MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

As many opinions, so many ways; In AI's path, each culture plays

Towards holistic cultural evaluation frameworks

Dr. Sougata Saha, Postdoc @MBZUAI

LLM and Society: The Current Landscape

- General-purpose LLMs should be equitable across cultures
  - Which are not
  - Their performance vary across cultures
  - Exhibit socio-demographic biases
- Biases might lead to cultural homogenization
  - Forces users to conform to the dominant culture to get service *[1]*
  - Erasure of underrepresented cultures in extreme cases
- What do we need?
  - Robust cultural evaluation frameworks

References: 1. Agarwal, D., Naaman, M., & Vashistha, A. (2024). Ai suggestions homogenize writing toward western styles and diminish cultural nuances. *arXiv preprint arXiv:2409.11360*.

Why is Cultural Evaluation Hard?

- Culture lacks a formal definition *[1]*
  - It arises due to distinctions in the "*way of life*" between groups *[2]*
  - An "*us versus them*" feeling *[3; 4]*
- Culture is an *individual* (undocumented) and a *social* construct (documented) *[5]* Ex: Robotics enthusiasts from Dabolim, Navajo tribe
- Cultural **evaluation frameworks must incorporate this dynamic essence of culture**

References:
1. Adilazuarda et.al., 2024. Towards measuring and modeling" culture" in llms: A survey.
2. Baldwin et.al. ,2006. A moving target: The illusive definition of culture.
3. Blake, 2000. On defining the cultural heritage. International & Comparative Law Quarterly.
4. Birukou et.al., 2013. A formal definition of culture. Models for intercultural collaboration and negotiation.
5. Spencer-Oatey et.al., 2012. What is culture. A compilation of quotations.

Issues with Current Evaluation Schemes

- Current methods **mainly test for cultural knowledge** *[3; 4]*

- Some test for **perceived alignment** along theoretical frameworks:
  - Hofstede's cultural dimensions *[1]*
  - World Values Survey *[2]*

- **Limited** to specific cultures

- We need something more:
  - Model-level: A **higher order objective** to optimize
  - System-level: Measuring their **real-world utility** across cultures

References:
1. Hofstede, 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.
2. Inglehart et.al., 2000. World values surveys and European values surveys, 1981-1984, 1990-1993, and 1995-1997.
3. Tanmay et.al., 2023. Probing the Moral Development of Large Language Models through Defining Issues Test.
4. Kharchenko et.al., 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions.

What we Propose?

- Optimizing for **Meta-cultural competency** *[1]* instead of only cultural competency
  - A higher order competency innate to humans
  - Enables intercultural communication. Comprises:
    - **Variational awareness**: self-awareness of cultural differences
    - **Explication & Negotiation Strategies**: conversational strategies that aim to reduce misinterpretations in cross-cultural settings

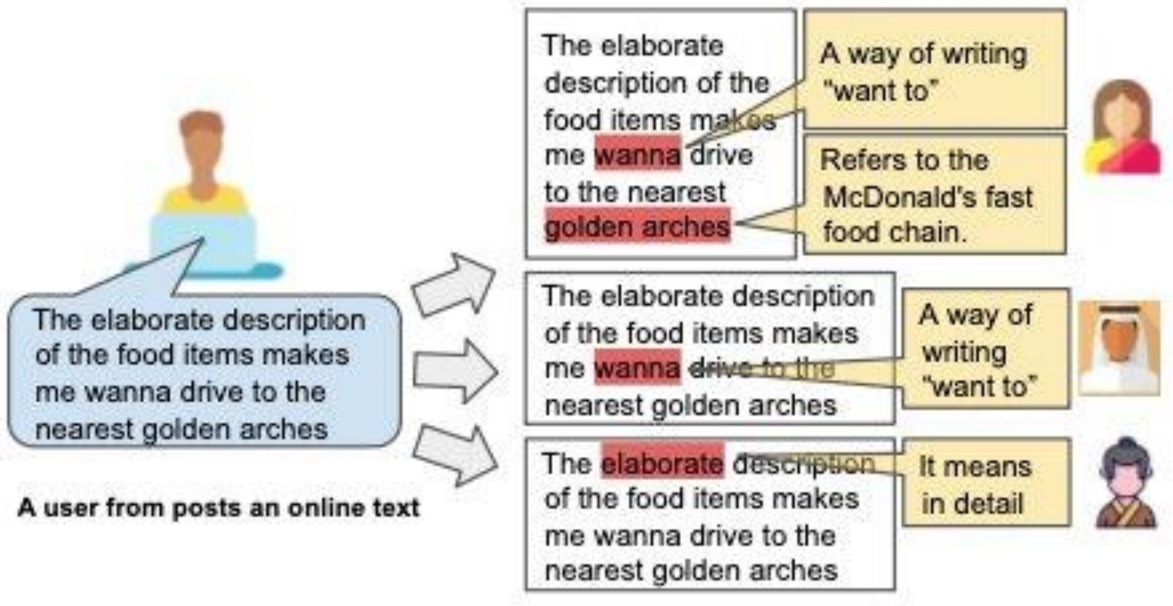
- **Functional and behavioral testing** instead of factual probing
  - Measure the utility and suitability of LLM-based tools across cultures

References:

1. Sharifian, 2013. Globalisation and developing metacultural competence in learning English as an International Language.

Culturally Yours (CY): LLMs as reading assistants

- CY *[1]* is an **online reading assistant**
- Preemptively **highlights and explains culture-specific items** (CSIs) that users might find difficult to understand due to their cultural background
- Uses **culture as a prior**- Country, age, genre preference, etc.
- Measure difference between model and human-identified CSIs as a measure of a model's cultural awareness.
- This approach is **free from test data leakage**, unlike probing for facts.



The CY Reading Assistant identifies culturally unfamiliar concepts and explains them as per the reader's cultural background.

References:

1. Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. [CULTURALLY 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.

Culturally Yours (CY): LLMs as reading assistants

- Prolific study with 50 participants from India, Mexico, and the USA
  - Highlight difficult to understand spans from reviews.
  - Associate level of unfamiliarity
  - Answer additional survey questions
- Measured:
  - How much do people not understand?
  - How much of the difficulty is due to culture? (GPT-4o as annotator)
- LLM as an agent:
  - Identify CSIs that a person from a given culture will not understand
  - Correlate agent responses with humans.
  - Measure equitability: Differences in overlap of CSIs between agents and humans from the same culture.

Culturally Yours (CY): Prolific Study Questions

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]

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.

Completely Unfamiliar 5 | Very Unfamiliar 6 | Somewhat Unfamiliar 7

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.

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]

Different ways of capturing understandability

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

Different ways of capturing the factors that affect understandability

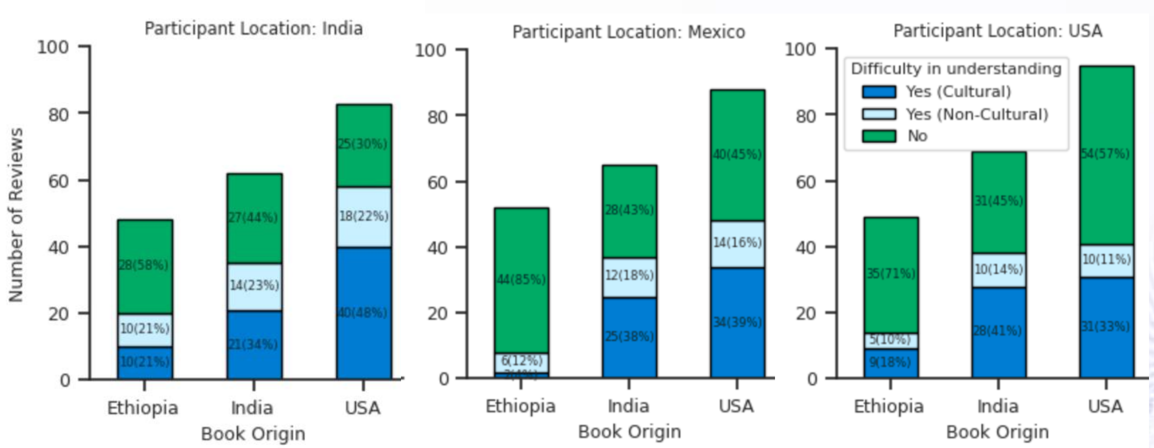
Question 8

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

- [ ] Yes
- [ ] No
- [ ] Maybe
- [ ] Can't say

Study Findings

- All reviews had at least 1 difficult-to-understand span
  - 83% (50) had culturally difficult spans
  - **Implication**: Cultural reading assistants might be beneficial.
- Inter-annotator agreements:
  - Review-level: **Intra-country agreement greater** than inter, except USA.
  - Span-level: **Lack of consensus** across all countries, denoting understandability is individual-specific. **Intra > inter** for CSIs
  - **Implications**:
    - CSIs are a set of harder-to-understand construct.
    - Good targets for priors for the cold-start problem.



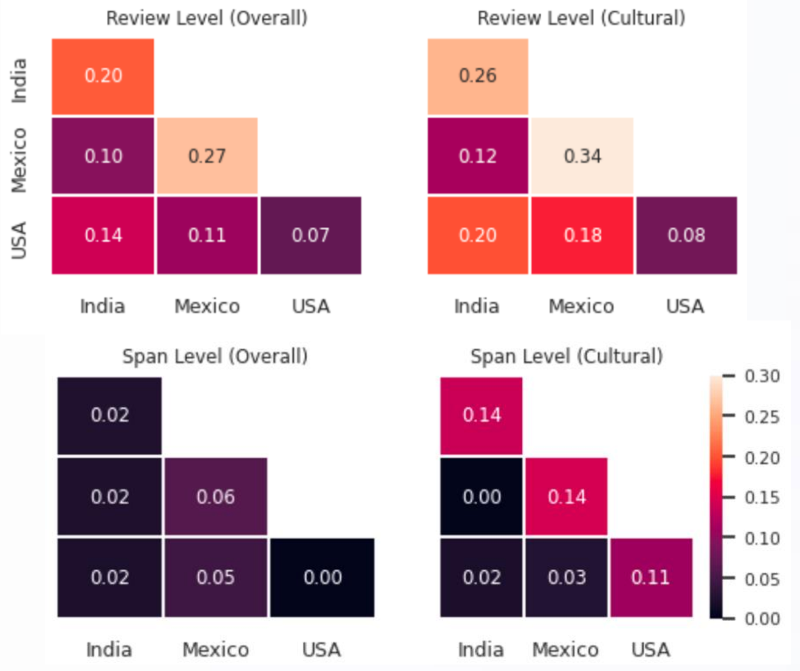


Figure 2: Inter Annotator Agreement at Review Level and Span Level across Countries.

GPT-4o Benchmarking

- 96/115 (83%) GPT-4o identified CSIs overlap with human-identified difficult spans
- 70/115 (60%) overlap with 116 user-identified CSIs
- 26 (22%) GPT-4o CSIs not cultural per users
  - Probably due to the GPT-4o post processing errors
  - 50 participants do not capture all variations
- GPT-4o **generalizes**: Low distinction in CSIs between fiction & non-fiction groups
- Recall higher than precision; **captures variety**
- **Implication**: GPT-4o equitably low-performing

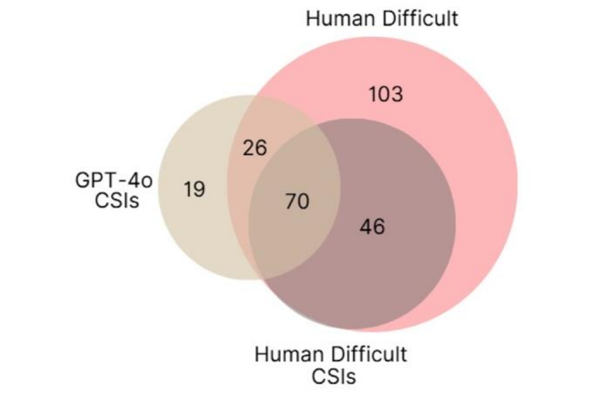


Figure 4: Overlap between Human-identified difficult spans and GPT-4o-identified CSIs

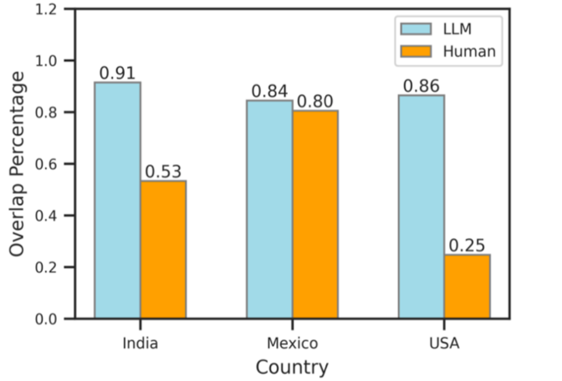


Figure 5: Overlap percentage of fiction and non-fiction spans across countries.

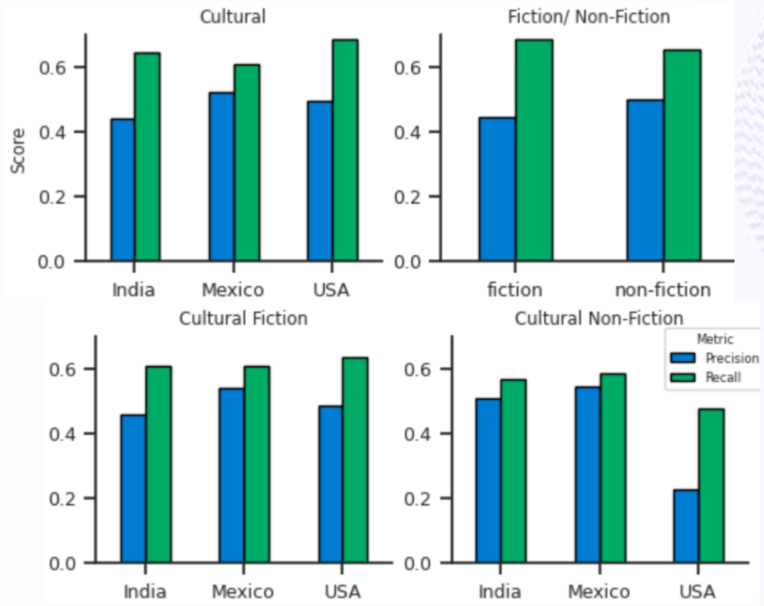


Figure: Precision and recall of the overlap between user and GPT-4o-identified CSIs

What we Propose?

- **Functional and behavioral testing** instead of factual probing
  - Measure the utility and suitability of LLM-based tools across cultures
- Optimizing for **Meta-cultural competency** *[1]* instead of only cultural competency
  - A higher order competency innate to humans
  - Enables intercultural communication. Comprises:
    - **Variational awareness**: self-awareness of cultural differences
    - **Explication & Negotiation Strategies**: conversational strategies that aim to reduce misinterpretations in cross-cultural settings

References:

1. Sharifian, 2013. Globalisation and developing metacultural competence in learning English as an International Language.

Optimize for Variational Awareness (VA)

- H*(Which side does Kenya drive?)* **<** H*(Which side does Asian countries drive?)* **>** H*(Which side does South Asian countries drive?)*
- Test model's **directionality of entropy change** across different cultural dimensions (proxies)
- Model can be *factually correct but directionally incorrect*
- Experiment with Llama-3.1-8B-Instruct on GeoMLAMA *[1]* dataset
- *C* = set of values of a demographic proxy Ex: countries
- *D* = set of values of a semantic domain. Ex: driving (left/right)
- Primary cultural knowledge: $f_k: C \rightarrow D$ and $f_v: P(C) \rightarrow [0, \log(|D|)]$

$$\Delta = \frac{1}{|C|} \sum_{c_i \in C} [\hat{f}_v(C) - \hat{f}_v(\{c_i\})]$$

References:
1. Yin et.al., 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models.

Measuring Variational Awareness: Results

- Accuracy and VA not correlated
- VA least for Iran. Most for India and USA.
- Wide variation of VA across semantic domains
- Low VA in color, measurement, food, indicating strong bias to certain cultures

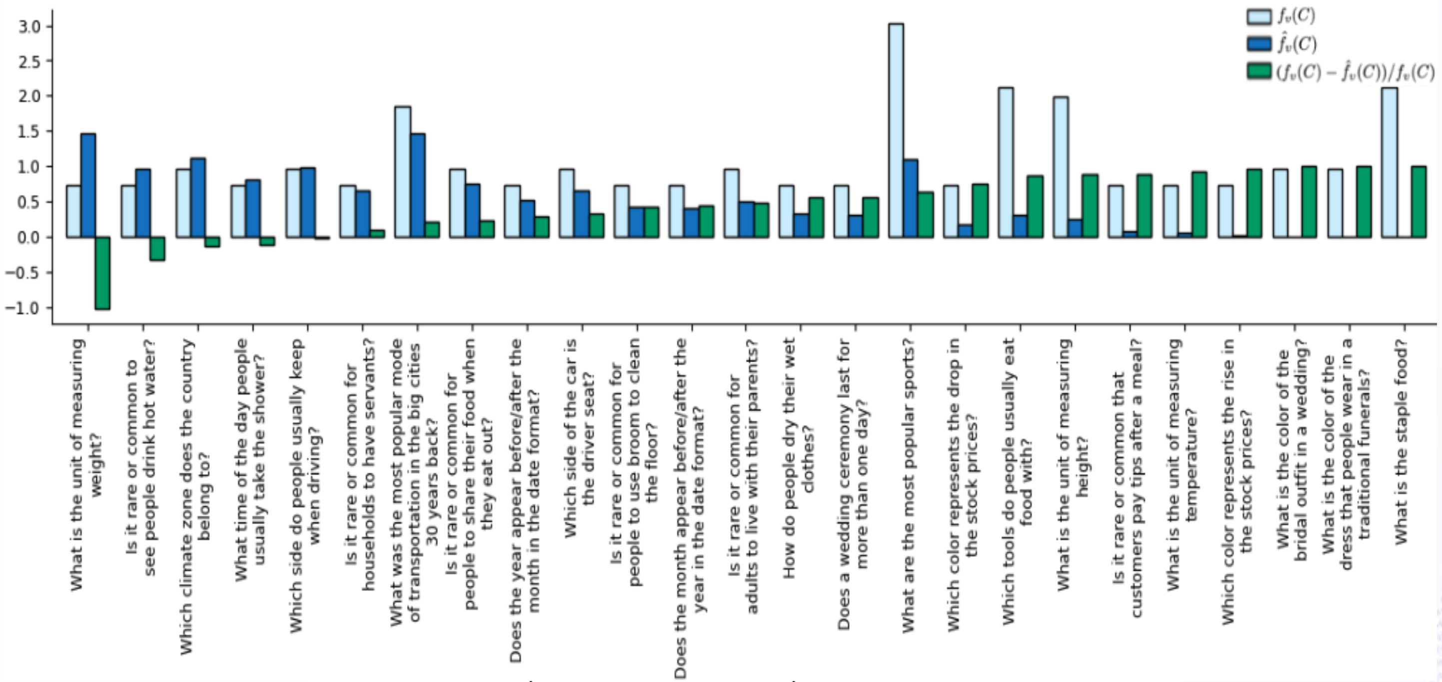


Figure 1: $f_v(C)$, $\hat{f}_v(C)$, and $(f_v(C) - \hat{f}_v(C))/f_v(C)$ for each question.

Metric	China	India	Iran	Kenya	USA
Δ_μ (Δ_σ)	-0.023 (0.494)	-0.049 (0.528)	-0.293 (0.605)	-0.114 (0.665)	0.094 (0.427)
Directionality	0.40	0.48	0.24	0.40	0.48
Knowledge	0.44	0.44	0.44	0.48	0.36

Table 1: Average (Δ_μ) and standard deviation (Δ_σ) of Δ, the fraction of questions with positive/correct directionality and accuracy of the model's response for Llama3.1-8B on GeoMLAMA dataset.

Few of Our Driving Questions

- How can AI/computational technology help in answering questions regarding the interaction between users and cultures?
- How can knowledge of this interaction help us build better and more equitable models and AI systems?
- How is cultural knowledge represented in Large Language Models?
- Can LLMs acquire cultural knowledge on-the-fly as it interacts with users?
- Can cultural knowledge be transferred across domains and regions?

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Thank you

Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, United Arab Emirates

mbzuai.ac.ae