XCR-Bench: Benchmarking Cross-Cultural Reasoning in LLMs via Culture-Specific Items and Hall's Triad

Anonymous ACL submission

Abstract

Cross-cultural competence in large language models (LLMs) requires understanding and adapting Culture-Specific Items (CSIs) across varying cultural contexts. However, progress in evaluating this capability remains limited by the lack of high-quality CSI-annotated corpora with parallel cross-cultural sentence pairs. We introduce XCR-BENCH, a Cross(**X**)-**C**ultural **R**easoning **Bench**mark containing **4.1k** parallel sentences and **1,098** CSIs across three reasoning tasks. XCR-BENCH integrates Newmark's CSI framework with Hall's Triad of Culture, enabling evaluation across levels of cultural visibility- from observable practices to implicit social norms and values. Experiments on eight multilingual LLMs show that state-of-the-art models exhibit consistent weaknesses in identifying and adapting specific categories of CSIs, revealing a gap between surface-level recall and explicit cultural reasoning. Performance declines significantly on culturally sensitive categories and deeper cultural levels ($p < 0.005$, 8/8 models), and adaptation quality varies systematically across target cultures and Bengali regional variants, indicating encoded regional and ethno-religious biases even within a single linguistic setting. We publicly release the corpus and code to support future research on cross-cultural NLP.[1]

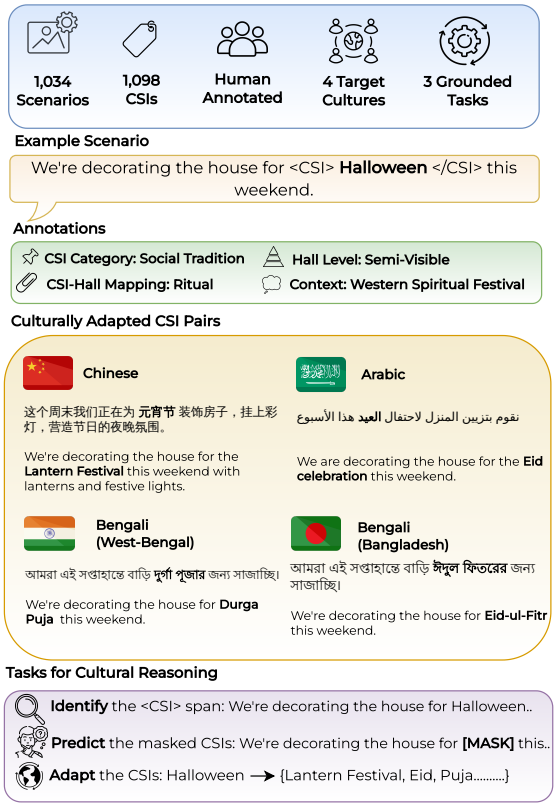


Figure 1: Overview of XCR-BENCH: Western-source scenarios annotated with Newmark's CSI categories and Hall's cultural levels, adapted across four cultures (flag of Saudi Arabia reflects annotator background) in both intra- and inter-lingual settings, supporting three tasks: CSI Identification, Prediction, and Adaptation.

1 Introduction

Large Language Models (LLMs) have been widely documented to exhibit a pronounced *Anglo-American*, or *Western* cultural bias, predominantly reflecting values and perspectives associated with the US and UK (Johnson et al., 2022; Dunn et al., 2024; Atari et al., 2023). This bias is often attributed to the West-centric composition of both their training data and the annotators involved in their development, making it a predictable artefact of their training process (Kabir et al., 2025c; Navigli et al., 2023). However, recent work has challenged a common implicit assumption in this line of research: the presence of cultural bias does not necessarily entail genuine cultural competence within the corresponding cultural context (Saha et al., 2025). This observation leads us to ask: **to what extent do current LLMs perform genuine culturally grounded reasoning, rather than surface-level representation matching driven by**

[1]Code & Corpus: https://anonymous.4open.science/r/xcr_bench-573E

Dataset	Domain	Task	CSI Annot.	Hall's Triad	Cross-Cult. CSI Pair
ApposCorpus (Kementchedjhieva et al., 2020)	Person, Org.	Trans.	✗	✗	✗
Adaption (Peskov et al., 2021)	Celebrity	Trans.	✗	✗	✗
IdiomKB (Li et al., 2024)	Idioms	Trans.	✗	✗	✗
CulturalRecipes (Cao et al., 2024)	Recipes	Trans.	✗	✗	✗
NormAd (Rao et al., 2025)	Soc. Etiq.	Reason.	✗	✗	✗
CAMT (Yao et al., 2024)	Cult. Cat.	Trans.	✓	✗	●
XCR-BENCH **(ours)**	Cult. Cat.	Reason.	✓	✓	✓

Table 1: Comparison of XCR-BENCH with prior cross-cultural datasets. Abbreviations: Cultural (*Cult.*), Category (*Cat.*), Annotation (*Annot.*), Translation (*Trans.*), Reasoning (*Reason.*). Symbols denote the level of support: ✓ indicates full support, ✗ indicates the absence of support, and ● denotes partial or limited support.

Western-centric training distributions, when interpreting and adapting Culture-Specific Items (CSIs) across cultures and levels of cultural depth?

Beyond performance within predominantly Western cultural settings, a more challenging problem is evaluating whether LLMs can reason about and adapt culturally grounded meaning across different societies and contexts (Hershcovich et al., 2022). Existing work mainly focuses on translation or narrow substitution tasks involving named entities, idioms, recipes, or dialogues (Peskov et al., 2021; Cao et al., 2024; Li et al., 2024), while recent benchmarks remain limited to partial CSI annotation or surface-level cultural categorisation (Yao et al., 2024). A key missing component is a model of cultural *depth*, like Hall's Triad of Culture (Hall, 1976) that distinguishes *Visible* (technical), *Semi-visible* (formal), and *Invisible* (informal) levels, where visible behaviours are shaped by underlying norms, values, and beliefs. Newmark's CSIs and Hall's levels can capture *orthogonal* dimensions: *what* cultural element is present versus *how deeply* its meaning is embedded, yet no existing typology or annotated resource situates CSIs within Hall's framework. As shown in Table 1, current datasets therefore lack the combination of fine-grained CSI annotations, theoretically grounded cultural depth modeling, and parallel cross-cultural CSI pairs. Consequently, existing evaluations **cannot assess whether LLMs move beyond surface-level representation matching to perform culturally grounded reasoning across different levels of cultural depth.**

To bridge this gap, we introduce the Cross(**X**)-**C**ultural **R**easoning **Bench**mark (XCR-BENCH), a benchmark for evaluating cultural reasoning in LLMs beyond translation-level accuracy. Grounded in Newmark's and Hall's cultural theories, XCR-BENCH consists of 4.1k parallel sentences and 1,098 unique CSIs across four target cultural settings: Chinese, Arabic, Bengali (West Bengal and Bangladesh). The corpus provides human annotations for CSI categories, CSI–Hall mappings, Hall's cultural depth levels, and culturally adapted CSI pairs from English into target cultures. Dataset construction follows a multi-stage human annotation and adjudication pipeline involving native speakers and expert validation, with substantial inter-annotator agreement for both CSI category and Hall-level annotations. XCR-BENCH supports three tasks: *CSI Identification*, *CSI Prediction*, and *CSI Adaptation*, together with task-specific evaluation metrics designed to measure culturally grounded reasoning rather than surface-level pattern matching.

We evaluate 8 state-of-the-art multilingual LLMs (4 closed-source, 4 open-weight). Across all three tasks, models handle surface-level cultural cues far more reliably than the implicit norms and values that require explicit reasoning. In summary, our contributions are:

- We introduce XCR-BENCH, a human-annotated benchmark linking Newmark's CSIs with Hall's Triad of Culture through CSI categories, cultural depth levels, and cross-cultural adaptation pairs.
- We formulate cultural competence as a culturally grounded reasoning problem through three tasks: *CSI Identification*, *CSI Prediction*, and *CSI Adaptation*, together with task-specific evaluation metrics beyond

translation-based settings.

- We provide a systematic evaluation of cross-cultural reasoning in multilingual LLMs across CSI categories, cultural depth levels, target cultures, and regional variants.

Beyond our core evaluation, XCR-BENCH provides a valuable resource for downstream applications including cross-cultural machine translation, cross-lingual transfer, and the design of culturally grounded evaluation protocols for LLMs.

2 XCR-Bench Corpus

Due to the nuanced nature of CSIs, we incorporate human annotation and validation throughout XCR-BENCH construction. The process comprises two phases: (1) building the base corpus with Western CSIs annotated by CSI categories and Hall's cultural levels; and (2) annotating adaptable CSI pairs for four target cultures (Chinese, Arabic, Bengali from West Bengal and Bangladesh) to form a parallel corpus.

2.1 Construction of Base XCR-Bench Corpus

Figure 2 provides an overview of the base XCR-BENCH construction pipeline. It includes: (1) CSI and cultural context extraction, (2) sentence generation with the extracted CSI, and (3) annotating the CSI to produce the category and Hall's level.

CSI & Cultural Context Extraction

The first step in corpus construction involves identifying and extracting Culture-Specific Items (CSIs) representative of Western cultures. To this end, we leverage two structured cultural knowledge resources: CANDLE[2] and Cultural Atlas[3]. CANDLE provides cultural commonsense assertions, while Cultural Atlas offers curated descriptions of multicultural norms (details in Appendix F). We crawl cultural facts from these platforms, primarily focusing on the US and UK, supplemented by a selected few highly relevant examples from Europe. We then manually extract CSIs by mapping these statements to the CSI categories proposed by Newmark (1988) (described in Appendix B.1). For each extracted CSI, we then retrieve a high-level cultural definition from Wikipedia to contextualize its usage. This process yields a set of CSIs paired with their corresponding cultural contexts.

[2]https://commonsense.scads.ai/candle/
[3]https://culturalatlas.sbs.com.au/

Sentence Generation Incorporating CSIs

In the second step, we generate realistic sentences that reflect daily communication by prompting GPT-4o, Claude-3.7-Sonnet, and DeepSeek-R1 with each CSI and its corresponding cultural context, as illustrated in Step 2 of Figure 2. Each model produces a candidate sentence, resulting in three alternatives per CSI. Two human annotators then evaluate these candidates based on three criteria: *accuracy*, *fluency*, and *realism*, and select the highest-scoring one. We explicitly emphasize *realism* in our prompt and selection, as real-world cross-cultural contexts are inherently complex and essential for assessing cross-cultural competence (Rauba et al., 2024). This process yields a set of culturally grounded, naturalistic sentences aligned with the original CSI-context pairs. For experimental convenience, we annotate the selected sentence by enclosing the CSI within <CSI> and </CSI> tags.

CSI Category and Hall's Level Annotation

Finally, we categorize each CSI based on its cultural context and the generated sentence from Steps 1 and 2, rather than at extraction time, as a CSI's category under Newmark's taxonomy can shift with contextual usage. For instance, *wearing jeans* may fall under Material Culture or Social Culture depending on whether the context emphasises attire or workplace norms. Although we use Newmark's CSI categories as a foundational framework, their broad scope often merges diverse cultural elements. To address this, we incorporate the cultural taxonomy from Liu et al. (2025), mapping its *ideational* and *social* dimensions to Newmark's Categories 3 (Social Culture) and 4 (Organizations, Customs, and Ideas), respectively. Annotations at this stage achieve a Cohen' s kappa of 0.68, indicating substantial agreement between the two annotators.

We further annotate each CSI with its corresponding Hall cultural level, independently of Newmark's categorization: for instance, food-related CSIs may encode formal or informal cultural meanings without functioning as material artefacts. This context-sensitive treatment aligns with Skopos Theory (Schäffner, 1998), which argues that the function of an element in the target communicative situation, rather than its surface form, determines its cultural role. Accordingly, we assign each CSI one of Hall's three levels of cul-

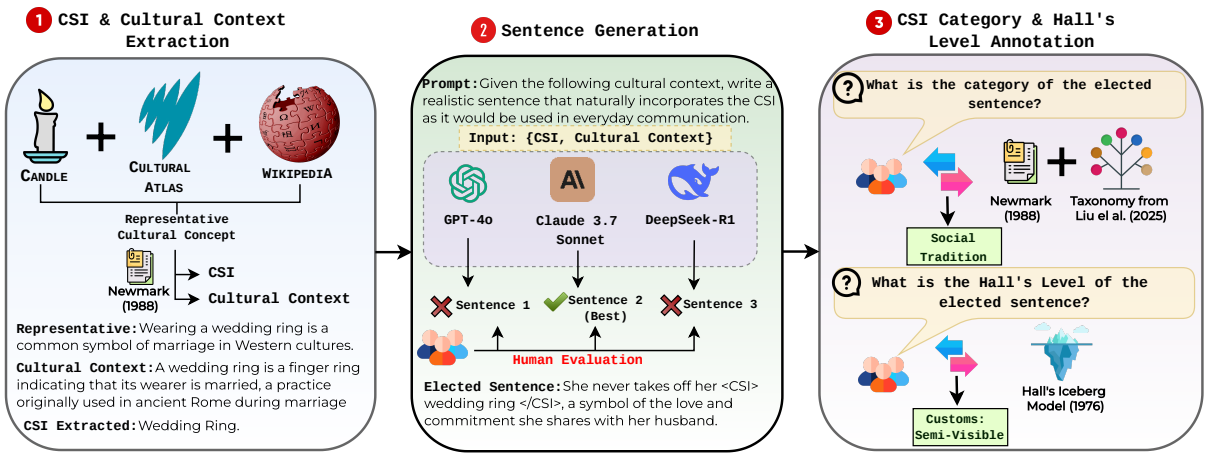


Figure 2: XCR-BENCH corpus creation and annotation pipeline. (1) Culture-specific items (CSIs) and their contexts are extracted from well-known cultural databases. (2) Sentences are generated using LLMs and filtered for accuracy and fluency. (3) Each sentence is annotated with a CSI category and mapped to Hall's cultural levels.

tural visibility (e.g. Visible, Semi-visible, Invisible). The mapping of cultural elements to Hall's Triad is provided in Table 8. Annotation of Hall's cultural elements yields a Cohen's kappa of 0.64 between the two annotators.

2.2 Construction of the Parallel Corpus

Each instance in our dataset is annotated with culturally adapted counterparts in four target cultures: Chinese, Arabic, and Bengali (two variants). The objective is to construct culturally equivalent instances of the Western CSIs from the base XCR-BENCH corpus. Because cultural equivalence is inherently subjective and cannot be treated as a categorical labeling problem, we adopt an adjudicated annotation protocol (Klie et al., 2024).

Since multiple culturally valid adaptations may exist for a single CSI, we interpret variation as cultural pluralism rather than disagreement (Berry, 1974). Consequently, conventional agreement-based metrics (e.g., IAA) are not appropriate for this setting.

Two native annotators per target culture independently produce adaptations following detailed guidelines (Appendix E.1), with disagreements resolved through discussion and final adjudication by an expert annotator with cultural and anthropological expertise (full procedure and qualifications in Section E). Following Newmark's framework for cultural equivalence (Newmark, 1988), annotators select one of four equivalence types: (1) a direct equivalent; (2) a functionally similar equivalent; (3) a neutral expression when no equivalent exists; or (4) *Non-transferable* when the CSI conflicts with local norms (e.g., <CSI>dating app</CSI> in certain Arab contexts). Annotators are also provided with relevant cultural information from CANDLE and Cultural Atlas, and encouraged to modify sentence structure to ensure coherent, culturally appropriate intra- and inter-lingual adaptations.

2.3 Dataset Analysis

The XCR-BENCH corpus comprises 4,136 parallel sentences spanning 4 distinct cultures and 2 regional variants. It contains 1,098 unique Culture-Specific Items (CSIs), distributed across 7 CSI categories. Table 1 situates XCR-BENCH in relation to existing datasets for culture-aware benchmarking. Most prior datasets in this area primarily provide translation pairs, with CAMT (Yao et al., 2024) being the closest in scope, as it includes cross-cultural CSI pairs for a subset of its data. In contrast, XCR-BENCH supports both reasoning-oriented and culturally aware adaptation evaluations by offering paired *intra-lingual* and *inter-lingual* annotations. Notably, XCR-BENCH is the first dataset aligned with Hall's triad of culture, with explicit emphasis on the *Semi-visible* and *Invisible* layers, enabling evaluation of cultural reasoning beyond translation-focused competence.

3 Cross-Cultural Reasoning Evaluation

3.1 CSI Identification

For the CSI Identification task, we remove the <CSI> tags from the base XCR-BENCH sentences and prompt LLMs to identify the term(s) that are specific to Western (specifically US/UK) culture

and salient for cross-cultural communication. To ensure a consistent interpretation of the task, we provide the definition of Culture-Specific Items (CSIs as defined by Newmark (1988)) within the prompt during inference.

3.2 CSI Prediction

For the CSI Prediction task, we replace the CSI spans in the base XCR-BENCH sentences with a [MASK] token enclosed within <CSI> tags, and prompt models to predict culturally appropriate terms that reflect Western (US/UK) cultural usage. As in the *CSI Identification* task, we include a brief definition of CSIs in the prompt to ensure a consistent interpretation of the task.

3.3 CSI Adaptation

In the CSI Adaptation task, we present LLMs with the base XCR-BENCH sentences containing explicitly marked <CSI> spans and prompt them to adapt these CSIs to Arabic, Bengali, or Chinese cultural contexts. For each instance, models are required to produce both *intra-lingual* (English) and *inter-lingual* (target language) adaptations. In addition, we ask models to explicitly indicate which of Newmark's twelve cultural adaptation strategies (e.g., *transference*, *equivalence*, etc. full details in Appendix B.3) they employ when adapting each CSI.

Our prompts used for each task are presented in Appendix I.

3.4 Evaluation Metrics

We evaluate model performance on the CSI Identification and Prediction tasks using a combination of strict ("hard") and lenient ("soft") metrics designed to capture different aspects of correctness. The hard metrics penalize any deviation from the ground truth, while the soft metrics provide graded credit for semantically or superficially similar outputs, which is crucial for culturally nuanced tasks. In the Cultural Adaptation task, direct string matching is insufficient, as valid adaptations may use contextually congruent alternatives rather than a literal match. We therefore evaluate at two complementary levels using BERTScore: CSI_{bert} assesses the semantic quality of individual adapted terms (within <CSI> tags), while $\text{SENT}_{\text{bert}}$ measures overall meaning preservation of the entire adapted sentence. A summary of all metrics is provided in Table 2 with detailed definitions and computation provided in Appendix G.

Task	Metric	Description
Identification	**HI-CSI**	Exact string match between gold and predicted spans.
	SI-CSI	Normalised Levenshtein similarity with Hungarian alignment.
Prediction	**HP-CSI**	Case-insensitive exact match.
	SP-CSI	Cosine similarity over Sentence-BERT embeddings.
Adaptation	$\textbf{CSI}_{\textbf{bert}}$	BERTScore F1 over aligned CSI spans.
	$\textbf{SENT}_{\textbf{bert}}$	BERTScore F1 over full sentences.

Table 2: Evaluation metrics across the three tasks. Hard metrics (HI/HP) enforce strict correctness; soft metrics (SI/SP, $\text{CSI}/\text{SENT}_{\text{bert}}$) give graded credit. Formal definitions in Appendix G.

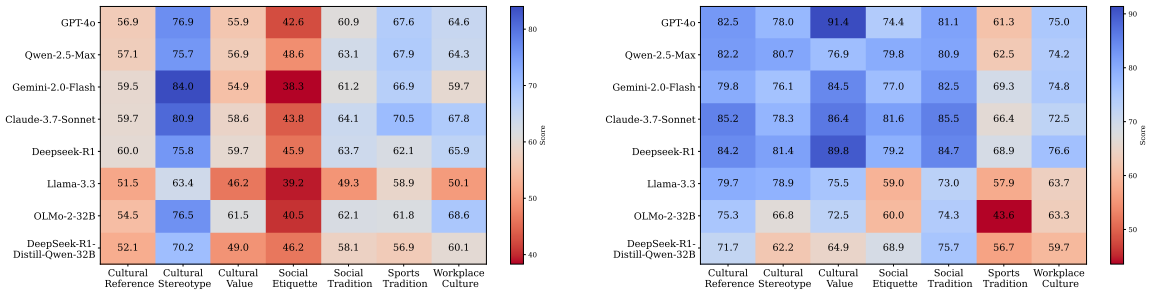
3.5 Models and Prompting Technique

We evaluate a diverse set of recent multilingual LLMs, including both proprietary and open-source models: GPT-4o, Qwen-2.5-Max, Gemini-2.0-Flash, Claude-3.7-Sonnet, DeepSeek-R1 (671B), Llama-3.3 (70B), OLMo-2-32B, and DeepSeek-R1-Distill-Qwen-32B. All models are evaluated across the full suite of tasks. For prompting, we adopt part of the Expert Prompting strategy proposed by Xu et al. (2023), which has been shown to elicit higher-quality responses than vanilla prompting.

4 Results and Analysis

4.1 CSI Identification and Prediction

Table 3 reports the results of the Identification and Prediction tasks. While the two tasks employ different evaluation metrics and are not directly comparable on absolute scores, the qualitative profile of model behaviour differs significantly. Prediction resembles a mask-filling task, which is a setting that is well-aligned with autoregressive pretraining over text containing Western CSIs. In contrast, CSI Identification requires explicit reasoning about which elements in a sentence are culturally specific and salient for cross-cultural communication, a substantially more demanding capability that all models struggle with. Qualitative inspection of model outputs across GPT-4o, Llama-3.3, Qwen-2.5-Max, and Gemini-2.0-Flash reveals a consistent failure mode: models tend to mark surface-level entities as CSIs while overlooking the culturally salient behaviour they participate in. For instance, given "He is the CEO, but everyone calls him by his <CSI>first name</CSI>



(a) CSI Identification (SI-CSI) (b) CSI Prediction (SP-CSI)

Figure 3: Performance of LLMs across CSI categories for CSI Identification and Prediction tasks. For clarity, only soft evaluation metrics are visualized. The results reveal consistent difficulties in CSI Identification across most CSI categories.

John", models correctly fill the masked span with *first name* in Prediction, yet mark *CEO* as the CSI in Identification, failing to recognise that addressing superiors by first name (rather than the title itself) is the culturally salient behaviour, common in Western workplaces but taboo in many others. The recurrence of this pattern across model families supports our claim that CSI Identification requires explicit cultural reasoning and judgment beyond surface lexical recall.

Model	CSI Identification		CSI Prediction	
	HI-CSI	SI-CSI	HP-CSI	SP-CSI
GPT-4o	50.94	57.74	43.70	78.47
Qwen-2.5-Max	<u>54.99</u>	59.75	43.70	79.30
Gemini-2.0-Flash	48.73	57.15	42.94	78.96
Claude-3.7-Sonnet	47.95	**60.47**	<u>46.76</u>	**81.76**
DeepSeek-R1	**59.05**	<u>60.01</u>	**47.33**	<u>81.68</u>
Llama-3.3	52.62	48.76	34.54	69.90
OLMo-2-32B	44.91	57.71	31.87	68.30
DeepSeek-R1-Distill-Qwen-32B	51.34	55.14	29.96	69.49

Table 3: CSI Identification and Prediction performance across models. Highest and second-highest scores are bolded and underlined, respectively. Values are percentages; higher is better.

Model performance is highly category-dependent, with systematic strengths and weaknesses that diverge across the two tasks. Figure 3 illustrates category-wise trends. In CSI Identification (Figure 3a), models perform best on *Cultural Stereotype* (mean $\approx$ 75) and worst on *Social Etiquette* (mean $\approx$ 43), with *Cultural Reference* and *Cultural Value* also posing consistent difficulty. In CSI Prediction (Figure 3b), the profile shifts: *Cultural Value* and *Cultural Reference* become the strongest categories, while *Sports Tradition* drops to the lowest. This divergence indicates that the categories LLMs find easiest to *predict* from context are not the same as those they reliably *identify* as culturally specific, suggesting that strong pretraining exposure to a category enables mask-fill recovery without conferring the explicit cultural-salience reasoning required for identification.

Statistical analysis. A Friedman test confirms significant performance differences across CSI categories for both Identification ($\chi^2(6) = 43.29$, $p < 0.001$) and Prediction ($\chi^2(6) = 37.55$, $p < 0.001$). Post-hoc Nemenyi tests (Table 12) reveal markedly different difficulty profiles: in Identification, *Cultural Stereotype* is the easiest and *Social Etiquette* the hardest, whereas in Prediction, *Sports Tradition* and *Workplace Culture* become the hardest while *Cultural Reference*, one of the hardest in Identification, becomes among the easiest. This reversal indicates that the two tasks tap qualitatively different capabilities.

Performance systematically declines from Visible to Invisible levels of Hall's cultural triad in both tasks. Figure 4 shows Visible-vs-Invisible decline for both tasks (full result in Table 10), and Table 11 reports per-category performance across Hall's levels. Models perform substantially worse on semi-visible and invisible elements such as *Appropriacy* and *Orientations*, which encode norms, values, and beliefs rather than observable practices. Together with the category-wise findings, these results indicate that LLMs handle surface-level cultural cues considerably more reliably than the deeper, norm- and value-driven aspects of culture that support genuine cross-cultural competence.

4.2 Cultural Adaptation

The CSI Adaptation task evaluates LLMs' ability to adapt culture-specific content in both intra-lingual and inter-lingual settings (Table 4). We focus on CSI_{bert}, as adaptation typically involves

Model	Arabic		Chinese		Bengali (WB)		Bengali (BD)	
	Intra	Inter	Intra	Inter	Intra	Inter	Intra	Inter
DeepSeek-R1	28.75	65.86	33.93	71.34	**37.23**	**48.44**	**35.40**	**47.41**
Gemini-2.0-Flash	27.78	67.09	32.74	70.33	26.44	44.20	25.94	43.13
GPT-4o	**35.21**	67.94	**38.85**	72.42	35.46	45.06	32.42	44.14
Llama-3.3	28.93	66.25	34.60	72.10	37.07	48.27	35.17	47.19
Qwen-2.5-Max	33.21	67.66	34.67	**72.74**	25.42	46.59	23.26	44.54
OLMo-2-32B	28.96	66.26	34.85	72.03	23.61	43.32	21.93	42.02
Claude-3.7-Sonnet	33.47	**68.69**	36.88	72.07	34.10	46.01	30.67	42.86
D-R1-D-Qwen-32B	23.96	59.49	34.48	69.73	28.25	36.54	26.13	35.46
Average	30.03	66.16	35.12	71.59	30.95	44.80	28.87	43.34

Table 4: CSI adaptation performance (%) across models measured using the CSI_{bert} metric, reported in both intra-lingual and inter-lingual settings across four target cultures. Sentence-level scores ($\text{SENT}_{\text{bert}}$) are reported in Table 14 in Appendix H. Abbreviations: Bangladesh (*BD*), West Bengal (*WB*).

localised substitutions that leave overall sentence structure largely unchanged, making $\text{SENT}_{\text{bert}}$ less discriminative. Across cultures, LLMs achieve highest adaptation performance for Chinese, followed by Arabic, with Bengali consistently lowest across models and CSI categories.

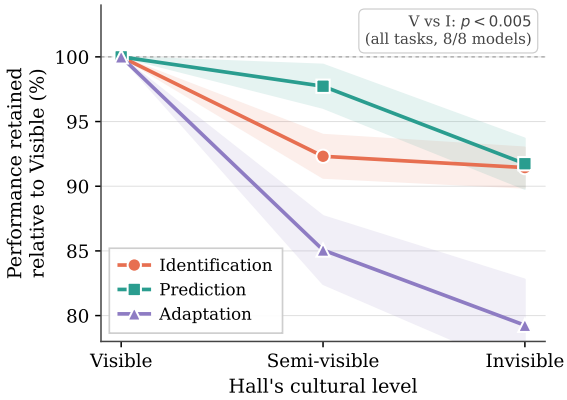


Figure 4: Within-task performance retention across Hall's cultural triad. For each model and task, Semi-visible and Invisible scores are normalised to its own Visible-level score (=100%) using the task's native metric (SI-CSI for Identification, SP-CSI for Prediction, CSI_{bert} for Adaptation); lines show mean across 8 LLMs with ± 1 SE. The figure compares relative degradation patterns, not absolute scores. Visible-vs-Invisible decline is significant in all tasks (one-sided Wilcoxon on raw scores, $p < 0.005$, 8/8 models).

LLMs consistently exhibit stronger CSI adaptation performance in inter-lingual settings than in intra-lingual settings. Across all four cultures, inter-lingual adaptation outperforms intra-lingual in 8/8 models (one-sided Wilcoxon signed-rank test, $p < 0.005$ for each culture), with mean CSI_{bert} gains of $+36.1$ (Arabic), $+36.5$ (Chinese), $+13.9$ (Bengali-WB), and $+14.5$ (Bengali-BD) percentage points. This pattern suggests that while

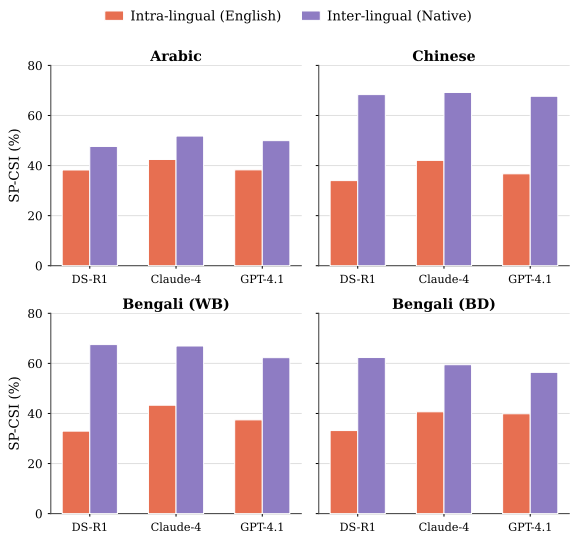


Figure 5: CSI Prediction performance (SP-CSI) under *Intra-lingual (English)*: English source; and *Inter-lingual (Native)*: native-language source. Models: DS-R1 (DeepSeek-R1), Claude-4 (Claude-4-Sonnet), GPT-4.1.

LLMs are relatively effective at adapting Western cultural content when operating directly to the target language, they encounter greater difficulty when adapting non-Western cultural concepts expressed in English. To better understand where these gains arise, we analyse performance differences at the CSI category level. Figures 10 and 11 show that inter-lingual adaptation yields the largest improvements in the *Cultural Stereotype* category and the smallest gains in *Cultural Reference*, suggesting that LLMs more readily adapt stereotypical or conventionalised cultural knowledge in the target language, while culturally grounded references tied to specific contexts or entities remain more challenging. A similar Visible-to-Invisible decline emerges through Hall's triad, though less pronounced in Bengali (Table 13).

Disentangling language and cultural context. To test whether inter-lingual gains reflect cultural reasoning rather than target-language fluency, we conduct a controlled ablation on a 5% sample for CSI Prediction across all four target cultures, evaluating DeepSeek-R1, GPT-4.1, and Claude-4-Sonnet.[4] As shown in Figure 5, native-language source inputs yield substantially stronger culturally grounded predictions in Chinese and both Bengali variants ($+20$ to $+30$ points on SP-CSI), with smaller gains for Arabic. We attribute this gap

[4]GPT-4o and Claude-3.7-Sonnet, used in our main experiments, were deprecated during the ablation period.

to a representational mismatch in intra-lingual prediction: when asked to insert culturally appropriate items into an English sentence, models often default to surface-level cultural markers such as transliterations or stereotypical borrowings (e.g., *ni hao*, *siheyuan*). Inter-lingual prediction, in contrast, lets models operate directly within the target language's cultural-pragmatic space, where culturally grounded expressions are more strongly represented in training data. A notable secondary observation arises in Bengali, which has no gendered third-person pronoun (only 'সে' for he/she): when the English source contains *he* or *she*, models often introduce a male-default term in the predicted CSI, but this gendered cue is lost in the inter-lingual setting where the source is already in Bengali.

Adaptation Procedure. During CSI adaptation, we prompt LLMs to indicate which of twelve adaptation strategies (Appendix B.3) is employed. Models frequently select *Couplet* combinations, with *Cultural Equivalent* the most common single strategy. Per-method effectiveness is reported in Table 15.

Regional Bias in Adaptation. Bengali constitutes a particularly informative case for evaluating regional and religious bias. Speakers are predominantly Muslim ($\sim$71%) with a Hindu minority ($\sim$28%), split between Bangladesh ($\sim$59% of speakers) and West Bengal, India ($\sim$38%) (Bangladesh Statistics Bureau, 2022), and the two populations differ in religion, dialect (*Bangal* vs. *Ghoti*), and culturally salient terminology.

Model	Intra-lingual ($\text{SENT}_{\text{bert}}$ / CSI_{bert})	Inter-lingual ($\text{SENT}_{\text{bert}}$ / CSI_{bert})
DeepSeek-R1	$0.003/0.005$	$-0.0002/-0.003$
Gemini-2.0-Flash	$0.002/0.003$	$0.001/0.012^{**}$
GPT-4o	$0.005/0.023^{*}$	$-0.0002/-0.001$
Llama-3.3	$0.003/0.006$	$-0.0002/-0.004$
OLMo-2-32B	$0.003/0.005$	$-0.0002/0.009$
Qwen-2.5-Max	$0.001/0.012$	$0.002/0.030^{**}$
Claude-3.7-Sonnet	$0.009/0.020^{*}$	$0.004/0.026^{*}$
DeepSeek-R1-Distill-Qwen-32B	$0.005/0.005$	$0.000/0.016$

Table 5: Mean regional performance differences in Bengali cultural adaptation (West Bengal − Bangladesh), aggregated across CSI categories. Statistical significance is evaluated using a one-sided Wilcoxon signed-rank test ($^{*}p < 0.05$, $^{**}p < 0.01$).

To assess whether such variation influences LLM-based cultural adaptation, we evaluate Bengali CSIs against both West Bengal and Bangladesh cultural corpora and compare adaptation performance across settings. As reported in Table 5, all statistically significant mean differences favour West Bengal over Bangladesh, indicating stronger model alignment with West Bengal–specific cultural knowledge, particularly for Claude and Gemini. While the effect sizes are not large, this reflects an inherent limitation of embedding-based metrics in low-resource settings: native speakers note that models systematically prefer lexical variants such as জল (Hindu-majority usage) over পানি (Muslim-majority usage), both denoting "water". Because these variants are semantically equivalent, Bangla BERT embeddings produce highly similar vectors, masking culturally meaningful preferences and compressing the observed effect size.

Native-speaker analysis confirms a consistent *Ghoti*-leaning pattern of LLMs across multiple domains: everyday vocabulary (ভোজ over খানা, both denoting 'feast'), kinship terms (দাদা over ভাই, both denoting 'brother'), festival names (পুজো over ঈদ when adapting Christmas). These choices systematically favour *Ghoti* lexical and cultural markers over *Bangal* alternatives, despite Bangladesh-Bengalis constituting the demographic majority within the broader Bengali-speaking population. And more surprisingly, this arises in a low-resource setting where overall Bengali performance is comparatively weaker than other languages (Table 4). Together with the quantitative results, these findings suggest that LLMs can encode subtle regional and ethno-religious biases within a single language, favouring culturally dominant or more frequently represented variants in their training data.

5 Conclusion

We introduce XCR-Bench, a human-annotated parallel corpus with three distinct reasoning tasks, designed to support the evaluation of cross-cultural reasoning in large language models (LLMs). By integrating Newmark's concept of CSIs with Hall's triad of culture, XCR-Bench enables fine-grained analysis for subtle cultural elements. These annotations further enable mechanistic analysis of how cultural information is encoded within model internals. Our evaluation shows that contemporary LLMs struggle to identify and adapt CSIs in cross-cultural contexts. We also observe that cultural adaptation responses often encode regional and ethno-religious biases within a single linguistic setting. Our work provides high-quality data and empirical insights to advance the study of cross-cultural reasoning in NLP.

Limitations

While our study introduces a novel benchmark for evaluating LLMs' cross-cultural adaptability, it is subject to several limitations.

Limited Coverage of Newmark Categories. Newmark's framework (Newmark, 1988) outlines five categories of CSIs, as detailed in Appendix B.1. As shown in Table 6, the CSIs in XCR-BENCH primarily span categories 2 and 3. This focus stems from our emphasis on the semi-visible and invisible levels of Hall's Iceberg model of culture, which encompass social values, beliefs, and norms. Consequently, our benchmark does not cover categories 1, 4, and 5 (e.g., ecology, material culture, and gestures), representing a gap that future work could address by expanding the CSI typology.

Limited Coverage of Cultures. Our corpus includes CSI annotations for only four distinct cultures (Western US/UK, Arabic, Bengali, and Chinese). While this selection enables targeted insights, incorporating additional languages and cultures would likely reveal broader patterns in LLMs' adaptation capabilities. This constraint was influenced by our available resources for collaboration and annotation.

Limited Coverage of Prompting Techniques. In our main experiments, we employed the expert prompting strategy from Xu et al. (2023). To assess whether our findings are sensitive to prompt formulation, we conducted a preliminary ablation on the CSI Prediction task with DeepSeek-R1, Claude-4-Sonnet, and GPT-4.1 under three prompt variants: expert prompting, chain-of-thought reasoning, and few-shot prompting with three examples (Appendix H, Figure 9). No prompting strategy consistently outperforms the others across models: few-shot prompting helps Claude-4-Sonnet, chain-of-thought marginally helps DeepSeek-R1 on the hard metric, while expert prompting remains strongest for GPT-4.1, and the qualitative patterns reported in our main results (e.g., the Visible-to-Invisible decline and category-wise weaknesses) hold across all three prompt formulations. This suggests that the cross-cultural reasoning gaps we observe are unlikely to be improved by prompt engineering alone, and instead reflect deeper limitations in how culturally grounded knowledge is represented during pretraining. A systematic study across all tasks, models, and a broader set of prompting strategies remains an important direction for future work, particularly because it may reveal how much of the cross-cultural reasoning gap can be closed through prompt-level interventions versus what requires model-intrinsic improvements such as targeted cultural pretraining or alignment.

Annotator-Bound Cultural Representation. While our annotators are native speakers with strong cultural familiarity, the adapted sentences inevitably reflect their specific cultural backgrounds and may not capture the full diversity within each target culture. This is particularly relevant for Arabic, which spans more than twenty countries with substantial inter-country variation in dialect, religious practice, and social custom, as well as intra-country diversity within individual nations (Naous et al., 2024). Our two Arabic annotators are based in Saudi Arabia, and their adaptations may therefore better reflect Gulf or Saudi-specific cultural norms than, for instance, North African or Levantine ones. Similar considerations apply to our Chinese and Bengali annotations, though we partially address regional variation in Bengali through separate West Bengal and Bangladeshi annotations. Broader coverage of intra-cultural diversity-through larger and more geographically distributed annotator pools—remains an important direction for future work.

Western-Centric Source Framing. XCR-BENCH is constructed around Western (US/UK) source CSIs adapted into non-Western contexts, and does not currently evaluate non-Western-to-Western or non-Western-to-non-Western adaptation. This design is deliberate: it tests LLMs' Western cultural knowledge as a baseline while simultaneously evaluating their ability to recontextualise that knowledge into divergent cultural settings—a more challenging and practically consequential capability than knowledge recall alone, motivated by recent evidence that Western training-data bias does not guarantee strong Western cultural competence (Saha et al., 2025). Our source resources (CANDLE and Cultural Atlas) further support this choice, offering well-structured Western cultural references with explicit cross-cultural mappings suitable for controlled evaluation. Nonetheless, this framing restricts cultural coverage, and findings may not fully generalise to settings where source items originate outside Western contexts. Extending the benchmark to non-Western-to-non-Western and non-Western-to-Western transfer is an important direction for future work.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Mohammad Atari, Mona Xue, Peter Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Bangladesh Statistics Bureau. 2022. Preliminary report on population and housing census 2022. Technical report, Bangladesh Statistics Bureau. English Version.

John W Berry. 1974. Psychological aspects of cultural pluralism: Unity and identity reconsidered. *Topics in culture learning*, 2:17–22.

Sjoerd Beugelsdijk and Chris Welzel. 2018. Dimensions and dynamics of national culture: Synthesizing hofstede with inglehart. *Journal of cross-cultural psychology*, 49(10):1469–1505.

Yong Cao, Yova Kementchedjhieva, 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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-trained language models represent some geographic populations better than others. *arXiv preprint arXiv:2403.11025*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shiwei Guo, Sihang Jiang, Qianxi He, Yanghua Xiao, Jiaqing Liang, Bi Yude, Minggui He, Shimin Tao, and Li Zhang. 2025. Do large language models truly understand cross-cultural differences? *arXiv preprint arXiv:2512.07075*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*.

Edmund T Hall and Edward T Hall. 1966. *The hidden dimension*, volume 609. Anchor.

Edward T Hall. 1973. *The silent language*. Anchor.

Edward T Hall. 1976. Beyond culture. *Garden City*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures et organisations: Nos programmations mentales*. Pearson Education France.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025a. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Mohsinul Kabir, Tasnim Ahmed, Md Mezbaur Rahman, Polydoros Giannouris, and Sophia Ananiadou. 2025b. Semantic label drift in cross-cultural translation. *arXiv preprint arXiv:2510.25967*.

Mohsinul Kabir, Tasfia Tahsin, and Sophia Ananiadou. 2025c. From n-gram to attention: How model architectures learn and propagate bias in language modeling. *arXiv preprint arXiv:2505.12381*.

Yova Kementchedjhieva, Di Lu, and Joel Tetreault. 2020. The apposcorpus: A new multilingual, multi-domain dataset for factual appositive generation. *arXiv preprint arXiv:2011.03287*.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When tom eats kimchi: Evaluating cultural bias of multimodal large language models in cultural mixture contexts. *arXiv preprint arXiv:2503.16826*.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb' s role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

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.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Peter Newmark. 1988. *A textbook of translation*, volume 66. Prentice hall New York.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A framework for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Abhinav Sukumar 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.

Paulius Rauba, Nabeel Seedat, Max Ruiz Luyten, and Mihaela van der Schaar. 2024. Context-aware testing: A new paradigm for model testing with large language models. *Advances in Neural Information Processing Systems*, 37:112505–112553.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can llms identify cross-cultural communication gaps? 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 8043–8067.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

C. Schäffner. 1998. Skopos theory. In *Routledge encyclopedia of translation studies*, pages 235–238.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Llms for intralingual cultural adaptation. *arXiv preprint arXiv:2406.14504*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Yuchen Song, Andong Chen, Wenxin Zhu, Kehai Chen, Xuefeng Bai, Muyun Yang, and Tiejun Zhao. 2025. Culture in a frame: C ^3 b as a comic-based benchmark for multimodal culturally awareness. *arXiv preprint arXiv:2510.00041*.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. 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.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen Meng. 2025. Socialcc: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

A Background Study

A.1 Measuring 'Culture' in LLMs

Culture is a multifaceted concept encompassing cultural heritage (art, music, food), interpersonal interaction norms, and the collective ways of life that distinguish groups from one another (Adilazuarda et al., 2024). Sociologically, it has been described as the abstract patterns of ideas and principles that guide human behavior in practically effective ways. Recent studies have predominantly relied on global surveys to assess the cultural knowledge of large language models (LLMs) regarding specific countries. In particular, multiple-choice questions drawn from surveys such as the World Values Survey (WVS) (Haerpfer et al., 2020), Hofstede's Cultural Dimensions (Hofstede et al., 2010), or PEW Global Surveys [5] are presented to LLMs to evaluate their understanding of particular cultures or countries (Zhao et al., 2024; Durmus et al., 2023; Kharchenko et al., 2024). Wang et al. (2024) uses Hofstede's Survey to construct a benchmark by forming natural scenario-based questions from the questionnaire itself using GPT-4. While these surveys provide an excellent means of gathering real-world human perspectives, their use for evaluating cultural knowledge in LLMs has several limitations: (i) LLMs are restricted to the predefined choices in multiple-choice formats, yet culture is a nuanced concept that often requires more sophisticated and open-ended responses (Beugelsdijk and Welzel, 2018); (ii) altering the order of multiple-choice options can lead LLMs to produce entirely different answers, which undermines reliability for a multi-faceted notion like culture (Pezeshkpour and Hruschka, 2024). In response, several studies have critiqued this approach and advocated for more open-ended evaluations to better capture the sophistication and intricacy of cultural understanding (Kabir et al., 2025a; Röttger et al., 2024). Other works have proposed specialized frameworks targeting specific cultural facets or proxies, such as food, etiquette, or regional differences, rather than attempting a holistic evaluation of culture (Naous et al., 2024; Adilazuarda et al., 2024).

[5] https://www.pewresearch.org/

A.2 Evaluation of Cross-Cultural Competence of LLMs

The ability of large language models (LLMs) to adapt meanings across cultural contexts is essential for broadening access to AI systems and mitigating harms from culturally maladapted outputs (Kabir et al., 2025b). To evaluate LLMs' proficiency in cross-cultural communication, recent studies have drawn on the concept of Culture-Specific Items (CSIs), as defined by Newmark (Newmark, 1988), emphasizing that identifying and adapting these items is critical for effective cross-cultural exchange.

Several works have explored CSI adaptation in specific domains. For example, Singh et al. (2024) adapted American CSIs to Indian equivalents using the Friends Dialogues corpus [6]. Similarly, Yao et al. (2024) introduced CAMT, a parallel corpus across six languages featuring paired CSIs, to assess LLMs' cultural adaptability. In a domain-specific approach, Cao et al. (2024) proposed tasks for translating and culturally adapting recipes between Chinese- and English-speaking cuisines. However, these studies primarily rely on machine translation evaluations and often lack explicit CSI annotations, limiting their ability to probe models' targeted capabilities in CSI span identification and sentence-level adaptation. More recent efforts have expanded the scope of cross-cultural evaluations. Guo et al. (2025) proposed a methodology using Cross-Cultural Core Concept Sets (CCCs) to assess LLMs' cross-cultural understanding in real-world scenarios, enabling more authentic capability assessments. Wu et al. (2025) introduced SocialCC, a benchmark for evaluating cultural competence through 3,060 multi-turn interactive intercultural scenarios spanning 60 countries. Incorporating multimodality, Song et al. (2025) developed C^3B, a benchmark with 2,000 images and over 18,000 QA pairs across three tasks of varying difficulty: basic visual recognition, higher-level cultural conflict understanding, and cultural content generation. In a related vein, Kim et al. (2025) created a benchmark to evaluate biases in multimodal LLMs across high- and low-resource cultural settings.

Our work complements these efforts by providing a culturally rich corpus of realistic scenarios, with explicit annotations for CSI categories, Hall's cultural levels, and cross-cultural CSI pairs, thereby

[6] https://convokit.cornell.edu/documentation/friends.html

advancing research in cross-cultural NLP.

B Utilizing Newmark's Theory of CSIs

B.1 Newmark's Definition of Culture and CSI Categories

Newmark (Newmark, 1988) defines culture as a shared way of life, encompassing the distinctive characteristics, beliefs, and practices of a particular social group. As outlined in the cultural proxies in Appendix A, these differences span multiple dimensions- from tangible elements like food and materials to abstract ones such as emotions and values (Thompson et al., 2020; Adilazuarda et al., 2024). Newmark (Newmark, 1988), in the 9th Chapter of his book *A textbook of Translation*, terms such concepts Culture-Specific Items (CSIs), whose identification and handling are vital for effective cross-cultural communication. He proposes five key categories of CSIs that are especially critical in cross-cultural translation:

Category 1: Ecology covers the relationship between a community and its natural environment, including indigenous flora, fauna, geography, and climate. *Example:* Translating the specific name for a seasonal desert wind, such as the *Simoom* in Arabic cultures, or the *Khamsin* in Egypt.

Category 2: Material Culture (Artefacts) includes all tangible, human-made objects that shape or reflect a society's way of living, such as tools, clothing, architecture, and food. *Example:* Translating terms like the *Shamagh* (the traditional checked men's headdress in the Gulf region) or *Iftar* (the meal to break the Ramadan fast).

Category 3: Social Culture refers to the patterns of social activity, including work, leisure, and communal practices that define daily and ceremonial life. *Example:* Translating the name of a traditional social gathering like the *Majlis* (a council or sitting room for community discussion in Arab cultures) or *Jiāzǐ* (a 60th birthday celebration in Chinese culture).

Category 4: Organizations, Customs, and Ideas encompasses the abstract and institutional aspects of culture, including its political, legal, religious, artistic, and customary systems. *Example:* Translating concepts such as *Apritismo* (a Peruvian political ideology) or *Barangay* (the smallest administrative division in the Philippines, rooted in pre-colonial history).

Category 5: Gestures and Habits involves non-verbal communication, body language, and routine behaviors whose meanings are culturally encoded and may not have direct equivalents. *Example:* Interpreting gestures like the *chin flick* (a sign of dismissal or negation in parts of Italy and France) or the *"Moutza"* (an offensive hand gesture in Greece).

B.2 CSI Categories Mapping

B.3 Adaptation Procedure

Newmark (1988) proposes a set of strategies for translating culture-specific terms, conditioned on contextual and audience-related factors. In this work, we reinterpret these strategies as *cultural adaptation methods* and explicitly prompt models to indicate which method is employed during adaptation. This allows us to analyze which adaptation strategies are most effective for handling CSIs across cultural contexts.

1. **Transference.** The original CSI is retained unchanged in the adapted output (e.g., *sari*, *kimono*).
2. **Cultural Equivalent.** A culturally analogous term from the target culture is substituted to convey a similar function or social meaning (e.g., adapting *Thanksgiving* as a local harvest festival).
3. **Neutralization.** The CSI is replaced with a description that explains its function or meaning in culturally neutral terms (e.g., rendering *kimono* as "a traditional Japanese robe").
4. **Literal Translation.** The CSI is translated directly on a word-by-word basis into the target language (e.g., translating *Bundestag* as "Federal Parliament").
5. **Labeling.** The CSI is retained or adapted and accompanied by a brief explanatory label that clarifies its cultural role or category.
6. **Naturalization.** The CSI is adapted to conform to the spelling or pronunciation conventions of the target language (e.g., *Pharisees*).
7. **Componential Analysis.** The CSI is decomposed into its constituent semantic components,

Newmark CSI Category	Category from Liu et al. (2025)	XCR-Bench Category	CSI Example
Material Culture			
Social Culture	Social.Relationship (e.g., Family, Fictive, etc.)	Social Etiquette	He <CSI> thanked </CSI> the stranger who held the elevator door open for him.
	Social.Context (e.g., Situational, Historical, etc.)	Sports Tradition	The family attends <CSI> Little League games </CSI> every Saturday during baseball season.
	Social.Communicative Goals (e.g., Greeting, Requesting, etc.)	Workplace Culture	At networking events, attendees often exchange <CSI> business cards </CSI> to facilitate future connections.
	Social.Demographics (e.g., Income, Education, etc.)		
Organisations, Customs and Ideas	Ideational.Concepts (e.g., Food, Holidays, etc.)	Social Tradition	They're hosting a <CSI> Hogmanay </CSI> celebration to welcome the New Year.
	Ideational.Knowledge (e.g., Factual, Common Sense, etc.)	Cultural Reference	He learned to use <CSI> Fahrenheit </CSI> instead of Celsius for temperature discussions.
	Ideational.Values (e.g., Bias, Hate, etc.)	Cultural Value	She donated to a group working to protect <CSI> wild horses </CSI>, seeing them as symbols of America's untamed spirit.
	Ideational.Norms (e.g., Descriptive, Normative, etc.)	Cultural Stereotype	They're <CSI> helicopter parents </CSI> who monitor their child's every move and decision.
	Ideational.artefacts (Art, Meme, Architecture, etc.)		Walking through the small town, she admired the charming <CSI> wooden houses </CSI> that lined the streets.
Ecology			
Gestures and Habits			

Table 6: Mapping between Newmark CSI categories, the taxonomy of Liu et al. (2025), and XCR-Bench categories, with illustrative CSI-tagged examples. Empty cells indicate mappings not specified in the current dataset.

each of which is explicitly explained (e.g., rendering *dacha* as "a summer house for wealthy people").

8. **Deletion.** The CSI is omitted entirely when it is non-essential to the meaning or when adaptation would introduce unnecessary complexity.
9. **Couplet.** Two adaptation strategies are combined, such as retaining the original CSI while also providing an explanatory description.
10. **Accepted Standard Translation.** A widely recognized and conventionally accepted translation is used (e.g., *Holy See* for *Saint-Siège*).
11. **Paraphrase or Gloss.** A longer explanatory paraphrase or gloss is added, either inline or as a footnote, to clarify the CSI's meaning and cultural context.
12. **Classifier.** A general category term is added to situate the CSI within a familiar conceptual class (e.g., "the Japanese game of Go").

C Hall's Triad of Culture

In Beyond Culture (1976), Edward T. Hall argued that culture operates largely through unconscious patterns that structure meaning, perception, time, space, and interaction (Hall, 1976). His distinction between observable practices and the underlying

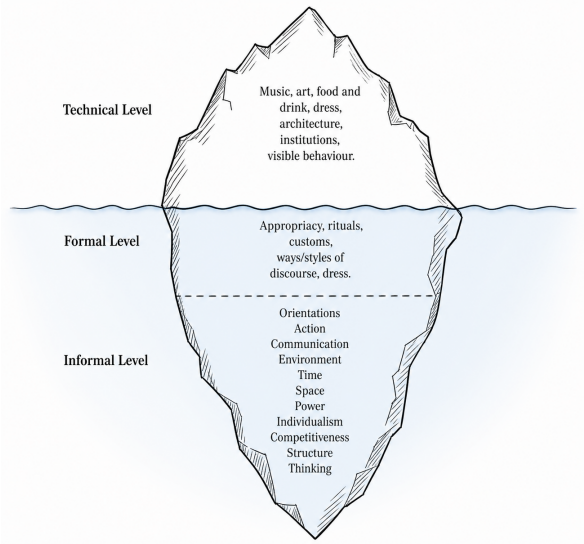


Figure 6: A depiction of Hall's Triad of culture, illustrating three hierarchical levels: Technical (visible artefacts and behaviours), Formal (social norms, rituals, and conventions), and Informal (underlying orientations and value systems). The waterline intersects the Formal level, indicating the boundary between observable and implicit cultural structures.

value systems that shape them provides the conceptual basis from which later intercultural scholars derived the now-popular "iceberg" or "triad" metaphor of culture. The metaphor draws on a striking property of icebergs: only about 10% of an iceberg is visible above the waterline at any given time, while the remaining 90% lies hidden beneath the surface. By analogy, the visible aspects of culture —language, dress, food, festivals, art —represent only a small fraction of what constitutes a cultural system, with the larger portion comprising values, beliefs, norms, and assumptions that operate beneath conscious awareness. The conceptual foundations of this metaphor can be traced to two of Hall's earlier works, The Silent Language (Hall, 1973) and The Hidden Dimension (Hall and Hall, 1966), which introduced related constructs such as high- and low-context communication, proxemics, and polychronic versus monochronic time orientations. Together, these works established a framework for understanding culture as a multi-layered system in which surface behaviours are governed by deeper, often invisible, structures of meaning. Despite its widespread influence, the iceberg analogy has notable limitations. As depicted in Figure 6, an iceberg is a static and fixed object, whereas culture is inherently dynamic, evolving across generations and reshaped through contact between communities. The analogy is therefore most useful for illustrating the layered structure of culture: its visible and invisible components, rather than its characteristic processes, such as transmission, contestation, and change. In this work, we adopt the iceberg metaphor in this restricted sense: as a heuristic for organising cultural elements by depth of visibility, while recognising that the underlying cultural system it represents is fluid rather than fixed.

D Necessity of Integrating Newmark's CSI taxonomy with Hall's Triad

Newmark's CSI taxonomy and Hall's cultural triad capture orthogonal dimensions of culture: CSI identifies what cultural element is present, while Hall's framework characterizes how deeply its meaning is embedded (visible vs. implicit norms). CSI categories alone do not distinguish surface artefacts from behavioral conventions or underlying beliefs. Hall's stratification enables analysis of whether LLM failures stem from (1) lexical gaps (visible CSIs), (2) pragmatic misunderstanding, or (3) reasoning deficits (semi-visible/invisible CSIs). Thus, Hall's triad provides independent explanatory value by assessing cultural reasoning depth rather than category coverage alone. Prior work, such as Singh et al. (2024) recognizes the need for Hall's framework in LLM cultural analysis but lacks the granular annotation required for systematic application. Our corpus and benchmark address this gap by operationalizing Hall's depth model within CSI-based evaluation.

E Annotation and Annotator Details

Annotation Training. The annotation process was preceded by structured training sessions conducted by the authors and the principal investigator (PI). For each language, we organized two online meetings (approximately one hour each) with both annotators and expert annotators. These sessions focused on clarifying the annotation objectives, discussing the notion of Culture-Specific Items (CSIs), and walking through the detailed annotation guidelines provided in Appendix E.1. Annotators were encouraged to ask questions and resolve ambiguities before beginning the annotation process.

Annotator Qualifications. Annotators were recruited through established academic collaborators in Bengali, Chinese, and Arabic. For each

language, the annotation team consisted of two annotators and one expert annotator. Expert annotators hold a PhD in Computational Linguistics or a closely related field and have substantial prior experience in corpus construction and linguistic annotation. The two annotators are either undergraduate or PhD students with prior experience in linguistic or NLP-related annotation tasks.

Specifically, the Arabic annotators are undergraduate students based in Saudi Arabia, the Chinese annotators are PhD students based in the UK, and the Bengali annotators are PhD students based in the UK or the USA, all working in NLP or closely related fields. The two annotators independently produced both intra-lingual and inter-lingual adaptations of the base sentences. In cases of disagreement, the expert annotator adjudicated and selected the most appropriate adaptation.

Compensation. Annotators are compensated for their work in accordance with the minimum wage standards of the countries in which they reside.

E.1 General Annotation Guidelines for Cross-Cultural CSI Adaptation

You will work with a dataset of English sentences in which culture-specific items (CSIs) which are predominantly drawn from Western (US/UK) cultural contexts. The CSIs are explicitly marked using <CSI> tags in the dataset samples. Your task is to adapt these CSI-marked terms into culturally appropriate counterparts for **Arabic/Chinese/Bengali** cultures. For each instance, you will produce: (i) an *intra-lingual* adaptation in English that reflects the target cultural context, and (ii) an *inter-lingual* adaptation in your own language.

Please ensure that adaptations maintain the original sentence's communicative purpose while being culturally appropriate and natural in the target context.

To maintain consistency across adaptations, kindly follow the guidelines provided below.

1. Purpose

The goal of annotation is to ensure that CSI adaptations:

- Reflect authentic cultural values, norms, and practices of the target culture.
- Use natural and idiomatic language in both intra- and inter-lingual adaptations.
- Respect cultural, religious, and social sensitivities, including regional variation where applicable.

2. Adaptation Rules

Following Newmark's cultural equivalence framework (Newmark, 1988), annotate each CSI using one of the following strategies. [For illustration, adaptation examples are given here in Arabic.]

A. Direct Equivalent Rule: Replace the CSI with a culturally identical or near-identical equivalent that exists in the target culture.

Example: Original: <CSI>pub gathering</CSI> Intra-lingual (Arabic culture): qahwa gathering Inter-lingual (Arabic): جلسة قهوة

Annotation tip: Ensure regional appropriateness (e.g., differences across Arab regions or between Bengali varieties).

B. Functionally Similar Equivalent Rule: When no direct equivalent exists, replace the CSI with a culturally different but functionally similar concept.

Example: Original: <CSI>trick-or-treating</CSI> Intra-lingual (Arabic culture): children receiving Eid gifts Inter-lingual (Arabic): عيدية الأطفال في العيد

C. Neutral Term Rule: If no clear cultural equivalent exists, use a neutral expression that fits social norms in the target culture without introducing foreign practices.

Example: Original: <CSI>friendly banter</CSI> Intra-lingual (Arabic culture): light-hearted teasing Inter-lingual (Arabic): مزاحًا خفيفًا

D. Non-transferable Rule: Label the CSI as *Non-transferable* if adapting it would conflict with cultural, religious, or social norms in the target culture.

Example: Original: They met through a <CSI>dating app</CSI>. Adaptation: Non-transferable

3. Cultural Sensitivity

While doing the annotation, please ensure to:

- Avoid culturally inappropriate or taboo content (e.g., alcohol, gambling, or sensitive relationships where applicable).
- Respect religious norms and social conventions.

4. Annotation Checklist

For each sentence, please verify:

- **Cultural accuracy:** Is the CSI adapted appropriately for the target culture?
- **Linguistic quality:** Is the adaptation natural and idiomatic?
- **Functional equivalence:** Does the adaptation preserve the original communicative intent?
- **Consistency:** Do the intra- and inter-lingual adaptations align?

5. Final Notes

- Consult the expert annotator for ambiguous or borderline cases.
- Flag creative or uncertain adaptations for further review.
- Use CANDLE and Cultural Atlas as reference sources for relevant cultural practices, rituals, and traditions.

F Properties of XCR-BENCH Corpus

F.1 Data Acquisition

CANDLE and Cultural Atlas are used as the primary knowledge bases for data acquisition to construct the XCR-BENCH base corpus. CANDLE provides high-quality cultural commonsense knowledge (CCSK), comprising 1.1M assertions organized into 60K coherent clusters, while Cultural Atlas offers curated descriptions of multicultural norms derived from extensive global community interviews and rigorous validation (Rao et al., 2024). Both resources present cultural knowledge in factual form (e.g., CANDLE states a Representative cultural concept as *"In the USA, people wear their wedding rings on the ring finger of their left hand"*), with Cultural Atlas further organizing information into thematic sections such as Core Concepts, Communication, and Business Culture.

F.2 Dataset Properties

As detailed in Section 2.3, XCR-BENCH is the first dataset to provide comprehensive annotations aligned with Hall's triad of culture, with particular emphasis on the *Semi-visible* and *Invisible* layers that encompass beliefs, values, and social norms. Figure 7 illustrates the mapping between CSI categories and Hall's cultural levels, highlighting a predominant focus on these *Semi-visible* and *Invisible* layers.

Table 7 shows an example sample from the XCR-BENCH corpus. Each sample contains its fine-grained annotation for CSI Category, CSI-Hall Mapping, Hall's Level and parallel inter-lingual and intra-lingual cross-cultural CSI pairs in four distinct cultures.

We present the mapping of Hall's cultural elements with our CSI categories in Figure 8. This shows how the Hall's cultural elements are mapped across our 7 distinct categories of CSIs. Some CSI categories and cultural elements are predominant in number because of their vast availability in the source where we extracted our CSIs from, namely CANDLE and Cultural Atlas.

We explicitly show the cultural element mapping with Hall's triads in Table 8 with instances from our XCR-BENCH corpus. We derive this mapping from Hall (1976). Notably, the Hall cultural level of CSIs is independent of Newmark's categorization and based on their context.

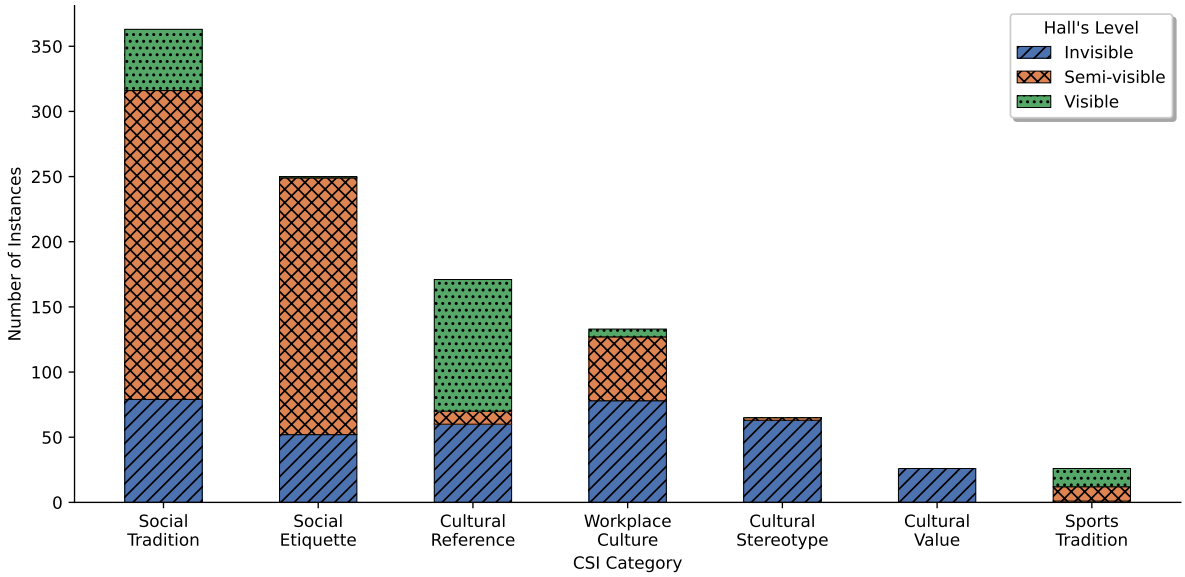


Figure 7: Mapping between CSI categories and Hall's cultural levels. The textured bar plots indicate a higher concentration of semi-visible and invisible cultural elements across all CSI categories.

Field	**Content**
CSI	Halloween
Cultural Context	Halloween is widely celebrated in the West with decorations, costumes, and trick-or-treating.
Sentence	We're decorating the house for <CSI> Halloween </CSI> this weekend.
CSI Category	Social Tradition
CSI-Hall Mapping	Rituals
Hall's Level	Semi-visible
Chinese	
Intra-lingual	We're decorating the house for the <CSI> Lantern Festival </CSI> this weekend with lanterns and festive lights.
Inter-lingual	这个周末我们正在为 <CSI> 元宵节 </CSI> 装饰房子，挂上彩灯，营造节日的夜晚氛围。
Arabic	
Intra-lingual	We're decorating the house for <CSI> Eid celebration </CSI> this weekend.
Inter-lingual	نقوم بتزيين المنزل لـ <CSI> احتفال العيد <CSI>.هذا الأسبوع
Bengali (West Bengal)	
Intra-lingual	We're decorating the house for <CSI> Durga Puja </CSI> this weekend.
Inter-lingual	আমরা এই সপ্তাহান্তে বাড়ি <CSI>দুর্গা পূজার</CSI> জন্য সাজাচ্ছি।
Bengali (Bangladesh)	
Intra-lingual	We're decorating the house for <CSI> Eid-ul-Fitr </CSI> this weekend.
Inter-lingual	আমরা এই সপ্তাহান্তে বাড়ি <CSI>ঈদুল ফিতরের</CSI> জন্য সাজাচ্ছি।

Table 7: An example sample from the XCR-Bench corpus along with the parallel Inter-lingual and Intra-lingual annotations in other cultures.

Hall's Level	Cultural Element	Instance from XCR-BENCH
Invisible	Action	He's an <CSI> adrenaline junkie </CSI> who loves skydiving and bungee jumping.
	Communication	He steered conversation away from <CSI> gun laws </CSI> to maintain pleasant dinner atmosphere.
	Competitiveness	He' s a <CSI> self-made </CSI> man who built his empire from scratch.
	Environment	She loves going to the <CSI> pub </CSI> for a pint after work.
	Individualism	She demonstrates <CSI> self-reliance </CSI> by managing projects independently.
	Power	The board member <CSI> endorses a bottom-up approach </CSI>, empowering employees to contribute to decision-making.
	Stereotype	He's the <CSI> all-American boy next door </CSI> with his football scholarship and volunteer work.
	Structure	Couples are <CSI> delaying parenthood </CSI> to establish <CSI> financial stability </CSI> first.
	Thinking	The building omitted floor <CSI> 13 </CSI> jumping from 12 to 14 due to <CSI> superstition </CSI>.
	Time	He made sure to arrive 10 minutes early for the meeting because <CSI> punctuality </CSI> is important in American culture.
Semi-visible	Appropriacy	Don't ask about <CSI> someone's weight or appearance </CSI> in personal conversations.
	Customs	She is used to <CSI> queue </CSI> politely and wait her turn, even in crowded places.
	Rituals	She always says <CSI> grace </CSI> before dinner, a tradition passed down through generations.
	Styles of Discourse	She said <CSI> 'cheers' </CSI> when the stranger held the door, a casual British thanks.
	Behavior	He crossed his <CSI> index and middle fingers </CSI> for <CSI> good luck </CSI> before the exam.
	Ways of Dress	He wore a tailored <CSI> suit </CSI> to the job interview but dressed in casual <CSI> jeans </CSI> and a t-shirt for the barbecue.
	Workplace Customs	We're implementing an <CSI> open-door policy </CSI> in management for better accessibility and transparency.
	Workplace Norms	The manager gave <CSI> constructive criticism </CSI> during the performance review.
Visible	Architecture	Their dining table, made of <CSI> White Oak </CSI>, has been in the family for generations, a testament to its durability.
	Art	When she visited the UK, she noticed that every British coin had a portrait of the reigning <CSI> monarch </CSI>, a tradition that felt like holding a piece of history.
	Dress	Everyone wore their team <CSI> merchandise </CSI> on game day to show support.
	Food and Drink	They enjoyed a <CSI> fish and chips supper </CSI> by the seaside, a classic British experience.
	Institutions	The startup has an <CSI> open office layout </CSI> without walls or dividers.
	Music	Listening to a British <CSI> brass band </CSI> at the local fair, she was struck by the rich, traditional sound.

Table 8: Mapping of cultural elements to Hall's Triad of Culture with instances from our constructed XCR-BENCH Corpus. The mapping is derived from Hall (1976).

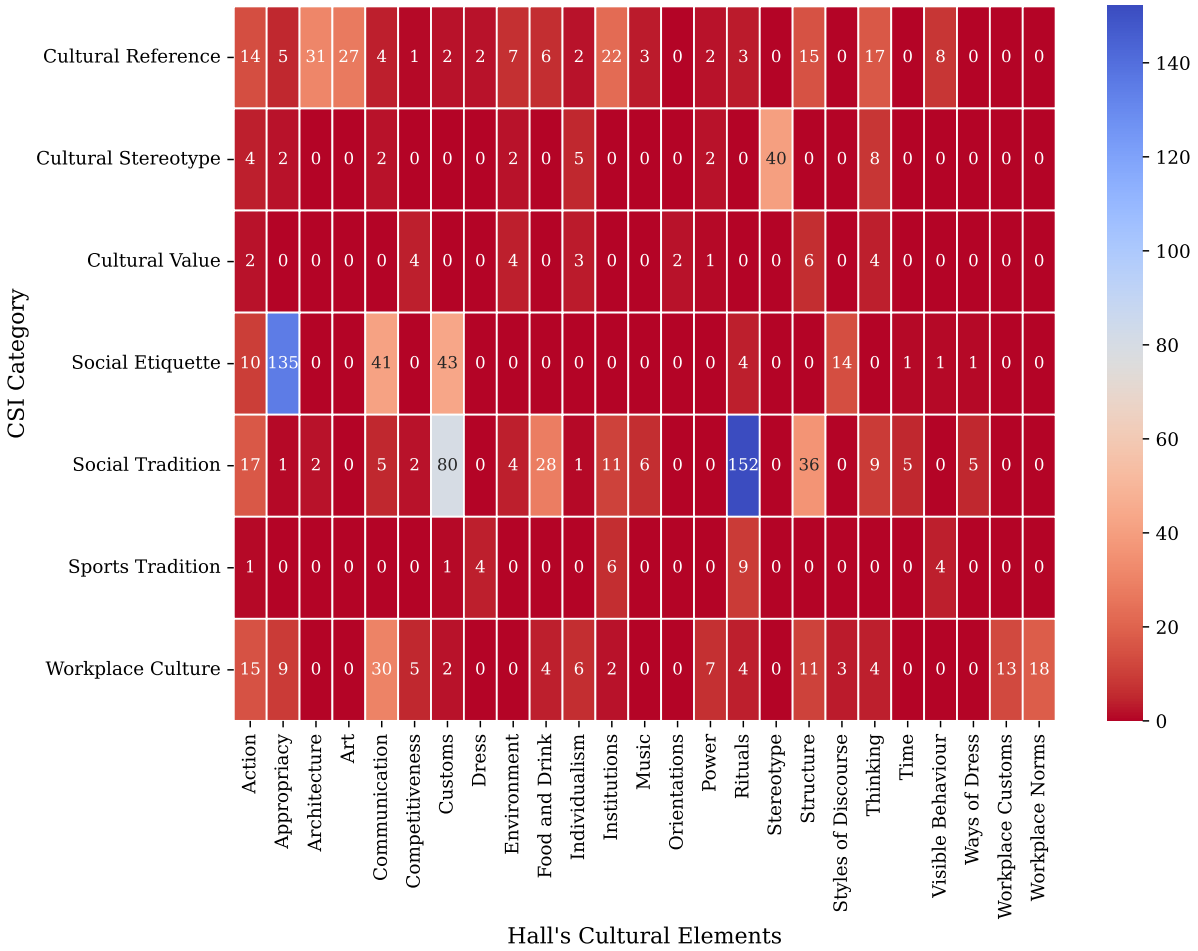


Figure 8: Mapping between CSI categories and Hall's cultural elements in the XCR-Bench Corpus.

G Details on Evaluation Metrics

A summary of the evaluation metrics used in this study is presented in Table 9.

Task	Metric	Description	Formula/Details
CSI Identification	Hard Identification of CSI (**HI-CSI**)	Strict matching using exact string equality.	$\text{sim}_{\text{exact}}(g, m) = \mathbb{I}[g = m]$, where g and m denote gold and model-predicted CSI spans.
	Soft Identification of CSI (**SI-CSI**)	Accounts for minor variations via normalized Levenshtein similarity, with optimal alignment using the Hungarian algorithm.	$s(g, m) = 1 - \frac{d_{\text{Lev}}(g,m)}{\max(\lvert g\rvert,\lvert m\rvert)}$; sentence-level soft F_1 based on alignment.
CSI Prediction	Hard Prediction of CSI (**HP-CSI**)	Case-insensitive exact match between gold and predicted spans. Span refers to text inside the enclosed <CSI> and </CSI> tags.	$\text{sim}_{\text{exact}}(g, p) = \mathbb{I}[\text{lower}(g) = \text{lower}(p)]$, where g and p denote gold and predicted spans.
	Soft Prediction of CSI (**SP-CSI**)	Cosine similarity between Sentence-BERT embeddings for paraphrases/ synonyms.	$\text{sim}_{\text{sem}}(g, p) = \cos\big(\mathbf{e}(g), \mathbf{e}(p)\big)$, using all-mpnet-base-v2 (Song et al., 2020).
CSI Adaptation	CSI-Level Semantic Similarity (**CSI_{bert}**)	CSI-level BERTScore F1 for aligned CSI spans.	$\text{CSI}_{\text{bert}} = \frac{1}{n}\sum_{i=1}^{n} \text{F1}_{\text{bert}}(g_i, o_i)$, where $\mathcal{G} = \{g_i\}$ and $\mathcal{O} = \{o_i\}$ are gold and predicted CSI spans extracted from <CSI> tags, respectively.
	Sentence-Level Semantic Similarity (**$\text{SENT}_{\text{bert}}$**)	Assesses sentence-level meaning preservation using BERTScore F1.	$\text{SENT}_{\text{bert}} = \text{F1}_{\text{bert}}\big(s^{\text{out}}, s^{\text{gt}}\big)$, where s^{out} and s^{gt} are model-generated and ground-truth sentences.

Table 9: Evaluation metrics for the three cross-cultural reasoning tasks: CSI Identification, Prediction, and Adaptation. Hard metrics enforce strict correctness, while soft metrics provide graded credit for semantically or superficially similar outputs.

G.1 CSI Identification

We evaluate CSI identification using two complementary metrics: a strict exact-match score and a soft similarity-based score.

Hard Identification of CSI (HI-CSI). HI-CSI enforces strict exact matching, rewarding only predictions that precisely replicate ground-truth CSIs. This metric is particularly suited for evaluating

models in high-stakes applications where partial identifications may not suffice, such as in content moderation systems requiring unambiguous detection. The sentence-level score HI-CSI is defined as follows:

- If $|G_s| = 0$:

$$\text{HI-CSI} = \begin{cases} 1 & \text{if } |P_s| = 0, \\ 0 & \text{otherwise.} \end{cases}$$

  This penalizes false-positive predictions when no CSIs are present in the ground truth.

- If $|G_s| > 0$:

$$\text{HI-CSI}_s = \frac{1}{|G_s|} \sum_{g \in G_s} \mathbb{I}(g \in P_s),$$

  where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 if the condition holds (i.e., an exact string match exists in P_s) and 0 otherwise. Note that if duplicate CSIs appear in G_s, each is evaluated independently, though matches are checked via membership, allowing a single matching prediction to satisfy multiple identical ground-truth entries if present.

The overall HI-CSI score for the dataset is the average across all sentences, scaled to a percentage for interpretability:

$$\text{HI-CSI} = \left(\frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \text{HI-CSI}_s \right)$$

This ranges from 0 (no matches) to 1 (perfect alignment across all sentences).

Soft Identification of CSI (SI-CSI). To account for minor surface-form variations, we compute a soft CSI matching score based on normalized Levenshtein similarity. For each sentence n, the extracted ground-truth CSI set $G^{(n)} = \{g_1^{(n)}, \ldots, g_{m_n}^{(n)}\}$ and the predicted CSI set $P^{(n)} = \{p_1^{(n)}, \ldots, p_{k_n}^{(n)}\}$ are compared using a similarity-based one-to-one matching procedure, where $m_n = |G^{(n)}|$ and $k_n = |P^{(n)}|$ denote the numbers of ground-truth and predicted CSIs, respectively.

Pairwise normalised Levenshtein similarities $s(g_i^{(n)}, p_j^{(n)}) = 1 - \frac{d_L(g_i^{(n)}, p_j^{(n)})}{\max(\ell(g_i^{(n)}), \ell(p_j^{(n)}))}$ are first computed between all ground-truth and predicted CSIs, and an optimal one-to-one alignment is then obtained using the Hungarian algorithm so that the total similarity of the matched pairs is maximized while preventing reuse of any CSI. Based on the optimal one-to-one alignment, a soft F1 score is computed for each sentence as

$$\text{SI-CSI} = \text{Soft (F1)}_n = \mathbf{1}[m_n = 0 \wedge k_n = 0] + \mathbf{1}[m_n > 0 \wedge k_n > 0] \; \frac{2}{m_n + k_n} \sum_{i=1}^{m_n} \sum_{j=1}^{k_n} s(g_i^{(n)}, p_j^{(n)}) \, X_{ij}^{*(n)},$$

where $X_{ij}^{*(n)}$ denotes the optimal one-to-one assignment variable, which takes the value 1 when $g_i^{(n)}$ is matched to $p_j^{(n)}$ and 0 otherwise, so that both missing ground-truth CSIs and over-generated predictions are penalized.

HI-CSI measures strict span identification accuracy, while SI-CSI provides graded credit for near-matches without relying on heuristic substring rules.

On the relative scores of HI-CSI and SI-CSI. A natural expectation might be that the soft metric (SI-CSI) should always score higher than the hard metric (HI-CSI), since soft matching is more lenient. However, our results show the opposite for several models. This arises from a fundamental difference in what the two metrics measure. HI-CSI is a recall-only metric that checks whether each gold CSI span is found in the model's predictions; extra (false-positive) predictions are ignored. SI-CSI, in contrast,

computes an F1 score using normalised Levenshtein similarity with optimal one-to-one alignment, and therefore penalises both missed gold spans (recall) and unjustified extra predictions (precision).

As a result, over-prediction lowers SI-CSI without affecting HI-CSI. For example, given the gold span "Christmas dinner" and a model prediction "Christmas dinner, turkey, stuffing", HI-CSI scores 1.0 (the gold span is recovered exactly), while SI-CSI's precision drops to $1/3$ due to the two unjustified extra spans, lowering the F1 score correspondingly. This pattern is most pronounced for models with high recall but low precision (e.g., Llama-3.3 in Table 3), which tend to over-tag spans as CSIs. The two metrics are therefore complementary: HI-CSI measures whether the model can locate culturally salient items at all, while SI-CSI additionally penalises models that flag culturally irrelevant content as CSIs.

G.2 CSI Prediction

We evaluate CSI prediction using two complementary metrics: a strict exact-match score and a soft semantic similarity score.

Hard Prediction of CSI (HP-CSI). HP-CSI requires strict exact matching (case-insensitive) for paired CSIs, rewarding only precise replications in the corresponding order. The sentence-level score HP-CSI$_s$ is defined as:

$$\text{HP-CSI}_s = \sum_{i=1}^{\min(n,m)} \mathbb{I}(g_i.\text{lower}() = p_i.\text{lower}()),$$

where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 if the condition holds and 0 otherwise. This sums the number of exact matches across the paired elements up to the minimum length. The overall HP-CSI score for the dataset is the average across all sentences, scaled to account for varying CSI counts (interpretable as a percentage when normalized by average CSI count per sentence):

$$\text{HP-CSI} = \left(\frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \text{HP-CSI}_s \right)$$

This value scales with the average number of CSIs per sentence; for datasets with one CSI per sentence, it ranges from 0 to 1.

Soft Prediction of CSI (SP-CSI). SP-CSI extends HP-CSI by incorporating semantic similarity for non-exact pairs, using cosine similarity on embeddings generated by the SentenceTransformer model ('all-mpnet-base-v2'). This allows partial credit for predictions that capture similar meaning despite lexical differences. The sentence-level score SP-CSI$_s$ is defined as:

$$\text{SP-CSI}_s = \sum_{i=1}^{\min(n,m)} m_i,$$

where for each paired index i,

$$m_i = \begin{cases} 1 & \text{if } g_i.\text{lower}() = p_i.\text{lower}() \text{ (exact match)}, \\ \cos(\mathbf{e}(g_i), \mathbf{e}(p_i)) & \text{otherwise}, \end{cases}$$

and $\mathbf{e}(\cdot)$ denotes the embedding function from the SentenceTransformer model, with $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\,||\mathbf{v}||}$ being the cosine similarity. The overall SP-CSI score is:

$$\text{SP-CSI} = \left(\frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \text{SP-CSI}_s \right)$$

This value scales with the average number of CSIs per sentence.

G.3 CSI Adaptation

We evaluate CSI adaptation along two complementary dimensions: (i) semantic fidelity of the adapted CSIs themselves, and (ii) overall sentence-level semantic preservation, as sentence structure may change during adaptation. We perform both intra-lingual (English-to-English) and inter-lingual (cross-language) evaluations.

Sentence-Level Semantic Similarity ($\text{SENT}_{\text{bert}}$). To assess whether the adapted sentence preserves the meaning of the original, we compute BERTScore F_1 between the model-generated sentence s^{out} and the ground-truth adapted sentence s^{gt}:

$$\text{SENT}_{\text{bert}} = \text{F1}_{\text{bert}}\big(s^{\text{out}}, s^{\text{gt}}\big).$$

BERTScore leverages contextual embeddings from BERT to compute F1 scores based on pairwise token similarities, making it particularly effective for assessing paraphrase quality and semantic preservation in natural language generation tasks. With baseline rescaling enabled to normalize scores relative to a random baseline, improving interpretability and cross-model comparability, the sentence-level full-sentence score is $\text{SENT}_{\text{bert},s} = \text{F1}'_{\text{BERT}}(a_s, g_s)$, using the entire strings a_s and g_s (where g_s is the ground-truth sentence and a_s the adapted output). The overall dataset-level metric is the average: $\text{SENT}_{\text{bert}} = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}} \text{SENT}_{\text{bert},s}$, where $\mathcal{D}$ denotes the dataset of sentence pairs.

CSI-Level Semantic Similarity (CSI_{bert}). Given aligned gold and predicted CSI spans $\mathcal{G} = \{g_i\}_{i=1}^n$ and $\mathcal{O} = \{o_i\}_{i=1}^n$ extracted from <CSI> tags, we compute BERTScore F_1 for each pair:

$$\text{sim}_{\text{CSI}}(g_i, o_i) = \text{F1}_{\text{bert}}(g_i, o_i).$$

The final CSI-level score is the average across all spans:

$$\text{CSI}_{\text{bert}} = \frac{1}{n}\sum_{i=1}^{n} \text{sim}_{\text{CSI}}(g_i, o_i).$$

CSIs are extracted from both ground-truth and adapted texts via regular expressions matching content within <CSI> tags, producing ordered lists of strings. We enforce that the number of extracted CSIs matches between ground-truth and output for each sentence, ensuring paired evaluation. Let $G_s = [g_1^{(s)}, g_2^{(s)}, \ldots, g_{k_s}^{(s)}]$ be the ordered list of ground-truth CSIs (where $k_s = |G_s|$) and $A_s = [a_1^{(s)}, a_2^{(s)}, \ldots, a_{k_s}^{(s)}]$ the corresponding list of adapted CSIs, with $|G_s| = |A_s| = k_s$. The sentence-level CSI score aggregates per-CSI F1 scores with baseline rescaling: $\text{CSI}_{\text{BERT},s} = \frac{1}{k_s}\sum_{i=1}^{k_s} \text{F1}'_{\text{BERT}}(a_i^{(s)}, g_i^{(s)})$, if $k_s > 0$; otherwise, it defaults to 0 (though datasets typically ensure CSIs exist). The overall dataset-level metric is the average: $\text{CSI}_{\text{BERT}} = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}} \text{CSI}_{\text{BERT},s}$. Category-wise scores can be derived by partitioning $\mathcal{D}$ into subsets $\mathcal{D}_c$ based on CSI categories and computing averages within each.

Intra- vs. Inter-Lingual Evaluation. For intra-lingual (English) adaptation, we compute BERTScore using bert-base-uncased. For inter-lingual adaptation, we employ language-specific BERTScore models: bert-base-arabic (Safaya et al., 2020) for Arabic, chinese-bert-wwm-ext (Cui et al., 2019) for Chinese, and bangla-bert-base (Sarker, 2020) for Bengali, to ensure that semantic similarity is evaluated within the appropriate linguistic and cultural representation space. For inter-lingual evaluation, we extend this framework using multilingual models (e.g., bert-base-multilingual-uncased) to handle cross-language pairs, computing analogous $\text{SENT}_{\text{BERT-multi}}$ and $\text{CSI}_{\text{BERT-multi}}$ scores.

Discussion and Rationale. These metrics emphasize semantic equivalence, with CSI_{BERT} isolating performance on culturally sensitive adaptations- critical for ensuring biases or sensitivities are appropriately mitigated without altering core meaning. The use of baseline rescaling maps scores to a more intuitive [0,1] range, where higher values indicate better alignment. Compared to n-gram-based metrics like BLEU, BERTScore better handles paraphrasing and synonyms, aligning with adaptation goals. Limitations include: (1) reliance on BERT's English-centric biases in intra-lingual settings and (2) computational cost

and latency from embedding computations, though batched inference mitigates this. In practice, metrics are implemented using the Hugging Face bert-score library, with time complexity dominated by BERT inference: $O(|\mathcal{D}| \times (L_s^2))$ per sentence for greedy matching in BERTScore, where L_s is token length (typically <512).

H Additional Results

H.1 CSI Identification & Prediction

Task	Metric	Mean ± SE (%)			Wilcoxon p-value (one-sided)		
		V	SV	I	V>SV	SV>I	V>I
Identification	Hard (HI-CSI)	58.07 ± 2.62	48.61 ± 1.68	51.81 ± 1.23	0.004^{**}	0.981	0.012^{*}
	Soft (SI-CSI)	61.20 ± 1.27	56.48 ± 1.43	55.99 ± 1.62	0.008^{**}	0.473	0.004^{**}
Prediction	Hard (HP-CSI)	40.66 ± 3.08	39.54 ± 2.35	40.59 ± 2.46	0.320	0.844	0.527
	Soft (SP-CSI)	79.18 ± 1.79	77.35 ± 2.05	72.64 ± 2.30	0.074	0.004^{**}	0.004^{**}

Table 10: Performance across Hall's cultural triad (V = Visible, SV = Semi-visible, I = Invisible) for the CSI Identification and Prediction tasks under hard and soft metrics. Mean ± standard error reported across 8 LLMs. One-sided Wilcoxon signed-rank tests assess whether higher-visibility levels yield higher scores. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Soft metrics consistently reveal a Visible-vs-Invisible decline; hard metrics show a less monotonic pattern, particularly for Prediction.

Hall's Cultural Element	DS-Distill		Claude		DS-R1		Gemini		GPT-4o		Llama-3.3		OLMo-2		Qwen-2.5	
	HI	SI	HI	SI	HI	SI	HI	SI	HI	SI	HI	SI	HI	SI	HI	SI
Action	51.43	54.91	67.14	73.51	64.29	63.74	58.57	63.64	65.71	66.35	61.43	51.43	54.29	59.68	61.43	66.04
Appropriacy	19.40	39.86	19.40	38.02	24.63	39.32	18.66	30.85	23.88	37.76	23.13	30.31	22.39	38.73	32.84	46.38
Architecture	38.89	48.55	38.89	50.65	72.22	67.84	55.56	63.23	55.56	57.61	41.67	45.70	33.33	46.40	55.56	65.57
Art	60.00	53.25	60.00	62.84	73.33	59.34	50.00	53.64	60.00	55.95	93.33	65.37	50.00	54.46	70.00	57.14
Communication	36.49	46.60	40.54	52.31	36.49	44.81	27.03	39.54	24.32	42.66	28.38	35.94	33.78	53.13	32.43	46.81
Competitiveness	68.75	66.23	62.50	68.65	62.50	68.63	56.25	66.99	56.25	66.99	62.50	63.80	56.25	69.04	62.50	66.96
Customs	43.22	55.94	38.14	56.83	50.85	58.50	38.98	51.66	40.68	53.54	42.37	46.72	29.66	44.61	47.46	55.98
Dress	25.00	42.42	50.00	56.10	50.00	56.10	25.00	41.51	25.00	44.91	50.00	50.93	25.00	41.44	50.00	56.94
Environment	88.89	65.43	77.78	76.94	88.89	78.40	88.89	82.41	66.67	65.24	55.56	48.87	77.78	85.89	77.78	69.88
Food and Drink	72.22	72.32	66.67	75.28	75.00	75.34	69.44	73.80	72.22	77.20	55.56	55.47	69.44	83.09	63.89	64.71
Individualism	38.89	53.71	33.33	53.35	55.56	56.64	33.33	43.08	55.56	49.34	55.56	55.20	50.00	64.64	55.56	63.23
Institutions	71.05	61.02	50.00	60.87	74.56	61.14	51.75	56.25	64.04	61.14	78.95	59.58	41.23	60.65	71.93	63.64
Music	40.00	49.13	60.00	72.83	70.00	82.40	70.00	84.44	30.00	61.39	70.00	62.17	50.00	74.72	70.00	82.40
Orientations	0.00	20.00	0.00	20.00	100.00	33.33	100.00	66.67	0.00	7.41	100.00	33.33	0.00	0.00	0.00	0.00
Power	42.86	37.97	28.57	57.46	57.14	60.85	42.86	55.25	42.86	56.75	57.14	45.75	42.86	61.50	42.86	62.10
Rituals	70.10	65.35	58.25	69.71	73.20	66.95	66.49	71.84	67.01	66.17	68.04	53.80	58.76	68.80	68.04	67.74
Stereotype	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**66.67**	**100.00**	**100.00**	**100.00**	**100.00**
Structure	25.00	31.02	31.25	45.24	45.31	46.68	25.00	36.77	34.38	45.12	40.62	35.45	31.25	44.24	37.50	42.60
Styles of Discourse	50.00	66.83	12.50	70.16	37.50	74.64	37.50	66.96	25.00	70.71	25.00	63.42	12.50	56.71	37.50	69.77
Thinking	45.83	48.08	50.00	57.83	62.50	59.56	50.00	58.93	45.83	53.24	58.33	51.07	54.17	60.37	54.17	55.82
Time	33.33	35.84	0.00	20.03	33.33	32.88	33.33	43.99	33.33	46.95	33.33	38.43	33.33	34.32	33.33	32.88
Visible Behavior	25.00	48.84	25.00	46.89	41.67	62.75	25.00	44.64	25.00	46.40	16.67	36.75	16.67	42.39	16.67	45.24
Ways of Dress	66.67	72.36	25.00	40.91	75.00	67.62	33.33	43.87	58.33	56.04	58.33	56.37	16.67	30.79	33.33	44.14
Workplace Customs	85.71	77.62	85.71	90.24	85.71	75.95	71.43	84.80	85.71	82.38	71.43	67.62	71.43	89.29	85.71	80.71
Workplace Norms	**94.44**	**88.50**	**94.44**	**87.04**	**94.44**	**88.89**	**83.33**	**81.48**	**94.44**	**96.30**	**72.22**	**69.05**	**83.33**	**85.19**	**94.44**	**88.89**

Table 11: Per-category (25) performance for HI-CSI (HI) and SI-CSI (SI) across Hall's cultural elements. Model abbreviations: DS-Distill = DeepSeek-R1-Distill-Qwen-32B; Claude = Claude-3.7-Sonnet; DS-R1 = DeepSeek-R1; Gemini = Gemini-2.0-Flash; OLMo-2 = OLMo-2-32B; Qwen-2.5 = Qwen-2.5-Max. Values are percentages; higher is better.

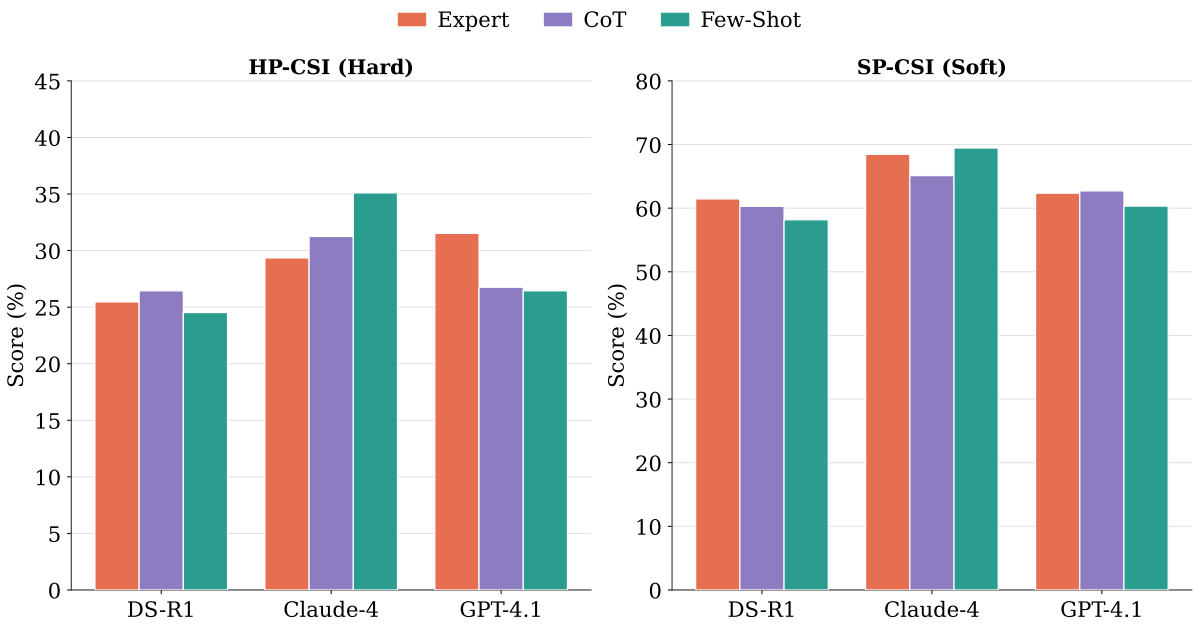


Figure 9: Prompt-sensitivity ablation on the CSI Prediction task across three prompting strategies (Expert, Chain-of-Thought, Few-Shot with 3 examples) for DeepSeek-R1 (DS-R1), Claude-4-Sonnet, and GPT-4.1. Hard (HP-CSI) and soft (SP-CSI) metrics are reported. A Friedman test detects no significant difference across prompts on either metric (HP-CSI: $\chi^2(2) = 0.67$, $p = 0.72$; SP-CSI: $\chi^2(2) = 0.67$, $p = 0.72$; $n = 3$ models). The absence of a consistent winner suggests that cultural reasoning gaps observed in the main results are not readily resolved by prompt engineering alone.

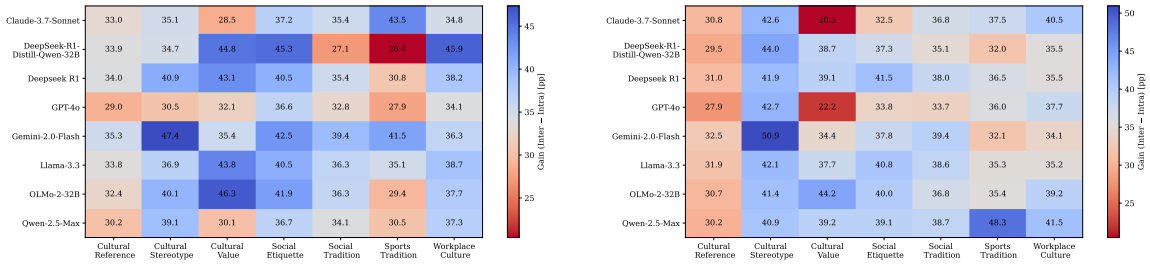
Category Pair	**Identification**	**Prediction**
Cultural Reference vs. Cultural Stereotype	0.006^{**}	0.513
Cultural Reference vs. Cultural Value	0.969	1.000
Cultural Reference vs. Social Etiquette	0.436	0.143
Cultural Reference vs. Social Tradition	0.910	1.000
Cultural Reference vs. Sports Tradition	0.436	$< 0.001^{***}$
Cultural Reference vs. Workplace Culture	0.513	0.014^{*}
Cultural Stereotype vs. Cultural Value	$< 0.001^{***}$	0.742
Cultural Stereotype vs. Social Etiquette	$< 0.001^{***}$	0.993
Cultural Stereotype vs. Social Tradition	0.186	0.669
Cultural Stereotype vs. Sports Tradition	0.669	0.080
Cultural Stereotype vs. Workplace Culture	0.592	0.742
Cultural Value vs. Social Etiquette	0.944	0.296
Cultural Value vs. Social Tradition	0.363	1.000
Cultural Value vs. Sports Tradition	0.058	$< 0.001^{***}$
Cultural Value vs. Workplace Culture	0.080	0.042^{*}
Social Etiquette vs. Social Tradition	0.030^{*}	0.237
Social Etiquette vs. Sports Tradition	0.002^{**}	0.363
Social Etiquette vs. Workplace Culture	0.003^{**}	0.984
Social Tradition vs. Sports Tradition	0.984	$< 0.001^{***}$
Social Tradition vs. Workplace Culture	0.993	0.030^{*}
Sports Tradition vs. Workplace Culture	1.000	0.864

Table 12: Post-hoc Nemenyi pairwise p-values for the CSI Identification and Prediction tasks across CSI categories. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The contrasting patterns of significance across the two tasks support the claim that Identification and Prediction tap qualitatively different model capabilities.

H.2 CSI Adaptation

Culture	Setting	Mean ± SE (CSI_{bert})			Wilcoxon p-value (one-sided)		
		V	SV	I	V>SV	SV>I	V>I
Arabic	Intra-lingual	0.353 ± 0.014	0.299 ± 0.013	0.279 ± 0.015	0.004^{**}	0.008^{**}	0.004^{**}
	Inter-lingual	0.682 ± 0.012	0.655 ± 0.015	0.661 ± 0.006	0.004^{**}	0.273	0.039^{*}
Chinese	Intra-lingual	0.442 ± 0.009	0.346 ± 0.010	0.314 ± 0.006	0.004^{**}	0.008^{**}	0.004^{**}
	Inter-lingual	0.745 ± 0.008	0.717 ± 0.005	0.700 ± 0.003	0.008^{**}	0.008^{**}	0.004^{**}
Bengali (BD)	Intra-lingual	0.345 ± 0.012	0.276 ± 0.017	0.279 ± 0.028	0.004^{**}	0.578	0.004^{**}
	Inter-lingual	0.457 ± 0.014	0.423 ± 0.010	0.436 ± 0.017	0.004^{**}	0.961	0.004^{**}
Bengali (WB)	Intra-lingual	0.387 ± 0.015	0.305 ± 0.019	0.280 ± 0.028	0.004^{**}	0.074	0.004^{**}
	Inter-lingual	0.494 ± 0.014	0.441 ± 0.012	0.436 ± 0.016	0.004^{**}	0.422	0.004^{**}

Table 13: CSI Adaptation performance (CSI_{bert}) across Hall's cultural triad (V = Visible, SV = Semi-visible, I = Invisible) for four target cultures, in both intra- and inter-lingual settings. Mean ± standard error reported across 8 LLMs. One-sided Wilcoxon signed-rank tests assess whether higher-visibility levels yield higher scores. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The Visible-vs-Invisible decline is significant in all eight (culture, setting) combinations; the Semi-visible-vs-Invisible step is less consistent, particularly for Bengali and the inter-lingual settings.



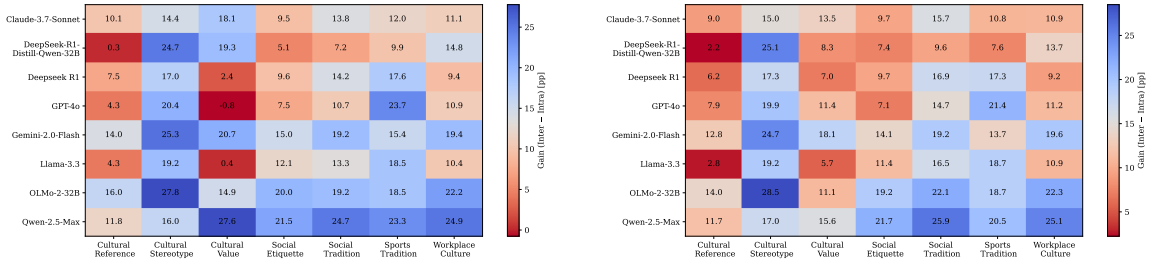
(a) Inter-Intra CSI Adaptation Gain for Arabic

(b) Inter-Intra CSI Adaptation Gain for Chinese

Figure 10: Inter–Intra CSI adaptation gain for Arabic and Chinese. The heatmap reports the difference between inter-lingual and intra-lingual CSI adaptation scores (Inter − Intra) under the CSI_{bert} metric across models and CSI categories. Across both cultures, all 56 (model, category) cells show positive gains (median gain ≈ 36 for Arabic and ≈ 37 for Chinese; one-sided Wilcoxon signed-rank test against zero, $p < 10^{-10}$ for each culture), indicating that LLMs adapt to Arabic and Chinese cultural content significantly more effectively in inter-lingual settings than in intra-lingual ones.

Model	Arabic		Chinese		Bengali (WB)		Bengali (BD)	
	Intra	Inter	Intra	Inter	Intra	Inter	Intra	Inter
DeepSeek-R1	69.02	<u>81.09</u>	72.46	<u>83.47</u>	**84.40**	**74.40**	**83.54**	**74.13**
Gemini-2.0-Flash	68.91	80.78	70.65	83.08	76.84	70.30	76.33	70.14
GPT-4o	**72.76**	81.08	**74.66**	<u>83.81</u>	81.69	70.49	80.56	70.16
Llama-3.3	68.95	80.90	72.74	83.59	<u>84.32</u>	<u>74.40</u>	<u>83.52</u>	<u>74.12</u>
Qwen-2.5-Max	70.54	80.51	67.44	83.75	64.68	70.60	64.02	70.42
OLMo-2-32B	69.01	81.09	72.71	83.56	76.11	70.80	75.27	70.61
Claude-3.7-Sonnet	<u>72.65</u>	**81.64**	73.88	**83.91**	82.49	73.01	80.92	72.24
D-R1-D-Qwen-32B	68.85	75.96	<u>74.15</u>	83.42	78.16	62.27	77.10	62.14
Average	70.08	80.38	72.34	83.58	78.59	70.78	77.66	70.50

Table 14: Sentence-level CSI adaptation performance (%) across models measured using $\text{SENT}_{\text{bert}}$, reported in both intra-lingual and inter-lingual settings across four target cultures. Best results per column are shown in bold, and second-best results are underlined. $\text{SENT}_{\text{bert}}$ scores remain uniformly high across all settings because adaptation typically involves localised substitutions that leave overall sentence structure largely unchanged; CSI-level scores (CSI_{bert}) in Table 4 are therefore the more discriminative metric. Abbreviations: Bangladesh (*BD*), West Bengal (*WB*).



(a) Inter-Intra CSI Adaptation Gain for Bn(West Bengal)

(b) Inter-Intra CSI Adaptation Gain for Bn(Bangladesh)

Figure 11: Inter–Intra CSI adaptation gain for Bengali (West Bengal and Bangladesh). The heatmap reports the difference between inter-lingual and intra-lingual CSI adaptation scores (Inter − Intra) under the CSI_{bert} metric across models and CSI categories. Positive gains dominate across both variants (55/56 cells for WB, 56/56 for BD; median gain ≈ 14 in both; one-sided Wilcoxon signed-rank test against zero, $p < 10^{-10}$ for each), confirming that adaptation in the target language consistently outperforms intra-lingual adaptation. Note that gain magnitudes are markedly smaller for Bengali than for Arabic or Chinese (Figure 10), reflecting weaker target-language proficiency in low-resource settings.

Language	Model	Best Method	Score	Worst Method	Score
Arabic	DeepSeek-R1	Couplet (Cultural Equivalent + Paraphrase)	1.00	Cultural Equivalent	0.50
Arabic	Gemini-2.0-Flash	Couplet (Cultural Equivalent + Classifier)	0.86	Couplet (Cultural Equivalent + Neutralization)	0.57
Arabic	GPT-4o	Cultural Equivalent	0.79	Cultural Equivalent + Componential Analysis	0.53
Arabic	Llama-3.3	Couplet (Cultural Equivalent + Accepted Standard Translation)	1.00	Couplet (Cultural Equivalent + Neutralization)	0.50
Arabic	OLMo-2-32B	Couplet (Cultural Equivalent + Literal Translation)	1.00	Cultural Equivalent + Labeling	0.50
Arabic	Qwen-2.5-Max	Transference	0.78	Paraphrase	0.57
Bengali	Claude-3.7-Sonnet	Couplet (Transference + Naturalization)	1.00	Componential Analysis	0.46
Bengali	DeepSeek-R1	Couplet (Cultural Equivalent + Accepted Standard Translation)	0.90	Cultural Equivalent	0.41
Bengali	Gemini-2.0-Flash	Cultural Equivalent + Labeling	0.80	Cultural Equivalent + Label	0.43
Bengali	GPT-4o	Literal Translation	0.88	Paraphrase	0.41
Bengali	Llama-3.3	Couplet (Cultural Equivalent + Literal Translation)	0.90	Cultural Equivalent	0.43
Bengali	Qwen-2.5-Max	Couplet (Cultural Equivalent + Paraphrase)	0.81	Transference	0.42
Chinese	DeepSeek-R1	Cultural Equivalent	1.00	Cultural Equivalent	0.50
Chinese	Gemini-2.0-Flash	Couplet (Transference + Labeling)	0.79	Couplet (Cultural Equivalent + Paraphrase)	0.53
Chinese	GPT-4o	Couplet (Cultural Equivalent + Accepted Standard Translation)	1.00	Couplet (Cultural Equivalent + Paraphrase)	0.50
Chinese	Llama-3.3	Cultural Equivalent	1.00	Cultural Equivalent	0.50
Chinese	OLMo-2-32B	Cultural Equivalent	1.00	Cultural Equivalent	0.51
Chinese	Qwen-2.5-Max	Couplet (Cultural Equivalent + Naturalization)	1.00	Couplet (Cultural Equivalent + Literal Translation)	0.51

Table 15: Best- and worst-performing adaptation methods (by average CSI score) for each language–model pair. "Couplet" denotes a combination of two translation strategies. Cultural Equivalent is abbreviated as CE in the main text.

I Prompts Used for Evaluation

I.1 CSI Identification

You are a highly skilled and culturally aware agent with expertise in identifying Culture-Specific Items (CSIs). CSIs are words, phrases, or concepts that are specific to a particular culture or carry significant cultural meaning. They depend on cultural context and can vary widely across cultures. CSIs encompass natural elements like animals, plants, and environmental features (ecology); physical objects like food, clothing, housing, and tools (material culture); activities related to work and leisure (social culture); political, social, legal, religious, and artistic concepts (organizations, customs, and ideas); and non-verbal communication or behaviors like gestures and habits.

Your task is to identify the CSI(s) in the given sentence and enclose them with '<CSI>' and '</CSI>' tags. The CSIs in this task will mostly pertain to American/Western/British Culture, so prioritize identifying items specific to these cultural contexts. Do not add any explanations, notes, or unnecessary text—only provide the revised sentence with the CSI(s) properly tagged.

Now process the following sentences exactly as instructed: Input:

Figure 12: Evaluation Prompt for CSI Identification Task.

I.2 CSI Prediction

You are a highly skilled and culturally aware agent with expertise in predicting Culture-Specific Items (CSIs). CSIs are words, phrases, or concepts that are specific to a particular culture or carry significant cultural meaning. They depend on cultural context and can vary widely across cultures. CSIs encompass natural elements like animals, plants, and environmental features (ecology); physical objects like food, clothing, housing, and tools (material culture); activities related to work and leisure (social culture); political, social, legal, religious, and artistic concepts (organizations, customs, and ideas); and non-verbal communication or behaviors like gestures and habits.

You' ll be given inputs where each sentence will contain a '[MASK]' tag, where the CSI has been masked. The CSI is surrounded with a beginning <CSI> tag and an ending </CSI> tag. Your task is to predict the most appropriate CSI to fill in the '[MASK]' based on the surrounding context. The CSIs in this task will mostly pertain to American/Western/British Culture, so prioritize predicting items specific to these cultural contexts.

Replace the [MASK] with the predicted CSI. Do not add any explanations, notes, or unnecessary text—only provide the revised sentence with the predicted CSI filled in.
Now process the following sentences exactly as instructed:
Input:

Figure 13: Evaluation Prompt for CSI Prediction Task.

I.3 CSI Adaptation

You are an expert in cultural adaptation and linguistic transformation with a deep understanding of Bengali/Chinese/Arabic culture, traditions, and social practices. You specialize in providing culturally sensitive adaptations that align with the values, norms, and etiquette of Bengali/Chinese/Arabic people. You will be given an input sentence containing Culture-Specific Items (CSIs) that mostly pertain to American/Western/British Culture. Your task is to adapt sentences containing <CSI> tags into the context of Bengali/Chinese/Arabic society, ensuring both cultural relevance and linguistic accuracy. You can use one or a combination of the following methods appropriate for adaptation:

1. Transference: Keep the original word unchanged in the adaptation.
2. Cultural Equivalent: Use a similar term from the target culture.
3. Neutralization: Explain what the term means or does.
4. Literal Translation: Translating the word directly to target culture.
5. Label: Add a brief explanation or tag to the term.
6. Naturalization: Adapt the word to fit the target language's spelling or sound.
7. Componential Analysis: Break the term into parts and explain each.
8. Deletion: Remove unnecessary words or phrases.
9. Couplet: Combine two adaptation methods
10. Accepted Standard Translation: Use a commonly accepted adaptation.
11. Paraphrase or Gloss: Give a longer explanation or footnote.
12. Classifier: Add a general category to clarify the term' s meaning.

Your task is to adapt the following sentences containing <CSI> tags into Bengali/Chinese/Arabic culture. Replace inside the <CSI> tags with culturally relevant practices, behaviors, or terms. Do not remove the <CSI> tags. In case there is no cultural equivalent in the target culture, return 'Non-transferable'.

Provide both Intra-lingual adaptation (Contextualized for Bengali culture but written in English) and Inter-lingual adaptation (Translated and adapted into Bengali/Chinese/Arabic for each input. Ensure cultural accuracy, politeness, and respect in all adaptations. Also output what method you used for the adaptation. The output format should be [Intra-lingual adaptation, Inter-lingual adaptation, Adaptation method used]. Do not add any explanation.
Now adapt the following sentences:
Input:

Figure 14: Evaluation Prompt for CSI Adaptation Task.