Beyond Monolithic Culture: Evaluating Understandability of Online Text Across Cultural Dimensions

Saurabh Kumar Pandey[1*], Harshit Gupta[2*], Sougata Saha[1], Monojit Choudhury[1]

[1]Mohamed Bin Zayed University of Artificial Intelligence, [2]Microsoft, India

[1]saurabh2000.iitkgp@gmail.com, [2]guptaharshi@microsoft.com, [1]{sougata.saha, monojit.choudhury}@mbzuai.ac.ae

Abstract

Culture shapes how people interpret language, especially in online reviews containing culture-specific items (CSIs). Yet, most existing evaluations treat culture as a monolithic construct, offering no insight into which cultural dimensions pose difficulty for readers, or how large language models (LLMs), which power AI reading assistants, perform across them. This gap limits our ability to obtain reliable, cross-cultural estimates of model performance. To address this, we analyze CSIs in English Goodreads reviews across Newmark's cultural dimensions (e.g., material, ecology, customs, habits, social) and evaluate six LLMs of varying sizes on their ability to identify CSIs within each dimension. We find that readers struggle most with CSIs from the material, customs, and social dimensions, while models underperform on more localized ones (e.g., habits), revealing systematic cultural blind spots. To support further research on culturally representative benchmarking, we release an expert-annotated dataset of CSIs labeled by cultural dimension. Empirical analysis shows our dataset as more challenging and of higher quality than existing cultural benchmarks, enabling finer-grained evaluation of cultural understanding in models.

1 Introduction

Online reviews often contain culture-specific items (CSIs) (Aixelá 1996), which, though intuitive to the writer, can hinder comprehension for readers from different cultural backgrounds. According to a study of Goodreads reviews by Saha et al. (2025), the prevalence of such CSIs in difficult-to-understand online text can be as high as 83%. Prior work has shown that LLM-based tools such as reading assistants and translation systems (Pandey et al. 2025) can help bridge these gaps, though they frequently exhibit cultural biases and stereotypes (Dwivedi, Lavania, and Modi 2023; Johnson et al. 2022). However, most existing studies treat culture as a single, monolithic construct, without systematically distinguishing its multiple facets (Hall 1976; Geertz 2017). This leaves us with a limited understanding of (i) how CSIs are distributed across different cultural dimensions in complex online text, and (ii) how LLMs perform across these dimensions.

In this work, we address this gap by adapting Newmark's taxonomy (Newmark 2003) of seven cultural dimensions, comprising Ecology, Material, Customs, Social, Habits, Linguistics, and Others, to analyze which categories of CSIs readers find difficult to understand[1]. Each dimension varies in its level of cultural universality (Brown 2004; Murdock et al. 1961): for example, Material and Ecology encompass broadly shared phenomena such as food, clothing, flora, and weather, whereas Social and Customs reflect localized practices, hierarchies, and rituals. To evaluate how LLMs perform as cross-cultural reading assistants, we benchmark six LLMs - GPT-4o, Llama-8B, Gemma-9B, Aya-8B, Mistral-7B, and Mixtral-7B - on their ability to identify CSIs within each of Newmark's dimensions, using country as a cultural proxy (Adilazuarda et al. 2024).

While prior work in translation studies has employed Newmark's taxonomy to assess cross-cultural transfer (Yao et al. 2024; Shi et al. 2024), such studies primarily evaluate whether models can translate cultural concepts between languages. In contrast, we use this taxonomy to examine the prevalence and interpretive difficulty of different cultural dimensions in human-generated text from the Goodreads-CSI dataset (Saha et al. 2025), and to systematically assess how LLMs detect and categorize CSIs. To enable robust evaluation, we standardize all spans and manually annotate them with multiple Newmark dimensions and universality levels.

Our analysis yields three key insights. First, most CSIs in the dataset pertain to the material, customs, and social dimensions, while habits and linguistic CSIs are comparatively rare. Second, models vary substantially across dimensions: GPT-4o and Mixtral perform best on more universal dimensions (e.g., Material, Ecology), whereas Llama-3.1 and Gemma perform better on less universal, context-dependent ones. Third, across all dimensions, models exhibit greater agreement with each other than with human annotators, suggesting systematic misalignment with human cultural understanding. Insights from our study can guide the development of culturally aware evaluation pipelines and LLM-based reading assistants (Pandey et al. 2025) that can accommodate diverse cultural dimensions rather than treat-

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]We extend Newmark's original five cultural dimensions (Ecology, Material, Customs, Social, and Habits) by adding two additional categories: Linguistics and Others, to capture language-based and any uncategorized cultural expressions.

ing culture as a homogeneous construct. Our contributions are as follows:

- We release a standardized, expert-annotated version of the Goodreads-CSI dataset (Saha et al. 2025), comprising 922 review-span combinations (671 unique spans, 159 CSIs) manually categorized by Newmark's cultural dimensions and universality levels.
- We systematically analyze the distribution of CSIs across Newmark's dimensions and share empirical insights on which types are most challenging for readers and LLMs, understanding which is critical for robust models.
- We evaluate GPT-4o and five open-weight LLMs (<8B parameters) using a socio-demographic prompting setup, benchmarking their ability to detect CSIs as cross-cultural reading assistants.

2 Related Work

Cross-cultural communication is a core aspect of language. Foundational work by Gudykunst (2003); Hurn et al. (2013) established how cultural differences shape interaction styles and conversational norms. Although these theories are well understood, their practical application has only recently gained traction (Singh, Patidar, and Vig 2024; Pandey et al. 2025). Cultural adaptation has been extensively explored in translation studies. Early research (Sperber, Devellis, and Boehlecke 1994; Trivedi 2008; Yao et al. 2024) showed that literal translation fails to capture cultural nuances, calling for a holistic approach that accounts for meaning, tone, and context (Bassnett 2007). While text difficulty studies often focus on linguistic factors like lexical familiarity and syntax (Wang and Lowie 2021; Jacob and Uitdenbogerd 2019; Leroy and Kauchak 2014), our work examines CSIs (Aixelá 1996), which are elements rooted in a culture that lack direct equivalents across languages (Zhang et al. 2024; Daghoughi and Hashemian 2016; Narváez and Zambrana 2014).

Newmark (2003) provided a structured taxonomy of CSIs, known as Newmark's cultural dimensions, for categorizing and analyzing cultural elements. While some works study LLMs through cultural theories such as Hofstede's framework (Li et al. 2024a; Kharchenko et al. 2024; Dawson et al. 2024; Hofstede 2001; Geert and Hofstede 2004), they rarely measure perceived difficulty. Our approach directly identifies which CSIs readers find challenging, rather than assuming cultural transparency. Studies on cultural familiarity and comprehension (Toti and Hamid 2022; Erten and Razi 2009) similarly show that people may struggle even with texts from their own culture.

Recent research has moved toward empirically evaluating the cultural knowledge encoded in LLMs, focusing on artifacts like food, art, and geography (Seth et al. 2024; Li et al. 2024a; Koto et al. 2024). These studies reveal both the strengths and limitations of current models, emphasizing the need for robust cultural benchmarks (Wang et al. 2024; Rao et al. 2024; Myung et al. 2024; Zhou et al. 2024; Putri et al. 2024; Wibowo et al. 2024; Owen et al. 2024; Chiu et al. 2024; Liu et al. 2024; Koto et al. 2024). Developing such benchmarks is considered crucial, as existing general-purpose benchmarks often fail to capture the subtle and complex nature of cultural understanding.

3 Analyzing CSIs in Online Texts

We use the Goodreads-CSI dataset by Saha et al. (2025) to analyze the types of CSIs people generally find difficult to understand in online texts. The original dataset comprises 245 crowd-sourced, unique, difficult-to-understand span annotations from 57 English Goodreads book reviews. The reviews pertain to books from Ethiopia, India, and the USA, and the annotations are performed by 50 users: 8 from India, 22 from Mexico, and 20 from the USA. Given a book review, each user identified text spans of varying lengths that they found difficult to understand. This dataset is interesting because it captures knowledge as behavior, which is a unique and cognitively more challenging aspect to measure than other forms of behavior, such as preferences (Dunbar 1995; Kuhl 2004). However, the dataset has the following several limitations:

1. Limited CSIs: Since the dataset contains CSIs identified by only 50 users across three countries, the actual space of all possible CSIs from the three countries (India, Mexico, and the USA) may be larger. Hence, we use LLMs to simulate behavior only for these three countries, which we discuss in detail in Section 4. Not only does this approach synthetically increase the coverage of CSIs, but it also provides a way to benchmark LLMs in simulating the behavior of users from these countries.

2. Non-standardized Spans: Since users can mark any contiguous text spans as difficult to understand, there is a high degree of variability in the annotations, making the dataset noisy. To handle this, the original dataset semantically clustered the spans using sentence transformers[2] and filtered out poor-quality annotations before conducting quantitative and qualitative analysis. However, this approach has several limitations:

(i) User-User Mismatch: The lengths of user annotations vary, where one user might identify multiple CSIs as a single contiguous span, whereas others might segment the same span into multiple non-contiguous spans. For example, the span *"from the Beats in On The Road to Ken Kesey's Merry Pranksters"* refers to two influential counter-cultural movements in American literature and history represented by *"the Beats in On The Road"* and *"Ken Kesey's Merry Pranksters"*. Someone unfamiliar with either of them might mark the entire span as difficult to understand, while others might partially mark the spans, signifying familiarity with either *"the Beat generation"* or *"the pranksters"*. Also, some might non-contiguously highlight both spans to indicate unfamiliarity with both of them. Such finer distinctions are lost in sentence-transformer-based semantic clustering.

(ii) User-Model Mismatch: User spans were longer than model spans, as users often grouped multiple CSIs into one, while models split them into more atomic units.

(iii) Model-Model Mismatch: Unlike Saha et al. (2025), we also evaluate multiple models (Section 4), which introduces span mismatches similar to user–user disagreements. To address this, we aggregated all CSI spans from users and models for each review and manually standardized them.

[2] sentence-transformers/all-MiniLM-L6-v2

3.1 Dataset Standardization

Through standardization, we aim to decompose compound annotations into their elementary discourse units (EDUs) (Mann and Thompson 1988). Extracting and utilizing EDUs is a reliable method for improving tasks like emotion classification (Zhu and Wu 2022) and abstractive summarization (Xiong, Racharak, and Nguyen 2022; Li, Wu, and Li 2020). Hence, we standardize the dataset to identify such EDUs, in the form of CSIs, for conceptual uniformity, measurement, and downstream processing. The below illustrated example shows the overlap scores between the user and two models, before and after standardization. Before standardization, both Models 1 and 2 score perfectly (1.0) using sentence transformers with a cosine similarity threshold of 0.5. However, after standardization, the user span is split into five unique spans (CSIs), and we obtain an overlap score of 0.4 for Model 1 and 0.2 for Model 2 using an exact match.

Before Standardization:
User: John Muir, Muir woods, Stickeen, The Moral Equivalent of War by William James
Model 1: John Muir? Sure, Muir woods,
Model 2: John Muir
Previous overlap scores (Cosine) - Model 1: 1.0 **Model 2:** 1.0

After Standardization:
User: John Muir#Muir woods#Stickeen#The Moral Equivalent of War#William James
Model 1: John Muir#Muir woods
Model 2: John Muir
Our overlap scores - Model 1: 0.4 **Model 2:** 0.2

The dataset contains 1,193 combinations of reviews and CSIs annotated by the users and generated by the models. Three Computer Science and Linguistics experts manually standardized all 1,193 spans in the dataset by converting them to their appropriate EDUs. Each annotator annotated 450 spans (avg 19 reviews per annotator), which were randomly sampled and had approximately 50 overlapping spans among them to facilitate calculating inter-annotator agreement (IAA) scores. The annotation guidelines for the standardization of spans were as follows:

- If a span contains multiple CSIs, split it into their EDUs by a "#" symbol.
- If a span contains only part of a named entity (such as a book title, proper noun), the span should be expanded to include the full entity.
- Correct any grammatical errors and formatting inconsistencies, wherever necessary.

Since each annotation involved splitting a span into multiple spans, we use an overlap metric to calculate agreement between the three expert annotators. Inspired by works like Braylan, Alonso, and Lease (2022); Passonneau (2006), we use Jaccard similarity, which measures the intersection over union (IoU) of span sets annotated by different annotators. Using this method, we obtained a high mean IoU score of 0.967 after the initial round of annotation. All remaining disagreements were resolved through an additional adjudication round, leading to the final gold standardized dataset.

Post standardization, we observe a decrease in the average number of words in user spans from 6.31 to 3.30, and from 5.22 to 3.36 in the model spans, indicating better consistency. The total review text and standardized span combinations were reduced to 922 (322 from users, 600 from models), compared to 1,193 (365 from users, 828 from models) in the non-standardized version. Overall, the dataset contains 671 unique standardized spans across all review texts, compared to the previous 1,122 non-standardized spans. The standardization process ensures the reliability of any subsequent span-based analysis, enabling robust comparisons.

Categorizing Spans by Newmark's Taxonomy: We designed a prompt (detailed in Appendix A.1) and used GPT-4o to perform two tasks: (i) identify whether a given span is cultural or non-cultural in the context of the review text, and (ii) classify cultural spans into the extended Newmark taxonomy - Customs, Ecology, Habits, Linguistic, Material, Social, and Other. Out of the 671 unique spans, 159 were categorized as cultural by GPT-4o. Two of the three expert annotators manually checked the categorization of all the standardized spans and deemed them to be correct, with an inter-annotator agreement (IAA) Krippendorff's alpha score of 0.96, indicating a high agreement. They further manually categorized these spans into Newmark's CSI categories. To ensure reliability, they first annotated 50 randomly sampled spans, achieving a Krippendorff's alpha score of 0.948 after two rounds of resolution. With this high agreement, they proceeded to annotate the remaining 113 spans. Given the subjective nature of span categorization, where a span can belong to multiple categories (e.g., "bar-hopping" reflecting both Habit and Social aspects), the third reviewer independently classified the entire dataset, assigning up to two categories per span when applicable. The IAA among all three reviewers was 0.88. The final dataset includes standardized spans with each span assigned to multiple labels based on Newmark's cultural dimensions. The dataset can serve as the gold standard for benchmarking and aligning LLMs for diverse cultural norms and epistemologies.

Cultural Universals and Newmark's Taxonomy: Formally introduced by Murdock et al. (1961) and expanded by Brown (1991, 2004), cultural universals are traits, practices, or institutions that are found in every human culture. Brown (2004) highlights that these elements are on a continuum, from widely shared concepts to culture-specific. This continuum is useful in analyzing CSIs along Newmark's taxonomy, as not all CSI categories are equally interpretable across cultures. We categorize Newmark's taxonomy into three *universality* levels: High, Medium, and Low, based on how likely they are to be recognized across different cultural contexts. Dimensions like Ecology and Material are assigned high universality, since they reference natural or physical elements that are conceptually familiar to most cultures, although named differently ("Muir Woods", "Smoky Mountains"). Medium universality includes dimensions like Social and Customs, where the underlying ideas ("belief in a higher power") are widespread, but their expression varies greatly across cultures. Low universality captures dimensions that are highly culture-bound, such as idioms like "had me at hello" or "home run", and behaviors such as "bar-

hopping", which rely heavily on specific cultural or linguistic context to be understood. This categorization helps assess the universality of Newmark's dimensions systematically.

3.2 Analysis: Distribution of CSIs

Table 1 presents the category-wise statistics and exemplar spans of the standardized dataset, with the associated **H**igh/**M**edium/**L**ow universality levels. The standardized dataset contains 159 unique cultural spans from users and models combined, which is dominated by Material (79), followed by Customs (45) and Social (43).

Category	Spans (#)	Avg len of spans	Spans with entities (#)	Example
Material (H)	79	2.38	48 (61%)	Devlok, oath of vayuputras
Ecology (H)	29	2.38	21 (72%)	Muir woods, Smoky Mountains
Customs (M)	45	2.78	19 (42%)	Mormanism, belief in God
Social (M)	43	3.12	26 (60%)	topper from engineering institute
Linguistic (L)	14	2.86	1 (7%)	had me at hello, home run
Habits (L)	1	1.00	0 (0%)	bar-hopping
Other (L)	4	4.75	1 (25%)	we are not like that

Table 1: Standardized dataset statistics and examples.

To further study and contrast the constituents of the CSI spans in each dimension, we analyze the distribution of named entities present in the CSI from each dimension using the Spacy[3] *en_core_web_sm* model. Named entities are real-world objects with a name, such as persons, organizations, locations, dates, and etc (Nadeau and Sekine 2007; Jurafsky and Martin 2020). For our purpose, we do not distinguish the named entity types and instead analyze the distribution of the number of named entities prevalent in the CSIs along each cultural dimension. As depicted in Table 1 (column "Spans with entities"), we find that Ecology has the highest percentage of named entities (72%), whereas Material and Social culture have 61% and 60% of entities, respectively. We also observe that dimensions that are highly universal (Material, Ecology) have a higher number of named entities as compared to other dimensions, such as Linguistic.

Figure 1 shows the distribution of the 159 unique standardized CSIs across Newmark's dimensions, segregated by users, models, and both. We observe that (i) User-identified CSIs mostly pertain to the Material, Social, and Customs dimensions. (ii) Models identify a lot of extra spans as CSIs, most of which belong to the Material (29) and Customs (23) dimensions of culture. Also, the majority of these additional CSIs belong to high and medium universal dimensions rather than the low universal ones.

In total, users from India identify 58 unique CSIs, Mexico 57, and the USA 61. Figure 3 presents the country-wise percentage distribution of CSIs across Newmark's dimensions. Across all three countries, Material CSIs account for the largest share of what users fail to understand, followed by Customs and Social dimensions. Indian users show the highest proportion of Material CSIs (41%), while the USA shows the lowest (37%). Conversely, for Customs and Social CSIs, users from the USA report the highest proportions.

[3]https://spacy.io/api/entityrecognizer

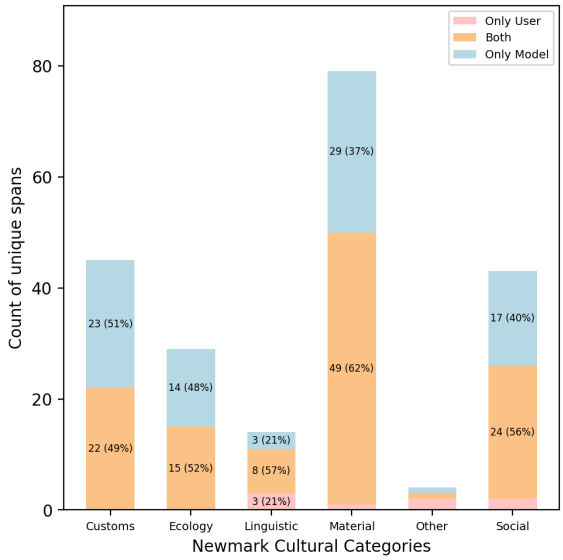

Figure 1: CSI distribution across Newmark's taxonomy.

Figure 4 further breaks down the dimension-wise proportions of CSIs by country. Here, USA users again dominate in Social and Customs CSIs, while Mexican users lead in the Linguistic and Other categories. Indian users, in contrast, show the highest proportions in Material and Ecology.

Similar to Saha et al. (2025), we plot the number of CSIs by user location and book origin, but extend the analysis by examining their distribution across Newmark's dimensions (Figure 11, Appendix A.2). As in prior work, the overall trends remain consistent even after span standardization: users from any country encounter more CSIs in reviews of books originating in the USA, while books from Ethiopia contain the fewest. Although the absolute number of CSIs increases with standardization, the relative patterns persist. Notably, the percentage distributions in Figure 2 show that users face greater difficulty with Material CSIs from foreign cultures than from their own. Moreover, users from India report no difficulty with low-universal categories, such as Linguistic and Others, when reading reviews of books from their own culture, which aligns with intuition. In contrast, USA users show difficulty across all dimensions, regardless of universality, even when engaging with texts from their own culture. Together, these findings reinforce that culture is not monolithic. Different dimensions emerge as more or less difficult to understand depending on context.

4 LLMs as Cultural Reading Assistants

Building on Saha et al. (2025)'s methodology, we pose six LLMs as cultural reading assistants (see Appendix A.1), and assess their performance in identifying CSIs from the standardized dataset using socio-demographic prompting (Li et al. 2024b; AlKhamissi et al. 2024; Wan et al. 2023).

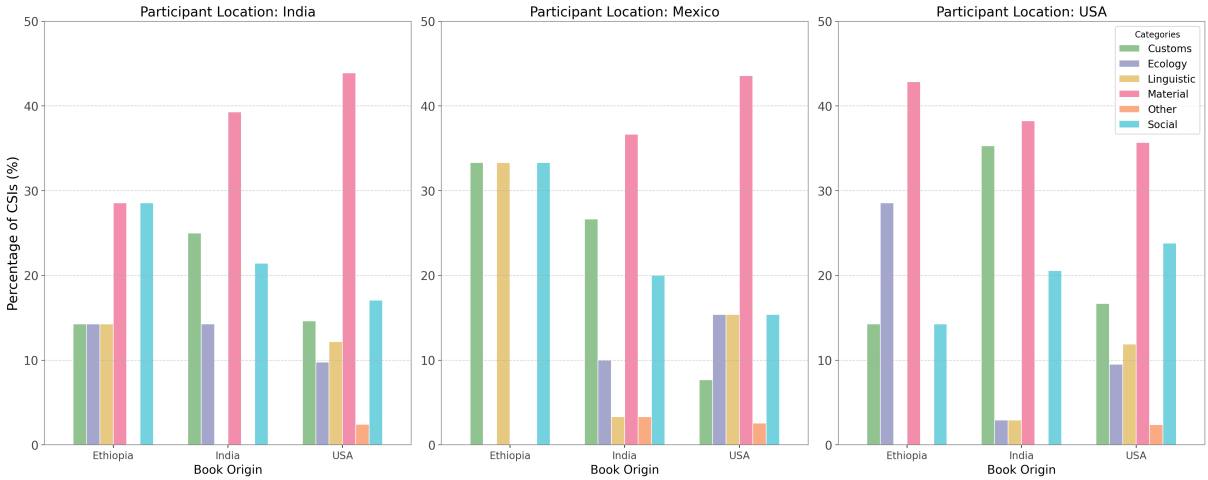

Figure 2: CSI distribution by participant location and book origin across Newmark's dimensions.

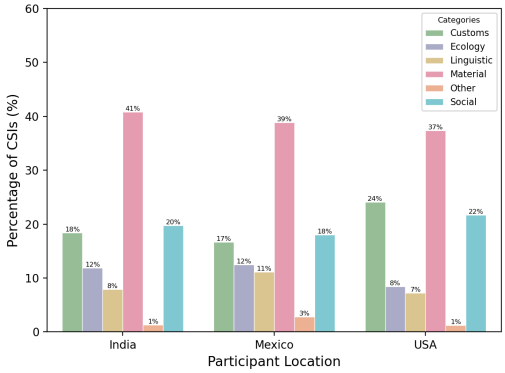

Figure 3: CSI distribution across countries and dimensions.

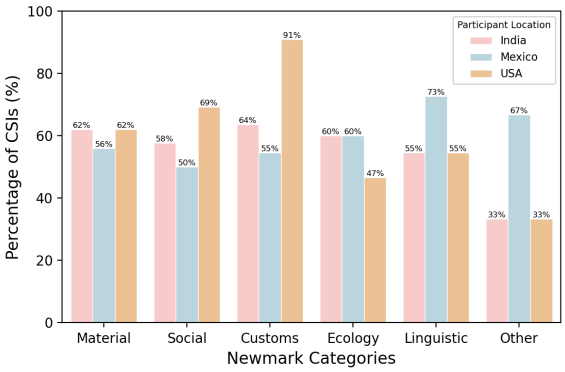

Figure 4: Category-wise distribution of CSIs by location.

4.1 Models

We experiment with the following LLMs: Llama-3.1-8B-Instruct (Dubey et al. 2024), Gemma-2-9B-it (Team et al. 2024), Aya-23-8B (Aryabumi et al. 2024), Mistral-7B-Instruct-v0.2 (Jiang et al. 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al. 2024), and gpt-4o-2024-05-01-preview (Achiam et al. 2023). We prompt all models using temperature 0.0 and generate the response in a JSON format.

4.2 Results

1. Impact of CSI Standardization: We compare the standardized CSIs identified by GPT-4o with user-identified standardized CSIs and calculate the precision, recall, and F1-score. We only measure the impact of span standardization using GPT-4o to maintain parity with the original study (Saha et al. 2025), which only evaluated GPT-4o's performance. As depicted in Figure 5, we observe that recall is consistently higher than precision across all countries, indicating the model's propensity to identify additional spans as CSIs. Although our findings are similar to prior work, we observe that standardization results in a decline in precision from 0.49 to 0.43 and recall from 0.65 to 0.63, indicating that **our standardized dataset is more challenging and higher-quality than the original dataset.**

2. LLM Benchmarking: Figure 5 also plots the country-wise Precision, Recall, and F1 score for all other LLMs against user-identified CSIs. We observe the following:

- All models, except Mistral, have higher recall than precision, indicating a broad and generic understanding of what people from different countries might find difficult to understand due to their culture.
- GPT-4o has the highest F1 score for all countries, whereas Llama-3.1 has the highest recall. The F1 scores for Llama-3.1 closely follow GPT-4o for India and Mexico, while Gemma-2 ranks second for the USA.
- The overall low F1 scores (between 0.4 and 0.55) depict the complexity of the CSI identification task.

We observe a few interesting cases: 'Home run', although a classic sports reference widely used in American culture, is marked as a hard-to-understand concept, even by users from the USA, which models fail to highlight. Similarly, '9/11', although a well-known tragic event in the USA, is contextually marked as difficult to understand by users from all countries. However, all models, except Llama-3.1, fail to capture

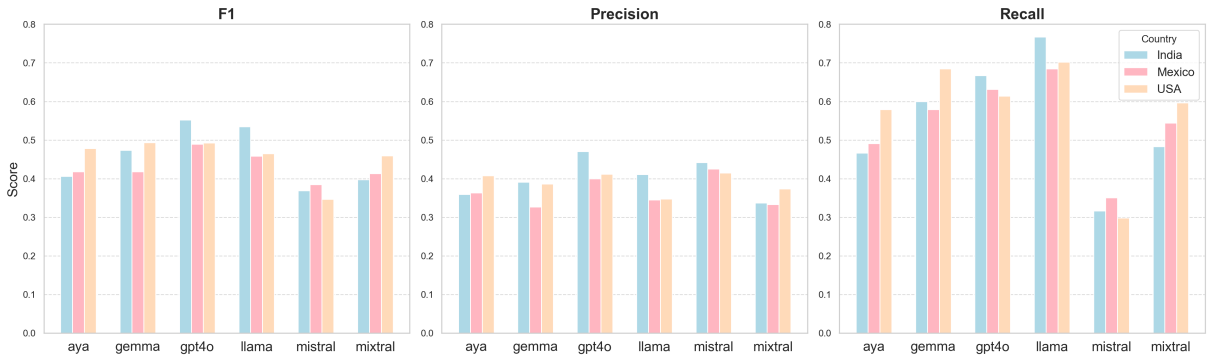

Figure 5: Model and country-wise Precision, Recall and F1-scores for CSI identification.

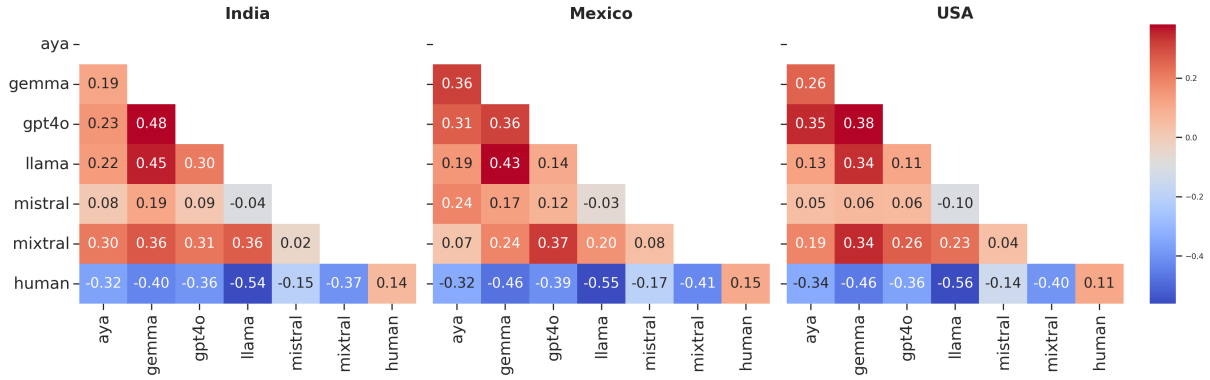

Figure 6: Inter-annotator agreement (Krippendorff's alpha) on standardized CSI spans across countries.

this as a CSI. These observations suggest that although user comprehension is pluralistic and context-dependent, LLMs risk stereotyping cultural knowledge, often missing nuanced difficulties that even native users experience. This highlights a gap between human understanding and model assumptions about cultural familiarity.

3. Overlap Analysis: To analyze the overlap of model and user-identified CSIs, we calculate Krippendorff's alpha IAA between all model-model and model-user pairs using the standardized spans and present the scores as a heatmap in Figure 6. Notably, we observe that **models tend to agree more with each other than with humans**, and vice versa. The agreement scores between users and models are notably low (between -0.56 and -0.15) across all countries. While Mistral shows a high IAA amongst all the models, it is primarily due to a lower number of CSIs identified, as indicated by the low recall in Figure 7. These results demonstrate a limited ability of models to identify CSIs in review texts, using country as the proxy.

In contrast, the high agreement scores among models suggest a consistent pattern in how they identify CSIs, likely influenced by similarities in their pre-training data sources, raising an important question: **Are LLMs learning nuanced cultural representations, or are they simply reinforcing stereotypes?** Since no member of a coherent socio-cultural group displays all prototypical behaviors of the group, but all members exhibit some prototypical behaviors (Leung and Cohen 2011; Morris et al. 2015; Hogg 2016), **the high intra-model agreement could reflect a reliance on oversimplified/stereotypical cultural patterns rather than a deep understanding of cultural nuances**.

4. Analysis along Newmark's Dimensions: Figure 7 depicts the category-wise Precision, Recall, and F1 score for all LLMs against user-identified CSIs. We observe:

- In most models except Mistral, recall exceeds precision for high and medium-universal categories, showing models over-identify CSIs. For low-universal categories (Linguistic, Others), precision exceeds recall, indicating models assume user knowledge that is often lacking.
- No single model dominates performance in all categories. Smaller models, such as Llama-3.1 and Gemma-2, achieve better F1 scores than GPT-4o in the Customs and Ecology categories, whereas GPT-4o fares better in the Material and Linguistic categories.
- Despite these results, overall F1-scores remain below 0.6 across most categories (Figure 8), indicating substantial room for improvement. Excluding the sparse Habits and Others categories, LLMs perform best on Social CSIs, followed by Customs, Material, Ecology, and Linguistic.

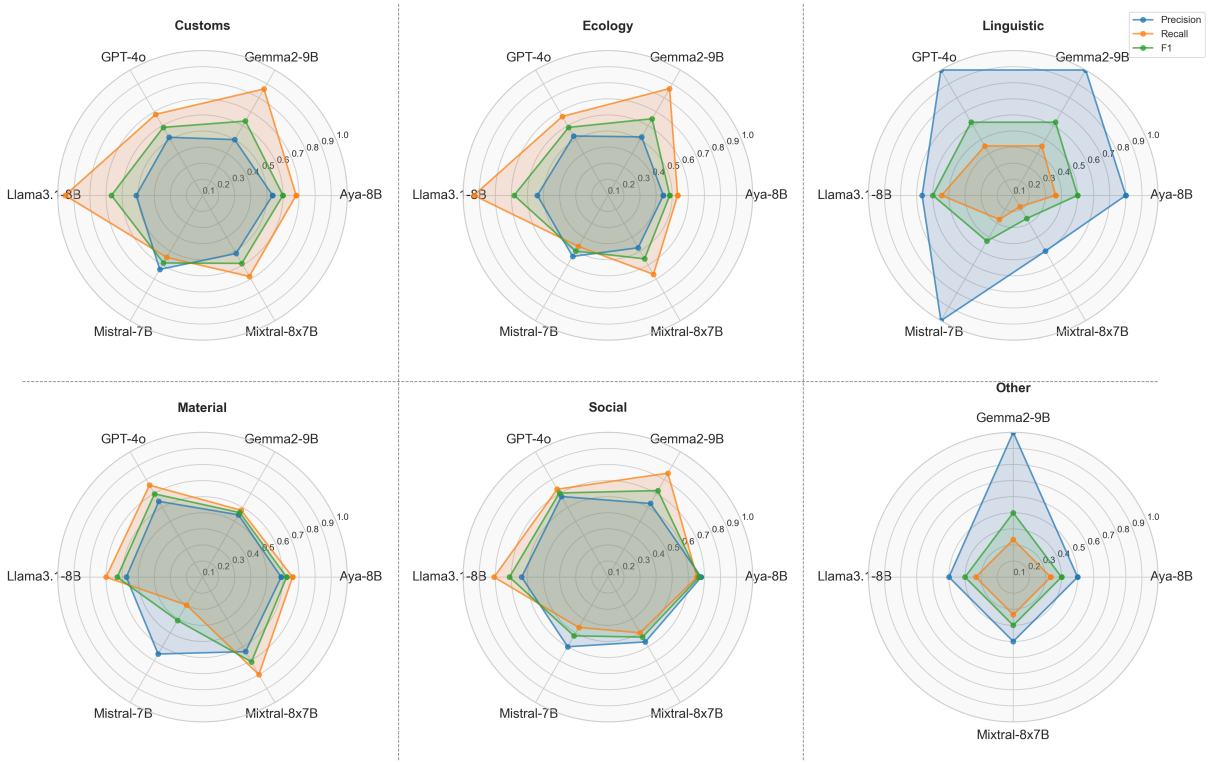

Figure 7: Category-wise overlap scores (all models) for CSI identification between LLMs and users.

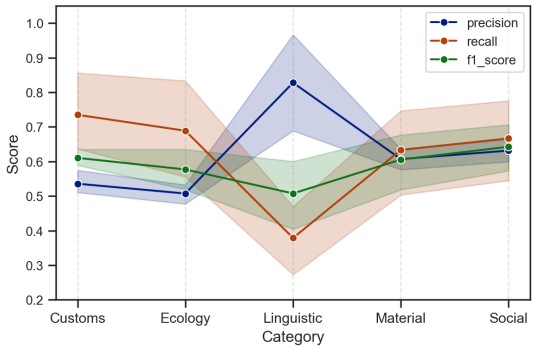

Figure 8: Models' performance by Newmark categories.

5 User and LLM Comparative Analysis

To assess alignment between LLM- and user-identified CSIs, Figure 9 plots the top 20 spans from Users and their rankings in LLMs, and vice versa, with left and right sides showing original and corresponding rankings.

What are the common concepts identified by LLMs and Users? A significant observation from the left-hand plot (Figure 9) is the conceptual mismatch between the CSIs identified by Users and LLMs. Users predominantly find concepts such as *"Heinlein female"* and *"Muir woods"*, which belong to the Material and Ecological dimensions, more difficult to understand, but models assume users would know. On the contrary, on the right-hand plot, we see that LLMs agreeably identify concepts such as *"Ethiopia"*, *"China"*, and *"traditional Arab society"*, which pertain to Ecology and Social dimensions. Moreover, many such concepts are not even identified as difficult by humans (indicated by missing links).

Additionally, as shown in Figure 10, GPT-4o and Mixtral perform best in highly universal categories such as Material and Ecology, which make up most spans, but their performance declines in medium and low ones. In contrast, Llama and Gemma perform worse on highly universal categories yet outperform all models in the medium and low ones.

Our findings indicate that: (i) There are local references in the highly universal Newmark's dimensions, which, although users might find difficult to understand, model rank them low. On the other hand, there are aspects in the less universal dimensions that, although they might be local, are understood by most people in today's globally connected world. Thus indicating that understandability is a very complex concept in a multi-cultural world. (ii) LLMs are unable to capture this tenet, raising questions about how we can align these models pluralistically.

6 Discussion and Conclusion

1. What do models learn about culture? From a psychological and epistemological perspective, models that perform well on Newmark's more universal categories may rely on epistemic projection (Camerer, Loewenstein, and Weber

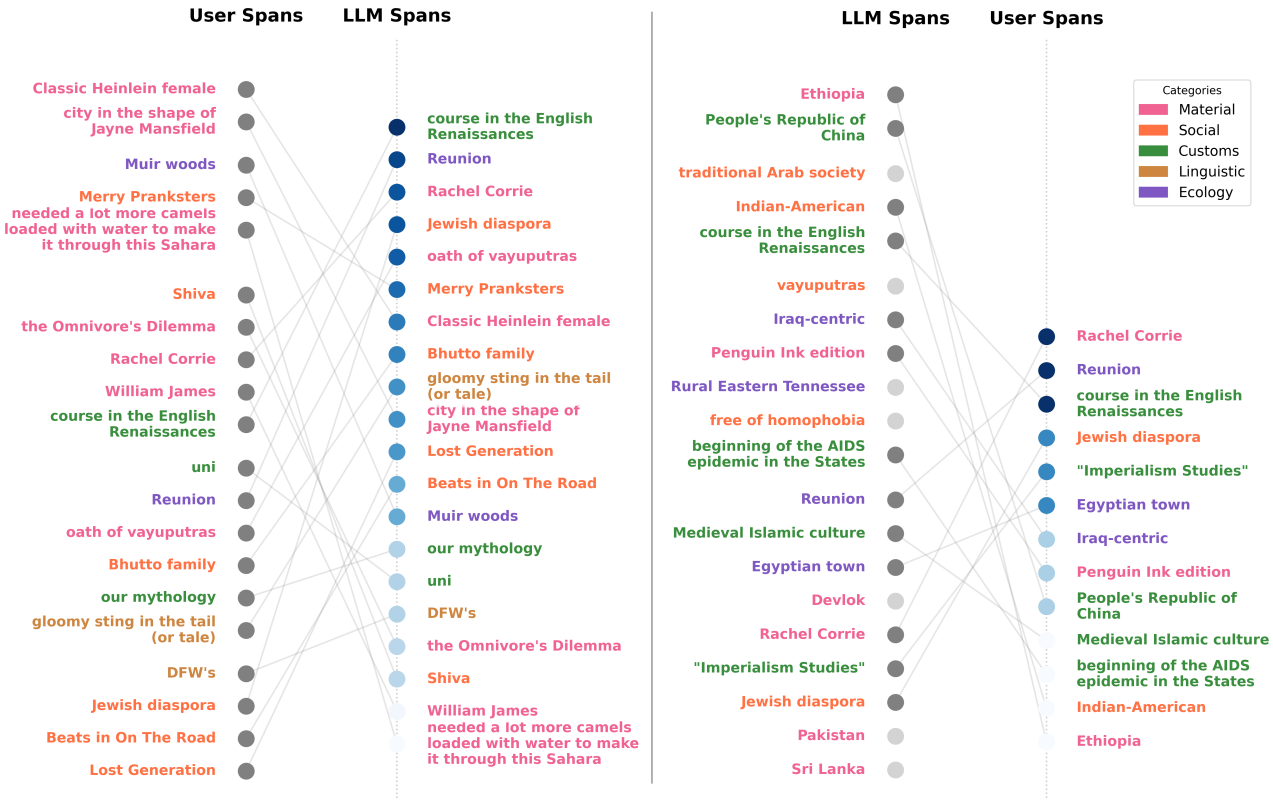

Figure 9: Visualization of the top 20 CSIs annotated by users and LLMs, ranked by frequency and shown with a color gradient.

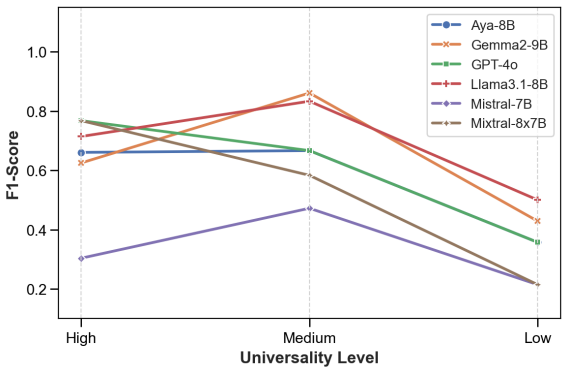

Figure 10: Model performance across universality levels.

1989; Nickerson 1999; Waytz, Cacioppo, and Epley 2010) - assuming that common patterns in their training data apply universally. This likely stems from exposure to dominant cultural knowledge. In contrast, models that perform better on low-universal categories appear more context-sensitive, grounding their predictions in culturally specific patterns. This highlights the challenge of training models that balance generalization with cultural specificity.

2. How can our analysis inform real-world cultural applications? Insights from our analysis can guide the development of better cross-cultural technologies (Pandey et al. 2025; Pokrivcakova 2019; Coenen et al. 2021; Chirkunov et al. 2025). Our findings show that some models overgeneralize, missing culture-specific details, while others capture local nuances shaped by their training data. This suggests that combining models with complementary strengths across Newmark's taxonomy could yield more effective systems - leveraging the universality of some models and the specificity of others to build pluralistically aligned AI that respects cultural diversity and reduces bias in cross-cultural applications.

In summary, unlike prior work that treats culture as a monolithic construct, we examine how distinct cultural dimensions affect the comprehension of online reviews. Using Newmark's taxonomy and the theory of cultural universals, we annotate and analyze the Goodreads-CSI dataset, revealing that understandability varies across cultural dimensions. Evaluations with socio-demographic prompting across different-sized LLMs show uneven performance, raising critical questions about what these models truly capture about culture. These findings underscore the need for ensemble approaches that leverage the strengths of individual models to create more robust cross-cultural AI assistants.

Acknowledgements

This research was supported by the Microsoft Accelerate Foundation Models Research (AFMR) Grant.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adilazuarda, M. F.; Mukherjee, S.; Lavania, P.; Singh, S. S.; Aji, A. F.; O'Neill, J.; Modi, A.; and Choudhury, M. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15763–15784. Miami, Florida, USA: Association for Computational Linguistics.

Aixelá, J. F. 1996. Culture-specific items in translation. *Translation, power, subversion*, 8: 52–78.

AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422. Bangkok, Thailand: Association for Computational Linguistics.

Aryabumi, V.; Dang, J.; Talupuru, D.; Dash, S.; Cairuz, D.; Lin, H.; Venkitesh, B.; Smith, M.; Campos, J. A.; Tan, Y. C.; et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Bassnett, S. 2007. Culture and translation. *A companion to translation studies*, 13–23.

Braylan, A.; Alonso, O.; and Lease, M. 2022. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, 1720–1730.

Brown, D. E. 1991. Human Universals.

Brown, D. E. 2004. Human universals, human nature & human culture. *Daedalus*, 133(4): 47–54.

Camerer, C.; Loewenstein, G.; and Weber, M. 1989. The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, 97(5): 1232–1254.

Chirkunov, K.; Alhafni, B.; Qwaider, C.; Habash, N.; and Briscoe, T. 2025. ARWI: Arabic Write and Improve. In Padmakumar, V.; Gero, K.; Wambsganss, T.; Sterman, S.; Huang, T.-H.; Zhou, D.; and Chung, J., eds., *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, 11–18. Albuquerque, New Mexico, US: Association for Computational Linguistics. ISBN 979-8-89176-239-8.

Chiu, Y. Y.; Jiang, L.; Lin, B. Y.; Park, C. Y.; Li, S. S.; Ravi, S.; Bhatia, M.; Antoniak, M.; Tsvetkov, Y.; Shwartz, V.; et al. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Coenen, A.; Davis, L.; Ippolito, D.; Reif, E.; and Yuan, A. 2021. Wordcraft: A human-AI collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*.

Daghoughi, S.; and Hashemian, M. 2016. Analysis of Culture-Specific Items and Translation Strategies Applied in Translating Jalal Al-Ahmad's" By the Pen". *English language teaching*, 9(4): 171–185.

Dawson, F.; Mosunmola, Z.; Pocker, S.; Dandekar, R. A.; Dandekar, R.; and Panat, S. 2024. Evaluating Cultural Awareness of LLMs for Yoruba, Malayalam, and English. *arXiv preprint arXiv:2410.01811*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dunbar, K. 1995. How scientists really reason: Scientific reasoning in real-world laboratories. *The nature of insight*, 18: 365–395.

Dwivedi, A.; Lavania, P.; and Modi, A. 2023. EtiCor: Corpus for Analyzing LLMs for Etiquettes. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6921–6931. Singapore: Association for Computational Linguistics.

Erten, İ. H.; and Razi, S. 2009. The effects of cultural familiarity on reading comprehension. *Reading in a foreign language*, 21(1): 60–77.

Geert, H.; and Hofstede, G. J. 2004. *Cultures and Organizations. Software of the Mind*, volume 2. ISBN 978-0071439596.

Geertz, C. 2017. *The interpretation of cultures*. Basic books.

Gudykunst, W. B. 2003. *Cross-cultural and intercultural communication*. Sage.

Hall, E. T. 1976. *Beyond culture*. Anchor.

Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, volume 41.

Hogg, M. A. 2016. Group members differ in relative prototypicality: Effects on the individual and the group. *Behavioral and Brain Sciences*, 39: e153.

Hurn, B. J.; Tomalin, B.; Hurn, B. J.; and Tomalin, B. 2013. *What is cross-cultural communication?* Springer.

Jacob, P.; and Uitdenbogerd, A. 2019. Readability of Twitter Tweets for Second Language Learners. In Mistica, M.; Piccardi, M.; and MacKinlay, A., eds., *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, 19–27. Sydney, Australia: Australasian Language Technology Association.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Johnson, R. L.; Pistilli, G.; Menédez-González, N.; Duran, L. D. D.; Panai, E.; Kalpokiene, J.; and Bertulfo, D. J. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*.

Jurafsky, D.; and Martin, J. H. 2020. Sequence labeling for parts of speech and named entities. *Speech and Language Processing*, 71.

Kharchenko, J.; Roosta, T.; Chadha, A.; and Shah, C. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Koto, F.; Mahendra, R.; Aisyah, N.; and Baldwin, T. 2024. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Transactions of the Association for Computational Linguistics*, 12: 1703–1719.

Kuhl, P. K. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11): 831–843.

Leroy, G.; and Kauchak, D. 2014. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, 21(e1): e169–e172.

Leung, A. K.-Y.; and Cohen, D. 2011. Within-and between-culture variation: individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of personality and social psychology*, 100(3): 507.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Li, H.; Jiang, L.; Hwang, J. D.; Kim, H.; Santy, S.; Sorensen, T.; Lin, B. Y.; Dziri, N.; Ren, X.; and Choi, Y. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Li, Z.; Wu, W.; and Li, S. 2020. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6191–6196.

Liu, C.; Koto, F.; Baldwin, T.; and Gurevych, I. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2016–2039. Mexico City, Mexico: Association for Computational Linguistics.

Mann, W. C.; and Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3): 243–281.

Morris, M. W.; Hong, Y.-y.; Chiu, C.-y.; and Liu, Z. 2015. Normology: Integrating insights about social norms to understand cultural dynamics. *Organizational behavior and human decision processes*, 129: 1–13.

Murdock, G. P.; et al. 1961. Outline of cultural materials.

Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R. A.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A. A.; et al. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *arXiv preprint arXiv:2406.09948*.

Nadeau, D.; and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26.

Narváez, I. C.; and Zambrana, J. M. V. 2014. How to translate culture-specific items: a case study of tourist promotion campaign by Turespaña. *The journal of specialised translation*, 21: 71–112.

Newmark, P. 2003. A textbook of translation.

Nickerson, R. S. 1999. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological bulletin*, 125(6): 737.

Owen, L.; Tripathi, V.; Kumar, A.; and Ahmed, B. 2024. Komodo: A Linguistic Expedition into Indonesia's Regional Languages. *arXiv preprint arXiv:2403.09362*.

Pandey, S. K.; Budhiraja, H.; Saha, S.; and Choudhury, M. 2025. CULTURALLY YOURS: A Reading Assistant for Cross-Cultural Content. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; Schockaert, S.; Mather, B.; and Dras, M., eds., *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, 208–216. Abu Dhabi, UAE: Association for Computational Linguistics.

Passonneau, R. 2006. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In Calzolari, N.; Choukri, K.; Gangemi, A.; Maegaard, B.; Mariani, J.; Odijk, J.; and Tapias, D., eds., *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Pokrivcakova, S. 2019. Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education*, 7(3): 135–153.

Putri, R. A.; Haznitrama, F. G.; Adhista, D.; and Oh, A. 2024. Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20571–20590. Miami, Florida, USA: Association for Computational Linguistics.

Rao, A.; Yerukola, A.; Shah, V.; Reinecke, K.; and Sap, M. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Saha, S.; Pandey, S. K.; Gupta, H.; and Choudhury, M. 2025. Reading between the Lines: Can LLMs Identify Cross-Cultural Communication Gaps? arXiv:2502.09636.

Seth, A.; Ahuja, S.; Bali, K.; and Sitaram, S. 2024. DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings*

of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 5323–5337. Torino, Italia: ELRA and ICCL.

Shi, W.; Li, R.; Zhang, Y.; Ziems, C.; Yu, S.; Horesh, R.; Paula, R. A. D.; and Yang, D. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4996–5025. Miami, Florida, USA: Association for Computational Linguistics.

Singh, P.; Patidar, M.; and Vig, L. 2024. Translating Across Cultures: LLMs for Intralingual Cultural Adaptation. In Barak, L.; and Alikhani, M., eds., *Proceedings of the 28th Conference on Computational Natural Language Learning*, 400–418. Miami, FL, USA: Association for Computational Linguistics.

Sperber, A. D.; Devellis, R. F.; and Boehlecke, B. 1994. Cross-cultural translation: methodology and validation. *Journal of cross-cultural psychology*, 25(4): 501–524.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Toti, U. S.; and Hamid, S. A. 2022. An Exploratory Study of Culturally Familiar or Unfamiliar Texts Contributing to Reading Comprehension in EFL Context. *Journal of Language Teaching & Research*, 13(5).

Trivedi, H. 2008. Translating culture vs. cultural translation. In *In translation–reflections, refractions, transformations*, 277–287. John Benjamins Publishing Company.

Wan, Y.; Zhao, J.; Chadha, A.; Peng, N.; and Chang, K.-W. 2023. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9677–9705. Singapore: Association for Computational Linguistics.

Wang, M.; and Lowie, W. 2021. Understanding advanced level academic writing on syntactic complexity. In *35th Pacific Asia Conference on Language, Information and Computation*.

Wang, Y.; Zhu, Y.; Kong, C.; Wei, S.; Yi, X.; Xie, X.; and Sang, J. 2024. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. In Prabhakaran, V.; Dev, S.; Benotti, L.; Hershcovich, D.; Cabello, L.; Cao, Y.; Adebara, I.; and Zhou, L., eds., *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, 1–16. Bangkok, Thailand: Association for Computational Linguistics.

Waytz, A.; Cacioppo, J.; and Epley, N. 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on psychological science*, 5(3): 219–232.

Wibowo, H.; Fuadi, E.; Nityasya, M.; Prasojo, R. E.; and Aji, A. 2024. COPAL-ID: Indonesian Language Reasoning with Local Culture and Nuances. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1404–1422. Mexico City, Mexico: Association for Computational Linguistics.

Xiong, Y.; Racharak, T.; and Nguyen, M. L. 2022. Extractive elementary discourse units for improving abstractive summarization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2675–2679.

Yao, B.; Jiang, M.; Bobinac, T.; Yang, D.; and Hu, J. 2024. Benchmarking Machine Translation with Cultural Awareness. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13078–13096. Miami, Florida, USA: Association for Computational Linguistics.

Zhang, Z.; He, X.; Iyer, V.; and Birch, A. 2024. Cultural Adaptation of Menus: A Fine-Grained Approach. In Haddow, B.; Kocmi, T.; Koehn, P.; and Monz, C., eds., *Proceedings of the Ninth Conference on Machine Translation*, 1258–1271. Miami, Florida, USA: Association for Computational Linguistics.

Zhou, L.; Karidi, T.; Liu, W.; Garneau, N.; Cao, Y.; Chen, W.; Li, H.; and Hershcovich, D. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*.

Zhu, Y.; and Wu, O. 2022. Elementary discourse units with sparse attention for multi-label emotion classification. *Knowledge-Based Systems*, 240: 108114.

A Appendix

A.1 Prompts

Prompt - CSI Identification

AI Rules
- Output response in JSON format
- Do not output any extra text.
- Do not wrap the json codes in JSON or Python markers
- JSON keys and values in double-quotes

You are a cultural mediator who understands all cultures across the world. As a mediator, your job is to identify and translate culturally exotic concepts from texts from an unknown source culture to my culture. I am a well-educated {genre} lover who grew up in {article_urban} urban {country}, which defines my culture. I came across a review of the book '{book}' by {author}, which belongs to the {book_genre} genre. Given my cultural background, perform the following tasks:

Task 1: Identify all culture-specific items (CSIs) from the review text that I might find hard to understand due to my cultural background. CSIs are textual spans denoting concepts and items uncommon and not prevalent in my culture, making them difficult to understand.

Task 2: For each CSI, identify its category from one of the following seven categories:
1. Ecology: Geographical features, flora, fauna, weather conditions, etc.
2. Material: Objects, artifacts, and products specific to a culture, such as food, clothing, houses, and towns.
3. Social: Hierarchies, practices, and rituals specific to a culture.
4. Customs: Political, social, legal, religious, and artistic organizations and practices. Customs, activities, procedures, and concepts.
5. Habits: Gestures, non-verbal communication methods, and everyday habits unique to a culture.
6. Linguistic: Terms unique to a specific language or dialect, including metaphors, idioms, proverbs, humor, sarcasm, slang, and colloquialisms.
7. Other: Anything not belonging to the above six categories.

Task 3: For each CSI, identify its familiarity from one of the following four levels:
1. Familiar: Most people from my culture know and relate to the concept as intended.
2. Somewhat familiar: Only some people from my culture know and relate to the concept as intended.
3. Unfamiliar: Most people from my culture do not know or relate to the concept.
4. Ambiguous: Most people from my culture know the concept, but its interpretation is varied or conflicting.

Task 4: For each CSI, identify its impact on the readability and understandability of the main point of the entire review text from one of the following three levels:
1. High: Greatly hinders the readability and comprehension of the review, making it difficult to convey its main points effectively.
2. Medium: It somewhat affects the readability and comprehension of the review, leading to only partial conveyance of its content.
3. Low: The review text's readability and comprehension will remain unaffected.

Task 5: Within 50 words, detail your reason for highlighting the span as CSI in Task 1 by correlating it with my background.

Task 6: Explain each CSI span within 20 words to make it more understandable to me. Provide facts, examples, equivalences, analogies, etc, if needed.

Task 7: Reformulate the entire text to make it more understandable to me. Keep the length similar to the original review text.

Format your response as a valid Python dictionary formatted as: {'spans': [List of Python dictionaries where each dictionary item is formatted as: {'CSI': <task 1: copy the CSI span from text>, 'category': <task 2: CSI category name>, 'familiarity': <task 3: familiarity level name>, 'impact': <task 4: impact level name>, 'reason': <task 5: reason within 50 words>, 'explanation': <task 6: explain the span within 20 words>}], 'reformulation': <task 7: reformulate entire review text>}. Respond with {'spans': 'None'} if you think I will not find anything difficult to understand.

Text: {review_text}

Prompt - CSI Classification and Newmark Categorisation

AI Rules
- Output response strictly in JSON format.
- Do not output any extra text or explanations outside the JSON.
- Do not wrap the JSON codes in JSON or Python markers.
- JSON keys and values should be enclosed in double quotes.

Annotators from various cultural backgrounds marked spans of the review text as unfamiliar or difficult to understand.
Your task is to classify these spans and provide an explanation.

Task 1: Determine whether each span is cultural (specific to a cultural context) or non-cultural (not specific to any culture).

Task 2: For each span, select the most appropriate category from the taxonomy below. Provide a concise explanation for your choice in no more than 50 words.

Taxonomy:
1. Ecology: Geographical features, flora, fauna, weather conditions, etc.
2. Material: Objects, artifacts, and products specific to a culture, such as food, clothing, houses, and towns.
3. Social: Hierarchies, practices, and rituals specific to a culture.
4. Customs: Political, social, legal, religious, and artistic organizations and practices. Customs, activities, procedures, and concepts.
5. Habits: Gestures, non-verbal communication methods, and everyday habits unique to a culture.
6. Linguistic: Terms unique to a specific language or dialect, including metaphors, idioms, proverbs, humor, sarcasm, slang, and colloquialisms.
7. Other: Anything not belonging to the above six categories.

Review text: {review_text_for_id}
Spans: {spans_string}

Format your answer as a JSON dictionary, where each key is a span from the input spans and the values are a nested dictionary with the format as follows:
{{ ”<span 1>”: {{ ”class”: ”<cultural/non-cultural>”, ”category”: ”<taxonomy category>”, ”explanation”: ”<reason for taxonomy categorisation>” }}, ”<span 2>”: {{ ”class”: ”<cultural/non-cultural>”, ”category”: ”<taxonomy category>”, ”explanation”: ”<reason for taxonomy categorisation>” }}, ... }}

A.2 Additional Analysis

id	book_title	author	country	genre	review_text	std_span	category	user_id	user_location	region	education	age	reading_habits	preferred_genre
2745	Job: A Comedy of Justice	Robert A. Heinlein	USA	fiction	I know I shouldn't, but I love Heinlein ,..., skip it.	Classic Heinlein female	['Material']	e450 bc49 eaa7	USA India USA	urban rural urban	high_school undergraduate college	40-60 18-25 18-25	moderate occasional moderate	fiction both non-fiction

Table 2: Sample Rows From the Standardized Dataset

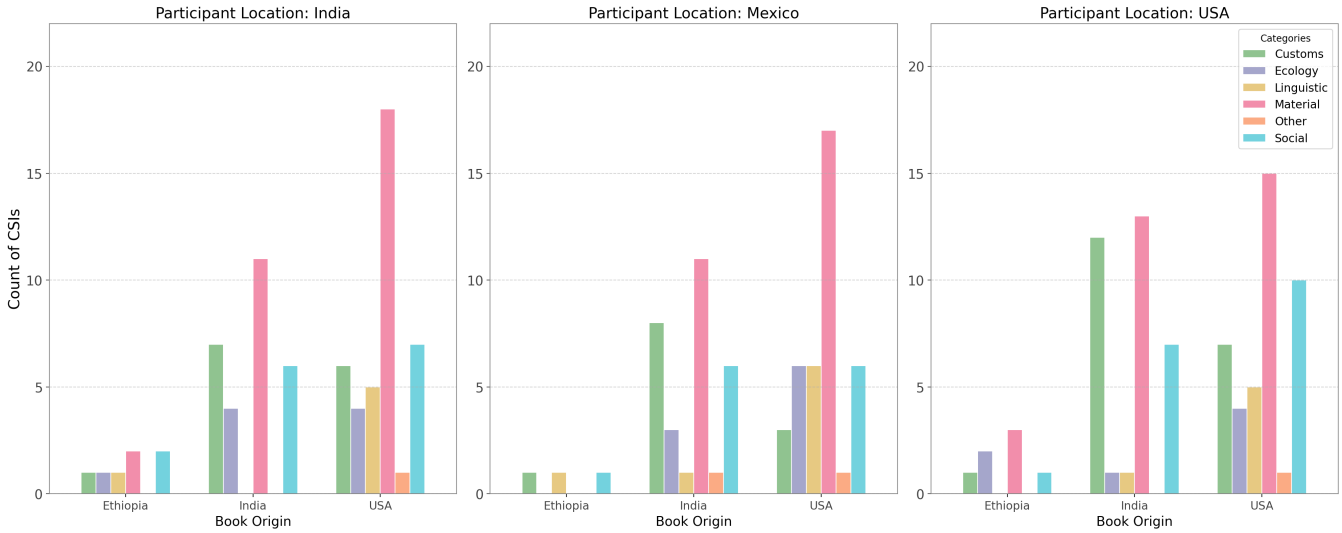

Figure 11: User location-wise distribution of CSIs across Newmark's dimensions for reviews by the book's country of origin.

B Open-Source and Closed-Source Models Used

Model Name	Hugging Face Link
Open Source Models	
Llama-3.1-8B-Instruct (Dubey et al. 2024)	meta-llama/Llama-3.1-8B-Instruct
Gemma-2-9B-it (Team et al. 2024)	google/gemma-2-9b-it
Aya-23-8B (Aryabumi et al. 2024)	CohereForAI/aya-23-8B
Mistral-7B-Instruct-v0.2 (Jiang et al. 2023)	mistralai/Mistral-7B-Instruct-v0.2
Mixtral-8x7B-Instruct-v0.1 (Jiang et al. 2024)	mistralai/Mixtral-8x7B-Instruct-v0.1
Closed Source Models	
GPT-4o (Achiam et al. 2023)	gpt-4o-2024-05-01-preview