Reading between the Lines: Can LLMs Help Bridge Cross-Cultural Communication Gaps?

Anonymous ACL submission

Abstract

In a rapidly globalizing and digital world, content such as book and product reviews created by people from diverse cultures are read and consumed by others from different corners of the world. In this paper, we investigate the extent and patterns of gaps in understandability of book reviews due to the presence of culturally-specific items and elements that might be alien to users from another culture. Our user-study on 60 book reviews from Goodreads reveal that 83% of the reviews had at least one culture-specific difficult-to-understand element. We also evaluate the efficacy of GPT-4o in identifying such items, given the cultural background of the reader; the results are mixed, implying a significant scope for improvement.

1 Introduction

Whether performing outdoor activities, household chores, or surfing the internet, our perception and understanding of the world are varied. Each of us has a distinct worldview, which causes us to understand and internalize information distinctly. Although our knowledge, and hence our worldview, are individual-specific (Collins and Gentner, 1987; Jonassen and Henning, 1999; Denzau et al., 1994), our culture does contribute to the variation by shaping our worldview and understandability (Bender and Beller, 2013; Cole and Packer, 2019). This effect of culture is also evident in the online world, where we might find text from distinct online sources hard to understand due to a lack of common ground between us and the writer of the text (Meyer, 2014; Korkut et al., 2018). For example, people unfamiliar with the Arabic culture might not understand the meaning of the dishes "Machboos" and "Luqaimat" from the review text "The Machboos was perfectly spiced, and the Luqaimat was a real treat." Or, someone unfamiliar with the US culture and not exposed to fast food might not understand that "golden arches" refers to "McDonald's" in the text "Let's go to the golden arches for a quick bite.".

The recent advancements in AI and NLP make us question whether and how Large Language Models (LLMs) can help overcome such barriers in online cross-cultural communication. Here, we consider a use case where LLMs act as cultural mediators by identifying, categorizing, and explaining text spans *(Culture-Specific Items)* from online reviews that users from distinct cultures might find hard to understand. Although numerous use cases of LLMs in cross-cultural communication are conceivable (Singh et al., 2024a; Liu et al., 2023), the following fundamental and overarching questions need answering before implementing them across any domain. Firstly, although studies discuss the importance of cross-cultural communication (Bourges-Waldegg and Scrivener, 1998), the need for LLMs and AI assistants in general for cross-cultural communication has not been studied. Secondly, the cross-cultural performance parity of LLMs in such settings is unclear and unmeasured. Although studies have shown LLMs to exhibit Anglo-centric biases (Johnson et al., 2022; Dwivedi et al., 2023), this does not mean they perform well in those cultures and poorly in others. We refer to this as the *equitability*, defined as a model's capability to maintain the parity between its performance and the user's need across cultures. Through our specific use-case of LLMs as a cultural mediator, we answer these fundamental questions, a prerequisite of using them to overcome cross-cultural communication barriers.

The contributions of this paper are as follows:

- We perform a user-study spanning participants from the USA, Mexico, and India to measure their difficulty in understanding Goodreads English reviews of books from the USA, India, and Ethiopia.
- We establish the nature and extent of cross-cultural communication gap for Goodreads

reviews that an AI reading assistant could help bridge.

- We evaluate GPT-4o as an *equitable* cultural reading assistant for identifying, categorizing, and explaining culture-specific items from the book reviews for urban educated users with different literary preferences from each country.
- We share our human (gold) and GPT-4o-annotated (silver) dataset that identifies cultural spans from book review texts.

2 Culture: A Primer

Culture is a complex concept, and is defined as the "Way of life of a collective group of people distinguishing them from other groups" (Blake, 2000; Monaghan et al., 2012; Parsons, 1972; Münch and Smelser, 1992), making it experiential (Geertz, 1973; Bourdieu, 1977). It provides a common ground to groups through shared experiences and creates distinct "worldviews" between people from different cultures (Collins and Gentner, 1987; Jonassen and Henning, 1999; Denzau et al., 1994; Bang et al., 2007; McHugh et al., 2008) by shaping their cognition (Mishra, 2001; Nisbett and Norenzayan, 2002; Bender and Beller, 2013; Cole and Packer, 2019). This difference in perception due to culture affects all aspects of life, where literary works composed by individuals from one culture might not be perceived similarly by individuals from a different culture and hamper cross-cultural communication (Meyer, 2014; Korkut et al., 2018). In NLP, these variations can broadly be due to differences in linguistic form and style, common ground, aboutness, and values (Hershcovich et al., 2022). Among demographies, these differences vary across dimensions or *semantic proxies*, ranging from physical items such as food and materials to abstract constructs like emotions and values (Thompson et al., 2020; Adilazuarda et al., 2024). Newmark (2003) terms such concepts as *Culture-Specific Items (CSIs)*, identifying and addressing which is crucial for cross-cultural communication. Thus, to technologically facilitate cross-cultural communication, it is essential to identify and gauge the extent of the variation of such CSIs, which can inform appropriate technological advancements for mitigation.

3 User Study

To measure the extent of the difficulty in comprehension due to culture and establish a need for cross-cultural tools, we conducted an online user study where we showed participants book reviews written in English and asked them to highlight text spans they did not understand. The length of the spans was unrestricted and could be individual words or even multiple sentences. Our goal was to quantify how much of things people do not understand and, out of that, how much is cultural (CSIs). Measuring this cultural knowledge gap can serve as an evaluation benchmark for tools that facilitate cross-cultural communication.

3.1 Dataset

The study was performed over 57 English review texts of books originating from Ethiopia, India, and the USA, which represent distinct cultures per Inglhart-Welzel's world cultural map (Inglehart and Welzel, 2010). The reviews were sampled from the Goodreads dataset collected by Wan and McAuley (2018); Wan et al. (2019) in 2017, a sizeable publicly available book-review dataset spanning 2.3 million books from multiple countries and languages. We sampled the 57 reviews in two steps. First, we prompted GPT-4o (Achiam et al., 2023) to list the top 20 famous authors of fictional and non-fictional books from each country and restricted to reviews of books from these authors with at least 50 reviews and review lengths between 50 and 200 words. We randomly sampled 3,100 reviews while ensuring at least 250 reviews from each country. Next, to ensure enough cultural diversity in our study, for each review text, we prompted GPT 3.5 to identify CSIs for people from our country of study, India, Mexico, and the USA, with varied literary preferences: fiction or non-fiction. Section 4.1 discusses the prompt in detail. The model identified CSIs in almost all review texts, with varied levels of familiarity. We randomly sampled 57 reviews containing at least one unfamiliar CSI as the final dataset for our study. Table 1 shares the data statistics after each filtering step.

3.2 Method

3.2.1 Pilot Studies

We conducted two small-scale internal and one external pilot study to determine the questionnaire.

Pilot 1: Similar to the task formulation for GPT

Country	Genre	# Reviews	Avg Length	Avg # CSI
Ethiopia	fiction	206 | 6	100.8 | 73.7	6.2 | 5.0
Ethiopia	non-fiction	51 | 8	102.3 | 101.4	6.2 | 6.8
India	fiction	393 | 7	101.9 | 98.6	5.9 | 5.6
India	non-fiction	209 | 11	100.9 | 96.4	6.6 | 6.1
USA	fiction	1487 | 15	104.5 | 91.1	6.0 | 5.8
USA	non-fiction	752 | 10	101.8 | 134.6	6.3 | 5.8

Table 1: Review statistics after each filtering step (1 | 2)

3.5 in the Step 2 filtering, in the first pilot, we showed a review text to six participants and asked them to identify spans mentioning CSIs that they did not understand or were unfamiliar with, given their cultural background. The participants were researchers from India, Ethiopia, and the USA. The key takeaways from the study were: (i) Although the participants cumulatively identified 13 CSIs (detailed in Table 2 Appendix A), the question forced ethnocentrism, where one had to internally assume a source culture of the review text before determining CSIs that might be common in the source culture but exotic to their culture. (ii) Such a formulation forced participants to generalize their culture, which they found difficult. (iii) It is difficult to distinguish the unknowability of CSIs borne out of cultural difficulty and otherwise.

Pilot 2: Incorporating their feedback, we reformulated the task in Pilot 2. Instead of asking what one does not understand due to their culture, we asked participants to identify spans they generally found hard to understand or were unfamiliar with. Additionally, we asked several other questions (detailed in Appendix A) with multiple choices to capture their understanding of the review from different aspects and provoke their thoughts around the task. Pilot 2 encompassed three internal participants, and each reviewed five texts. Overall, we observed from the study that (i) The participants found the task more natural than before as the formulation did not enforce taking any cultural viewpoint. (ii) Since the questions revolved around general understandability, they highlighted spelling mistakes, complex words, and vocabulary as things they did not understand. (iii) As depicted in Table 3, the length of the spans highlighting the same concept varied by participant.

Pilot 3: Incorporating the feedback from Pilot 2 participants, we refined a few questions and conducted a pilot study in Prolific with 13 participants from India and the USA. We used Google Forms for the survey, comprising a single review text. We observed that (i) Similar to Pilot 2, the highlighted span length varied between annotators for the same concept, where some annotators marked the entire review text instead of only highlighting spans. (ii) The annotations varied greatly. Some annotators did not find any concepts difficult, and some marked the entire text as hard. A few participants highlighted the author's name and book title in the review text as difficult-to-comprehend concepts.

Since performing span-level annotations is difficult in Google Forms, we transitioned to Label Studio as the annotation interface to ease the annotation difficulty in the actual experiments.

3.2.2 Actual Study

We used Label Studio (Tkachenko et al., 2020-2022) with Prolific[1] for the actual study. Our goal was to collect annotated text spans from review texts that the participants found difficult to understand or were unfamiliar with. In addition to this, we asked follow-up questions about the review texts to gather further insights. We hosted multiple instances of Label Studio, each secured with individual login credentials, and provided them to annotators through Prolific. Once participants logged into their assigned instance, they were directed to the annotation interface, as shown in Appendix D.1.

The study first explicitly sought consent from the participants, allowing only those who agreed to proceed. Following this, they completed screener validation questions designed to assess their backgrounds, including where they spent most of their time before the age of 18 and their highest level of education. Next, the study asked demographic and preference questions like their age, reading habits, and literary genre preferences. The consent forms, screener validation questions, and demographic and preference forms are in Appendix D.2. The survey questions were carefully constructed by adhering to established methodologies and best practices by Beck (2023) that ensure comprehensive coverage of relevant variables and a high-quality data annotation.

Next, participants were shown a review text and asked if they had read the reviewed book and if they were familiar with the author and their other works. Next, the study asked them to mark the spans of text they did not understand, using different levels of unfamiliarity. Additionally, the study asked five more questions related to the review text to capture their understanding of the review text from diverse aspects. The span-marking task and the seven questions about the review text are in Appendix D.3. After completing the annotations, the study directed participants to the end of the survey section (depicted in Appendix D.4), which shared a survey code to collect their reward in Prolific. Participants were also provided with a detailed guideline document to assist them in performing the annotations, ensuring they followed all necessary principles following Prolific's and our requirements.

---

[1] www.prolific.com

The study was completed by 50 participants: 8 from India, 22 from Mexico, and 20 from the USA. To prevent fatigue, which could reduce the quality of responses, we split the data annotation tasks into manageable batches of 12 data points per batch. Participants were paid 7.5 to 8 GBP per batch, and the median time to complete a batch of annotations was approximately 30 to 40 minutes. This payment structure aligns with Prolific's and country-wise standard wage guidelines, ensuring fair compensation for their time and effort. We also ensured that each data point received annotations from multiple annotators, ensuring the reliability and diversity of the collected data. Appendix C shares additional demographic details of the participants.

3.3 Post Processing

We collect a total of 761 responses across all three countries. Although crowdsourcing platforms make it easy to conduct user studies at scale for specific demography, the quality of the annotations is sometimes questionable (Hsueh et al., 2009; Huynh et al., 2021; Eyal et al., 2021; Douglas et al., 2023). We perform the following quality checks and filter out bad-quality answers by removing responses where (i) People mention they do not understand or are not familiar with certain aspects of the review text but do not mark any spans. (ii) People mention they understand everything and are familiar with everything but still highlight spans. The filtration step removes 11% of cases, yielding 668 responses.

3.3.1 LLM-assisted post processing

Since the study only asked participants to identify spans they found hard to understand, we processed the results using GPT-4o to categorize them as CSIs (cultural spans) or non-CSIs. We first created a taxonomy by interactively prompting ChatGPT using the template outlined in Appendix B.1. We first provided a few review texts and their highlighted spans and instructed ChatGPT to generate categories and subcategories as the initial taxonomy. Then, we iteratively fed the model its proposed taxonomy along with marked spans from an additional 20 annotators, enabling it to continuously update and refine the taxonomy based on this new input. This iterative process allowed the model to enhance the taxonomy effectively.

The final taxonomy classified each span into "Cultural and Linguistic," "Cultural and Non-Linguistic," "Non-Cultural and Linguistic," "Non-Cultural and Non-Linguistic," and "Poor Annotation", listed in Appendix B.2. Adhering to the human-in-the-loop validation process by Shah et al. (2023), we consulted two linguistic experts to review and validate the generated taxonomy. This step ensured clarity and robustness, preventing over-generalization and ambiguity in the definitions and confirming that our taxonomy effectively captures the diversity and complexity of the spans while maintaining consistency across annotations.

We used GPT-4o to annotate each identified span using our taxonomy. The model inputs each review text and a list of all spans highlighted as non-understandable by participants from any country and associates each span with one of the five classes. Also, since the pilot studies indicated that human annotations tend to be noisy, with some marked spans being superfluous while some reflecting low-quality annotations, we instructed the model to cluster the spans semantically based on the context. This clustering process helped filter out low-quality and irrelevant annotations while resolving ambiguities related to span lengths, allowing us to focus on meaningful data. Appendix B.2 depicts the prompt used to post-process using GPT-4o.

3.3.2 Evaluation

The GPT-4o annotations were evaluated by two human experts comprising graduate students from Linguistics and Computer Science. Each expert evaluated 120 random annotations and had 54 overlapping samples. They assessed the span type and the cluster assigned by GPT-4o by marking the correctness with binary flags. The evaluators agreed on 48 out of 54 (89%) cases for span type and 52 out of 54 (96%) for the cluster assignment. For the 54 overlapping annotations, at least one evaluator found the span type labels correct in 94% of cases and the cluster labels correct in all cases. Although there is an approximate 10% margin of error for incorrect span type classification and sub-

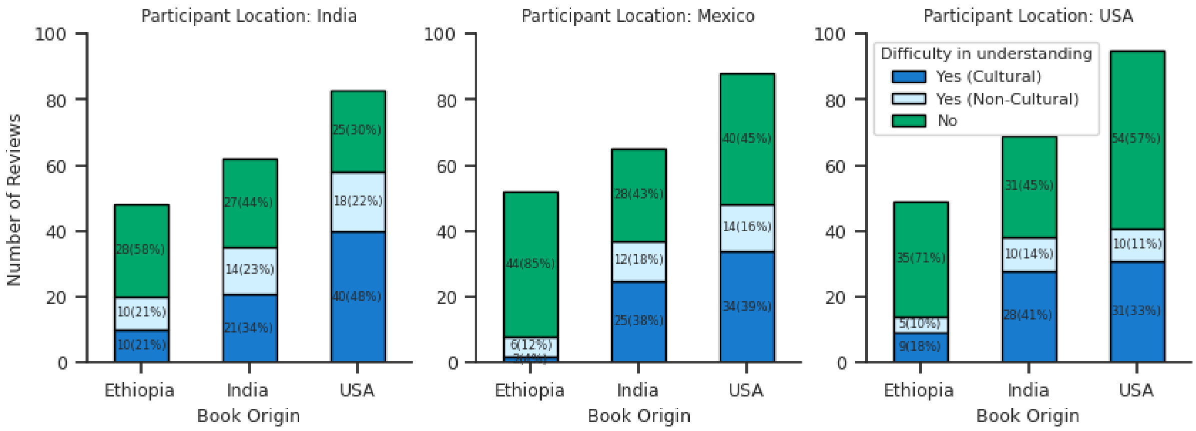


Figure 1: Participant location-wise difficulty in understanding book reviews by country of origin of the book.

sequently 15% error margin for cluster assignment, the high agreement scores indicate that GPT-4o's annotations are reliable. Hence, we cluster all spans identified by the participants from the study as per the clusters assigned by the model. Henceforth, we perform all span-level analysis on its cluster representative.

3.4 Analysis

1. What is the overall and cultural difficulty of understanding concepts from reviews? Out of 611 total evaluations encompassing all participants from all countries, 312 (51%) evaluations had no difficult-to-understand spans highlighted. The remaining ones had at least one difficult-to-understand span, and 200 (33%) had some culturally difficult spans. At a review level, all 60 reviews had at least one difficult-to-understand element, where 50 reviews (83%) had culturally difficult-to-understand elements, while others were non-cultural. On average, a participant evaluated 12 reviews, of which six were hard to understand. Four of the six hard-to-understand reviews had cultural elements.

2. What is the overall difficulty by each culture and how much of it is cultural? Figure 1 plots the proportions of hard-to-understand evaluations by a participant's location and the book's country of origin. We observe the following: (i) Participants from all locations found the reviews from Ethiopia easier to understand and did not contain difficult spans, and evaluators from Mexico did not find such spans in 85% of cases. Also, in each country, the proportion of the evaluations highlighting culturally difficult spans was lower than the reviews of books from other countries. (ii) Considering participants and books from India and the USA, culturally difficult spans are lowest for the participants from their own country. For evaluators from India, 34% of unfamiliar spans from the reviews of Indian books were cultural, compared to 48% for USA book reviews. For the USA, 41% of unfamiliar spans from the reviews of Indian books were cultural, compared to 33% for USA book reviews, (iii) For evaluators from Mexico, the proportions of evaluations comprising non-difficult spans and culturally and non-culturally difficult spans were similar for books originating from India and the USA. (iv) Overall, the highest proportion of evaluations highlighting culturally difficult spans were for US-origin books by evaluators from India (48%), which was followed by India-origin book reviews by evaluators from the USA (41%).

This cross-cultural gap in understanding indicates that people from all locations might benefit from tools such as cultural reading assistants and translation systems that enable communication across cultures. Also, the cultural difficulty in understanding book reviews from one's own country indicates the dynamic nature and long tail of culture, where culture varies within a country, which also motivates the need for such tools.

3. How much of the reviews and spans are cultural? We measure the review and span-level agreements between participants from the same country and compare them against the agreements when the participants are from different countries. We create all combinations of inter and intra-country annotators and compute Krippendorf's alpha to measure the inter-annotator agreement within a country and between each pair of countries. Figure 2 captures the agreement between annota-

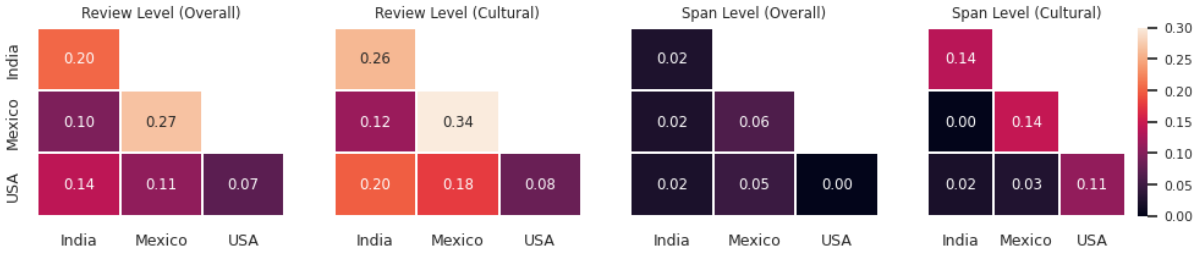


Figure 2: Inter Annotator Agreement at Review Level and Span Level across Countries.

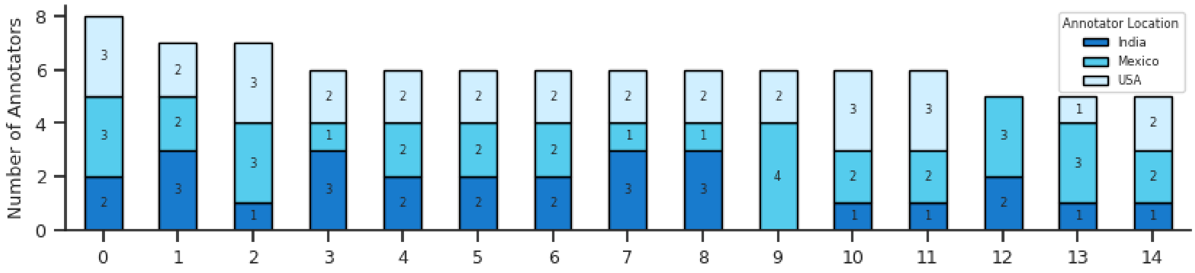


Figure 3: Top 15 Span Clusters identified by participants across countries. Corresponding span text in Table 4.

tors within a country and across countries. We observe the following at a review level: (i) Annotators from the same country agree more amongst themselves on which reviews are generally hard to understand than annotators across countries except for the USA. We observe a similar trend for reviews containing CSIs, where within a country, annotators agree more on which reviews contain difficult CSIs than inter-country. (ii) For generally difficult and CSI-containing reviews, participants from Mexico agree more among themselves (0.27/0.34) than with reviewers from India and the USA. (iii) Similarly, Indian participants agree more amongst themselves (0.2/0.26) than from other countries. However, they align more with the USA (0.14/0.2) reviewers than Mexicans (0.10/0.12). (iv) Interestingly, reviewers from the USA agree more with Indian and Mexcian participants than among themselves. (v) At a span level, there is a lack of consensus across all countries, which testifies that understandability is individual-specific, which was already indicated by our pilot studies.However, annotators from the same country seem to agree more on culturally difficult spans than other kinds of spans, which indicates that although different annotators might find different things hard to understand, there is a level of consensus on CSIs that they find hard to understand. This also implies that CSIs are a set of harder-to-understand constructs in text and, therefore, are good targets for circumventing the cold-start problem (Hu et al., 2008) of personalization for user-facing systems.

4. What are the most common spans identified by people from each culture? Figure 3 plots the top 15 spans frequently highlighted by participants as hard to understand. Except for a few cases, participants from all countries find these spans hard to understand. Table 4 lists the corresponding span texts of the figure. Table 6 lists the most frequent unfamiliar text spans by each participant location. Interestingly, cultural spans referring to "Al Gore" and "Hemmingway" are frequently marked as unfamiliar by participants from India, whereas participants from the USA and Mexico find concepts like "topper from engineering institute," which are colloquial terms frequently used in India, as hard to understand. Given that the understandability of text varies by region and culture, we next experiment with GPT-4o and benchmark its performance as a cultural assistant.

4 Benchmarking GPT-4o as a Cultural Assistant

We designed a prompt to evaluate GPT-4o (Achiam et al., 2023) as a cross-cultural reading assistant by assigning it the role of a cultural mediator who understands all cultures in the world. We tasked the model to identify, categorize, adapt, and explain CSIs from book reviews from unknown source cultures to the target culture of the reader defined by their country and book-genre preference.

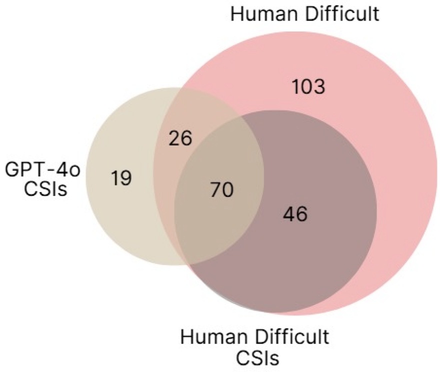


Figure 4: Overlap between Human-identified difficult spans and GPT-4o-identified CSIs

4.1 Experiment Setup

We carefully constructed a prompt encompassing the desired attributes of a cross-cultural assistant. As depicted in Appendix F.1, we task GPT-4o with the following tasks: (i) Identifying and explaining all CSIs from the review text that may be difficult to understand due to a user's cultural background. (ii) For each CSI, identify its category from one of the five categories by Newmark (2003). Following Singh et al. (2024b), we additionally added two extra categories: linguistic and others. (iii) For each CSI, identify its familiarity as one of the following four levels: (a) Familiar- Known by most people, (b) Somewhat familiar- Known by some people, (c) Unfamiliar- Unknown to most people, and (d) Ambiguous- Known but in an opposite way. (iv) Identify the impact on the comprehension of the main review point and reformulate the entire text per the user's background. Since prompting LLMs to generate explanations of their response has shown success (Wang et al., 2022, 2023), we additionally prompt the model to reason their identification of each CSI.

We use the Azure OpenAI platform for our experiments and set the model's temperature to 0. We prompt GPT-4o to obtain outputs in the required format as detailed in Figure F.1. We perform post-processing on the obtained response. We remove any JSON or Python markers in the response and convert the response to a dictionary with the required keys. If the post-processing does not yield the response in the required format, we again prompt GPT-4o to format the response.

4.2 Analysis

We cluster the GPT-4o-identified spans by matching their embedding-based cosine similarity scores (Reimers and Gurevych, 2019) with the span clusters from the user study (similarity threshold=0.5 and measured their overlap. As depicted in Figure 1, combining all the CSIs identified by GPT-4o across all countries and genre preferences, 96 (83%) overlap with the 245 unique spans marked as difficult to understand by humans, and 70 (60%) overlap with the 116 user-identified culturally difficult spans. Interestingly, 26 (22%) of the GPT-4o-identified CSIs, although considered difficult, are not deemed cultural by humans. We hypothesize this to be confounded by two things. First, as discussed in Section 4, the GPT-4o-based cultural classification of the user-identified spans has an error margin of approximately 15%. Second, since the user study encompassed 50 participants, their identified spans most likely do not capture all the difficult spans for someone from the same culture, contrary to the CSIs identified by GPT-4o, which are more encompassing. This generalization of the model is further evident from Figure 5, which measures the overlap of spans between fiction and non-fiction readers from each country. For all countries, the overlap between GPT-4o-identified spans for both groups is consistently higher than those identified by the users. Table 5 lists a few such spans.

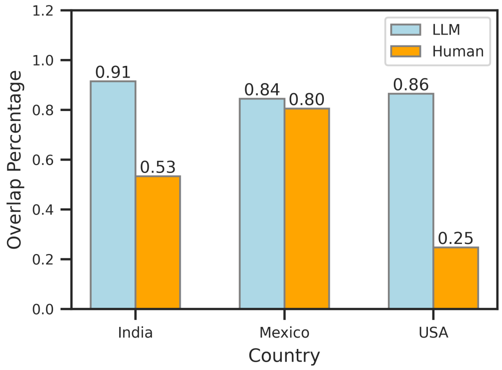


Figure 5: Overlap percentage of fiction and non-fiction spans across countries.

Considering the user-identified cultural spans as the ground truth, we compute the precision and recall scores of GPT-4o-identified CSIs at different levels of country and genre-preference combinations. Figure 6 plots the scores, where we observe that (i) For all countries, GPT-4o's recall (approx. 0.65) is higher than the precision (approx. 0.49), signifying that the model is generally able to capture a variety of CSIs. Also, the recall for all countries is comparable. (ii) We see a similar

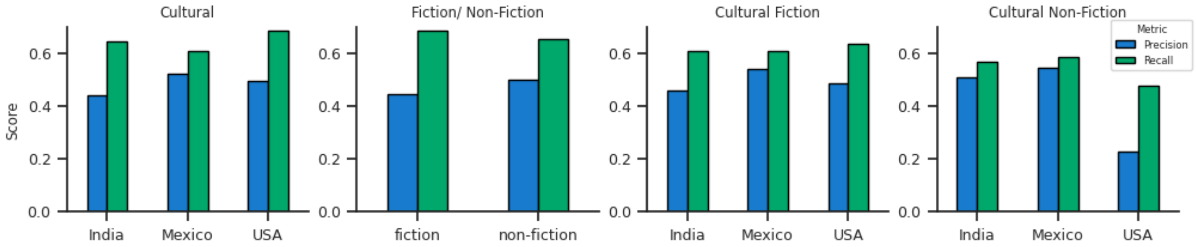


Figure 6: Precision and recall of the overlap between user-identified culturally difficult spans and GPT-4o-identified CSIs

trend across genre preferences and combinations of country and genre, where the recall values are higher than precision, signifying the model's variational awareness. However, the model's performance drops for non-fiction readers from the USA.

Our results indicate that GPT-4o seems equitably aligned with users from all countries. However, with an overall precision of 0.49 and recall of 0.65 the magnitude of the alignment is low, indicating much room for improvement.

5 Related Work

Communicating across cultures is crucial in today's global world. Although several theories and frameworks exist that define the nature of cross-cultural communication (Gudykunst, 2003; Tannen, 1983; Hurn et al., 2013; Gardner, 1962), practical applications implementing such concepts are still nascent. Computationally, a considerable amount of work has been done in HCI in defining the considerations of cross-cultural tools and systems (Bourges-Waldegg and Scrivener, 1998; Kyriakoullis and Zaphiris, 2016; Heimgärtner, 2018). Cultural adaptation is also extensively studied in machine translation (Bassnett, 2007; Trivedi, 2007; Sperber et al., 1994), Aixelá (1996) introduced the term "Culture-Specific Items" (CSIs), and Newmark (2003) further presented a taxonomy of CSIs and proposed methods to tackle each type of CSI. Recently, Zhang et al. (2024) introduces the ChineseMenuCSI dataset where they integrate the translation theories to create an annotated dataset with CSI vs Non-CSI labels. Singh et al. (2024b) leverages open-sourced LLMS to adapt CSIs from the USA TV series Friends, for an Indian audience.

Advances in LLMs have garnered studies that detect biases in LLMs (Tao et al., 2023; Kharchenko et al., 2024a; Duan et al., 2023; Lin et al., 2024). Using curated cultural datasets, most methods probe LLMs and test their knowledge and reasoning capabilities in culture-specific settings (Nadeem et al., 2021; Nangia et al., 2020; Wan et al., 2023; Jha et al., 2023; Li et al., 2024b; Cao et al., 2023; Tanmay et al., 2023; Rao et al., 2023; Kovač et al., 2023). Some methods (Kharchenko et al., 2024b; Li et al., 2024a; Dawson et al., 2024) also analyze the model-generated responses along theoretical frameworks such as Hofstede's cultural dimensions (Hofstede, 2001; Geert and Hofstede, 2004) and measure their proximity with cultures, where high proximity indicates better value alignment between the nearby cultures and the values portrayed by the model's response. Most of these methods necessitate constructing cultural-specific test beds (Wang et al., 2024; Rao et al., 2024; Myung et al., 2024; Zhou et al., 2024; Putri et al., 2024; Davani et al., 2024; Wibowo et al., 2024; Owen et al., 2024; Chiu et al., 2024; Liu et al., 2024; Koto et al., 2024).

6 Conclusion

Although cultural knowledge is necessary for LLMs to operate in cross-cultural settings, its impact on practical cross-cultural settings is unmeasured. Hence, determining LLMs' cross-cultural equitability primarily based on their factual knowledge about cultures is limiting. To address this gap, here we take a distinct approach where we perform a thorough user study of the cross-cultural understandability of Goodreads reviews to understand the need for cross-cultural AI tools. We then evaluate GPT-4o as a cultural reading assistant that can identify and adapt culture-specific items to the culture of a user. Our evaluations reveal that there are indeed lots of cultural things in Goodreads reviews that might not be comprehensible to people from distinct cultures. Also, GPT-4o performs equitably across the studied cultures. However, with overall low precision and recall scores, the amount of alignment is low, indicating room for improvement.

Limitations

In this work, we perform a user study on 57 review texts from Goodreads. We limit our study to demographics of India, USA, and Mexico, in the English language. We consider genre preference as the only semantic proxy under study. A large-scale multilingual study across more diverse cultures worldwide and at a larger intersection of demographic and semantic proxies would help strengthen the findings of our work. We limit our evaluation to using GPT-4o since it is known to be better than other models, a full-fledged assessment of other closed-source and open-source models needs to be performed.

Ethical Considerations

We do not capture any personally identifiable information regarding the users involved in user study. We follow standard ethical guidelines (Rivers and Lewis, 2014) and do not attempt to track users across sites or deanonymize them. We also adhere to the minimum wage policies of all demographics involved in the user study. We filter out all the offensive/adult content from the review texts presented to the users. We plan to release our data and the annotations to foster further research on developing AI assistants for efficient cross-cultural communication.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. *Preprint*, arXiv:2403.15412.

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

Megan Bang, Douglas L Medin, and Scott Atran. 2007. Cultural mosaics and mental models of nature. *Proceedings of the National Academy of Sciences*, 104(35):13868–13874.

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

Jacob Beck. 2023. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17.

Andrea Bender and Sieghard Beller. 2013. Cognition is… fundamentally cultural. *Behavioral Sciences*, 3(1):42–54.

Janet Blake. 2000. On defining the cultural heritage. *International & Comparative Law Quarterly*, 49(1):61–85.

Pierre Bourdieu. 1977. *Outline of a Theory of Practice*. Cambridge Studies in Social and Cultural Anthropology. Cambridge University Press.

Paula Bourges-Waldegg and Stephen AR Scrivener. 1998. Meaning, the central issue in cross-cultural hci design. *Interacting with computers*, 9(3):287–309.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

Michael Cole and Martin Packer. 2019. Culture and cognition. *Cross-cultural psychology: Contemporary themes and perspectives*, pages 243–270.

Allan Collins and Dedre Gentner. 1987. How people construct mental models. *Cultural models in language and thought*, 243(1987):243–265.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. *Preprint*, arXiv:2404.10857.

Fiifi Dawson, Zainab Mosunmola, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2024. Evaluating cultural awareness of llms for yoruba, malayalam, and english. *Preprint*, arXiv:2410.01811.

Arthur T Denzau, Douglass C North, et al. 1994. Shared mental models: ideologies and institutions. *KYKLOS-BERNE-*, 47:3–3.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

Yucong Duan, Fuliang Tang, Kunguang Wu, Zhendong Guo, Shuaishuai Huang, Yingtian Mei, Yuxing Wang, Zeyu Yang, and Shiming Gong. 2023. Ranking of large language model (llm) regional bias.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. Eticor: Corpus for analyzing llms for etiquettes. *arXiv preprint arXiv:2310.18974*.

Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, pages 1–20.

George H Gardner. 1962. Cross cultural communication. *The Journal of social psychology*, 58(2):241–256.

Hofstede Geert and Gert Jan Hofstede. 2004. *Cultures and Organizations. Software of the Mind*, volume 2.

Clifford Geertz. 1973. *The interpretation of cultures*. Basic books.

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

Rüdiger Heimgärtner. 2018. Culturally-aware hci systems. *Advances in culturally-aware intelligent systems and in cross-cultural psychological studies*, pages 11–37.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

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

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee.

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

Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *Preprint*, arXiv:2111.05241.

Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. *World Values Survey*.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

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*.

David H Jonassen and Philip Henning. 1999. Mental models: Knowledge in the head and knowledge in the world. *Educational technology*, pages 37–42.

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

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

Perihan Korkut, Mustafa Dolmacı, and Burcu Karaca. 2018. A study on communication breakdowns: Sources of misunderstanding in a cross-cultural setting. *Eurasian Journal of Educational Research*, 18(78):139–158.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *Preprint*, arXiv:2404.01854.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *Preprint*, arXiv:2307.07870.

Leantros Kyriakoullis and Panayiotis Zaphiris. 2016. Culture and hci: a review of recent cultural studies in hci and social networks. *Universal Access in the Information Society*, 15:629–642.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Huihan Li, Liwei Jiang, Jena D. Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *Preprint*, arXiv:2309.08591.

Anna P McHugh, Jennifer L Smith, and Winston R Sieck. 2008. Cultural variations in mental models of collaborative decision making. *Macrocognition and naturalistic decision making*, pages 141–158.

Erin Meyer. 2014. *The culture map: Breaking through the invisible boundaries of global business*. Public Affairs.

Ramesh C Mishra. 2001. Cognition across cultures. *The handbook of culture and psychology*, pages 119–135.

Leila Monaghan, Jane E Goodman, and Jennifer Robinson. 2012. *A cultural approach to interpersonal communication: Essential readings*. John Wiley & Sons.

Richard Münch and Neil J Smelser. 1992. *Theory of culture*. University of California Press Berkeley.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Peter Newmark. 2003. A textbook of translation.

Richard E Nisbett and Ara Norenzayan. 2002. Culture and cognition. John Wiley & Sons.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia's regional languages. *Preprint*, arXiv:2403.09362.

Talcott Parsons. 1972. Culture and social system revisited. *Social Science Quarterly*, pages 253–266.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *Preprint*, arXiv:2402.17302.

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

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3:38.

Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. 2023. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*.

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

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

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

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

Deborah Tannen. 1983. Cross-cultural communication.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

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.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Harish Trivedi. 2007. Translating culture vs. cultural translation. *Benjamins translation library*, 71:277.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

J. Wang, Qiushi Sun, Nuo Chen, Xiang Lorraine Li, and Ming Gao. 2023. Boosting language models reasoning with chain-of-knowledge prompting. *ArXiv*, abs/2306.06427.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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. *Preprint*, arXiv:2311.16421.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2024. Copal-id: Indonesian language reasoning with local culture and nuances. *Preprint*, arXiv:2311.01012.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural adaptation of menus: A fine-grained approach. *arXiv preprint arXiv:2408.13534*.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *Preprint*, arXiv:2404.06833.

A Appendix

Identified CSI	Customs	Habits	Linguistic	Material	other	Social
crazy damsel in distress			4			
kickass model			4			
purple for my taste			4			
gateway book			2			
paranormal/ vampire element	1		2		2	
Harlequins			1	3		
leave much room			1			
Minus			1			
romance for the ages			1			
slow and atmospheric		1	1			
weak female-strong male trope					2	1

Table 2: Pilot 1 results

Review Text	Evaluator	Span	Familiarity
I always look for the range of landscape and plant descriptions in books. Gargash brought the realities of inland and coastal climate differences to life - especially as related to seasonal changes and shelter -living in agricultural shelter inland and in urban shelter on the coast. Do you want air conditioning or not? Is there such a thing as green air conditioning? Is there such a thing as sustainable air conditioning? Good questions, no?	1	living in agri - cultural shelter inland	Unfamiliar
	1	green air conditioning	Unfamiliar
	1	sustainable air conditioning	Unfamiliar
	2	sustainable air conditioning	Unfamiliar
	2	agricultural shelter	Ambiguous
	2	urban shelter	Ambiguous
	3	I always ... shelter on the coast.	Unfamiliar
	3	Do you ... Good questions, no?	Unfamiliar

Table 3: Pilot 2 identified spans

Id	Span Type	Spans
0	Cultural	{’Muir woods’ ’John Muir?’, ’John Muir? Sure, Muir woods,’}
1	Non-Cultural	{’logical meaning ofost’, ’ofost’}
2	Cultural	{’"the oath of vayuputras"’, ’the title of the book is "oath" of vayuputras! \n what oath?’}
3	Cultural	{’"People\’s Republic of China".’, ’former state head of "People\’s Republic of China".’}
4	Cultural	{’Vlad the Impaler, a city in the shape of Jayne Mansfield’, ’Vlad the Impaler, a city in the shape of Jayne Mansfield’}
5	Cultural	{"Beats in On The Road to Ken Kesey’s Merry Pranksters.", "the Beats in On The Road to Ken Kesey’s Merry Pranksters."}
6	Cultural	{’("Stickeen"). "The Moral Equivalent of War",’, ’("Stickeen"). "The Moral Equivalent of War", by William James,’}
7	Cultural	{"’Infinite Jest’.", ’Penguin Ink edition’, "attempting ’Infinite Jest’."}
8	Cultural	{’curmudgeon’, ’old-fashioned curmudgeon’, ’Dalrymple is an old-fashioned curmudgeon’}
9	Non-Cultural	{’abridged’, "I should’ve went for the abridged version.", ’abridged version.’}
10	Cultural	{’The female character is such a classic Heinlein female.’, ’classic Heinlein female.’, ’a classic Heinlein female.’}
11	Cultural	{’Nature and Selected Essays by Ralph Waldo Emerson’}
12	Non-Cultural	{’amphibians,’, ’ocean acidification, invasive species, ecosystem fragmentation, poaching,’, ’ecosystem fragmentation, poaching’}
13	Cultural	{’DFW intro,’, ’DFW’}
14	Cultural	{’- "Evil" had risen in Meluha, Shiva had to stop it.’, ’"Evil" had risen in Meluha, Shiva had to stop it.’}

Table 4: Span texts of the top 15 span clusters identified by the participants across all countries

Spans
’tennis lingo’, ’western diseases’, ’Imperialism Studies’, ’poaching’, ’anthropologic’, ’topper’, ’engineering institute’, ’bogged down’, ’Iraq-centric’, ’courtroom dramas’

Table 5: Spans which humans identify as difficult to understand (non-cultural) whereas LLM identifies it as cultural.

Participant Location	Book Country	Span	Span Type
India	India	{’"People\’s Republic of China".’, ’former state head of "People\’s Republic of China".”}	Cultural
	USA	{"I understand that Al Gore’s new movie ends on a slightly more upbeat tone,", "Al Gore’s new movie"}	Cultural
	USA	{"Hemingway’s Lost Generation...and", "Hemingway’s Lost Generation...and a similar lack of purpose."}	Cultural
	USA	{’Drum Skin: Fragments on R.A. Harris:’, ’Drum Skin: Fragments on R.A. Harris: "All Art is Junk."’}	Cultural
	India	{’curmudgeon’, ’old-fashioned curmudgeon’, ’Dalrymple is an old-fashioned curmudgeon’}	Cultural
	USA	{"’Infinite Jest’.", ’Penguin Ink edition’, "attempting ’Infinite Jest’."}	Cultural
	USA	{"her cockatiel’s fame", "cockatiel’s fame"}	Non-Cultural
	India	{’logical meaning ofost’, ’logical meaning ofost of the question that comes to our mind about our mythology.’}	Non-Cultural
Mexico	India	{’Fisssssss...ooouiuu.....!’}	Cultural
	India	{’eye of a chauffeur.’, ’chauffeur.’}	Cultural
	India	{’"the oath of vayuputras"’, ’"oath"’, ’the title of the book is "oath" of vayuputras! \n what oath?’, ’oath?’}	Cultural
	USA	{’DFW intro,’, ’DFW’}	Cultural
	USA	{’amphibians,’, ’ocean acidification, invasive species, ecosystem fragmentation, poaching,’}	Non-Cultural
	USA	{’Muir woods,’, ’John Muir?’,’John Muir? Sure, Muir woods,’}	Cultural
	USA	{"’easiest’ thing of DFW’s that I’ve read,", "DFW’s", "Probably the ’easiest’ thing of DFW’s that I’ve read,"}	Cultural
	USA	{’abridged’, "I should’ve went for the abridged version.", ’abridged version.’}	Non-Cultural
	India	{’I myself is a topper from engineering institute’, ’I myself is a topper’,}	Non-Cultural
USA	India	{’"the oath of vayuputras"’, ’"oath"’, ’the title of the book is "oath" of vayuputras! \n what oath?’, ’oath?’}	Cultural
	USA	{’Nature and Selected Essays by Ralph Waldo Emerson’}	Cultural
	India	{’Medieval Islamic culture,’}	Cultural
	USA	{ ’sentimentality about War Being Bad.’, ”militaristic vision’}	Non-Cultural
	USA	{’Muir woods,’, ’John Muir?’,’John Muir? Sure, Muir woods,’}	Cultural
	India	{’especially the parts about Sri Lanka, Pakistan & Reunion.’, ’Sri Lanka, Pakistan & Reunion.’}	Cultural
	USA	{’The female character is such a classic Heinlein female.’, ’classic Heinlein female.’, ’a classic Heinlein female.’}	Cultural
	India	{’I started reading it long after watching the movie 3 idiots’, ’the movie 3 idiots’, ’3 idiots’}	Cultural
	India	{’a topper from engineering institute’, ’topper from engineering’, ’topper from engineering institute’}	Non-Cultural

Table 6: Top difficult to understand spans by each participant country

B Taxonomy Generation and Span Categorization

B.1 Initial Taxonomy Generation Template

Prompt

```
### CONTEXT
Annotators were asked to mark spans of text that they found unfamiliar. Below is a list of categories and subcategories. I will provide you with a list of spans, and for each span, please specify the corresponding category and subcategory under which it falls.
{current taxonomy generated by the model}
### FULL TEXT
{source text on which spans were marked}
### LIST OF SPANS
{list of spans marked by the annotator}

### OUTPUT FORMAT
[[Span 1, Category; Subcategories], ..., [Span X, Category; Subcategories]]
```

B.2 Final Span Categorization And Clustering Template

Prompt

Please categorize the following text spans based on the provided taxonomy and cluster similar spans.
Annotators from different demographics marked these spans as unfamiliar or non-understandable from a review text.
Please carefully classify each span and provide a brief explanation for your classification.
Also, please cluster similar spans.

Taxonomy:
1. Cultural, Linguistic: Differences related to language use, idioms, dialects, or phrasing specific to a culture.
2. Cultural, Non-linguistic: Cultural references such as items, traditions, customs, or social norms unrelated to language.
3. Non-Cultural, Linguistic: General language issues such as grammar, vocabulary, or syntax that are not culturally specific.
4. Non-Cultural, Non-linguistic: Non-cultural factors like complex concepts, technical jargon, or other knowledge-related gaps.
5. Poor Quality Annotations: Vague or Ambiguous Annotation, Superfluous Annotation, Marked majority span.

Task 1: For each span from the review text below, select the appropriate category from the above taxonomy and briefly explain your reasoning behind the classification in 20 words.
Task 2: Cluster the spans which are similar.

Review text: {text}
Spans: {spans}

Format your answer as a Python dictionary as follows:
{{"Task 1": {<span number>: {'Type': <taxonomy number>, 'Explanation': <explain classification reason in 20 words>}}, "Task 2": [<python list of lists of span numbers>]}}

C Demographic of Annotators from USA, Mexico, India

C.1 Demographic of Annotators from USA

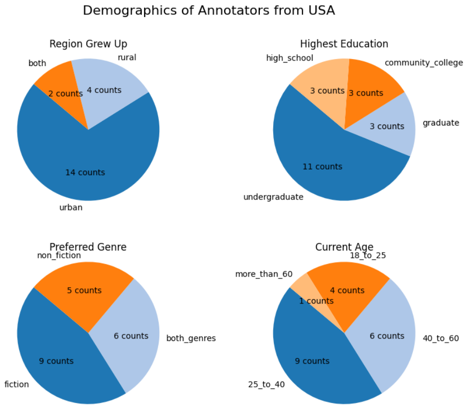


Figure 7

C.2 Demographic of Annotators from Mexico

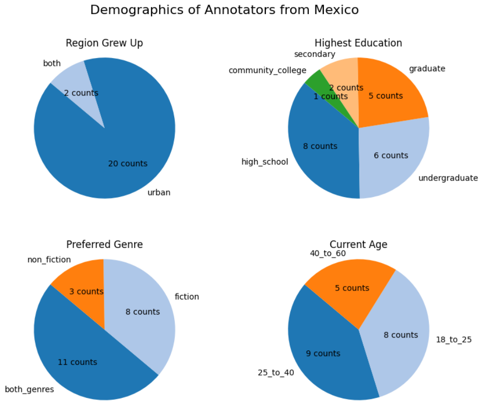


Figure 8

C.3 Demographic of Annotators from India

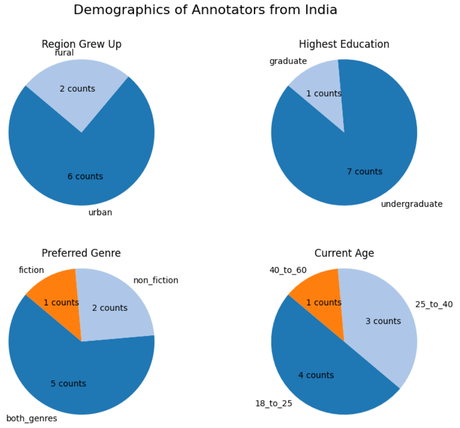


Figure 9

D Label Studio Setup

D.1 Annotation Interface After Signing In

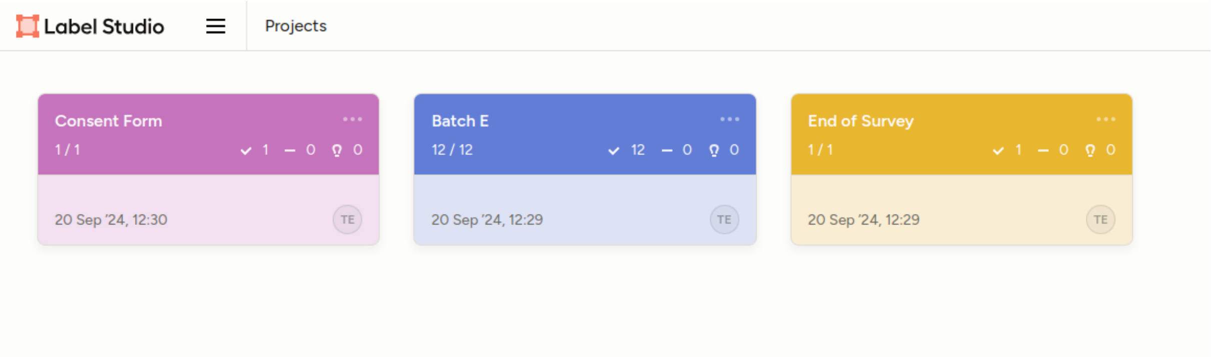
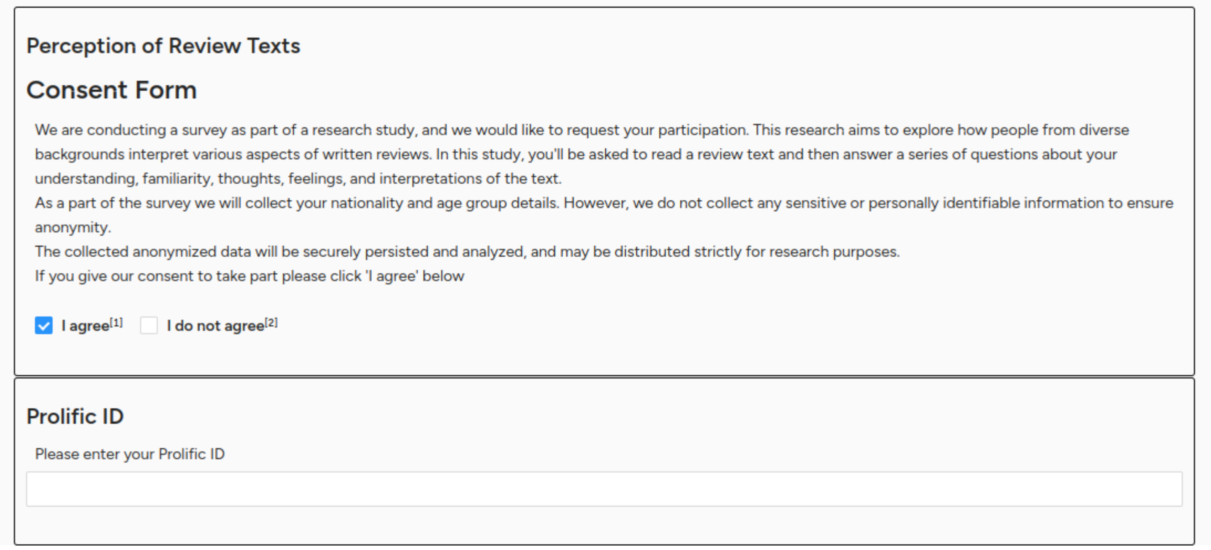

Label Studio
Projects
Consent Form
1 / 1
✓ 1 – 0 0
20 Sep '24, 12:30
TE
Batch E
12 / 12
✓ 12 – 0 0
20 Sep '24, 12:29
TE
End of Survey
1 / 1
✓ 1 – 0 0
20 Sep '24, 12:29
TE

Figure 10

D.2 Consent, Screener Validation and Demographic Interface



Perception of Review Texts

Consent Form

We are conducting a survey as part of a research study, and we would like to request your participation. This research aims to explore how people from diverse backgrounds interpret various aspects of written reviews. In this study, you'll be asked to read a review text and then answer a series of questions about your understanding, familiarity, thoughts, feelings, and interpretations of the text.

As a part of the survey we will collect your nationality and age group details. However, we do not collect any sensitive or personally identifiable information to ensure anonymity.

The collected anonymized data will be securely persisted and analyzed, and may be distributed strictly for research purposes.

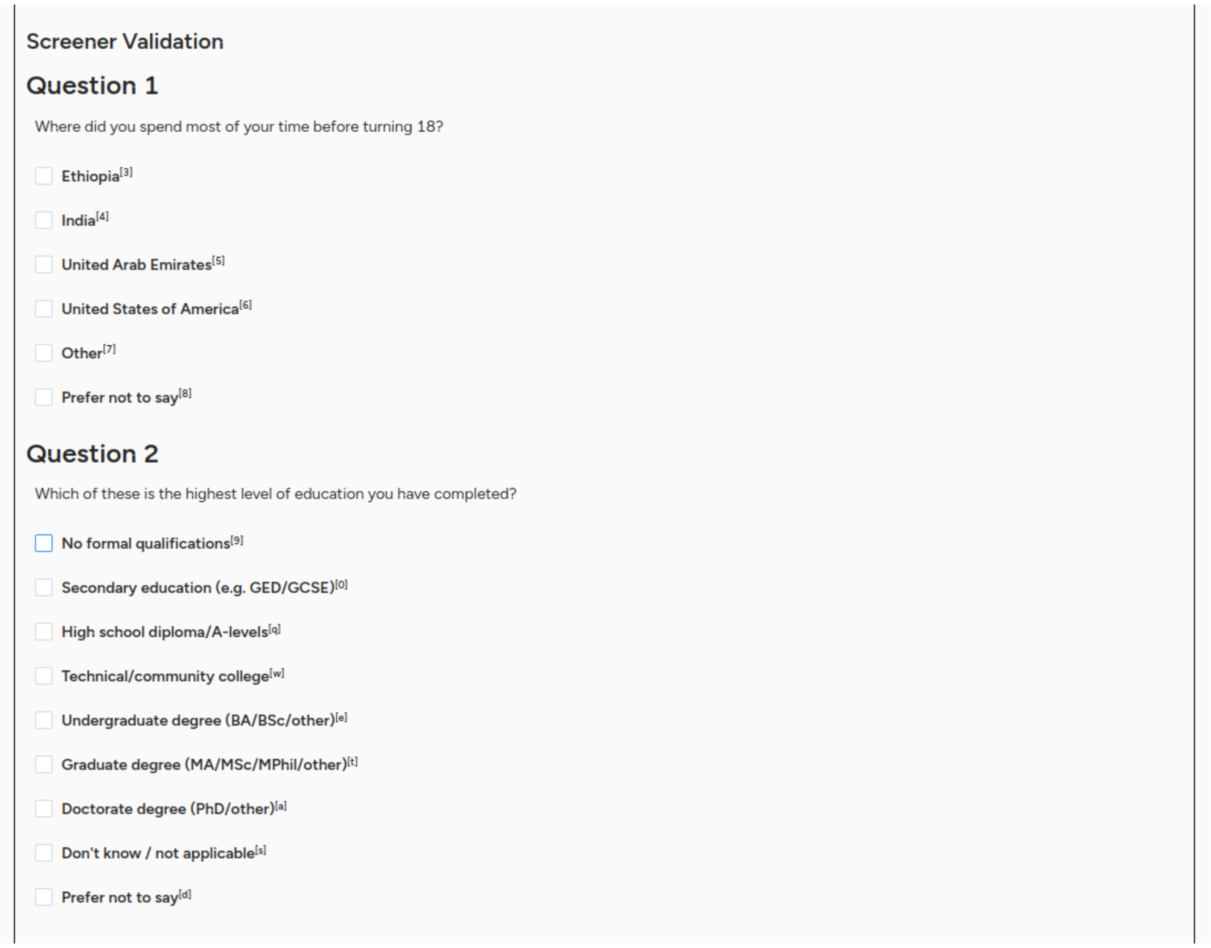
If you give our consent to take part please click 'I agree' below

☑ I agree[1] ☐ I do not agree[2]

Prolific ID

Please enter your Prolific ID

Figure 11



Screener Validation

Question 1

Where did you spend most of your time before turning 18?

☐ Ethiopia[3]

☐ India[4]

☐ United Arab Emirates[5]

☐ United States of America[6]

☐ Other[7]

☐ Prefer not to say[8]

Question 2

Which of these is the highest level of education you have completed?

☐ No formal qualifications[9]

☐ Secondary education (e.g. GED/GCSE)[0]

☐ High school diploma/A-levels[q]

☐ Technical/community college[w]

☐ Undergraduate degree (BA/BSc/other)[e]

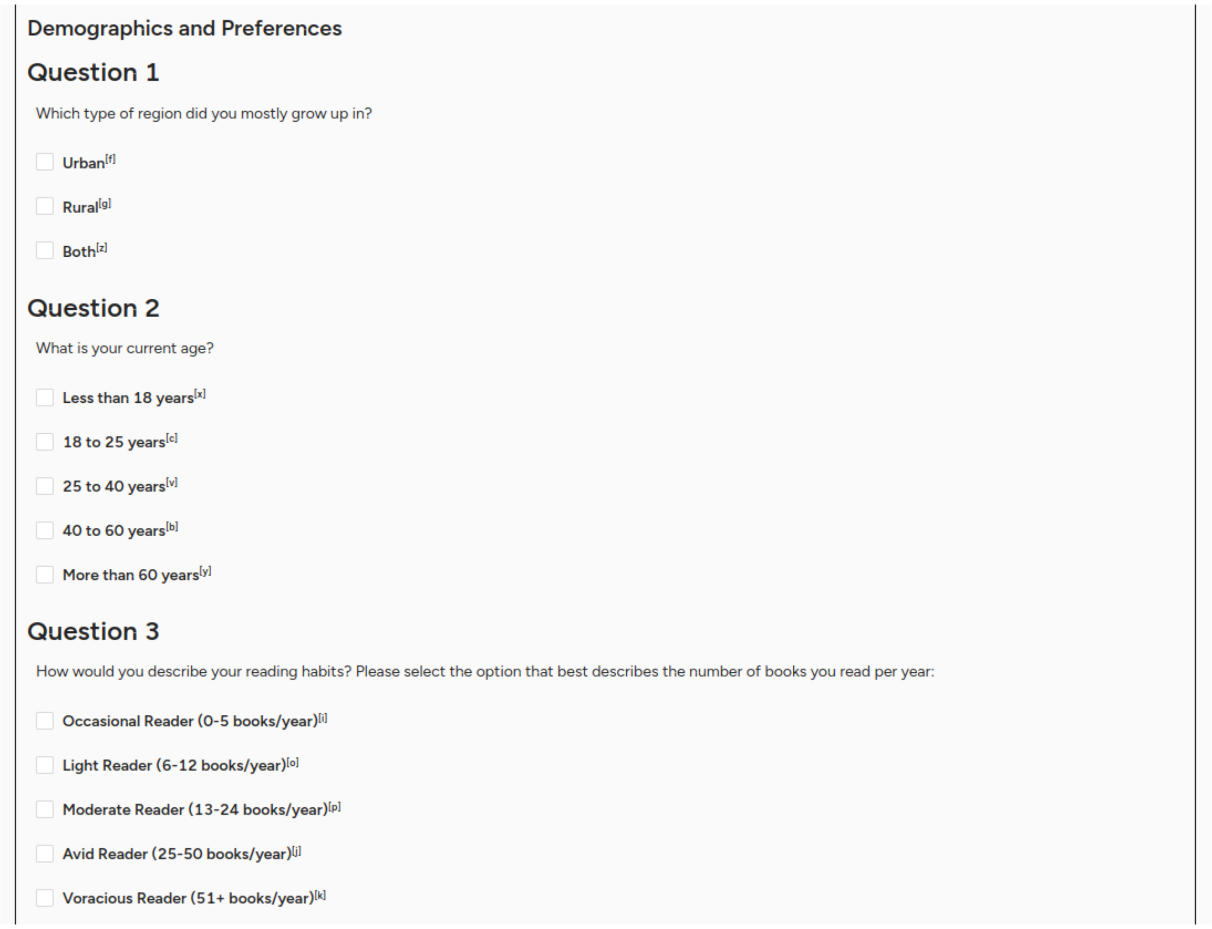
☐ Graduate degree (MA/MSc/MPhil/other)[t]

☐ Doctorate degree (PhD/other)[a]

☐ Don't know / not applicable[s]

☐ Prefer not to say[d]

Figure 12



Demographics and Preferences

Question 1

Which type of region did you mostly grow up in?

☐ Urban[f]

☐ Rural[g]

☐ Both[z]

Question 2

What is your current age?

☐ Less than 18 years[x]

☐ 18 to 25 years[c]

☐ 25 to 40 years[v]

☐ 40 to 60 years[b]

☐ More than 60 years[y]

Question 3

How would you describe your reading habits? Please select the option that best describes the number of books you read per year:

☐ Occasional Reader (0-5 books/year)[i]

☐ Light Reader (6-12 books/year)[o]

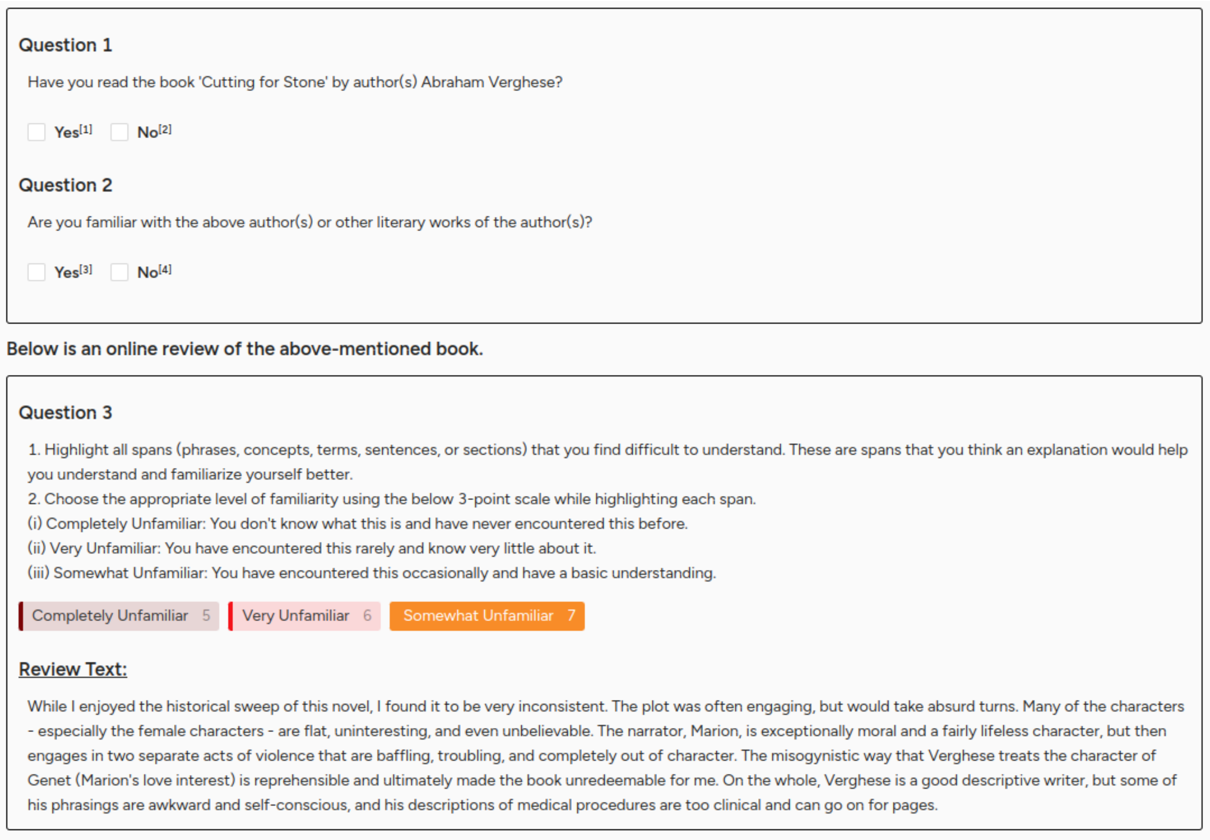
☐ Moderate Reader (13-24 books/year)[p]

☐ Avid Reader (25-50 books/year)[j]

☐ Voracious Reader (51+ books/year)[k]

Figure 13

D.3 Unfamiliarity Span Marking and Other Relevant Questions Interface



Question 1

Have you read the book 'Cutting for Stone' by author(s) Abraham Verghese?

☐ **Yes**[1] ☐ **No**[2]

Question 2

Are you familiar with the above author(s) or other literary works of the author(s)?

☐ **Yes**[3] ☐ **No**[4]

Below is an online review of the above-mentioned book.

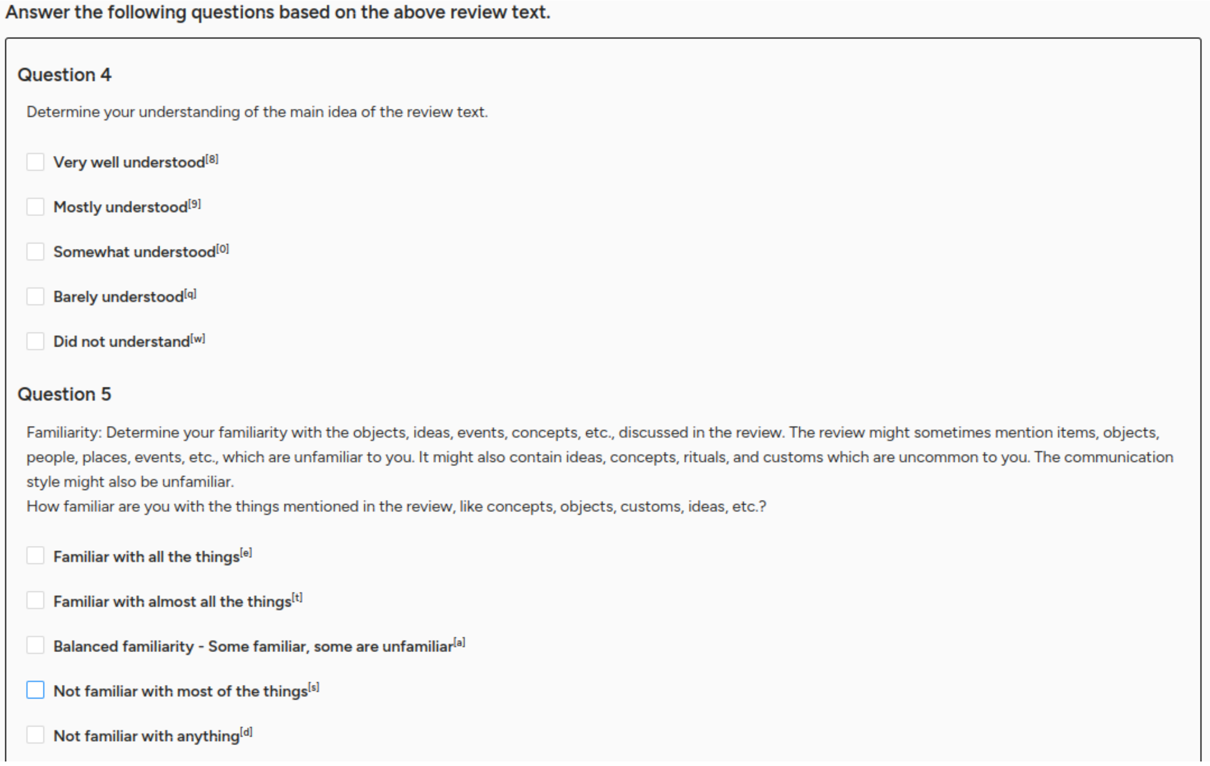
Question 3

1. Highlight all spans (phrases, concepts, terms, sentences, or sections) that you find difficult to understand. These are spans that you think an explanation would help you understand and familiarize yourself better.
2. Choose the appropriate level of familiarity using the below 3-point scale while highlighting each span.
(i) Completely Unfamiliar: You don't know what this is and have never encountered this before.
(ii) Very Unfamiliar: You have encountered this rarely and know very little about it.
(iii) Somewhat Unfamiliar: You have encountered this occasionally and have a basic understanding.

Review Text:

While I enjoyed the historical sweep of this novel, I found it to be very inconsistent. The plot was often engaging, but would take absurd turns. Many of the characters - especially the female characters - are flat, uninteresting, and even unbelievable. The narrator, Marion, is exceptionally moral and a fairly lifeless character, but then engages in two separate acts of violence that are baffling, troubling, and completely out of character. The misogynistic way that Verghese treats the character of Genet (Marion's love interest) is reprehensible and ultimately made the book unredeemable for me. On the whole, Verghese is a good descriptive writer, but some of his phrasings are awkward and self-conscious, and his descriptions of medical procedures are too clinical and can go on for pages.

Figure 14



Answer the following questions based on the above review text.

Question 4

Determine your understanding of the main idea of the review text.

☐ **Very well understood**[8]

☐ **Mostly understood**[9]

☐ **Somewhat understood**[0]

☐ **Barely understood**[q]

☐ **Did not understand**[w]

Question 5

Familiarity: Determine your familiarity with the objects, ideas, events, concepts, etc., discussed in the review. The review might sometimes mention items, objects, people, places, events, etc., which are unfamiliar to you. It might also contain ideas, concepts, rituals, and customs which are uncommon to you. The communication style might also be unfamiliar.
How familiar are you with the things mentioned in the review, like concepts, objects, customs, ideas, etc.?

☐ **Familiar with all the things**[e]

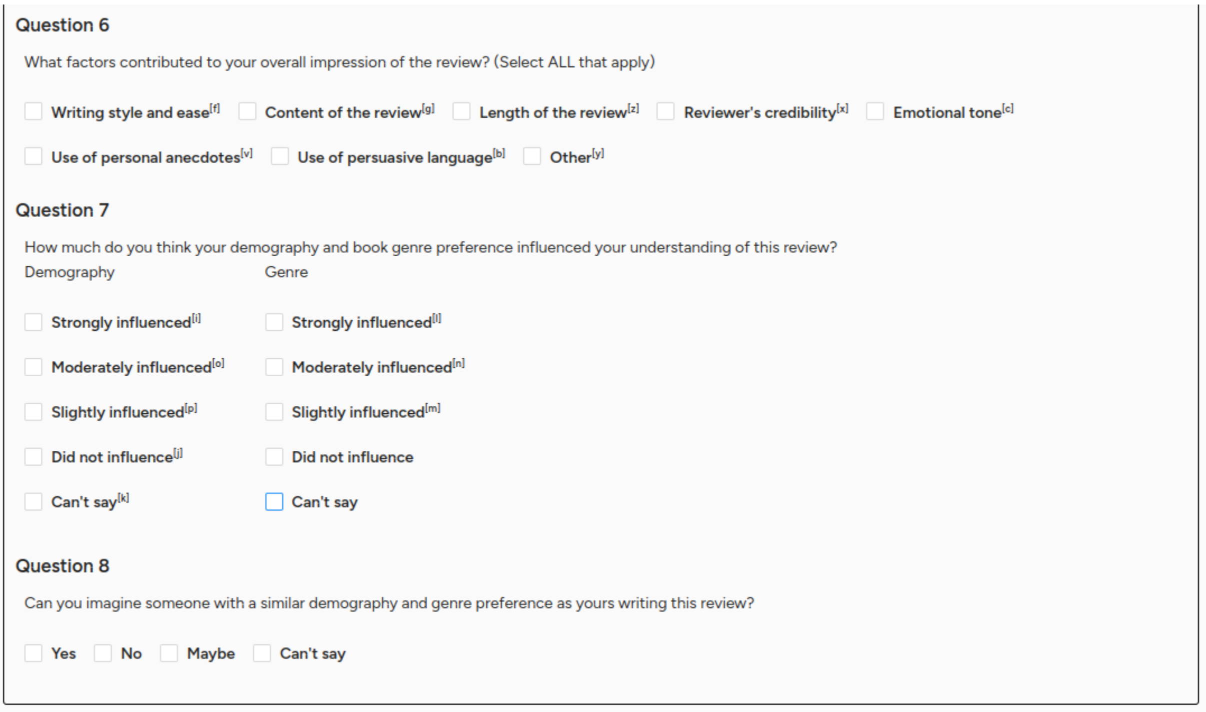
☐ **Familiar with almost all the things**[t]

☐ **Balanced familiarity - Some familiar, some are unfamiliar**[a]

☐ **Not familiar with most of the things**[s]

☐ **Not familiar with anything**[d]

Figure 15



Question 6

What factors contributed to your overall impression of the review? (Select ALL that apply)

☐ Writing style and ease[f] ☐ Content of the review[g] ☐ Length of the review[z] ☐ Reviewer's credibility[x] ☐ Emotional tone[c]

☐ Use of personal anecdotes[v] ☐ Use of persuasive language[b] ☐ Other[y]

Question 7

How much do you think your demography and book genre preference influenced your understanding of this review?

Demography	Genre
☐ Strongly influenced[i]	☐ Strongly influenced[l]
☐ Moderately influenced[o]	☐ Moderately influenced[n]
☐ Slightly influenced[p]	☐ Slightly influenced[m]
☐ Did not influence[j]	☐ Did not influence
☐ Can't say[k]	☐ Can't say

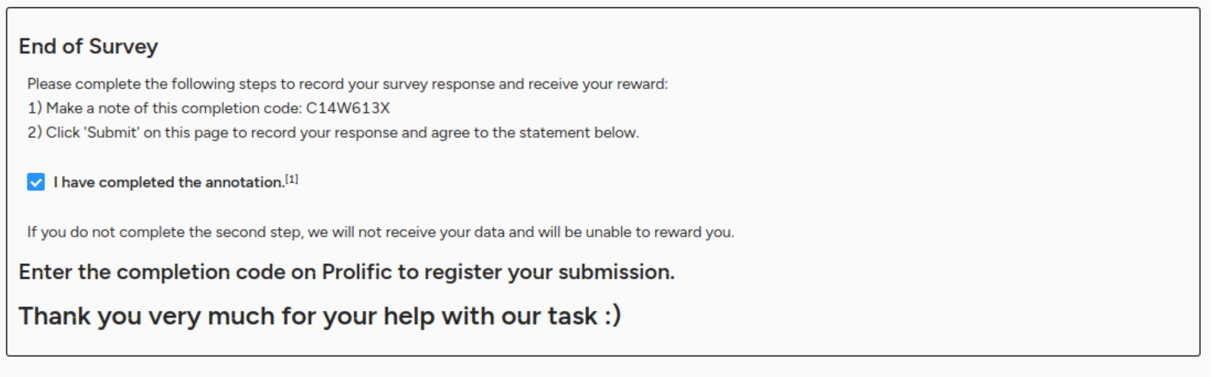
Question 8

Can you imagine someone with a similar demography and genre preference as yours writing this review?

☐ Yes ☐ No ☐ Maybe ☐ Can't say

Figure 16

D.4 End Of Survey Interface



End of Survey

Please complete the following steps to record your survey response and receive your reward:
1) Make a note of this completion code: C14W613X
2) Click 'Submit' on this page to record your response and agree to the statement below.

☑ **I have completed the annotation.**[1]

If you do not complete the second step, we will not receive your data and will be unable to reward you.

Enter the completion code on Prolific to register your submission.

Thank you very much for your help with our task :)

Figure 17

E Additional Analysis

As depicted in Table 7, participants from India analyzed 26.1, Mexico reviewed 10.5, and the USA reviewed 11.3 reviews on average. Although Indians found more unfamiliar spans (57%) compared to Mexicans and Americans (45% each), Americans had the highest proportion of culturally unfamiliar spans (73%), followed by Mexicans (67%) and Indians (64%). Overall, participants from all cultures encountered unfamiliar spans in 50% of their reviews, and more than two-thirds of them were cultural. This gap in understanding motivates a need for tools that enable cross-cultural communication and provide a headspace for such tools to operate.

Underlined in Table 7, for Indians, the highest proportion of unfamiliar spans (70%) were for books originating from the USA. Mexicans found the highest proportion of unfamiliar spans, almost equally, from books in India and the USA (approximately 55%). Americans found unfamiliar spans in 55% of the Indian book reviews. For participants from all countries, most of the unfamiliarity was for reviews of books from different cultures, which further motivates the need for cross-cultural communication tools.

Participant Country	**Review Country**	**# Reviews**	**# Difficult**	**# Cultural**
India	Ethiopia & UAE	8.00	3.29 (41%)	1.86 (57%)
	India	7.75	4.38 (57%)	2.63 (60%)
	USA	10.38	7.25 (70%)	5.00 (69%)
	Total	**26.13**	**14.92 (57%)**	**9.49 (64%)**
Mexico	Ethiopia & UAE	3.28	0.72 (22%)	0.39 (54%)
	India	3.25	1.85 (57%)	1.25 (68%)
	USA	4.00	2.18 (55%)	1.55 (71%)
	Total	**10.53**	**4.75 (45%)**	**3.19 (67%)**
USA	Ethiopia & UAE	3.11	1.00 (32%)	0.68 (68%)
	India	3.45	1.90 (55%)	1.40 (74%)
	USA	4.75	2.05 (43%)	1.55 (76%)
	Total	**11.31**	**4.95 (44%)**	**3.63 (73%)**

Table 7: Participant location-wise statistics of unfamiliar spans

F Benchmarking LLMs

F.1 Prompts

Prompt

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}