To Generate or Discriminate? Methodological Considerations for Measuring Cultural Alignment in LLMs

Saurabh Kumar Pandey[*]**, Sougata Saha**[*]**, Monojit Choudhury**
Mohamed bin Zayed University of Artificial Intelligence
saurabh2000.iitkgp@gmail.com, {sougata.saha, monojit.choudhury}@mbzuai.ac.ae

Abstract

Socio-demographic prompting (SDP) - prompting Large Language Models (LLMs) using demographic proxies to generate culturally aligned outputs - often shows LLM responses as stereotypical and biased. While effective in assessing LLMs' cultural competency, SDP is prone to confounding factors such as prompt sensitivity, decoding parameters, and the inherent difficulty of generation over discrimination tasks due to larger output spaces. These factors complicate interpretation, making it difficult to determine if the poor performance is due to bias or the task design. To address this, we use inverse socio-demographic prompting (ISDP), where we prompt LLMs to discriminate and predict the demographic proxy from actual and simulated user behavior from different users. We use the Goodreads-CSI dataset (Saha et al., 2025b), which captures difficulty in understanding English book reviews for users from India, Mexico, and the USA, and test four LLMs: Aya-23, Gemma-2, GPT-4o, and LLaMA-3.1 with ISDP. Results show that models perform better with actual behaviors than simulated ones, contrary to what SDP suggests. However, performance with both behavior types diminishes and becomes nearly equal at the individual level, indicating limits to personalization.

1 Introduction

"It is the novel bits of behaviour, the acts that couldn't plausibly be accounted for in terms of prior conditioning or training or habit, that speak eloquently of intelligence; but if their very novelty and unrepeatability make them anecdotal and, hence, inadmissible evidence, how can one proceed to develop the cognitive case for the intelligence of one's target species?" - Dennett (1988)

Human behavior, broadly defined as the preference for certain values, artifacts, knowledge, etc (Hogg, 2016), is inherently complex and highly variable among individuals (Brunswik, 1955; Henrich et al., 2010; Markus and Kitayama, 2014). However, patterns do emerge in certain aspects when behavior is aggregated among user groups, which are loosely defined by a *demographic proxy* (Adilazuarda et al., 2024) such as combinations of country, religion, etc. Such patterns constitute a proxy's *prototypical* behavior, which are the most typical or frequent behaviors, values, or norms observed in the group, and also allow for variation and exceptions (Rosch, 1975; Holland and Quinn, 1987). Nonetheless, assessments of LLMs' cultural biases (Bender et al., 2021; Masoud et al., 2023) often reduce behavior to *stereotypes*, which are grossly oversimplified and often exaggerated beliefs about the traits or behaviors of members of a demographic proxy (Tajfel, 1979; Lippmann, 2017). Probing models using SDP (Li et al., 2024b; AlKhamissi et al., 2024; Wan et al., 2023), such studies usually test for specific sociocultural knowledge through specially curated datasets (Nguyen et al., 2023; Dwivedi et al., 2023; Fung et al., 2024). They assume that certain knowledge is central to and therefore known to most members of a demographic proxy, which the models must therefore know (Nguyen et al., 2023, 2024; Shen et al., 2024; Naous et al., 2023; Kotek et al., 2023; Shrawgi et al., 2024), which is faulty and does not optimally measure models' cultural awareness (Saha et al., 2025a; Zhou et al., 2025). Several studies have also highlighted SDP's prompt sensitivity, which could lead to potentially misleading results (Mukherjee et al., 2024; Beck et al., 2024). Also, computationally, SDP is inherently complex as the output space of generative tasks is usually large (Ng and Jordan, 2001; Li et al., 2008).

We argue that apart from knowing a culture's *stereotypical behavior*, a model's cultural awareness should also encompass the *broader under-*

---

[*]Both authors contributed equally to this paper.

standing and knowledge of the *prototypical behavior*. However, since SDP is constrained by assumptions and limiting factors, it becomes difficult to measure the broader form of cultural awareness. As a solution, we propose ISDP, where we reverse the SDP task by providing the user behavior and asking the model to guess the probable membership of the user across different demographic proxies. We hypothesize that *if indeed models only understand stereotypes, then they should be able to guess the demographic proxy better from simulated behaviors rather than actual user behaviors*.

It is crucial to distinguish stereotypical behavior in SDP versus ISDP setups. In the ISDP context, a stereotypical behavior can be understood as the model either always associating a particular behavior (text span) with a culture(s) or never associating it, regardless of the surrounding context or user history. Whereas, in SDP, a model's stereotypical behavior would translate to systematically generating the same behavior (response) irrespective of the context.

We test our hypothesis on the Goodreads-CSI dataset (Saha et al., 2025b), which captures incomprehensibility of English book reviews by users from the USA, Mexico, and India, where country is the demographic proxy and incomprehensibility is the behavior. We use SDP to simulate user behavior with four LLMs: Aya, Gemma, GPT-4o, and Llama, and then use them as discriminators in ISDP. To our surprise, our hypothesis turned out to be partially false, where GPT-4o is better at predicting the country from actual user behavior rather than simulated behavior. On the contrary, other LLMs perform better with simulated behavior, except for a few instances.

2 Methodology

Figure 2 illustrates our study's overall setup. We discuss the details in this section.

2.1 Dataset

User Behavior: We use the Goodreads-CSI dataset (Saha et al., 2025b), which contains Culture-Specific Item (CSI) annotations of 57 Indian, Ethiopian, and US English book reviews by 50 users, from India, Mexico, and the USA. CSIs (Aixelá, 1996) are difficult to understand spans that people agreeably do not understand from a culture, where the incomprehensibility can be attributed to their distinct cultural background.

Simulated Behavior: Similar to Saha et al. (2025b), we simulate user behavior on the dataset using SDP, where an LLM is tasked to assume the role of a cultural reading assistant. Defining a user at the intersection of two proxies - *country* and *genre preference*, we use the prompt in Appendix A.1 to simulate behavior using GPT-4o, Aya-23, Gemma-2, and Llama-3.1.
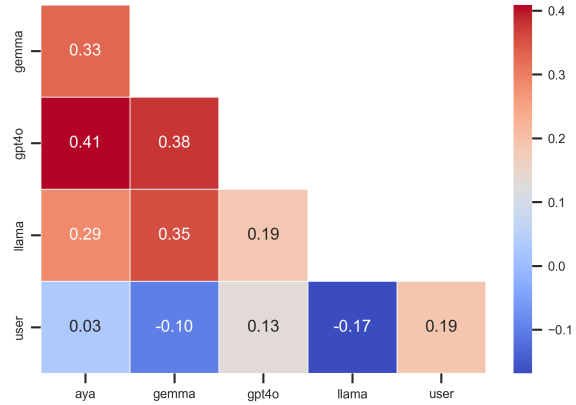

Figure 1: IAA between models and humans for SDP.

Figure 1 depicts the Inter Annotator Agreement (IAA) calculated using Krippendorff's alpha between models and users. We observe that the user-user and model-model IAA scores are significantly higher than the user-model scores, implying that models and users agree more among themselves on what is a CSI and exhibit similar behaviors among themselves rather than with each other. The model-model IAA values are also higher in comparison to the user-user IAA, which indicates that although models generate similar (possibly stereotypical) behavior for a demographic proxy, the actual user behavior is much varied. For example, irrespective of user history, all models consistently flagged the span "FDA" (Food and Drug Administration, USA) as a CSI, even for users whose histories suggested familiarity with American culture. Conversely, in cases where user histories indicated unfamiliarity with American culture, models sometimes overlooked spans like "home run" (a baseball reference), which should have been identified as a CSI. **Behavior Groups**: We also aggregate the user behavior (original and simulated) at three different levels, to capture a model's ISDP capacity with different amounts of behavior: **(i) Review level:** Aggregates all CSIs (user behaviors) for a review at a country level. **(ii) User + Review level:** Default level pertaining to individual user behavior for each review for each country. **(iii) User level:**

Aggregates all CSIs across all reviews for a user.

2.2 Method

Hypothesis: We test the hypothesis that given two models, M_1 and M_2, if M_1 is tasked to simulate the behavior of a user from a group g (in this case, country) using SDP, and M_2 is tasked to guess g from this simulated behavior b_{sim} using ISDP, then models will be able to predict g better from b_{sim} rather than from real user's behavior b. This hypotheses arise from the observation that models trained on similar datasets are likely to generalize in comparable ways and will have similar biases.

Models: We experiment with Aya-23-8B (Aryabumi et al., 2024), Gemma-2-9B-it (Team et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), and gpt-4o-2024-05-01-preview (Achiam et al., 2023), setting temperature to 0.

Task: Given a set of behavior ($s \in \{b, b_{sim}\}$), we prompt the models to rank three countries $c \in$ {India, Mexico, USA}, according to how likely is the behavior from an individual or a group of users from each country: $Rank(P(g = c|s))$. Each prompt is repeated 5 times, where the ordering of the countries is randomized in each prompt to account for any positional bias. The country with rank 1 indicates individuals or groups from that country would most likely not understand the CSI spans. The country ranked 3 is where the CSIs are least likely to be misunderstood. Performance is evaluated using the Mean Reciprocal Rank (MRR).

3 Results and Analysis

Figure 3 plots the MRR for all models and all levels of behavior aggregation. The spread of each box depicts the MRR variance due to repeated prompting. The red dotted line (Y=0.61) is the *random baseline*, where each option is likely with a 1/3 probability. Statistical significance of the results using paired t-test is presented in Tables 1, 2, and 3 in Appendix A.3. We observe the following for the ISDP task:

1. For **user-generated spans**, across **all levels** of behavior aggregation, the average MRR of almost all models is greater than the *random baseline*, indicating "understanding" of prototypical behavior to an extent. GPT-4o has the highest average MRR at the **Review and User+Review** levels, whereas Gemma has the highest score at the **User** level, for such spans. Interestingly, for **Review and User+Review** levels, GPT-4o consistently per-

forms worse with model-simulated behavior (including itself), compared to user-generated spans, proving our hypothesis wrong.

2. Contrasting GPT-4o, Gemma, Llama, and Aya perform better with simulated behavior rather than user-generated behavior. Gemma and Llama consistently perform better with **GPT-4o-generated spans**; Their scores are further amplified at the **User+Review** level than the **Review** level. Although this behavior aligns with our hypothesis, it raises questions about the source and nature of the data they are trained on. Furthermore, since the difference between GPT-4o's scores with user vs. its own generated spans increases at the **User+Review**, we conjecture that GPT-4o (and other models) possibly generates stereotypes during SDP.

3. At the **User** level, the average MRR scores across models are less varied and their performance is near baseline. Interestingly, GPT-4o performs best on spans generated by itself, which sharply contrasts the trend observed at **Review** and **User+Review** levels. We conjecture that non-GPT models are possibly trained on stereotypical data, which enables their performance with simulated data at the **Review and User+Review** levels. However, since each individual's behavior across all reviews is aggregated at the **User** level, even if the review-level behaviors were stereotypical, the aggregate possibly reflects the prototypical behavior better, causing all discriminators to perform poorly.

4. Plotting the average MRR scores between different pairs of behavior generators and discriminators in Figure 4 shows our hypothesis does not hold; There are cases, apart from GPT-4o, where models are better discriminators when the behavior is generated by users rather than LLMs.

5. We plot the average MRR's trend across the three levels of behavior aggregation for each model in Figure 5. Similar to the previous findings, the trends indicate that GPT-4o is possibly better at personalization to users, and possibly models user behavior better than other models. This finding aligns with Saha and Choudhury (2025), where they show the empirical limits of LLMs' personalization capacity (Fan and Poole, 2006; Lury and Day, 2019) to different sizes of user groups, ranging from a country-level to an individual. Furthermore, this also indicates that other models are essentially learning stereotypes, which further raises questions about the nature of their training data. This might also indicate that the other (smaller) models are trained on GPT or other LLM-generated
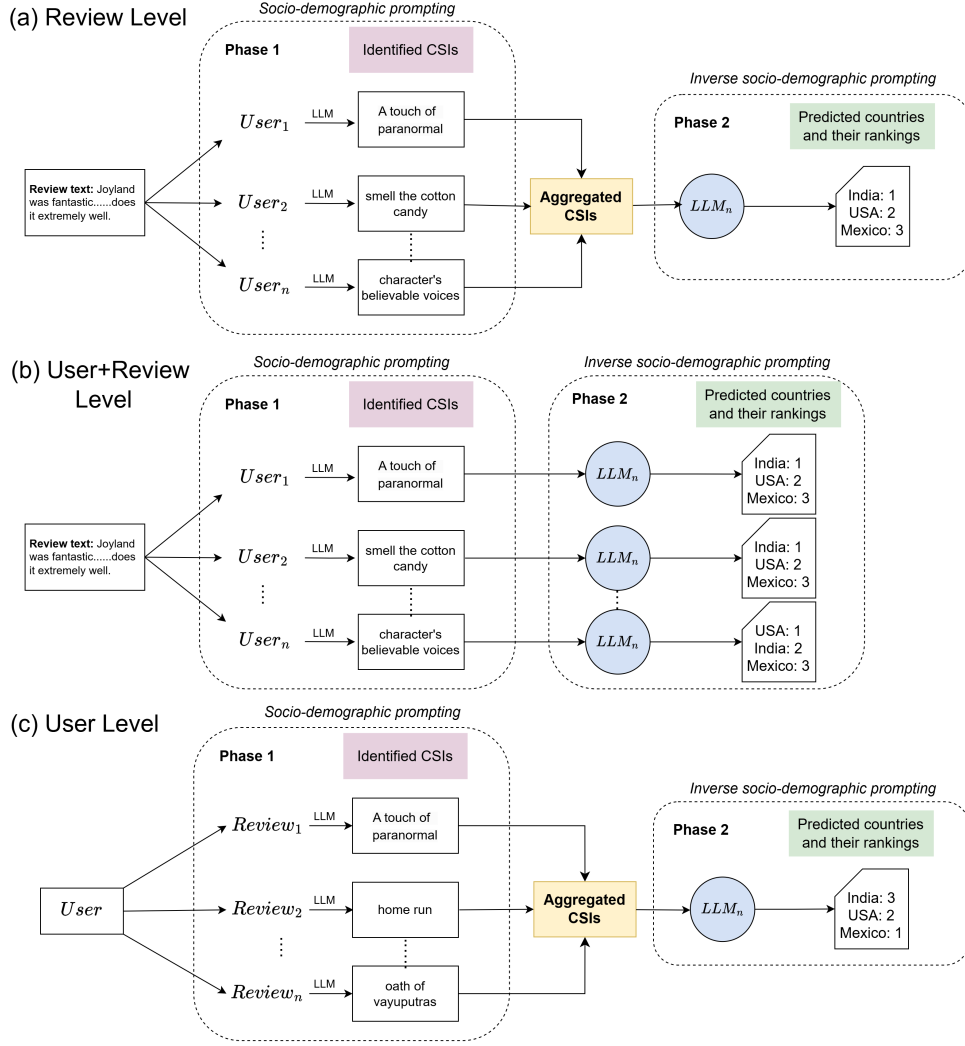
Figure 2: Experimental setup for all three levels of behavior aggregation. Phase 1: SDP and Phase 2: ISDP.

data, causing them to model user behavior differently than evident in the real world, ascertaining which we leave as future work.

4   Discussion and Conclusion

Since LLMs are being used in several everyday applications, strong alignment of their generated behavior to over-represented norms or *typical* behavior poses serious risks, not only for the under-represented communities (say, Mexico or India), but also for users from over-represented groups (say, USA). This is because, as our study shows, every user exhibits some behavior that is not part of the norm. This was demonstrated by (Agarwal et al., 2024) in their study on writing style, where they showed that LLM-assisted writing results in convergence of styles for users from both the USA and India. However, the degree of loss of style diversity was greater for users from India (a fur-

ther underrepresented group) than for those from the USA. Our study sheds light on why this might be the case, not only for writing styles but for any other aspect of user behavior that LLM-driven applications are expected to replicate or interpret.

A key question remains: *when should we use SDP versus ISDP to measure a model's cultural competency?* Although SDP is widely used to study cultural bias, it is confounded by prompt sensitivity, decoding choices, and the greater difficulty of generation relative to discrimination due to the much larger output space. ISDP mitigates these issues by reframing the task as discrimination: given a behavior, the model selects from a small candidate set of countries (e.g., three in our experiments). Standardizing the input spans further reduces stylistic variance, such as wording, grammar, and other non-content factors. In our experiments, these constraints make ISDP a more
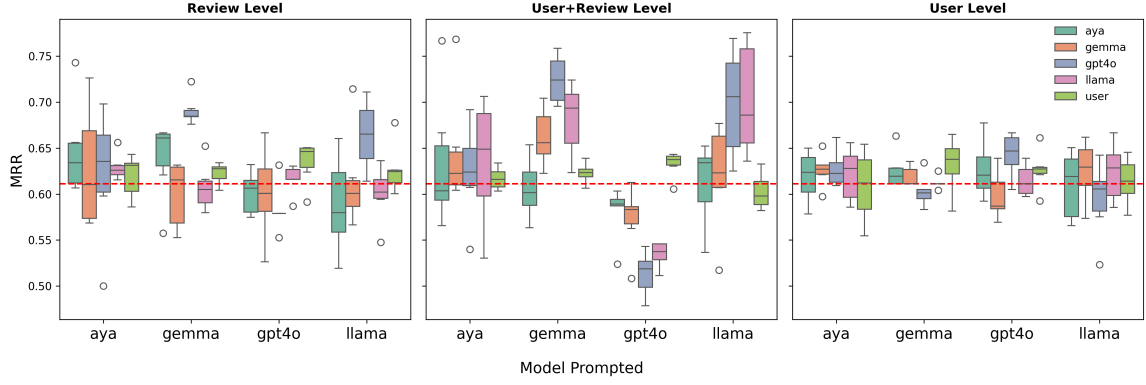
Figure 3: Model-wise MRR scores across five different sources of behavior and three levels of behavior aggregation. The x-axis represents the discriminators, whereas the legend represents the generators used in the experiments.
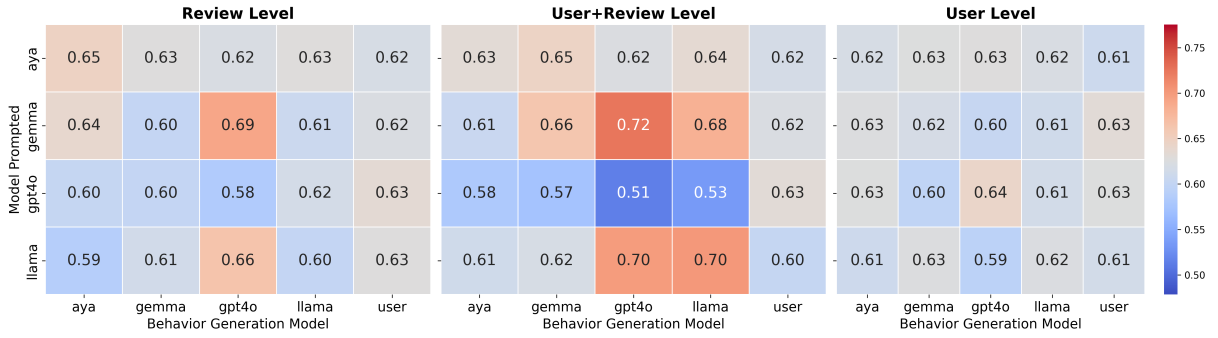

Figure 4: Average MRR scores for different combinations of generators and discriminators.

reliable indicator of cultural competency. Nevertheless, ISDP is not a replacement for SDP. Rather, it is a complementary method that, alongside SDP, provides a fuller picture of a model's competency. We leave a systematic validation of this complementarity to future work.

Limitations

While our study provides critical insights into the cultural competency of LLMs, it has several limitations. First, we rely only on the Goodreads-CSI dataset to test our hypothesis, which might hurt the generalizability of the results. Nonetheless, the results are useful because, as indicated by Saha et al. (2025b), the dataset contains CSIs pertaining to different cultural dimensions of Newmark's taxonomy (Newmark, 2003), which cover most aspects of culture. They also justify the dataset's size and argue why it suffices as an evaluation benchmark. Furthermore, the dataset captures knowledge as behavior, which is a unique and cognitively more challenging aspect than other forms of behavior, such as preferences (Dunbar, 1995; Kuhl, 2004). Second, we restrict to English reviews while limiting users from only three demographics: India,

Mexico, and the USA and which might not be representative of user behavior from all demographics. Finally, all models used in our study were developed in the West. A study incorporating regionally developed models in regional languages would be valuable.

Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784,

Miami, Florida, USA. Association for Computational Linguistics.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2024. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. *Preprint*, arXiv:2409.11360.

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

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Taluparu, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Egon Brunswik. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193.

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 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Peng Cui, Huan Liu, Charu Aggarwal, and Fei Wang. 2016a. Computational modeling of complex user behaviors: Challenges and opportunities.

Peng Cui, Huan Liu, Charu Aggarwal, and Fei Wang. 2016b. Uncovering and predicting human behaviors. *IEEE Intelligent Systems*, 31(2):77–88.

Shekoufeh Daghoughi and Mahmood Hashemian. 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.

Daniel C Dennett. 1988. The intentional stance in theory and practice.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Haiyan Fan and Marshall Scott Poole. 2006. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

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

Dorothy Holland and Naomi Quinn. 1987. *Cultural models in language and thought*. Cambridge University Press.

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.

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

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

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

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

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

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. *arXiv preprint arXiv:2404.10199*.

Ming Li, Paul Vitányi, and 1 others. 2008. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer.

Walter Lippmann. 2017. *Public opinion*. Routledge.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Celia Lury and Sophie Day. 2019. Algorithmic personalization as a mode of individuation. *Theory, Culture & Society*, 36(2):17–37.

Hazel Rose Markus and Shinobu Kitayama. 2014. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*, pages 264–293. Routledge.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *ArXiv*, abs/2309.12342.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv 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.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Isabel Cómitre Narváez and José María Valverde Zambrana. 2014. How to translate culture-specific items: a case study of tourist promotion campaign by turespaña. *The journal of specialised translation*, 21:71–112.

Peter Newmark. 2003. A textbook of translation.

Andrew Ng and Michael Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1774–1784.

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

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTUR-ALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

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

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Sougata Saha and Monojit Choudhury. 2025. User behavior prediction as a generic, robust, scalable, and low-cost evaluation strategy for estimating generalization in llms. *arXiv preprint arXiv:2507.05266*.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025a. Meta-cultural competence: Climbing the right hill of cultural awareness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8025–8042.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025b. Reading between the lines: Can llms identify cross-cultural communication gaps? *arXiv preprint arXiv:2502.09636*.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857.

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.

Henri Tajfel. 1979. Individuals and groups in social psychology. *British Journal of social and clinical psychology*, 18(2):183–190.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Exploring large language models' cognitive moral development through defining issues test. *arXiv preprint arXiv:2309.13356*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

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.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. 2024a. A survey on human-centric llms. *arXiv preprint arXiv:2411.14491*.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural adaptation of menus: A fine-grained approach. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA. Association for Computational Linguistics.

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

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Appendix

A.1 Prompts - SDP

Prompt - Socio Demographic Prompting

AI Rules
- Output response in JSON format
- Do not output any extra text.
- Do not wrap the outputs 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.
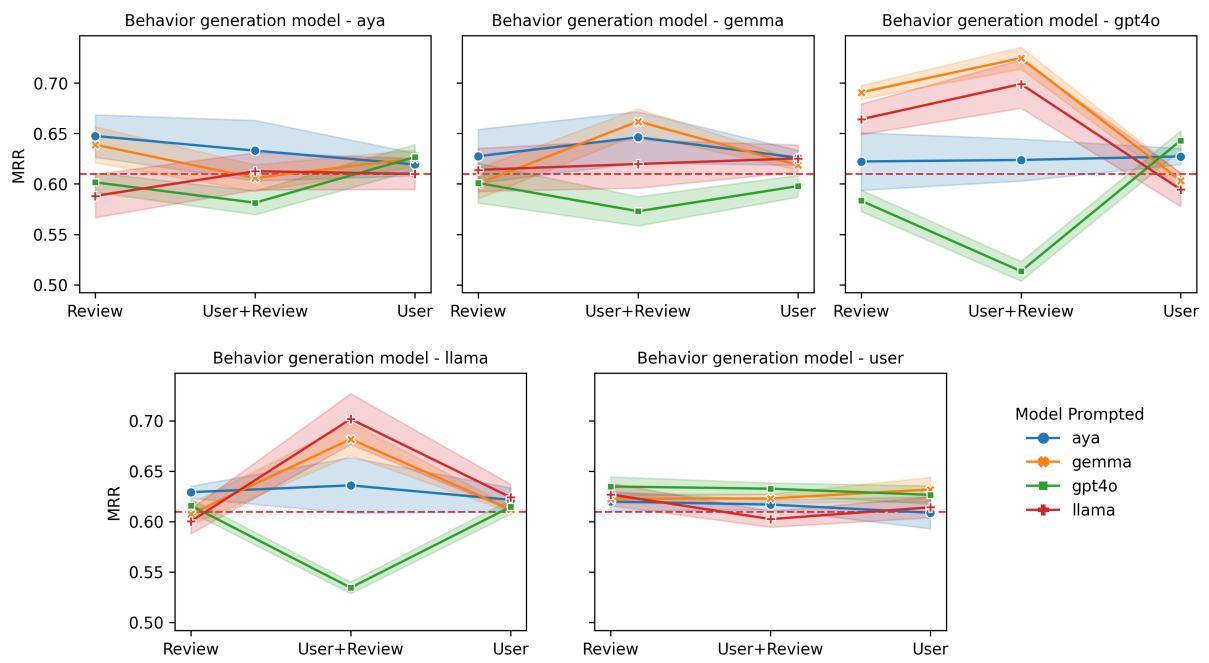
Text: {review_text}

Figure 5: Average MRR scores for different levels of behavior.

A.2 Prompts - ISDP

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

You are a cultural mediator who understands all cultures across the world. As a mediator, your job is to identify the cultural background of the users. The user came across a review of the book {book} by {author}, which belongs to the {book_genre} genre. Culture-specific items (CSIs) are textual spans denoting concepts and items uncommon and not prevalent in a culture, making them difficult to understand for an individual from a given culture. Given the review text and all the CSI spans that the users found hard to understand due to their cultural background, perform the following task.

Task 1: Based on the CSI spans that the users found difficult to understand due to their cultural background, provided as a tuple of (CSI span, number of users who found the span difficult to understand), identify the users' likely country of origin. Rank the countries in decreasing order of how unfamiliar the CSI spans are in that country. The possible countries are India, USA, and Mexico. The country ranked highest is the one where the CSIs are least common (most unfamiliar), and the lowest-ranked country is where the CSIs are most common (least unfamiliar).

Task 2: Explain within 20 words the reason behind the ranking of the countries in Task 1.

Format your response as a valid Python dictionary formatted as: {'country': <task 1: dictionary of all possible countries with their rankings formatted as {'country': ranking}>, 'reason': <task 2: reason within 20 words>}

Review Text: {review_text}

CSIs: $\{(\text{span}_1, n_1), (\text{span}_2, n_2), \ldots, (\text{span}_N, n_N)\}$

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

You are a cultural mediator who understands all cultures across the world. As a mediator, your job is to identify the cultural background of a user. Culture-specific items (CSIs) are textual spans denoting concepts and items uncommon and not prevalent in a culture, making them difficult to understand for an individual from a given culture. The user came across reviews of books. Given the review text and all the CSI spans that the user found difficult to understand due to their cultural background, perform the following task.

Task 1: Based on the CSI spans that the users found difficult to understand due to their cultural background, provided as a tuple of (CSI span, number of users who found the span difficult to understand), identify the users' likely country of origin.. Rank the countries in decreasing order of how unfamiliar the CSI spans are in that country. The possible countries are India, USA, and Mexico. The country ranked highest is the one where the CSIs are least common (most unfamiliar), and the lowest-ranked country is where the CSIs are most common (least unfamiliar).

Task 2: Explain within 20 words the reason behind the ranking of the countries in Task 1.

Format your response as a valid Python dictionary formatted as: {'country': <task 1: dictionary of all possible countries with their rankings formatted as {'country': ranking}>, 'reason': <task 2: reason within 20 words>}

Given below is the review of the book {book} by {author}, which belongs to the {book_genre} genre.
Review Text1: {review_text}
CSIs: $\{(\text{span}_1, n_1), (\text{span}_2, n_2), \ldots, (\text{span}_N, n_N)\}$

Given below is the review of the book {book} by {author}, which belongs to the {book_genre} genre.
Review Text: {review_text}
CSIs: $\{(\text{span}_1, n_1), (\text{span}_2, n_2), \ldots, (\text{span}_N, n_N)\}$
.
.
.
.
Given below is the review of the book {book} by {author}, which belongs to the {book_genre} genre.
Review Text: {review_text}
CSIs: $\{(\text{span}_1, n_1), (\text{span}_2, n_2), \ldots, (\text{span}_N, n_N)\}$

A.3 Results: Statistical Significance

span_generator	discriminator1	discriminator2	t_statistic	p_value
aya	llama	aya	-5.326117	0.003124
aya	llama	gemma	-1.347977	0.235514
aya	llama	gpt4o	-0.450854	0.670974
aya	aya	gemma	0.227204	0.829263
aya	aya	gpt4o	1.572355	0.176675
aya	gemma	gpt4o	3.152963	0.025294
gemma	llama	aya	-0.740100	0.492505
gemma	llama	gemma	0.413237	0.696558
gemma	llama	gpt4o	0.467127	0.660058
gemma	aya	gemma	0.769376	0.476424
gemma	aya	gpt4o	0.912667	0.403278
gemma	gemma	gpt4o	-0.029062	0.977939
gpt4o	llama	aya	1.083931	0.327873
gpt4o	llama	gemma	-1.355720	0.233205
gpt4o	llama	gpt4o	4.574447	0.005978
gpt4o	aya	gemma	-2.422944	0.059899
gpt4o	aya	gpt4o	1.034381	0.348370
gpt4o	gemma	gpt4o	8.260590	0.000424
user	llama	aya	0.354713	0.737275
user	llama	gemma	0.343953	0.744874
user	llama	gpt4o	-0.405212	0.702077
user	aya	gemma	-0.251829	0.811197
user	aya	gpt4o	-1.647885	0.160292
user	gemma	gpt4o	-1.140894	0.305595
llama	llama	aya	-3.120119	0.026249
llama	llama	gemma	-0.835801	0.441365
llama	llama	gpt4o	-1.177835	0.291865
llama	aya	gemma	4.332499	0.007482
llama	aya	gpt4o	1.412226	0.216988
llama	gemma	gpt4o	-0.618236	0.563491

Table 1: Paired t-test results for review level experiments. Rows colored orange show that they are statistically significant (p_value < 0.05).

span_generator	discriminator1	discriminator2	t_statistic	p_value
aya	llama	aya	-0.512273	0.630278
aya	llama	gemma	0.355522	0.736705
aya	llama	gpt4o	1.397820	0.221018
aya	aya	gemma	0.648904	0.545023
aya	aya	gpt4o	1.443309	0.208528
aya	gemma	gpt4o	1.843266	0.124623
gemma	llama	aya	-0.569640	0.593560
gemma	llama	gemma	-1.727200	0.144713
gemma	llama	gpt4o	1.799847	0.131783
gemma	aya	gemma	-0.486936	0.646898
gemma	aya	gpt4o	2.736133	0.040980
gemma	gemma	gpt4o	6.344467	0.001436
gpt4o	llama	aya	3.633479	0.015006
gpt4o	llama	gemma	-0.968320	0.377358
gpt4o	llama	gpt4o	8.703857	0.000331
gpt4o	aya	gemma	-4.606414	0.005806
gpt4o	aya	gpt4o	5.604685	0.002499
gpt4o	gemma	gpt4o	12.269834	0.000064
user	llama	aya	-1.301253	0.249904
user	llama	gemma	-4.421119	0.006885
user	llama	gpt4o	-3.433467	0.018566
user	aya	gemma	-0.779455	0.470977
user	aya	gpt4o	-1.842872	0.124686
user	gemma	gpt4o	-2.003821	0.101443
llama	llama	aya	2.007264	0.100998
llama	llama	gemma	0.598369	0.575666
llama	llama	gpt4o	6.860223	0.001006
llama	aya	gemma	-1.150194	0.302086
llama	aya	gpt4o	4.116861	0.009202
llama	gemma	gpt4o	7.476294	0.000676

Table 2: Paired t-test results for user+review level experiments. Rows colored orange show that they are statistically significant (p_value < 0.05).

span_generator	discriminator1	discriminator2	t_statistic	p_value
aya	llama	aya	-0.466986	0.660152
aya	llama	gemma	-0.800898	0.459539
aya	llama	gpt4o	-0.755307	0.484104
aya	aya	gemma	-0.679548	0.526970
aya	aya	gpt4o	-0.314223	0.766040
aya	gemma	gpt4o	-0.063718	0.951664
gemma	llama	aya	-0.036038	0.972647
gemma	llama	gemma	0.553196	0.603952
gemma	llama	gpt4o	1.631805	0.163649
gemma	aya	gemma	0.951767	0.384923
gemma	aya	gpt4o	2.026568	0.098541
gemma	gemma	gpt4o	2.157585	0.083441
gpt4o	llama	aya	-2.018184	0.099601
gpt4o	llama	gemma	-0.724634	0.501154
gpt4o	llama	gpt4o	-2.326693	0.067489
gpt4o	aya	gemma	2.038030	0.097112
gpt4o	aya	gpt4o	-1.258016	0.263937
gpt4o	gemma	gpt4o	-3.157931	0.025154
user	llama	aya	0.379708	0.719756
user	llama	gemma	-1.288302	0.254034
user	llama	gpt4o	-1.117324	0.314649
user	aya	gemma	-1.075421	0.331318
user	aya	gpt4o	-1.512893	0.190718
user	gemma	gpt4o	0.382935	0.717507
llama	llama	aya	0.117893	0.910742
llama	llama	gemma	0.872386	0.422905
llama	llama	gpt4o	1.005745	0.360700
llama	aya	gemma	0.910761	0.404190
llama	aya	gpt4o	0.502238	0.636832
llama	gemma	gpt4o	-0.395285	0.708934

Table 3: Paired t-test results for user level experiments. Rows colored orange show that they are statistically significant (p_value < 0.05).

A.4 Related Work

Predicting and modeling human behavior has always been a challenging task (Cui et al., 2016b,a; Wang et al., 2024a). Measurement of cultural awareness in LLMs inherently requires modeling and/or prediction of human behavior, which is typically conducted using SDP on specifically curated datasets under culture-specific settings (Nguyen et al., 2023; Dwivedi et al., 2023; Fung et al., 2024; Shi et al., 2024; Nadeem et al., 2021; 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). Studies have also evaluated LLMs' knowledge of cultural artifacts such as food, art forms, clothing, and geographical markers (Seth et al., 2024; Li et al., 2024a; Koto et al., 2024). However, many of these methods argue that there is a need for the development of robust evaluation benchmarks that can test the cultural understanding in LLMs (Wang et al., 2024b; Rao et al., 2024; Myung et al., 2024; Zhou et al., 2024; Putri et al., 2024; Mostafazadeh Davani et al., 2024; Wibowo et al., 2024; Owen et al., 2024; Chiu et al., 2024; Liu et al., 2024; Koto et al., 2024). The Goodreads-CSI dataset, recently introduced by (Saha et al., 2025b), serves as a robust benchmark as they capture CSIs- a term introduced by Aixelá (1996) and further explored in various works (Pandey et al., 2025; Zhang et al., 2024; Daghoughi and Hashemian, 2016; Narváez and Zambrana, 2014; Sperber et al., 1994; Trivedi, 2008), which depicts things that people would not understand due to their culture.

B Dataset standardization

The dataset contains human annotations of 57 English reviews of books originating from India, UAE, and the USA. The annotations are performed by 50 users, where 8 are from India, 22 are from Mexico, and 20 are from the USA. Given a book review, each user identifies text spans of varying lengths that they found difficult to understand.

Span standardization: Since users can mark any contiguous text spans as difficult to understand, there is a high degree of variability in the annotations. To handle this, the original dataset (Saha et al., 2025b) semantically clustered the spans using sentence transformers[1] and filtered out poor-quality annotations before conducting quantitative

---

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

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 countercultural 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*. Some might non-contiguously highlight both spans to indicate unfamiliarity with both of them. Such fine-level distinctions are lost in sentence transformers-based semantic clustering. **(ii) User-Model Mismatch:** User-annotated spans were generally longer than those generated by the model, primarily due to differences in word boundary recognition. Users often group multiple CSIs into a single span, whereas the model-generated responses tend to break them down into more atomic CSIs. **(iii) Model-Model Mismatch:** Unlike Saha et al. (2025b), since we also evaluate multiple models here, there might be scenarios where the model-highlighted spans do not have a 1:1 match, similar to the user-user mismatches. To alleviate these issues, we collected all the CSI spans annotated by users and models for a given review text and manually standardized the spans for each review text, ensuring consistent discourse segmentation.

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
Overlap scores. 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
Overlap scores. Model 1: 0.4 **Model 2:** 0.2

Above is an illustrative example, which shows the overlap scores for two models before and after the standardization process. Before standardization, both models 1 and 2 attain a perfect score using sentence transformers with a similarity threshold of 0.5. However, after standardization, the user span

is split into five unique spans depicting different CSIs. Using an exact match, we obtain an overlap score of 0.4 for Model 1 and 0.2 for Model 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 elementary discourse units (EDUs). Each annotator annotated 450 spans (avg 19 reviews per annotator), which were randomly sampled and had ~50 overlapping spans among them to facilitate calculating inter-annotator agreement (IAA) scores. The annotation guidelines for the standardization of spans are as follows.

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

Since each annotation either involved correcting the errors in the span or splitting a span into multiple spans, we use a mean-based IAA metric to calculate agreement between the three expert annotators. We assign a weighted score to the agreements and disagreements while calculating IAA and assign a perfect score if the two annotators agree on an annotation and a score of 0.7 otherwise. After the first round of annotation, we obtain a mean agreement of 0.967. The disagreements were discussed and resolved by an additional round of annotation. 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. Post standardization, the total review text and span combinations were 922 (322 from users, 600 from models), compared to 1,193 (365 from users, 828 from models). Overall, the dataset contains 671 unique spans across all review texts, compared to 1,122 spans. The standardization process ensures the reliability of any following span-based analysis, enabling more robust comparisons between humans and models.[2]

---

[2]We will release the updated dataset with standardized spans upon acceptance.