failure. Stereotypes (C5.2) are manually flagged for representational harms on the human-validated subset. The overall plausibility score is computed as $C_5 = S_{Ov}/10$.

³Weights ($\alpha=0.35, \beta=0.25, \gamma=0.20, \delta=0.10, \zeta=0.10$) were set via grid search on a 500-item split to maximize Pearson correlation with human judgments.

⁴GPT-4o as reverse model in ablation study in Appendix E.

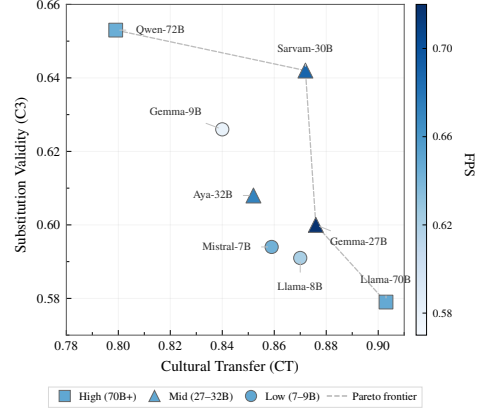


Figure 2: **The Pareto frontier** (best trade-off boundary) shows that models substituting the most (high CT) also make more local errors (lower C3). Yet, models at opposite behavioral extremes (Qwen2.5-72B and Llama-3.3-70B) converge to similar contextual function preservation (FPS ≈ 0.65).

5.7 The Aggregate CRAFT Score

A good adaptation requires cohesion across scope (C2), validity (C3), function (C4), and robustness (C5). We define CRAFT Score as harmonic mean of scores C_2 through C_5 . We treat C1 (Integrity) as a strict prerequisite and define integrity gate $G = 1[S_{Itg} < 2]$. The aggregate score is calculated as:

$$\text{CRAFT}_{\text{Score}} = G \cdot \frac{4}{C_2^{-1} + C_3^{-1} + C_4^{-1} + C_5^{-1}}$$

6 Results

6.1 Adaptations Generation Setup

We generate 191,968 target-culture adaptations from the CULTURALADAPT-6K dataset using eight open-weight, instruction-tuned models (greedy decoding throughout). The high-parameter tier includes Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and Qwen2.5-72B-Instruct (Yang et al., 2024). The mid-parameter tier includes Aya-Expanse-32B (Dang et al., 2024), Gemma-3-27B-it (Gemma Team et al., 2025) and Sarvam-30B (Sarvam Foundation Models Team, 2026). The lower-parameter tier consists of Llama-3.1-8B-Instruct, Gemma-2-9B-it and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We include Sarvam-30B as a target-culture-native (Indian) model, incorporating culturally relevant training.

To isolate whether adaptation failures are intrinsic (model-internal) or prompt-dependent, we choose four prompting categories of increasing guidance. These are: minimal-instruction (MI), constraint-aware rewriting (CR), one-shot

Category	N	Adapt(%)	Permit(%)	Retain(%)	Sal.
Unit / currency	10,174	69.7	6.9	23.4	0.557
Food & drink	9,788	9.1	56.4	34.5	0.534
Place / brand	6,647	11.7	47.6	40.6	0.475
Personal name	2,433	4.7	8.5	86.8	0.290
Pop culture	1,544	26.9	29.0	44.1	0.540
Holiday / ritual	1,204	47.1	19.6	33.3	0.636
Social norm	1,125	13.2	59.2	27.6	0.448
Idiom / figurative	634	14.0	70.0	15.9	0.532
Total	33,549	30.1	33.9	36.0	0.513

Table 4: Source CSI distribution, necessity, and salience. Green and red highlight best and worst values, respectively.

exemplar-guided (1S), and structured chain-of-thought (CoT).⁵ The source culture is set to American and the target to Indian with an explicit English-output constraint to prevent code-switching.⁶ We also present a study extending our framework to Eastern Indian (West Bengal) and Western Indian (Gujarat and Maharashtra) cultures in Appendix D.

6.2 Source CSI Extraction

All source CSIs are not uniform adaptation targets (Table 4). Only 30.1% are labeled ADAPT, necessitating change for the target culture. 36.0% are RETAIN, meaning they should stay unchanged. Rest 33.9% are PERMIT, meaning either retention or adaptation can be acceptable. Thus, unnecessary localization would cause C2 Scope errors, since many CSIs are not mandatory adaptation targets.

6.3 Output Integrity Check (C1)

Of the 191,968 generated adaptations, 61.5% pass the integrity gate ($G = 1$; Table 5). Failures concentrate in two cases. Firstly, the MI prompt lowers pass rates to 20.1% overall, mainly through meta-commentary. Secondly, both Aya-Expanse-32B and Sarvam-30B show structural failures for most prompt categories, mainly through non-English generation, despite explicit prompting. The other models maintain pass rates averaging above 80% under the CR, 1S, and CoT prompt categories.

6.4 CSI Scope Check (C2)

Across integrity-passed adaptations, the framework assigns 852,330 CSI decisions ($\sim 660k$ source, $\sim 192k$ insertions). Annotators label 2,000 items ($\kappa = 0.76$), confirming substantial agreement.

Model analysis. Table 6 shows a trade-off in foreignness, transfer, and saturation. Llama-3.3-70B

⁵Full prompt templates are provided in Appendix B.

⁶The one-shot condition contributes 47,968 outputs after excluding 32 cases for exemplar (1S) leakage control.

	MI	CR	1S	CoT	Overall
Qwen2.5-72B	14.4	95.8	95.9	96.3	75.6
Llama-3.3-70B	9.9	95.6	95.7	95.5	74.2
Aya-Expanse-32B	27.6	3.3	21.2	73.2	31.3
Sarvam-30B	14.7	4.5	9.6	12.8	10.4
Gemma-3-27B	2.5	95.9	96.1	94.9	72.4
Gemma-2-9B	29.6	94.7	94.9	92.2	77.9
Llama-3.1-8B	12.8	79.5	94.2	91.6	69.5
Mistral-7B	49.5	85.5	94.7	91.7	80.3
<i>Overall</i>	20.1	69.4	75.3	81.0	61.5

Table 5: Integrity pass rates (%) by model and prompt type.

Model	CF↓	CT↑	CS↓	Loc%	Ret%
Qwen2.5-72B	0.122	0.799	0.313	57.5	39.2
Llama-3.3-70B	0.055	0.903	0.660	77.0	19.4
Aya-Expanse-32B	0.074	0.852	0.997	67.2	25.4
Sarvam-30B	0.053	0.872	0.289	63.8	28.7
Gemma-3-27B	0.055	0.876	0.637	71.7	21.4
Gemma-2-9B	0.073	0.840	0.515	65.3	27.1
Llama-3.1-8B	0.077	0.870	0.700	71.9	24.1
Mistral-7B	0.081	0.859	0.857	69.4	24.9
<i>Human annot.</i>	0.031	0.882	0.134	—	—

Table 6: C2 scope metrics by model. Loc% and Ret% are behavioral rates. Human annot. reflects ground-truth labels.

localizes aggressively yet appropriately (CT 0.903, CF 0.055). Qwen2.5-72B over-retains at 39.2% (over twice Llama’s 19.4%). Aya-Expanse-32B has high CS score (0.997), meaning nearly one unsupported target-culture insertion per source CSI.

Prompt analysis. Minimal-instruction (MI) prompting increases unsupported target-culture content (CS 1.207), with six of eight models exceeding $CS > 1.0$, suggesting models fill guidance gaps by over-localizing. CoT and CR prompts produce the most balanced scope decisions, while 1S prompting keeps saturation lowest (CS 0.345).

CSI Category analysis. Unit/currency CSIs are the main source of under-adaptation: 30.1% are retained despite 69.7% requiring adaptation. Yet social norms and idioms show high retention (60.7%, 57.4%) given few mandatory adaptation targets (13.2%, 14.0%). Localization remains below 50% for both categories, showing models are least willing to adapt softer, contextual CSIs (Table 7).

6.5 Substitution Validity Check (C3)

Mean C3 is 0.610 across eligible substitutions. The primary failures are affordance violations (C3.1) and scale/numeric transfer errors (C3.2). Models often choose culturally plausible replacements, but mostly fail to preserve real-world constraints.

Model analysis. Table 8 shows that higher substitution rate does not always imply better substitution quality. Qwen2.5-72B has the highest C3 score as its retention bias reduces risky substitutions. Llama-3.3-70B shows the opposite pattern: it localizes more strongly, which improves cultural transfer but creates more scale and affordance errors. Saliency weighting reduces all model scores without rank change, confirming that failures concentrate on culturally significant CSIs. Two annotators scored 1,000 substitutions ($\kappa_w = 0.72$; $r = 0.74$ against automated C3). Human judges penalize affordance violations more strongly, suggesting C3 underestimates contextual misfit.

Prompt and CSI Category analysis. CoT and CR prompting produce the most reliable substitutions, while MI prompting scores higher mainly by attempting fewer risky substitutions. Unit/currency/scale is the hardest category, showing the lowest mean and highest variance. This reflects all-or-nothing outcomes: a single unchanged price, unit, or quantity can invalidate the entire adaptation.

6.6 Function Preservation Check (C4)

C4 requires wiki-frames for assessing functionality match. The system anchors 82% of source CSIs to Wikidata entities, enabling factually grounded frame generation⁷. Frame labels were validated by five annotators on 500 items (Fleiss’s $\kappa = 0.84$). FES and FBS are computed on the generated frame descriptions. Their ranges (0.51 to 0.57 and 0.57 to 0.63 respectively) reflect that most models identify culturally similar replacements semantically. However, FPS reveals whether these replacements fulfill the original source frame’s contextual role.

Model analysis. Gemma-3-27B and Gemma-2-9B have the highest and lowest FPS (0.718, 0.581)

⁷For the remaining 18% without a Wikidata link, frames are generated in the same schema without KB-attribute injection in prompt, and retained only after human validation.

Category	Loc% $\uparrow$	Ret% $\downarrow$	Trans%	Om% $\downarrow$	Must%*
Place / institution / brand	68.8	23.2	1.3	5.1	11.7
Food & drink	65.9	28.8	1.7	2.2	9.1
Holiday / event / ritual	62.0	34.2	1.3	1.8	47.1
Unit / currency / scale	61.1	30.1	5.4	2.4	69.7
Personal name / identity	60.9	36.6	0.9	1.2	4.7
Pop culture / sport / media	59.8	34.7	2.7	1.5	26.9
Social norm / address	33.2	60.7	3.2	1.4	13.2
Idiom / figurative	28.0	57.4	9.3	1.3	14.0

Table 7: Adaptation strategies across various CSI categories.

Model	C3 $\uparrow$	SW-C3 $\uparrow$	Δ	C3.1% $\downarrow$	C3.2% $\downarrow$
Qwen2.5-72B	0.653	0.619	-0.035	16.0	32.0
Llama-3.3-70B	0.579	0.561	-0.018	9.8	23.9
Aya-Expanse-32B	0.608	0.582	-0.025	14.2	21.8
Sarvam-30B	0.642	0.614	-0.028	5.4	26.6
Gemma-3-27B	0.600	0.580	-0.019	9.3	17.3
Gemma-2-9B	0.626	0.600	-0.027	11.3	20.0
Llama-3.1-8B	0.591	0.571	-0.021	12.4	15.7
Mistral-7B	0.594	0.572	-0.022	14.7	18.7
Global	0.610	0.586	-0.024	11.8	21.8

Table 8: Substitution validity by model. SW-C3: saliency-weighted C3; Δ : SW-C3 minus unweighted C3.

Model	FES-C $\uparrow$	FES-H $\uparrow$	FBS-C $\uparrow$	FBS-H $\uparrow$	FPS $\uparrow$
Qwen2.5-72B	0.533	0.559	0.599	0.602	0.645
Llama-3.3-70B	0.536	0.556	0.599	0.601	0.648
Aya-Expanse-32B	0.558	0.574	0.614	0.617	0.672
Sarvam-30B	0.562	0.581	0.618	0.619	0.689
Gemma-3-27B	0.571	0.589	0.628	0.631	0.718
Gemma-2-9B	0.512	0.531	0.574	0.573	0.581
Llama-3.1-8B	0.524	0.543	0.587	0.586	0.621
Mistral-7B	0.539	0.558	0.598	0.600	0.643
Global	0.542	0.561	0.602	0.604	0.652

Table 9: Function preservation by LLM. FES: frame embed. similarity (C: contextual, H: holistic); FBS: BERTScore-F1.

respectively. Notably, Qwen2.5-72B and Llama-3.3-70B converge on nearly identical FPS (0.645, 0.648). This shows that localization strategy, whether conservative or aggressive, does not predict its functional quality. Models score higher on functional equivalence than affordance coverage in FPS sub-scores (0.774, 0.673), indicating meaning is easier to preserve than real-world contextual fit.

Metric validity. FES and FBS correlate moderately ($r=0.66$) but weakly with FPS ($r\approx 0.11$ to 0.20). Six annotators scored 800 CSI pairs (Krippendorff’s $\alpha = 0.69$; $r = 0.79$ against automated FPS). Human judges flag subtle pragmatic mismatches that embedding scores miss, averaging 0.09 points below automated FPS on such items.

	Surface	CSI Alignment				LLM Judge (0–10)			
	BERT-F1 $\uparrow$	Recall $\uparrow$	Leak $\downarrow$	Miss $\downarrow$	Mean $\uparrow$	Str. $\uparrow$	Detail $\uparrow$	Ov. $\uparrow$	
Qwen2.5-72B	0.861	0.721	0.009	0.293	8.44	9.01	5.04	7.68	
Llama-3.3-70B	0.868	0.726	0.008	0.290	8.57	9.14	5.26	7.82	
Aya-Expanse-32B	0.840	0.694	0.017	0.314	7.98	8.65	4.55	7.24	
Sarvam-30B	0.846	0.701	0.014	0.309	8.09	8.70	4.68	7.31	
Gemma-3-27B	0.858	0.714	0.011	0.295	8.42	8.96	5.09	7.64	
Gemma-2-9B	0.852	0.689	0.010	0.298	8.18	8.77	4.92	7.50	
Llama-3.1-8B	0.841	0.695	0.014	0.322	8.16	8.72	4.57	7.34	
Mistral-7B	0.835	0.714	0.020	0.319	7.89	8.59	4.44	7.17	
Global avg.	0.850	0.707	0.013	0.305	8.22	8.82	4.82	7.46	
Human annot.	—	—	—	—	8.38	8.97	4.71	7.53	

Table 10: Round-trip consistency. Two annotators scored 500 items (inter-annotator $r=0.88$, annotator-judge $r=0.91$).

	Gate	Scope (C2)					Substitution (C3)			Function (C4)		Plausibility (C5)			Aggregate		
Model	C1%↑	Loc%	Ret%	CT↑	CF↓	CS↓	C2↑	C3↑	3.1%↓	3.2%↓	FES↑	C4↑	BF1↑	Rec↑	C5↑	HM↑	CRAFT↑
Qwen2.5-72B	75.6	57.5	39.2	0.799	0.122	0.313	0.520	0.619	16.0	32.0	0.559	0.645	0.861	0.721	0.768	0.626	0.473
Llama-3.3-70B	74.2	77.0	19.4	0.903	0.055	0.660	0.518	0.561	9.8	23.9	0.556	0.648	0.868	0.726	0.782	0.612	0.454
Aya-Expanse-32B	31.3	67.2	25.4	0.852	0.074	0.997	0.279	0.582	14.2	21.8	0.574	0.672	0.840	0.694	0.724	0.489	0.153
Sarvam-30B	10.4	63.8	28.7	0.872	0.053	0.289	0.674	0.614	5.4	26.6	0.581	0.689	0.846	0.701	0.731	0.674	0.070
Gemma-3-27B	72.4	71.7	21.4	0.876	0.055	0.637	0.502	0.580	9.3	17.3	0.589	0.718	0.858	0.714	0.764	0.623	0.451
Gemma-2-9B	77.9	65.3	27.1	0.840	0.073	0.515	0.510	0.600	11.3	20.0	0.531	0.581	0.852	0.689	0.750	0.599	0.466
Llama-3.1-8B	69.5	71.9	24.1	0.870	0.077	0.700	0.443	0.571	12.4	15.7	0.543	0.621	0.841	0.695	0.734	0.573	0.398
Mistral-7B	80.3	69.4	24.9	0.859	0.081	0.857	0.350	0.572	14.7	18.7	0.558	0.643	0.835	0.714	0.717	0.529	0.425
Global avg.	61.5	68.0	26.3	0.859	0.074	0.621	0.475	0.587	11.8	21.8	0.561	0.652	0.850	0.707	0.746	0.591	0.361

Table 11: CRAFT scores by model. BF1: BERTScore-F1 surface recovery. Rec: CSI recall. FES: frame embedding similarity. HM: harmonic mean of C2-C5 (quality on integrity-passed items). The metrics contributing to CRAFT score are highlighted.

6.7 Round-Trip Plausibility Check (C5)

Round-trip results show unreverted adaptations (Leakage) from y to $\tilde{x}$ to be near zero (1.3%), confirming the reverse pass reliably replaces Indian CSIs with source-culture equivalents. 30.7% source CSIs are not recovered after the overall round trip. Table 10 reports results for the three metric groups. While structure and meaning scores are high (8.82, 8.22), detail drops to 4.82, making it the primary bottleneck. In the educational domain, the CSIs often return, but numerical values are lost. Leaked CSIs concentrate in food/drink and unit/currency (73% combined), matching the highest forward-pass substitution categories.

6.8 CRAFT Score

Table 11 reports CRAFT scores (ranks in Figure 3). Harmonic mean (HM) reflects adaptation quality on integrity-passed outputs. Sarvam-30B has the highest HM score yet lowest CRAFT score (0.674, 0.070), with Aya-Expanse-32B (0.153) due to lower integrity pass rates. Qwen2.5-72B ranks highest in HM and CRAFT scores (0.626, 0.473).

Human evaluation. To validate HM, four annotators rated 3,200 integrity-passed adaptations (100 per model per domain) on overall cultural adaptation quality in $[0, 10]$, which we normalize to $[0, 1]$. Human scores correlate substantially with HM (Spearman $\rho=0.79$, Pearson $r=0.71$), with both rankings agreeing on the top three models. Human scores average 0.07 points above HM scores.

Holistic LLM Baseline. We also prompt GPT-4o to rate the same 3,200 adaptations on the same $[0, 10]$ scale. These zero-shot LLM scores correlate weakly with human judgments (Spearman $\rho=0.51$, Pearson $r=0.48$) compared to HM. Furthermore,

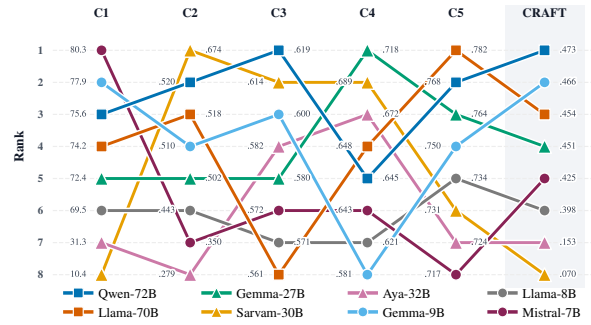


Figure 3: Rank shifts of models across evaluation stages.

the holistic LLM consistently over-scores outputs, averaging 0.21 points above human ratings.

6.9 Validation and Reliability

Human Annotation. We recruited eight annotators through university research networks, representing five Indian states across North, South, East, West, and Central regions. All hold at least a bachelor’s degree, are native Indian English speakers, and had no prior involvement in prompt design. They were compensated above local rates through a research fund. Automated scores correlate substantially with human judgments ($r \geq 0.71$), though judges penalize subtle contextual misfits more strictly. IAA is lowest for stereotype and authenticity flags ($\kappa = 0.73$), where boundary cases between *retained*, *globalized*, and *inauthentic* substitutions show the highest disagreement.

LLM Inter-Judge Agreement. The two LLM judges exhibit strong scalar agreement (Pearson $r = 0.84$ on C3, 0.81 on FPS, and 0.88 on round-trip) and substantial categorical agreement (Cohen’s $\kappa = 0.71$). Disagreements are below 12%, not concentrating in any specific model or domain.

7 Conclusion

Cultural adaptation requires more than vocabulary substitution. It demands preserving the contextual roles that cultural items represent. To quantify this, we introduced CULTURALADAPT-6K, CRAFT, and a decomposable evaluation framework. While models easily generate culturally specific terms, they routinely fail to respect the real-world constraints and functions attached to them. Substituting more does not mean substituting well. Global frontier models of large size lead the performance on the task, although they too have a huge scope for improvement. No model, neither frontier nor Indic, has a consistent high rank across all evaluation stages indicating the non-triviality of the task.

Although we focused on American to Indian English, the methodology generalizes. We envisage extending it to new cultural pairs, interlingual settings, and open-ended generation. Until evaluations measure underlying cultural affordances, surface-level adaptations will continue to give the illusion of multicultural competence while obscuring the cultures they attempt to represent.

Limitations

This work has some important limitations. First, we operate the framework exclusively intralingually (specifically, American to Indian English). While the methodology is designed to generalize, our empirical findings may not transfer directly to interlingual adaptation or other cultural pairs. Second, we acknowledge that "Indian culture" is not monolithic. Our evaluation instructs judges to accept all plausible Indian cultural contexts, but regional, linguistic, and religious variation within India is vast, and a single-label target compresses this diversity. Third, our evaluation relies on LLM-based judges, which introduces potential evaluation circularity and self-preference bias. To mitigate this, we explicitly pair open-weights models as adaptation generators with closed-weights frontier models as judges, further anchoring the evaluation with rule-based gates, Wikidata entity grounding, and extensive human validation. Fourth, for this work, we assume that the contextual function can be captured in a short description and compared via embeddings. For highly implicit cultural roles, this representation may be lossy. Finally, we evaluated eight instruction-tuned, open-weights models. Commercial APIs, fine-tuned models, and decoding variation are beyond our current scope.

Acknowledgements

The authors thank the eight annotators who volunteered their time and cultural expertise across the pipeline. Their judgments were crucial for validating every stage of the framework. We also thank IIT Hyderabad for providing computational resources for this work.

Ethical Considerations

We acknowledge the use of generative AI tools, such as Large Language Models (LLMs), to assist with manuscript writing, code implementation, and data preprocessing. All AI-generated outputs were thoroughly reviewed by the authors, who assume full responsibility for the content of this work.

Our work involves cultural content that can reflect stereotypes, biases, or insensitive portrayals. We deliberately analyze such failures as objects of evaluation rather than suppressing them, as diagnosing stereotypical output is a necessary step toward mitigating it. All source texts used during dataset creation are drawn from publicly available, previously published datasets. Human annotators were informed of the nature of the task, including the possibility of encountering stereotyped or culturally inappropriate output. We note that our framework, while designed to detect cultural harms, should not be treated as a substitute for community-centered evaluation by people of the target culture. Automated and LLM-based judgments remain proxies of human cultural experience.

References

- 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.
- Kirti Bhagat, Shaily Bhatt, Athul Velagapudi, Aditya Vashistha, Shachi Dave, and Danish Pruthi. 2025. [Tales: A taxonomy and analysis of cultural representations in llm-generated stories](#). *Preprint*, arXiv:2511.21322.
- Soumedhik Bharati, Subhrajit Mukherjee, and Shibam Mandal. 2026. [The mirage of diversity: Unmasking the cultural vocabulary ceiling in LLMs](#). In *Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026)*, pages 101–107, San Diego, California, United States. 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.
- Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. [RecipeNLG: A cooking recipes dataset for semi-structured text generation](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28, Dublin, Ireland. Association for Computational Linguistics.
- Yong Cao, Yova Kementchedjhiya, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. [Cultural adaptation of recipes](#). *Transactions of the Association for Computational Linguistics*, 12:80–99.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- John Dang, Madeline Smith, Marzieh Fadaee, Ahmet Üstün, Beyza Ermis, Sara Hooker, and 1 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Gemma Team and 1 others. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- 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, Suzhou, China. Association for Computational Linguistics.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Jing Huang and Diyi Yang. 2023. [Culturally aware natural language inference](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.
- Shulin Huang, Linyi Yang, and Yue Zhang. 2025. [Mceval: A dynamic framework for fair multilingual cultural evaluation of llms](#). *Preprint*, arXiv:2507.09701.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, and Maite Melero. 2025. [Culture-aware machine translation: the case study of low-resource language pair Catalan-Chinese](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 150–161, Geneva, Switzerland. European Association for Machine Translation.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#). *Preprint*, arXiv:2303.16634.
- 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.
- 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.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and evaluation framework for deeper understanding of commonsense stories](#). *Preprint*, arXiv:1604.01696.

- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. [Having beer after prayer? measuring cultural bias in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. 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.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- 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.
- E. Sapir. 1929. [The status of linguistics as a science](#). *Language*, 5(4):207–214.
- Sarvam Foundation Models Team. 2026. Introducing sarvam’s sovereign models. <https://www.sarvam.ai/blogs/sarvam-30b-105b>. Accessed: 2026-04-05.
- Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. [Translating across cultures: LLMs for intralingual cultural adaptation](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, Miami, FL, USA. Association for Computational Linguistics.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. [An evaluation of cultural value alignment in llm](#). *Preprint*, arXiv:2504.08863.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024a. [Not all countries celebrate thanksgiving: On the cultural dominance in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. [CDEval: A benchmark for measuring the cultural dimensions of large language models](#). In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

Appendix

A Data Construction Details

A.1 Step 1: Domain-Specific Candidate Filtering

The initial filtering criteria are adjusted on a domain basis. **RecipeNLG** (2,231,142 items) undergoes the most rigorous filtering. Our *Structural checks* require 5-8 directions, 5-10 ingredients, and ≤ 40 words per step. *Cultural checks* require ≥ 2 US measurement markers (*cup, ounce, pound, table-spoon, teaspoon, °F*) and ≥ 1 American brand or anchored ingredient (*Bisquick, Velveeta, Campbell's*). We assign an aggregate score based on these criteria. Overall, deduplication and rank-by-score selection yield 50,000 candidates. This ensures that every surviving recipe presents both a conversion and a substitution challenge to the adaptation task.

Our filtering criterion for **ROCStories** (98,161 items) requires exactly 5 canonical sentences, 32-58 words, and ≥ 1 strong American anchor (holiday, sports league, retail brand or dollar amount). 91% of stories are culturally generic and fail the marker check (98,161 $\rightarrow$ 8,960). **GSM8K** (8,792) and **DailyDialog** (13,118) pass Step 1 intact: GSM8K problems embed the American context by construction. DailyDialog is structurally clean. Both receive only a minor quality normalization prior to Step 2.

A.2 Step 2: Cultural Sensitivity Identification

Step 2 identifies whether each candidate contains CSIs genuinely requiring adaptation for an Indian reader, through a four-tier pipeline. **Tier 0** applies high-precision regex patterns covering US currency, federal holidays, retail brands, and measurement strings. These matches receive an unconditional YES without LLM evaluation (71.1% of the final benchmark). **Tier 1** runs two independent systems: Qwen2.5-7B produces a binary verdict with confidence (HIGH/MEDIUM/LOW) and a detected marker list, while spaCy `en_core_web_trf` extracts GPE, LOC, NORP, FAC, and PERSON spans as a model-independent grounding signal. **Tier 2** is a deterministic conflict detector that escalates items where: (a) the LLM says NO but NER finds a non-person entity, (b) confidence is MEDIUM/LOW, (c) the LLM says YES but NER finds zero entities, or (d) only PERSON spans appear without geographic support (38-48% of items per domain). **Tier 3** reruns the LLM with the NER entity list injected in the prompt, serving as a grounding signal.

Domain	Verif.	T0	T1	T3↑	Final	High
GSM8K	2,844	1,361	128	8	1,500	1,496
ROCStories	1,561	1,162	277	39	1,500	1,467
DailyDialog	1,629	645	748	86	1,500	1,427
RecipeNLG	16,913	1,099	334	67	1,500	1,468
Total	23,947	4,267	1,487	246	6,000	5,858

Table 12: Filtration for CULTURALADAPT-6K. **Verif.**: Step 2 verified items. **T0/T1/T3↑**: resolution tier of selected items (T3↑: upgrades only; remaining items resolved at Tier 2 re-confirmation). **High**: high-confidence verdicts.

A.3 Step 3: Diversity-Aware Sampling

K -means ($K=1,500$, k -means++ initialization) partitions each domain’s Step 2 pool in 768-dimensions sentence-transformer space. One item is selected per cluster by $S_i = 0.65 \cdot P_i + 0.35 \cdot R_i$, where P_i is centroid proximity and R_i aggregates marker density, confidence tier, and NER entity count. ROCStories enters Step 3 with only 1,561 verified items against 1,500 target clusters, so diversity gains are minimal for this domain.

Table 12 reports the filtration steps from raw corpus to final dataset with per-domain tier resolution. Tier 1 and Tier 3 use the prompts below. It includes one domain-matched positive and one negative few-shot example (educational shown).

A.4 Dataset Statistics

We identified a total of 33,549 cultural spans in CULTURALADAPT-6K. The overall mean is 5.59 CSIs per item (median 5). 83.9% of instances contain three or more distinct CSIs. Per-domain mean CSI counts are 8.41 for RecipeNLG, 6.21 for DailyDialog, 4.34 for GSM8K, and 3.40 for ROCStories.

The overall mean density is 7.27 CSIs per 100 words (median 5.41). RecipeNLG has the highest density at ~ 13.5 CSIs per 100 words. DailyDialog and ROCStories average between 5 and 7 CSIs per 100 words, while GSM8K has the lowest density at approximately 3.7 CSIs per 100 words.

The CSIs are distributed across multiple categories, such as Unit/Currency/Scale (10,174 spans), Food & Drink (9,788 spans), and Place/Institution/Brand (6,647 spans). 8,525 CSIs (25.4%) are marked scale-sensitive. Figurative language and cultural idioms constitute 482 CSIs (1.4%).

B Prompts and Annotation Instructions

We share the detailed prompts used in dataset construction, adaptation generation, evaluation framework, and annotation. We include all four prompt

modes and the specific exemplar items for the domains covered in this work. All conditions use the message structure $\langle \text{SYSTEM}, \text{USER} \rangle$. P2 to P4 share the same system prompt, while P1 uses a minimal role assignment. Domain-specific one-shot examples are shown in Table 13. Reverse-pass prompts and exemplar items are also included. For annotators, the same instructions were utilized, presented using an annotation interface with guidelines.

Tier 1 | CSI Detection

SYSTEM

You are a Culturally Sensitive Item (CSI) detector for cross-cultural NLP research. Detect ONLY items undeniably, heavily grounded in a specific culture. A text is culturally sensitive ONLY if it contains markers that would REQUIRE active adaptation when transplanting to a different culture.

Inclusion: (1) Unambiguous person names (not sufficient alone), (2) Specific place names, (3) Currency with amounts, (4) US customary measurements, (5) US-specific holidays, (6) American cultural products/brands, (7) American sports.

Exclusion: Trivial/incidental markers, generic/universal items, pronouns or dialogue labels, culturally neutral text.

Decision rule: A single weak marker is not enough. YES requires clear, non-trivial grounding forcing meaningful American→Indian adaptation.

Format: MARKERS_FOUND: [...] | REASONING: [...] | CONFIDENCE: [HIGH/MEDIUM/LOW] | VERDICT: [YES/NO]

USER

[Few-shot, e.g. Educational:]

Q: “Weng earns \$12/hr for babysitting...” → **YES** (\$12 = currency anchor)

Q: “A library received 50 visitors...” → **NO** (no cultural markers)

Now classify: {text}

Tier 3 | CSI Reconsideration

SYSTEM

You are a CSI detector performing a rigorous second-pass review. An initial LLM pass produced a classification; an independent NER system detected named entities.

Rules: Names count only as unambiguous proper nouns; places only as real geographies; measurements only with numeric context; holidays only if distinctly American; foods/brands only if iconic to US culture. If any doubt → reject.

Format: MARKERS_FOUND: [...] | REASONING: [...] | CONFIDENCE: [HIGH/MEDIUM/LOW] | VERDICT: [YES/NO]

USER

Initial verdict: {tier1_verdict} ({confidence}) | NER entities: {entity_list}

Text: {text}

Provide your FINAL classification.

P1 | Minimal Instruction

SYSTEM

You are a cultural adaptation specialist.

USER

{domain_label}: {text}

P1 | Minimal Instruction (cont.)

Adapt this from {source_culture} to {target_culture} context.

P2 | Constraint-Aware Rewriting

SYSTEM

You are a cultural adaptation specialist.

Replace every culture-specific item (CSI) in the text with a natural {target_culture} equivalent. The source text is grounded in {source_culture} cultural context. CSIs include: ecology, material culture (food, clothing, objects), social practices, institutions, currency, measurements, and names.

Rules:

1. Preserve all numeric values, logical dependencies, and causal structure exactly.
2. Output must be entirely in the same language as the input.
3. Do not solve, answer, or explain. Output the adapted text only.

USER

{domain_label}: {text}

Write the culturally adapted version. Output the adapted text only.

P3 | Chain-of-Thought

SYSTEM

You are a cultural adaptation specialist.

Replace every culture-specific item (CSI) in the text with a natural {target_culture} equivalent. The source text is grounded in {source_culture} cultural context. CSIs include: ecology, material culture (food, clothing, objects), social practices, institutions, currency, measurements, and names.

Rules:

1. Preserve all numeric values, logical dependencies, and causal structure exactly.
2. Output must be entirely in the same language as the input.
3. Do not solve, answer, or explain. Output the adapted text only.

USER

{domain_label}: {text}

Adapt this text step by step.

STEP 1 IDENTIFY: List each CSI and its type (ecology / material culture / social practice / institution / currency / measurement / name / other).

STEP 2 MAP: For each CSI, state the {target_culture} equivalent replacing the {source_culture} original.

STEP 3 VERIFY: Confirm all numeric values and logical dependencies remain valid after substitution.

Finally, write ONLY the final adapted text after the marker below. No explanations, no notes.

ADAPTED TEXT:

P4 | One-Shot Exemplar

SYSTEM

You are a cultural adaptation specialist.

Replace every culture-specific item (CSI) in the text with a natural {target_culture} equivalent. The source text is grounded in {source_culture} cultural context. CSIs include: ecology, material culture (food, clothing, objects), social practices, institutions, currency, measurements, and names.

Rules:

Phase 3 | Adaptation Audit (CSI Alignment) (cont.)

YOUR TASK: For each source CSI, determine what happened to it in the adapted text. Also identify any NEW target-culture items not traceable to any source CSI.

For each SOURCE CSI, return:

```
{
  "source_span": "<original CSI span>",
  "adapted_span": "<corresponding span in adapted text, or null>",
  "adapted_char_start": <int or null>,
  "adapted_char_end": <int or null>,
  "strategy": "<one of the strategies below>",
  "validity_note": "<1-sentence note>"
}
```

STRATEGY labels:

- retained: source CSI appears unchanged or near-unchanged
- localized: replaced with a specific Indian-culture equivalent
- globalized: replaced with a culture-neutral or universal alternative
- transformed: substantially restructured; function preserved, form very different
- omitted: CSI removed with no replacement
- distorted: replacement exists but clearly does not serve the same function

Also return a list of **UNSUPPORTED ADDITIONS**:

```
{
  "added_span": "<new Indian-culture span not traceable to source CSI>",
  "char_start": <int>, "char_end": <int>,
  "category": "<CSI category>",
  "justified": <boolean>
}
```

Return the complete result as:

```
{"csi_alignments": [...], "unsupported_additions": [...],
"source_csis_total": <int>, "retained_count": <int>,
"localized_count": <int>, "globalized_count": <int>,
"transformed_count": <int>, "omitted_count": <int>,
"distorted_count": <int>}
```

IMPORTANT RULES:

- csi_alignments must have exactly one entry per source CSI, in the same order.
- adapted_char_start/adapted_char_end refer to offsets in the ADAPTED TEXT (0-indexed).
- If a source CSI is entirely absent, set adapted_span=null and strategy="omitted".
- Use "distorted" only if a replacement clearly fails.
- unsupported_additions should only include spans carrying new Indian cultural content not directly mappable to any source CSI.
- All counts must sum to source_csis_total.

Return ONLY valid JSON, no other text.

Phase 4, C2 | Substitution Validity Judge

SYSTEM

You are evaluating whether cultural substitutions in an American-to-Indian adaptation are VALID.

USER

SOURCE TEXT: {source_text}
ADAPTED TEXT: {adapted_text}
DOMAIN: {domain}

You will evaluate up to 3 substitutions. For EACH substitution, score it on 5 dimensions (0.0–1.0) and return a JSON object.

SUBSTITUTIONS TO EVALUATE: {substitutions_json}

Each substitution has: source_span, adapted_span, strategy, contextual_function, required_affordances, category.

For EACH substitution return:

```
{
  "source_span": "<echo back>",
```

Phase 4, C2 | Substitution Validity Judge (cont.)

```
"affordance_validity": {"score":<0-1>,
"rationale": "<...>"},
"scale_transfer": {"score":<0-1>,
"rationale": "<...>"},
"dependency_propagation": {"score":<0-1>,
"rationale": "<...>"},
"identity_marker_fit": {"score":<0-1>,
"rationale": "<...>"},
"figurative_equivalence": {"score":<0-1>,
"rationale": "<...>"},
"overall_c2_score": <weighted>,
"c2_subtype": "<...>", "c2_severity": "<...>"
}
```

Weights: $\text{affordance} \times 0.35 + \text{scale} \times 0.30 + \text{dependency} \times 0.15 + \text{identity} \times 0.10 + \text{figurative} \times 0.10$

C2 SUBTYPE OPTIONS:

- none: substitution is valid
- C2.1_affordance_violation: replacement cannot do the same job
- C2.2_scale_transfer_error: numbers/units/prices not converted properly
- C2.3_dependency_propagation_failure: linked elements not updated
- C2.4_identity_marker_mismatch: name/title/kinship does not fit Indian context
- C2.5_figurative_equivalence_failure: idiom/metaphor not adapted equivalently

SEVERITY: none (≥ 0.8), surface (0.5–0.79), substantive (0.3–0.49), critical (< 0.3)

SPECIAL RULES:

- If adapted_span is null and strategy is "omitted": score all dimensions 1.0.
- If strategy is "retained" and necessity is "must_adapt": set overall_c2_score=0.0, subtype="C2.1_affordance_violation", severity="critical".
- If strategy is "globalized": affordance_validity should reflect whether the neutral replacement preserves the function.

Return ONLY a JSON array with one object per substitution, in the same order as input.

Phase 4, C3 | Meaning Preservation Judge

SYSTEM

You are evaluating whether an adapted text preserves the meaning, logic, and communicative purpose of the original.

USER

SOURCE TEXT (American context): {source_text}
ADAPTED TEXT (Indian context): {adapted_text}
DOMAIN: {domain}

Evaluate on these 5 dimensions. Score each 0.0–1.0.

```
{
  "core_intent": {"score":<0-1>, "rationale": "<same primary purpose?>"},
  "causal_temporal_logic": {"score":<0-1>,
"rationale": "<sequence intact?>"},
"unwarranted_elaboration": {"score":<0-1>,
"rationale": "<extra content? 1.0=none>"},
"pragmatic_register": {"score":<0-1>, "rationale": "<tone preserved?>"},
"character_role_preservation": {"score":<0-1>,
"rationale": "<roles intact?>"},
"overall_c3_score": <weighted>,
"c3_subtype": "<...>", "c3_severity": "<...>"
}
```

Weights: $\text{core_intent} \times 0.35 + \text{causal} \times 0.25 + \text{elaboration} \times 0.15 + \text{register} \times 0.15 + \text{character} \times 0.10$

DOMAIN-SPECIFIC GUIDANCE:

- gsm8k: core_intent MUST check the adapted text is still a solvable math problem with identical logical structure. If destroyed or unsolvable, core_intent=0.0. Verify currency/unit conversions yield internally consistent arithmetic.

Phase 4, C3 | Meaning Preservation Judge (cont.)

- rocstories: Check that the 5-sentence narrative structure survives. Events must follow logically. Character motivations must remain intact.
- dailydialog: Check each conversational turn maintains its communicative function (question stays question, refusal stays refusal). Pragmatic force must be preserved.
- recipenlg: Check ingredients and steps are internally consistent. If an ingredient is changed, every step referencing it must also update.

C3 SUBTYPE OPTIONS:

- none
- C3.1_core_intent_shift: adapted text serves a different purpose
- C3.2_causal_temporal_disruption: events reordered, steps broken, logic severed
- C3.3_unwarranted_elaboration: substantial content added not in source
- C3.4_pragmatic_force_drift: tone or register shifts significantly
- C3.5_character_role_shift: character roles, relationships, or agency altered

SEVERITY: none (≥ 0.8), surface (0.5–0.79), substantive (0.3–0.49), critical (< 0.3)

Return ONLY valid JSON, no other text.

Phase 4, C4 | Indian Cultural Plausibility Judge

SYSTEM

You are an Indian cultural consultant evaluating whether an adapted text reads as culturally plausible and authentic in an Indian context.

CRITICAL INSTRUCTION: An adaptation is valid if it represents ANY plausible Indian cultural context. India is diverse. Do NOT penalize for regional specificity (South Indian food is as valid as North Indian food; Mumbai settings are as valid as Delhi settings). Only flag something as implausible if it would be strange in ALL Indian contexts, or if different Indian cultural markers are mixed incoherently.

USER

ADAPTED TEXT: {adapted_text}

DOMAIN: {domain}

NOTE: This text was adapted from an American original. Your job is NOT to compare it to the source. Read it as an Indian reader would and judge whether it feels natural, plausible, and non-stereotyped.

Evaluate on these 5 dimensions. Score each 0.0–1.0.

```
{
  "internal_coherence": {"score":<0-1>,
    "rationale": "<elements fit together?>"},
  "norm_appropriateness": {"score":<0-1>,
    "rationale": "<behaviours plausible?>"},
  "stereotype_check": {"score":<0-1>, "rationale": "<no clichés? 1.0=none>"},
  "contextual_plausibility": {"score":<0-1>,
    "rationale": "<prices/settings realistic?>"},
  "perspective_authenticity": {"score":<0-1>,
    "rationale": "<insider voice?>"},
  "overall_c4_score": <weighted>,
  "c4_subtype": "<...>", "c4_severity": "<...>",
  "stereotype_examples": [... ]
}
```

Weights: coherence $\times$ 0.25 + norms $\times$ 0.20 + stereotype $\times$ 0.20 + plausibility $\times$ 0.20 + authenticity $\times$ 0.15

C4 SUBTYPE OPTIONS:

- none
- C4.1_internal_cultural_incoherence: Indian markers mixed incoherently
- C4.2_norm_violation: social behaviours implausible in Indian context
- C4.3_stereotypical_portrayal: culture reduced to clichés
- C4.4_contextual_implausibility: prices/settings/objects unrealistic for modern India

Phase 4, C4 | Indian Cultural Plausibility Judge (cont.)

- C4.5_perspective_inauthenticity: reads as outsider stereotyping

SEVERITY: none (≥ 0.8), surface (0.5–0.79), substantive (0.3–0.49), critical (< 0.3)

Return ONLY valid JSON, no other text.

Phase 5 | Function Frame Analysis

SYSTEM

You are a cultural adaptation specialist. For each Culture-Specific Item (CSI) extracted from an American text, define its FUNCTIONAL ROLE in context.

USER

SOURCE TEXT: {source_text}

DOMAIN: {domain}

CSIs TO ANALYZE: {csi_list_json}

For each CSI, return a JSON object with these fields:

```
{
  "span": "<the CSI span>",
  "denotation": "<what it literally is, 1 phrase>",
  "cultural_role": "<American cultural anchor, 1 phrase>",
  "contextual_function": "<what JOB it does in THIS text, 1 sentence>",
  "required_affordances": [<"property that must survive">, ... ],
  "numeric_constraints": "<conversion rule if scale_sensitive, else null>",
  "register_constraints": "<formality/tone requirement, else null>"
}
```

RULES:

- contextual_function is the most important field. It must describe WHAT THE CSI DOES HERE, not what it is in general. GOOD: “Serves as the everyday grocery store where the protagonist buys ingredients, anchoring the scene in routine domestic life”

BAD: “Safeway is an American supermarket chain”

- required_affordances should list 2–5 properties that ANY valid replacement must have.

Example for “Safeway”: [“grocery store”, “familiar to general public”, “walkable/drivable for routine shopping”]

- numeric_constraints only for units, currency, quantities. Be specific:

Example: “1 pound = 0.4536 kg; must convert value, not just label”

[4 few-shot examples follow the same schema as the CSI extraction examples.]

Return ONLY a JSON array of objects, one per CSI, in the same order as input.

Phase 7 | Round-Trip LLM Judge

SYSTEM

You are evaluating whether two texts in an American cultural context convey the same meaning after a cultural round-trip (American $\rightarrow$ Indian $\rightarrow$ American).

USER

TEXT A (original American):

{text_a}

TEXT B (re-Americanized after Indian cultural adaptation):

{text_b}

Score Text B relative to Text A on these dimensions (integer 0–10 each):

meaning_preservation : Same facts, events, logical relationships?

structural_preservation: Same sentence and discourse flow?

Phase 7 | Round-Trip LLM Judge (cont.)

```
detail_preservation : Specific numbers, names, brands,
dates retained?
overall : Holistic round-trip fidelity (may
be non-integer)

Also give a one-sentence rationale describing the main
difference, if any.

Return ONLY a JSON object with exactly these keys:
{
  "meaning_preservation": <int 0-10>,
  "structural_preservation": <int 0-10>,
  "detail_preservation": <int 0-10>,
  "overall": <float 0-10>,
  "rationale": "<one sentence>"
}
```

C Hardware and Setup Details

All experiments were run on a workstation with 2 Intel Xeon Gold 6542Y processors (48 cores, 96 threads, up to 4.10 GHz), 251 GiB RAM, and three NVIDIA RTX 6000 Ada GPUs (48 GB VRAM each). Open-source models were served via vLLM with BF16 precision. The total compute cost is estimated to be \$2,100 USD, with \$1,480 in closed-source API calls, and \$620 in local GPU electricity.

C.1 API Model Versions and Reproducibility.

To ensure reproducibility, we specify the exact API model snapshots used. For our LLM-as-judge evaluation, we utilized OpenAI’s gpt-4o (standard) API endpoint and Google’s gemini-2.5-pro API endpoint. All API-based queries and evaluations were executed between February 2026 and May 2026 using the official Python SDKs.

C.2 Operational Footprint

To apply our framework to other contexts, we outline the exact operational footprint required to process a single source text x and evaluate a single adaptation candidate y generated by a model M .

The pipeline begins with the static setup phase, which is executed only once per source text x regardless of how many models are evaluated. First, a single API call extracts the culture-specific items $C(x)$. A second API call then generates the Wikidata-anchored function frames $\{f_i^{\text{src}}\}$ using the extracted entities. Once established, these reference frames serve as the ground truth against which any candidate adaptation of x is compared.

The evaluation phase initiates with a rule-based integrity check (C1) on the candidate model’s output y . This check requires zero API calls. If y fails the integrity gate, the evaluation terminates, sparing resources. For outputs that pass, the framework initiates the forward evaluation of Scope (C2),

Substitution Validity (C3), and Function Preservation (C4). These assessments are consolidated into two API calls per frontier LLM judge (GPT-4o and Gemini 2.5 Pro) to minimize token consumption and API latency, resulting in four API calls. An adjudication call is executed only if a categorical disagreement occurs between the two judges.

The final phase evaluates Round-Trip Plausibility (C5). To generate the reconstructed source text $\tilde{x}$, the candidate model M performs its own reverse pass, which requires a single local inference call rather than an external API call. This ensures that the plausibility score measures the model’s internal consistency. GPT-4o is employed to generate the reverse adaptation $\tilde{x}$ solely during as ablation, which replaces the local inference with one API call. Once $\tilde{x}$ is generated, the comparison of the original text x and the reconstruction $\tilde{x}$ is evaluated via two API calls, one for each frontier LLM judge.

In total, evaluating a single completed adaptation requires two static reference API calls, one local inference pass by the evaluated model, and six dynamic evaluation API calls, with approximately 0.3 additional adjudication calls expected due to occasional judge disagreements.

D Pilot Study: Regional Cultural Adaptation

To test whether the evaluation pipeline generalizes beyond a fixed, pan-Indian target, we conduct a regional pilot study targeting two distinct cultural corridors: Eastern India (Bengali) and Western India (Maharashtra and Gujarat). For both targets, we sample the same 500 texts from CULTURALADAPT-6K (125 per domain) and generate adaptations using three models (Llama-3.3-70B, Qwen2.5-72B, Gemma-3-27B) under three structured prompt conditions. All pipeline stages evaluate the outputs identically to the main study. Table 14 reports the comparative metrics across both regions alongside the main pipeline averages.

Eastern India (Bengali). Target-culture instructions specify Bengali cultural context: Kolkata-centric geography, Bengali cuisine (e.g., *ilish*, *mishti doi*), regional festivals (e.g., *Durga Puja*, *Poila Boishakh*), and Bengali naming conventions (Bangladeshi Bengali culture is excluded from the scope). All pipeline stages produce interpretable scores without modification. However, Wikidata coverage drops 13.6 points because regional entities are sparser in the knowledge base. We observe

Domain	American	Indian
Educational	Charlotte is making a pie. She can't decide if she will make cherry or blueberry, so she makes whichever is cheapest. Both pies need \$2 for flour, \$1 for sugar, and \$1.50 for eggs and butter. The blueberry pie needs 3 lb in 8 oz containers for \$2.25; the cherry pie needs 4 lb at \$14 per four-pound bag. What is the total price to make the cheapest pie?	Ananya is making a pie. She can't decide if she will make guava or mango, so she makes whichever is cheapest. Both pies need Rs. 160 for flour, Rs. 80 for sugar, and Rs. 120 for eggs and butter. The guava pie needs 1.5 kg in 250 g packs for Rs. 180; the mango pie needs 2 kg at Rs. 1,120 per two-kilogram box. What is the total price to make the cheapest pie?
Stories	Nashville is Robyn's favorite place to visit. She was raised on country music and loved it still. When there she always tried to go to the Ryman and 2nd Ave. One time she actually got up and sang at one of the watering holes! She eventually decided to relocate to Nashville.	Jaipur is Priya's favorite place to visit. She was raised on Rajasthani folk music and loved it still. When there she always tried to go to Hawa Mahal and Johari Bazaar. One time she actually got up and sang at one of the chai stalls! She eventually decided to relocate to Jaipur.
Dialogue	A: We need to figure out how much we've spent. B: Why? A: We should save more – if we save more we can retire earlier. B: OK, get the receipts. A: We spent \$700 on the home loan, \$300 on groceries, \$75 on utilities, \$250 on gas, \$100 on books, and \$400 on entertainment. B: That's \$1,825. [...]	A: We need to figure out how much we've spent. B: Why? A: We should save more – if we save more we can retire earlier. B: OK, get the receipts. A: We spent Rs. 25,000 on the home loan EMI, Rs. 12,000 on groceries, Rs. 3,000 on utilities, Rs. 8,000 on petrol, Rs. 4,000 on books, and Rs. 15,000 on entertainment. B: That's Rs. 67,000. [...]
Procedural	Holiday Cheese Tree. 2 pkg. (8 oz. ea.) Philadelphia Cream Cheese; 1 pkg. (8 oz.) Kraft Sharp Cheddar; 1 Tbsp. red bell pepper; 1 Tbsp. onion; 2 tsp. Lea & Perrins Worcestershire Sauce; 1 tsp. lemon juice; dash ground red pepper; toasted bread or crackers. [...]	Festive Cheese Tree. 2 pkg. (225 g ea.) Amul Cream Cheese; 1 pkg. (225 g) Amul Shredded Cheddar; 1 Tbsp. red bell pepper; 1 Tbsp. onion; 2 tsp. Weikfield Worcestershire Sauce; 1 tsp. lemon juice; dash ground red pepper; toasted bread or mathri. [...]

Table 13: One-shot examples for P4. Forward adaptation (American→Indian) reads left-to-right; the same pairs serve as reverse examples (Indian→American) by reading right-to-left. IDs: gsm8k_train_6713, rocstories_train_46523, dailydialog_train_8500, recipenlg_train_2012030. All examples were manually authored. Full text in released code.

that this cascades into higher cultural foreignness (CF; models retain source items) and lower substitution validity (C3). Affordance violations (C3.1) rise by 5.1 points, driven by the strictness of Bengali food and ritual specifics, while scale-transfer error rates (C3.2) hold constant since unit and currency norms are shared nationally. Function preservation (C4 FPS) is similar: models locate semantically parallel Bengali referents but satisfy contextual affordances less reliably than for pan-Indian targets. Three Bengali-native annotators independently audited 200 stratified outputs ($\kappa = 0.82$), with agreement profiles matching the main study.

Western India (Maharashtra/Gujarat). Instructions here specify Mumbai/Pune-centric geography, Western Indian cuisine (e.g., *vada pav*, *dhokla*, *puran poli*), regional festivals (e.g., *Ganesh Chaturthi*, *Navratri*), and Marathi/Gujarati conventions. Wikidata coverage recovers 5.7 points over Bengali, as Mumbai and Pune have denser Wikidata representation. Cultural foreignness (CF) and substitution validity (C3) both sit between the main and Bengali values. The affordance violation rate (C3.1) is 2.4 points higher than the main pipeline (vs. 5.1 for Bengali). This reflects that Western Indian food and festival items overlap more with pan-Indian defaults than Bengali-specific ones do. Scale-transfer rates are constant (+0.5). Function preservation is marginally higher than Bengali (0.627 vs. 0.618), tracking the Wikidata coverage. Crucially, the rank ordering of models is preserved identically across all metrics in both regional pilots. Three Western Indian-native annotators au-

dated 200 stratified outputs ($\kappa = 0.78$).

Metric	Main	Bengali	Δ_{Bng}	Western	Δ_{Wst}
Wikidata cov. (%)	82.0	68.4	-13.6	74.1	-7.9
C1 Integrity pass (%)	61.5	58.1	-3.4	59.3	-2.2
C2 CF ↓	0.074	0.098	+0.024	0.085	+0.011
C2 CT ↑	0.859	0.827	-0.032	0.841	-0.018
C3 Validity ↑	0.610	0.571	-0.039	0.588	-0.022
C3.1 Affordance (%) ↓	11.8	16.9	+5.1	14.2	+2.4
C3.2 Scale (%) ↓	21.8	22.1	+0.3	22.3	+0.5
C4 FPS ↑	0.652	0.618	-0.034	0.627	-0.025
C5 Overall RT ↑	7.46	7.08	-0.38	7.22	-0.24
Human κ ($n=200$)	0.76	0.82	+0.06	0.78	+0.02

Table 14: Regional Pilots vs. Main Pipeline. Evaluated on 3 models, 3 structured prompts, and 500 source texts. Main values are global averages from main text tables. Δ : pilot minus main; red highlights structurally meaningful degradation.

E Round-Trip Fixed-Reverser Ablation

The main round-trip evaluation uses the same model for both the forward and reverse adaptations, which may introduce a potential confound: model's generation biases may systematically inflate or suppress its own reverse-pass fidelity. To isolate this, we replace the same-model reverser with a fixed external frontier GPT-4o model applied uniformly in Table 10. All other conditions are kept constant. We report the complete surface, CSI-alignment and semantic layers in Table 15 and compare directly with the values of the main-study.

Setup. For each of the forward-adapted outputs, GPT-4o receives the adapted Indian-context text and the same reverse-adaptation prompt used in

Model	Surface			CSI Alignment				LLM Judge (0–10)				ΔOv
	BLEU $\uparrow$	RG-L $\uparrow$	BS-F1 $\uparrow$	Rec $\uparrow$	Sub $\downarrow$	Lk $\downarrow$	Ms $\downarrow$	Mn. $\uparrow$	St. $\uparrow$	Dt. $\uparrow$	Ov. $\uparrow$	
Llama-3.3-70l	0.641	0.769	0.876	0.744	0.661	0.006	0.271	8.63	9.19	5.31	7.88	+0.06
Gemma-2-9B	0.638	0.751	0.861	0.712	0.609	0.008	0.281	8.22	8.83	4.97	7.55	+0.05
Llama-3.1-8B	0.581	0.724	0.849	0.714	0.701	0.011	0.306	8.21	8.78	4.62	7.40	+0.06
Mistral-7B	0.513	0.692	0.843	0.731	0.672	0.016	0.301	7.94	8.64	4.49	7.22	+0.05
Global avg.	0.593	0.734	0.857	0.725	0.661	0.010	0.290	8.25	8.86	4.85	7.51	+0.05
Main study	0.574	0.717	0.849	0.706	0.677	0.013	0.307	8.20	8.81	4.80	7.46	—

Table 15: Round-trip consistency with GPT-4o as the fixed reverse-pass model, across 95,984 pairs (same outputs as Table 10). **Surface**: string overlap vs. original American source. **CSI Alignment**: Rec = original CSI returned; Sub = different American CSI; Lk = Indian CSI persists; Ms = absent. **LLM Judge**: Mn. = meaning, St. = structural, Dt. = detail, Ov. = overall (0–10). ΔOv : difference in overall score relative to the same-model main study (Table 10).

the main study, with the source and target cultures swapped. The resulting reverse outputs are evaluated identically to the main pipeline. To avoid self-evaluation bias, we chose not to use GPT-4o as an LLM-as a judge in this specific experiment.

Results. Table 15 shows that replacing the same-model reverser with GPT-4o produces uniformly small upward shifts across all metrics and all four models. Surface scores improve marginally (BLEU +0.018–+0.023, ROUGE-L +0.016–+0.019, BERTScore +0.008–+0.009). This is consistent with GPT-4o’s stronger text reconstruction capability relative to the smaller open-source reversers. CSI recovery improves modestly (Rec +0.017–+0.023) while leakage decreases (Lk −0.002–−0.004), showing that a more capable reverser model strips Indian cultural spans more reliably. This confirms the forward-pass bottleneck.

Specifically, semantic scores shift by at most +0.06 points in any individual dimension (detail, Dt) and $\leq$ +0.06 in the general equivalence (Ov) in all models. The global mean shift in the overall score is +0.05. The model rankings are fully preserved: Llama-3.3-70B, followed by Gemma-2-9B, Llama-3.1-8B, and Mistral-7B, thus yielding Spearman $\rho = 1.0$ against the main-study ranking.

F CRAFT Subtype Definitions

The proposed CULTURAL REWRITING AND ADAPTATION FAILURE TAXONOMY (CRAFT) assigns each failure to the earliest pipeline stage where it originates. Severity ratings reflect the overall functional impact and usability of each adaptation. *Critical* errors make the output arithmetically wrong, functionally unusable, or semantically inverted. *Substantive* errors produce culturally inappropriate or misleading text that retains surface plausibility. *Surface* errors reduce naturalness or precision without invalidating the core adaptation task. Table 16 illustrates all subtypes.

C1: Integrity. This category marks degenerate outputs that prevent evaluation at any subsequent stage. **C1.1 Output Integrity Failure** covers outputs that are entirely non-functional or non-evaluable. This includes task refusals, meta-commentaries instead of adaptations, code-mixed or wrong-language responses, and truncated or repetitive outputs. Severely affected items are excluded from all downstream scoring, and moderately affected items are retained but penalized.

C2: Scope. This category evaluates whether the model identified the correct cultural spans (CSIs) and applied the right degree of change. **C2.1 Residual Source Retention** occurs when a required cultural marker survives the adaptation unchanged, even when it should have been adapted. The source element remains where it would be unfamiliar or inappropriate to a reader of the target-culture. **C2.2 Cultural Saturation** is the inverse. The output introduces target-culture markers beyond what the source supports. Extra content with no source counterpart is added, inflating cultural detail without a textual basis (oversaturation). **C2.3 Over-Globalization** arises when a cultural marker is replaced by a generic or neutral term instead of a target-culture equivalent. The adaptation is not wrong, but carries no local identity.

C3: Substitution. This category evaluates the local validity of each replacement. A replacement may be culturally plausible in isolation, yet violate real-world affordances, numeric conventions, or figurative norms of the target culture. **C3.1 Affordance Violation** occurs when the replacement does not exist or function in the target culture. The product is not available locally, the institution does not operate there, or the practice is not feasible. **C3.2 Scale/Numeric Transfer Error** occurs when a numeric value (price, weight, distance) is copied from the source after only the unit label is changed. The

resulting figure is economically or physically implausible. **C3.3 Dependency Propagation Failure** applies when a cultural marker is correctly replaced at its first mention but not carried through to subsequent references, derived quantities, or dependent calculations. **C3.4 Identity-Marker Mismatch** flags a name, demographic label, or group identifier that conflicts with the cultural context set by the surrounding replacements. **C3.5 Figurative Equivalence Failure** covers idioms, proverbs, or culturally embedded phrases that are kept verbatim or mapped to a non-equivalent expression instead of an idiomatic target-culture counterpart.

C4: Preservation. This category evaluates functional equivalence. Surface-level substitutions can silently shift intent, alter scope, or change pragmatic force. Such failures emerge only when the replacement is evaluated in context. **C4.1 Core Intent Shift** occurs when the adapted text no longer serves the same function as the source. A question becomes a statement, a reasoning problem becomes a narrative, or an instruction becomes a description. **C4.2 Scope/Granularity Distortion** denotes a change in scale, quantity, or extent that has no textual justification and meaningfully alters the problem space. **C4.3 Pragmatic/Register Drift** applies when the register or speech-act type shifts across the adaptation. The text functions differently as a communicative act even though its factual content may be preserved.

C5: Plausibility. This category evaluates holistic authenticity. Unlike earlier stages, these failures arise from the interaction of choices in an adaptation, rather than from any single isolated replacement. **C5.1 Internal Cultural Incoherence** occurs when multiple replacements are each plausible on their own but co-occur in a combination that is impossible or contradictory within the target culture. **C5.2 Stereotypical Portrayal** marks adaptations that reduce the target culture to a reductive representation, collapsing internal diversity into a single default frame. **C5.3 Perspective Inauthenticity** denotes outputs narrated from an external observer's viewpoint rather than from within the target culture. Signals include other language or explanatory language aimed at an outside audience.

Code	Sev.	Adapted Excerpt	Discussion
C1.1	***	[Model returns: "I cannot adapt this text as it involves meat products which may be culturally sensitive."]	The model refuses to produce an adapted output. The result is non-functional and cannot be evaluated at any subsequent stage.
C1.1	***	"Priya ne 1.5 kg keema kharida DMart se Rs. 450/kg mein. Usne aadha hissa Independence Day ke liye kebabs banane mein use kiya. Bache hue keema ka kina paisa laga?"	The model produces output in transliterated Hindi instead of English. The task required English-language adaptation, not translation or code-mixing.
C2.1	**	"Priya bought 3 pounds of keema at Safeway for \$4.99/lb . She used half to make kebabs for the Fourth of July barbecue. How much did the unused keema cost?"	Only the name and food item were changed. The unit (pounds), store (Safeway), currency (dollars), and holiday (Fourth of July) all remain from the American source.
C2.2	**	"Ananya bought 1.5 kg of tandoori gosht at DMart for Rs. 450/kg for the Navratri feast to make seekh kebabs for the entire mohalla . How much did the unused gosht cost?"	Valid replacements are present, but extra insertions (mohalla, seekh kebabs, tandoori) have no counterpart in the source. The output adds cultural detail that was never in the original.
C2.3	*	"A person bought 1.36 kg of minced meat at a local supermarket for the regional price per unit. They used half to make patties for a public holiday gathering. How much did the unused meat cost?"	Every cultural marker is replaced with a maximally generic term. The output could describe any country and carries no Indian cultural identity.
C3.1	***	"Priya bought 1.5 kg of ground beef at Reliance Fresh for Rs. 450/kg. She used half to make kebabs for the Eid barbecue. How much did the unused beef cost?"	Ground beef is not commonly stocked at mainstream Indian grocery chains and carries religious restrictions for a large portion of shoppers. The replacement does not reflect what is actually available in the target culture.
C3.2	***	"Priya bought 1.5 kg of mutton at Rs. 4.99/kg . She used half to make kebabs for Independence Day. How much did the unused mutton cost?"	The unit label was changed from lb to kg, but the numeric value was copied directly. Rs. 4.99/kg is off by orders of magnitude from real mutton prices, making the arithmetic answer meaningless.
C3.3	**	"Priya bought 1.5 kg of mutton at Rs. 450/kg. She used half to make kebabs for Independence Day. The remaining 1.5 pounds of mutton cost how much?"	The weight was converted at first mention (3 lb → 1.5 kg), but the dependent reference later reverts to the source value and unit (1.5 pounds). The two mentions are internally inconsistent.
C3.4	**	" Sarah bought 1.5 kg of mutton at BigBazaar for Rs. 450/kg. She used half to make kebabs for the Eid celebration. How much did the unused mutton cost?"	Store, price, unit, and holiday are all plausible Indian replacements, but the Western name Sarah clashes with the otherwise Indian cultural frame.
C3.5	*	"Priya bought 1.5 kg of mutton at DMart for Rs. 450/kg. She hoped her kebabs would be a real home run at the Independence Day party. How much did the unused mutton cost?"	The baseball idiom "home run" has no cultural basis in India. A cricket-based equivalent like "hit for a six" would fit the target culture.
C4.1	***	"Priya bought mutton at the market. She prepared kebabs for Independence Day. The family enjoyed the meal together."	The arithmetic question has been removed entirely. The core purpose of the original, testing mathematical reasoning, is lost.
C4.2	**	"Priya bought 15 kg of mutton at Rs. 450/kg for the Independence Day celebration for the entire neighbourhood . She used half to make kebabs. How much did the unused mutton cost?"	The quantity is inflated 10× without any narrative reason. A single household buying 15 kg makes the scenario implausible and changes the problem's scale.
C4.3	**	" Take 1.5 kg of mutton and marinate half for Rs. 450 worth of kebabs for Independence Day. Use the remaining portion for the gravy."	The register shifts from a third-person math problem to a second-person recipe instruction. The question is gone and the text now reads as a cooking directive.
C5.1	**	"Priya bought 1.5 kg of beef at a temple market for Rs. 450/kg. She used half to make kebabs for the Navratri feast. How much did the unused beef cost?"	Beef at a temple market during Navratri (a Hindu festival commonly observed with vegetarian food) creates a combination that no single Indian cultural context supports.
C5.2	**	"Ananya bought some spicy curry meat at the colourful bazaar for a few rupees. She used half to make curry for a festival, as Indians always cook elaborate curries for every occasion . How much did the unused meat cost?"	"Spicy curry meat," "colourful bazaar," and the sweeping generalization reduce Indian food culture to a single stereotype rather than reflecting actual practices.
C5.3	*	"In a typical Indian household , Priya would buy 1.5 kg of mutton, which Indians commonly purchase for their colourful festivals . She used half to make kebabs. How much did the unused mutton cost?"	The explanatory aside describes Indian practices as if for an outside audience. A culturally grounded version would narrate from within the culture without explaining it.

Table 16: Extended CRAFT illustration. All subtypes demonstrated on a single GSM8K source item (*Sarah bought 3 pounds of ground beef at Safeway for \$4.99/lb. She used half to make cheeseburgers for the Fourth of July barbecue. How much did the unused beef cost?*). Row shading groups by pipeline stage: C1 Integrity, C2 Scope, C3 Substitution, C4 Preservation, C5 Plausibility. Severity: *** Critical, ** Substantive, * Surface.