# EtiCor++: Towards Understanding Etiquettical Bias in LLMs

**Ashutosh Dwivedi*** **Siddhant Shivdutt Singh*** **Ashutosh Modi**
Indian Institute of Technology Kanpur (IIT Kanpur)
{ashutoshd20,siddhss20}@iitk.ac.in ashutoshm@cse.iitk.ac.in

## Abstract

In recent years, researchers have started analyzing the cultural sensitivity of LLMs. In this respect, Etiquettes have been an active area of research. Etiquettes are region-specific and are an essential part of the culture of a region; hence, it is imperative to make LLMs sensitive to etiquettes. However, there needs to be more resources in evaluating LLMs for their understanding and bias with regard to etiquettes. In this resource paper, we introduce **EtiCor++**, a corpus of etiquettes worldwide. We introduce different tasks for evaluating LLMs for knowledge about etiquettes across various regions. Further, we introduce various metrics for measuring bias in LLMs. Extensive experimentation with LLMs shows inherent bias towards certain regions.

## 1 Introduction

In recent times, Large Language Models (LLMs) have shown drastic improvements across almost all NLP tasks involving language understanding and generation (Chang et al., 2024; Patra et al., 2023; Dong et al., 2022; Zhong et al., 2024), resulting in wide-spread adoption in real life applications such as using LLM as personal digital assistants where the LLM is used for querying about various kinds of information including those related to cultural aspects of human societies. Consequently, the NLP research community has recently started focusing on evaluating and improving cultural understanding (and possible biases) of LLMs (Hershcovich et al., 2022; Abrams and Scheutz, 2022; Li et al., 2024b). It has resulted in the need to develop new culture-centric tasks and datasets. Culture is a multi-faceted topic and has been studied in the NLP community via various proxies (Adilazuarda et al., 2024). One aspect of culture is Etiquettes.[1] Etiquettes can be generic (common across the majority of societies/regions) as well as localized (specific to a society/region). LLMs have been trained on almost the entire internet's data (Villalobos et al., 2024) and have very likely picked up information about etiquettes in various societies. However, it remains to be evaluated if LLMs are able to understand intricate and subtle differences in social norms across cultures and are possibly biased towards certain cultures. Since LLMs are increasingly being used for seeking information, potential etiquettical biases can have detrimental consequences for the user. Hence, there is a need for evaluation and understanding of inherent biases. In this resource paper, we attempt to achieve this goal. In a nutshell, we make the following contributions:

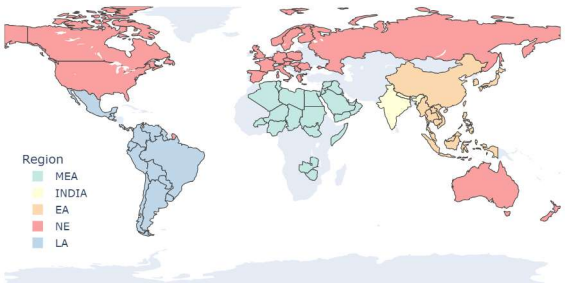



Figure 1: Regions covered under **EtiCor++**

- In this paper, we introduce a new language resource: **EtiCor++**, a large English corpus of $48K$ etiquettes that cover the majority of regions across the globe as shown in Fig. 1.
- We perform an in-depth analysis of **EtiCor++**. Though etiquettes vary from region to region, there are some commonalities. We develop an algorithm (Algorithm 1) for measuring the correlation between etiquettes of different regions to check for similarities and differences.
- In addition to the existing task of Etiquette Sensitivity (Dwivedi et al., 2023), we propose two new tasks (Region Identification and Etiquette Gener-

*Equal Contribution

[1]In this work, we follow the previous definition of etiquette as defined in Dwivedi et al. (2023): a set of social norms/conventions or rules that tell how to behave in a particular social situation.



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 9355–9376

ation) to evaluate LLMs for the understanding of etiquettes across regions.

- We propose seven new algorithmic metrics (§4) for measuring Etiquettical bias in LLMs: Preference Score, Bias For Region Score, Pairwise Regions Bias Score, Generation Alignment Score, Odds Ratio, and two variants of Incremental Option Testing.
- We conduct an extensive set of experiments on five popular LLM models (Llama3.1, Phi-3.5-mini, Gemma2, Gemini, and GPT-4o); our experiments show that LLMs tend to prefer certain regions more than others when it comes to social norms. We release dataset and code via GitHub: `https://github.com/Exploration-Lab/Eticor-Plus-Plus`

## 2 Related Work

**Culture-centric Research in NLP Community:** With the aim to deploy NLP technologies (e.g., LLMs) in human societies, recent research in the NLP community has focused on ethics and culture-centric techniques and models (Adilazuarda et al., 2024; Ziems et al., 2023; Agarwal et al., 2024). Various works have been proposed covering social reasoning (Jiang et al., 2021), cross cultural understanding (Pandey et al., 2025), social dimensions (Hershcovich et al., 2022), CultureLLM (Kovac et al., 2023), cultural corpora (Nguyen et al., 2023; Cao et al., 2024; Ammanabrolu et al., 2022), LLM alignment (AlKhamissi et al., 2024), and inter alia. Due to space constraints we provide more details in App. A.

**Comparison with `EtiCor`:** We are inspired by Dwivedi et al. (2023), where the authors create a corpus of etiquettes (**`EtiCor`**) from major regions of the world and propose the task of Etiquette sensitivity. Since **`EtiCor`** is available under open-source license, our work **`EtiCor++`** takes **`EtiCor`** as the starting point (removes some of the noisy samples by manually analyzing it) and extends **`EtiCor`** significantly from $35K$ etiquette text (each text roughly equivalent to a sentence) to $48K$. We have included many more diverse cultures around the world, such as the Aborigines of Australia, Maori in New Zealand, China, Russia, and southern parts of Africa, which were missing in the previous dataset. The corpus coverage per region has been expanded from a set of few countries to several nearby countries. The consensus of joining the countries to this list was based on the idea of common inclusion. We perform an in-depth analysis and propose a new algorithm for measuring the correlation between etiquettes belonging to different regions. Previous work had only one task for measuring etiquettical knowledge; we have included new tasks and metrics along with evaluation using the latest LLMs.

**Measuring Bias and Stereotypes:** There has been extensive research on measuring biases and stereotypes in deep models and LLMs (Gallegos et al., 2024; Shrawgi et al., 2024). This paper highlights only the relevant works (details in App. A). Researchers have addressed stereotypical biases in models (Koch et al., 2016; Cao et al., 2022; Nadeem et al., 2021; Nangia et al., 2020; Jha et al., 2023; Dev et al., 2024; Das et al., 2023; Palta and Rudinger, 2023), persona bias (Wan et al., 2023b), effect of cultural bias on NLU (Wan et al., 2023a; Huang and Yang, 2023).

Example

**Sentence 1:** `Living with your parents after the age of 18 is considered a bad practice.`
**Sentence 2:** `Young people choose to live with their parents even after the age of 18 and it's considered okay.`
**Sentence 3:** `Independent living after a certain age is considered more appropriate in this culture.`

**Motivation for New Metrics:** Existing works have very little coverage (mostly restricted to sentence-level semantic similarity) for evaluating the generative capabilities of LLMs in the context of culture and, in particular, in the context of etiquettes. Consider the example shown above. As per sentence similarity models (sentence-transformers/all-mpnet-base-v2), the first and second sentences are more similar (0.792) than the first and third (0.501), even though the first two convey opposite values. To take care of nuanced responses and their alignment, we propose new metrics inspired by the NLI task (Storks et al., 2019). Most works focus on a sensitivity-based bias analysis where a model is evaluated for culture based on food, names, gender, or other proxies (Adilazuarda et al., 2024). We wanted to have a metric that quantifies the bias of an LLM in mapping etiquettes to cultures/regions to cover broader use cases mentioned in §4.

| Etiquette | Group | Region | Label |
|---|---|---|---|
| Observing seniority and rank are extremely important in business. | Business | EA | Positive |
| It's not alright to eat beef in front of the people. | Dining | India | Positive |
| Hand your tip to the waiter, do not leave it on the table. | Dining | LA | Positive |
| A small burp signifies satisfaction. | Visits | MEA | Positive |
| Familiarize yourself with any posted visitor guidelines or rules before entering a European location. | Travel | NE | Positive |
| Women and men generally eat together. | Dining | MEA | Negative |
| The business meal is generally the time to make business decisions. | Business | LA | Negative |
| It's okay to stand with your hands on your hips while talking with someone. | Visits | EA | Negative |

Table 1: **EtiCor++** corpus examples.

## 3 EtiCor++

**EtiCor++** contains $47,720$ region-specific etiquette texts in English. As done in previous work (Dwivedi et al., 2023) we intentionally do not have a multi-lingual corpus due to reasons related to maintaining compatibility across regions and the possibility of introducing biases during translation (see Limitations for details). We have categorized etiquettes into five regions. Table 1 shows examples of sentences belonging to different regions. Each region is sub-categorized into one of the 4 four social activities (Dining, Travel, Visits, Business). Further, each etiquette is assigned a label: "Positive" (acceptable in the region) or "Negative" (not acceptable in the region). Table 2 shows the corpus statistics. We created **EtiCor++** by scraping, manually cleaning, and refining content from authentic government websites and travel blogs/websites (details in App. B).

**Regions in EtiCor++:** We categorize the etiquettes collected across the globe into five regions (East Asia (EA), Middle East and Africa (MEA), India Subcontinent (IN), Latin America (LA), and North America and Europe (NE) (also see Fig. 1). Compared to **EtiCor**, the region names have also been updated since several new countries were added. Consequently, the region-wise categorization is different from **EtiCor**. Countries that share culture and several other aspects, such as religion, dining, and history, are brought under one region. For example, Russia is included in the North-America-Europe (NE) region due to similarities in dining habits and shared history. We also include some countries in a common region, even though they are geographically far away, such as European countries, Australia, and New Zealand. Note that social norms are very often common in geographically close countries. However, this was not the reason to club countries into one region (details in App. B).

| Region | # Travel | # Business | # Visits | # Dining | Total |
|---|---|---|---|---|---|
| **EA** | 1190 | 2960 | 4878 | 1100 | 10128 |
| **MEA** | 1776 | 3448 | 6984 | 1300 | 13508 |
| **IN** | 364 | 1104 | 2252 | 450 | 4170 |
| **LA** | 1044 | 2058 | 3166 | 996 | 7264 |
| **NE** | 1330 | 3548 | 6284 | 1488 | 12650 |
| **Total** | 5704 | 13118 | 23564 | 5334 | **47720** |

Table 2: Distribution of different etiquette types

Following regions are created:

**a) East Asia (EA):** This region includes Japan, Korea, Taiwan, China, and all the other Southeast Asian countries, e.g., Indonesia, Malaysia, Philippines, Thailand, Vietnam, etc. There is a significant overlap in these countries' cultural and social values; hence, to maintain harmony, they are in one region. Nevertheless, country information is maintained along with the etiquette.

**b) Middle East and Africa (MEA):** We studied the information collected for countries in the Middle East and Africa and excluded texts very niche to certain religious and tribal practices. It was done to maintain consistency across etiquettes in the MEA region. Africa could not be separated from the Middle East due to the lack of data, and a detailed study of the contrast of regions is required. Furthermore, the North Africa and Middle-East regions shared more cultures and practices than Southern Africa. Thus, we only included some Southern African countries with common etiquettes. In the future, once we have more data available, we plan to create a separate region (and sub-regions) for Africa.

**c) Indian Subcontinent (IN):** We created a separate region for India (and its neighboring countries) due to its vibrant sociocultural diversity. We also include Nepal in this region due to a high overlap in the social practices between the two countries. We use the terms India and Indian Subcontinent interchangeably.

**d) Latin America (LA):** This region has a large geographical area covering diversity in etiquettes. After an in-depth study of the cultural similarities, Cuba and Colombia are included in this region.

**e) North America and Europe (NE):** This region, due to prominent social and cultural commonalities, includes the U.S.A., Canada, Australia, New Zealand, and Russia. Even though these countries are geographically apart, they have high cultural similarity as well as historical alignment, hence these are clubbed together.

**Inter-Region Analysis:** We analyzed the correlation between etiquettes across five regions to study the similarities and differences in social norms globally. Given the complex sociocultural nature

**Algorithm 1** Inter-Region Correlation

**Input:** $\{E_i^{(R_j)} \forall i \in \{1,\ldots,n_{R_j}\}, j \in \{1,\ldots,5\}\}$ : Etiquettes from each of the five regions.
$G \in \{Dining, Travel, Business, Visits\}$
**Output:** Corr$(R_j, R_k)$ $\forall j,k \in \{1,\ldots,5\}; j \neq k$
Start:
Initialize Corr$(R_j, R_k)$ = [5][5]
Calculate embedding for each etiquette using SBERT:
$S_{E_i^{(R_j)}} = SBERT(E_i^{(R_j)})$
**for** $j$ in $\{1,\ldots,5\}$ **do**
  **for** $i$ in $\{1,\ldots,n_{R_j}\}$ **do**
    Initialize $CorrList(n_{R_j}, 5)$ = [][]
    **for** $k$ in $\{1,\ldots,5\}$; $k \neq j$ **do**
      Initialize $SimList^{(j)}(i)$ = []
      **for** $l$ in $\{1,\ldots,n_{R_k}\}$ **do**
        **if** $G(E_i^{(R_j)}) \neq G(E_l^{(R_k)})$
          **continue**
        $sim(i,l) = CosSim(S_{E_i^{(R_j)}}, S_{E_l^{(R_k)}})$
        append $sim(i,l)$ in $SimList^{(j)}(i)$[]
      **end for**
      $m = argmax(SimList^{(j)}(i))$
      Determine relationship ($\mathcal{R}$) using MNLI model
      $\mathcal{R} \in$ {Supportive, Contrastive} $\in \{+1, -1\}$
      $\mathcal{R}(i,m)$ = RoBERTa-MNLI$(E_i^{(R_j)}, E_m^{(R_k)})$
      $Corr(i^{(j)}, m^{(k)}) = \mathcal{R}(i,m) * sim(i,m)$
      Append $Corr(i^{(j)}, m^{(k)})$ in $CorrList[i][]$
    **end for**
  **end for**
  Corr$(R_j, R_k)$ = $mean(CorrList[:][k])$
  Append Corr$(R_j, R_k)$ in Corr[j, k]
**end for**

of etiquettes, it is not straightforward to measure correlation; however, in this paper, we adopted the simplest possible approximate measure based on semantic similarity and NLI (also see Limitations section). We propose Algorithm 1. Note, as explained above, this algorithm only serves as a proxy for measuring correlation between regional etiquettes. In Algorithm 1, for finding the correlation between a region with others, first the similarity (using SBERT model (Reimers and Gurevych, 2019a)) between an etiquette is compared with etiquettes belonging to the same group (dining, travel, business, and visits); next among all the etiquettes of other regions (with which similarities were calculated), the one with maximum value is selected and compared (via RoBERT-MNLI model (Liu et al., 2019)) with the original etiquette to find out if it supports it or contradicts it. We wanted to use a pre-trained off-the-shelf NLI model for our experiments. Consequently, we went with the most readily available model: RoBERTa-MNLI. Correlation is approximated by taking the product of similarity and NLI score. The process is repeated for each of the etiquettes in the original region.

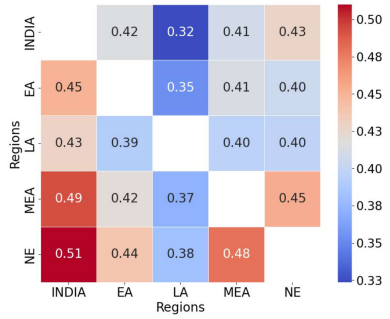



Figure 2: Region wise Correlation

Fig. 2 shows the correlation between $R_j$ and $R_k$ with $R_j$ at x-axis. Note (as can be inferred from Algorithm 1), Corr$(R_j, R_k) \neq$ Corr$(R_k, R_j)$. The important point to note here is that the correlations are not diagonally symmetric because the number of data points in each region is not same. Hence, a region R1 can have a high overlap with the R2 region, but the percentage of points considered could make up only 20% of the total points of R2, thus enabling R2 to have a higher match with other regions. It also gives an idea of the global distribution of data points available on the internet and its effect on LLM during training. As can be observed, the LA region has the least similarity with the rest of the regions (possibly because of geographical distances) and the highest similarity with Europe (since Europeans colonized it). IN has large similarities with other regions (possibly because they were colonized at various times in history). We also calculate the group-wise correlation between regions (see App. C). We also provide statistics related to General Etiquettes in App. B.4.

## 4 Tasks and Bias Metrics

We use **EtiCor++** to check LLMs for cultural bias. For this, we created various tasks and metrics for measuring bias, as outlined below.

**Etiquette Sensitivity (ES) Task:** This task is similar to the one introduced in Dwivedi et al. (2023). Given an etiquette, the task is to predict whether the etiquette is acceptable or unacceptable for a region. We evaluate LLMs via zero-shot setting (App. D provides the prompt). This task is useful as we need the models to be sensitive to different cultures and not discriminate against any of them. The models should not deem some cultural values acceptable while others are unacceptable. ES is measured using the standard metric of Accuracy and F1 score.

**Region Identification (RI) Task:** This newly introduced task aims to test if a model can correctly

identify the region corresponding to an etiquette. The model is provided with an etiquette text and asked to identify the region from a list of regions (see App. D for the prompt). We created this task by keeping the following use cases in mind. Let's say that a person asked the LLM to suggest a gift for their friend's wedding. However, the LLM is unaware of the friend's cultural belonging. Suppose the model responded with "`You can gift them white flowers, as it represents purity and peace`" but then you respond with "`In our culture, White is an ominous color for us. Please suggest a different gift.`" Now, the model can actually guess what culture is involved here (e.g., East Asian) and respond accordingly. There are many other use cases, such as asking the model to assist you in writing a speech at somebody's funeral or promotion, which can involve etiquette regarding first and last names, etc. We devise three metrics to evaluate a model.

***1. Preference Score ($\mathbf{PS(R)}$)*** for a region $\mathbf{R} \in$ {EA, IN, MEA, LA, NE} calculates how often the model prefers to select the region across all etiquettes in the corpus, i.e.,

$$\mathbf{PS(R)} = \frac{\sum_{i=1}^{N} \mathbb{I}_{\mathbf{R}==RI(E_i)}}{N}$$

where, $\{\mathbb{I}_a = 1 \text{ if } a = \texttt{True}\}$ is the indictor function, $N$ is total number of etiquettes in the corpus and $RI(E_i)$ is the answer generated by the model for the RI task query. A higher value of $\mathbf{PS(R)}$ than expected is indicative of a model's bias towards a region. The expected value for each region is their share in the actual data distribution (NE - 26.50%, IN - 8.73%, EA - 21.22%, LA - 15.22%, MEA - 28.30%). To estimate this deviation, we calculate

$$(\mathbf{PS(R)}\% - \mathbf{D(R)}),$$

where $\mathbf{D(R)}$ is the percentage share of the region $\mathbf{R}'s$ data in the whole dataset. We also calculate standard deviation ($\sigma_{PS(R)}$) for each model (§E.1).

***2. Bias For Region Score ($\mathbf{BFS(R)}$)*** for a region $\mathbf{R}$ is calculated by iterating over all etiquettes not in $\mathbf{R}$ and checking how often the RI query returns $\mathbf{R}$ as the answer, i.e.,

$$\mathbf{BFS(R)} = \frac{\sum_{\mathbf{R}' \neq \mathbf{R}} \sum_{i=1}^{N_{\mathbf{R}'}} \mathbb{I}_{RI(E_i)==\mathbf{R}}}{\sum_{\mathbf{R}' \neq \mathbf{R}} N_{\mathbf{R}'}}$$

Using this score, we look at the defaulting behavior of the LLMs. We are trying to quantify that whenever a model is wrong about the region of an etiquette, what region does it prefer in those cases. A high $\mathbf{BFS(R)}$ score indicates model bias for a region. Similar to $\mathbf{PS(R)}$, we calculate standard deviation ($\sigma_{BFS(R)}$) for each model (§E.2).

***3. Pairwise Regions Bias Score ($\mathbf{BSP(R, R')}$)*** Given RI predictions for etiquettes in region $\mathbf{R}$, $\mathbf{BSP(R, R')}$ assess how often the incorrect predictions are confused for region $\mathbf{R}'$, i.e.,

$$\mathbf{BSP(R, R')} = \frac{\sum_{i=1}^{N_{\mathbf{R}}} \mathbb{I}_{RI(E_i)==\mathbf{R}'}}{\sum_{i=1}^{N_{\mathbf{R}}} \mathbb{I}_{RI(E_i)\neq\mathbf{R}}}$$

BSS is not a symmetric metric (higher the score more the bias).

**Etiquette Generation (EG) Task:** Restricting the LLM to specified options, as in previous tasks, might prevent us from observing the generational biases of the model. We propose the Etiquette Generation (EG) task where a model is provided with an etiquette of one region in one context (e.g., group or etiquette type) and is asked to generate an etiquette for the other regions in the same context (see App. D for the prompt template). We propose this task as we want the real-world LLM response to be non-stereotypical and non-contradictory. A possible use-case may involve a person (belonging to a region) simply wanting to know about the traditions, norms, values, and etiquette of any other culture in a particular context. We want to qualitatively and quantitatively assess these properties. We propose two new metrics (Generation Alignment Score (GAS) and Odds Ratio):

***1. Generation Alignment Score (GAS):*** For all the generated etiquettes for a particular region $\mathbf{R}$, we would like to measure the alignment and consistency of generated responses for other regions. For this, we first calculate the embeddings for each generated etiquette using *sentence-transformers/all-mpnet-base-v2* model (Reimers and Gurevych, 2019b; Song et al., 2020) and then filter out etiquettes that have similarity less than a threshold (selected as 0.55 via initial experiments). However, this also resulted in etiquette that had contradictory stances being selected. Consequently, we used Natural Language Inference (NLI) to calculate the entailment and contradiction scores (a threshold of 0.90 was used to filter out). The GAS score is defined as:

$$\frac{\#entailment}{\#entailment + \#contradictions}$$

GAS helps us gauge the robustness and confidence of the model. GAS score lies between 0 (worst) and 1 (best).

***2. Odds Ratio:*** Inspired by the work of Naous et al. (2024), we apply the Odds Ratio test to identify

the dominating themes of the generated etiquettes. In particular, we analyze frequent Nouns, Verbs, and Adjectives in the responses generated for each pair of regions. This qualitative metric aims to investigate the generation of stereotypes for certain regions.

**Incremental Option Testing:** The Region Identification task involves providing a query with one correct and a set of incorrect choices. However, it does not provide the means to evaluate the stability and confidence of the model about a set of choices. We propose the task of **Incremental Option Testing** for this purpose. We intend to create metrics that also map the stability of the model concerning the etiquette on which they are making the decisions. It helps us to understand the randomness they might exhibit when new data is presented and how it changes their decisions, ultimately resulting in changes in their bias in light of new information. In this task, a query is posed to the model along with options to select an answer (MCQA style). Initially, two options are provided and the model's response is recorded. Subsequently, the same query is posed again but with the addition of one more choice. Again, the model's response is recorded. The consistency within the sequence of predictions made by the model is observed. Algorithm 2 gives details. We consider two possible types of increments.

**Algorithm 2** Incremental Option Testing

**Input:** $\mathcal{Q} = \{Q^{(i)} \mid i = 1, \ldots, |\mathcal{Q}|\}$: Set of questions
**Output:** Choices, $\{C_j^{(i)} \mid i = 1, \ldots, |\mathcal{Q}|; j = 1, \ldots, m\}$: Model predictions for each question and iteration $(j)$. Given $(m+1)$ total number of options (regions).
**for** $i = 1, \ldots, |\mathcal{Q}|$ **do**
  Present question $Q^{(i)}$
  Present initial options $[O_1^{(i)}, O_2^{(i)}]$
  Let the model's predicted choice be $C_1^{(i)}$
  Initialize $j = 2$
  **while** additional options remain to be tested **do**
    Introduce a new option $O_{j+1}^{(i)}$
    Query the model, yielding choice $C_j^{(i)}$
    Increment $j \leftarrow j + 1$
  **end while**
  Record set of predictions $\{C_j^{(i)} \mid j = 1, \ldots, m\}$
**end for**

***1) Correct Option at the Start Increment:*** In this method, the correct choice is introduced initially at index 0, and subsequently incorrect choices are introduced (in decreasing order of correlation based on Fig. 2). The expected behavior of the model is to select the correct option at the first choice and not waiver from this decision even when new options are added to the list (prompt in App. D). This variant is evaluated using two metrics: *Accuracy:* We have the accuracy of the models for each set of options for a particular etiquette. The general trend shows a decline in accuracy with increase in number of options (§5). Furthermore, some models perform decently in the initial step which suggest that given the limited number of choices and possibility of their bias source lacking, they will have high accuracy (§5). *Distancing:* It measures distance between true and predicted choice (see App. E.3 Algorithm 4 for calculation details). It gauges the increase in bias of the model as more choices are presented. An increase in magnitude of negative score states that the model is moving toward biased opinion and a sharper fall indicates a greater tendency to move towards the least possible options. We want the model to be as close to zero as possible.

***2) Correct Option at the End Increment:*** In this method, the choices are introduced one at a time, with the correct option introduced at the end. The incorrect options are added in the increasing order of region-wise correlation (Fig. 2). The typical behavior expected from the model is to pick the newest option from the choices. This approach is evaluated with three metrics:

*a) Closeness:* This metric help us to understand how much is the bias of a model near the optimal value. Algorithm 3 gives the details of the calculation.

*b) Consistency:* A consistency score determines the direction of the decision the model is making under the change of available information. Through this score we try to measure how consistent the models are when they are probed for the etiquette. Based on Algorithm 3,

$$\text{Consistency Score}^{(j)} = \frac{\sum_{i=1}^{N} \mathbb{I}(S_i^{(j)} = -1)}{N},$$

where,

$$\mathbb{I}(S_i^{(j)} = -1) = \begin{cases} 1 & \text{if } S_i^{(j)} = -1, \\ 0 & \text{otherwise.} \end{cases}$$

*c) Option Sensitivity:* This score helps us to understand the extent to which the model is bothered by addition of a new information. We have evaluated this score in cases where the model is inconsistent and moved away from the correct option.

$$\text{Sensitivity Score}^{(j)} = \frac{\sum_{i=1}^{N} \mathbb{I}(S_i^{(j)} = -2)}{N},$$

| Region | ChatGPT-4o | | Gemini-1.5 | | Llama-3.1 | | Gemma-2 | | Phi-3.5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc.(%) | F1-Score | Acc. | F1-Score | Acc. | F1-Score | Acc. | F1-Score | Acc. | F1-Score |
| NE | 42.0 | 0.59 | 43.5 | 0.61 | 46.4±1.24 | 0.64±0.02 | 45.5±2.05 | 0.61±0.012 | 45.5±2.88 | 0.62±0.015 |
| INDIA | 44.0 | 0.61 | 41.0 | 0.58 | 45.2±2.03 | 0.62±0.025 | 43.2±1.94 | 0.60±0.009 | 45.9±2.86 | 0.63±0.024 |
| EA | 41.1 | 0.58 | 37.6 | 0.54 | 41.3±1.22 | 0.60±0.021 | 40.7±1.76 | 0.57±0.005 | 42.2±1.62 | 0.59±0.015 |
| LA | 42.5 | 0.59 | 39.1 | 0.56 | 41.8±1.14 | 0.61±0.01 | 38.9±2.0 | 0.54±0.012 | 41.2±1.96 | 0.58±0.006 |
| MEA | 42.7 | 0.59 | 38.6 | 0.55 | 43.8±2.23 | 0.63±0.011 | 42.6±1.78 | 0.58±0.016 | 42.8±2.51 | 0.60±0.013 |
| **Average** | 42.3 | 0.59 | 40.0 | 0.57 | 43.9 | 0.62 | 42.1 | 0.58 | 43.5 | 0.60 |

Table 3: Region-wise Performance of LLMs on the Etiquette Sensitivity task

| Region | ChatGPT-4o | | Gemini-1.5 | | Llama-3.1 | | Gemma-2 | | Phi-3.5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PS(%) | BFS(%) | PS | BFS | PS | BFS | PS | BFS | PS | BFS |
| NE | 30.6(10.6) | 38.1 | 48.1(28.1) | 65.0 | 23.7±1.33 (-2.8) | 25.5±1.12 | 30.0±2.28 (3.49) | 35.4±1.64 | 32.7±1.33 (6.19) | 57.3±1.24 |
| INDIA | 6.4(-13.6) | 2.7 | 7.5(-12.5) | 1.8 | 13.9±1.03 (5.16) | 22.1±3.71 | 26.5±1.21 (17.7) | 47.2±1.57 | 19.8±1.93 (11.06) | 25.6±1.55 |
| EA | 22.6(2.6) | 20.4 | 16.7(-3.3) | 13.6 | 31.2±1.48 (9.97) | 35.6±0.50 | 12.7±2.53 (-8.52) | 5.5±1.45 | 15.2±2.55 (-6.02) | 8.0±1.41 |
| LA | 12.6(-7.4) | 3.1 | 11.7(-8.3) | 0.5 | 13.1±0.96 (-2.12) | 5.0±0.69 | 11.5±1.81 (-3.72) | 3.1±0.46 | 12.6±1.65 (-2.62) | 3.2±0.55 |
| MEA | 27.9(7.9) | 35.8 | 15.9(-4.1) | 19.1 | 13.9±1.70 (-14.4) | 13.6±0.77 | 12.7±1.26 (-15.60) | 9.3±0.25 | 16.5±3.06 (-11.8) | 7.9±0.49 |
| **Std Dev** | 9.19 | 15.3 | 14.4 | 23.5 | 8.27 | 10.42 | 11.39 | 17.8 | 8.26 | 19.9 |

Table 4: Performance of LLMs on the Region Identification task using PS score and BFS score. The colouring is according to the excess score of the model compared to the expected score (PS(R) - D(R)), indicated in the brackets beside the PS. Last row corresponds to standard deviation $\sigma_{PS(R)}$ and $\sigma_{BFS(R)}$ as described in App. E.

**Algorithm 3** Closeness Metric Calculation

**Input:** $N$: Number of questions.
$M$: Number of choice iterations (ITRs).
$\{C_i^{(j)}\}$: Choice for question $Q_i$ at ITR $j$, $\forall i \in \{1, \ldots, N\}$, $\forall j \in \{0, \ldots, M-1\}$.
$\{O_j\}$: Latest Options introduced at each ITR $j$, $\forall j \in \{0, \ldots, M-1\}$.
**Output:** $\{\text{Closeness}^{(j)}\}$: Closeness value for each phase $j$, $\forall j \in \{0, \ldots, M-1\}$.
**for** $j$ in $\{0, \ldots, M-1\}$ **do**
  **for** $i$ in $\{1, \ldots, N\}$ **do**
    Initialize score $S_i^{(j)} = 0$
    **if** $C_i^{(j)} = O_j$ **then**
      $S_i^{(j)} = 0$
    **else if** $C_i^{(j)} = O_{j-1}$ **then**
      $S_i^{(j)} = -1$
    **else if** $C_i^{(j)} = O_k$ where $k < j-1$ **then**
      $S_i^{(j)} = -2$
    **end if**
  **end for**
  Calculate $\text{Closeness}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} S_i^{(j)}$
**end for**
**Return** $\{\text{Closeness}^{(j)}\ \forall j \in \{0, \ldots, M-1\}\}$

where

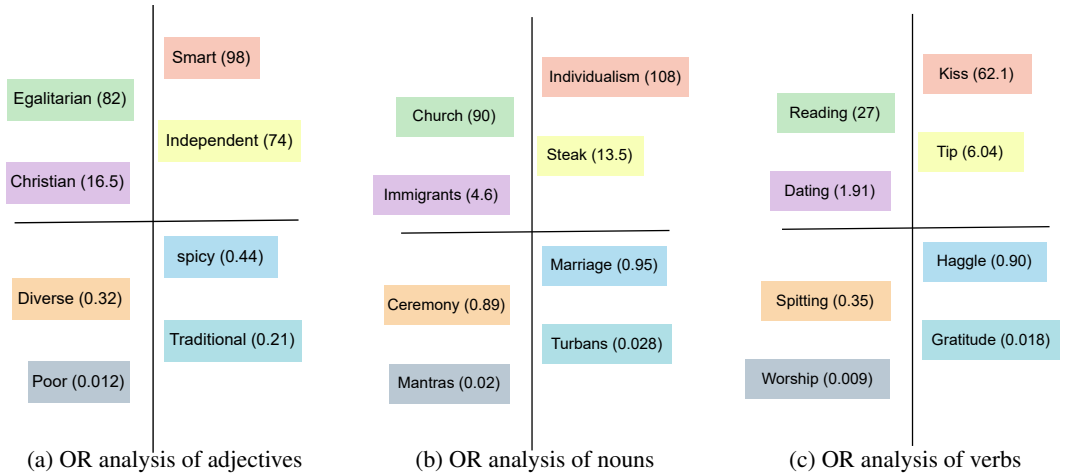
$$\mathbb{I}(S_i^{(j)} = -2) = \begin{cases} 1 & \text{if } S_i^{(j)} = -2, \\ 0 & \text{otherwise.} \end{cases}$$

## 5 Experiments and Results

We experimented with 5 LLMs (mix of closed and open-weights models): `GPT-4o` (OpenAI, 2024), `gemini-1.5-flash` (Gemini, 2024), `Llama-3.1-8B-Instruct` (Meta, 2024), `gemma-2-9b-it` (Google, 2024) and `Phi-3.5-mini-instruct` (Microsoft, 2024). Except for GPT-4o and Gemini, we conducted the quantitative experiments three times (temperature = 0.3 and top-p = 0.9) to account for output variability. Due to high cost of experiments, for GPT-4o and Gemini, we took 200 samples for each region (1000 samples in total) for each of the experiments. Note we did not experiment with the models used in **`EtiCor`** since these are outdated/unavailable and have in general shown to have poorer performance than the more recent models used in this paper (details in §F.2).

**Etiquette Sensitivity (ES):** The results are presented in Table 3, some examples are given in App. Table 16. As can be observed, in general, most of the models have higher performance in the NE region; this may be due to the internet data used to train these models coming heavily from countries in this region. All models show poor performance on cultures (such as LA, MEA, and EA) that have low resources available online. This demonstrates a presence of bias arising from a lack of knowledge regarding these cultures. Overall, the `Llama` model has the best performance on average and across regions. Another surprising observations is that large and competent models like `ChatGPT-4o` and `Gemini-1.5` are not able to beat extremely small models such as `Phi` and `Llama` Please note that LLMs sometimes tend to abstain in some cases where they do not understand the etiquette fully or if they find the etiquette content controversial. We don't include these in our calculations (see §F.1 for details).

**Region Identification Task:** Table 4 shows the results with $\mathbf{PS(R)}$ and $\mathbf{BFS(R)}$ metrics. We measure performance using deviation from the expected values of scores. We find that `Phi` and




(a) OR analysis of adjectives (b) OR analysis of nouns (c) OR analysis of verbs

Figure 3: Odds Ratio analysis of etiquettes generated by Llama-3.1 for Europe vs India. The figure shows the words followed by their Odds Ratio.

| Region ($R/R'$) | NE | INDIA | Phi-3.5 EA | LA | MEA |
|---|---|---|---|---|---|
| NE | - | 61.4±2.4 | 20.2±1.65 | 5.1±1.4 | 13.3±1.5 |
| INDIA | 78.8±2.5 | - | 14.6±0.5 | 4.5±0.7 | 2.1±0.8 |
| EA | 68.0±3.7 | 21.2±0.7 | - | 3.3±0.7 | 7.5±1.3 |
| LA | 54.6±1.8 | 20.7±0.9 | 12.4±1.6 | - | 12.3±1.1 |
| MEA | 51.9±0.7 | 33.8±3.3 | 7.5±0.8 | 6.8±1.0 | - |

Table 5: Bias Score Pairwise(%) for Phi-3.5

| Region | ChatGPT | Gemini | Llama | Gemma | Phi |
|---|---|---|---|---|---|
| NE | 0.26 | 0.10 | 0.41±0.011 | 0.40±0.04 | 0.25±0.02 |
| INDIA | 0.60 | 0.36 | 0.84±0.019 | 0.64±0.009 | 0.23±0.014 |
| EA | 0.32 | 0.15 | 0.72±0.016 | 0.66±0.018 | 0.37±0.009 |
| LA | 0.34 | 0.13 | 0.34±0.014 | 0.45±0.015 | 0.17±0.015 |
| MEA | 0.53 | 0.24 | 0.80±0.01 | 0.52±0.03 | 0.34±0.023 |
| **Average** | 0.40 | 0.20 | 0.62 | 0.53 | 0.27 |

Table 6: Performance comparison of LLMs on the etiquette generation task by region using GAS.

Llama have comparably lower deviations ($\sigma_{PS(R)}$ and $\sigma_{BFS(R)}$) than other models. We also see that the models rarely prefer Latin America, Middle East Africa, or East Asia as answers and underestimate when compared to their expected scores. The results show a preference for models towards Western countries (NE region) and bias against under-represented regions. Pairwise Resion Bias Score ($\mathbf{BSP(R, R')}$) for Phi are shown in Table 5 (results for other models are in App. Table 15); we see that all the models select the high-resource regions as their answers and tend to neglect others. In some rare cases, we see models like Llama and Gemma to be biased for regions like EA or INDIA. The low-resource regions, such as LA and MEA, still suffer from low representation. On the other hand, when the model is incorrect for these regions, it overwhelmingly selects NE as its answer. The bias against low-resource regions is a common trend across all models. This metric uncovers that the bias for NE is even more prevalent in scenarios where the model hallucinates.

**Etiquette Generation Task:** The generation alignment score is presented in Table 6. It shows how much consistency the model has while generating etiquettes for a region. A score of above 0.5 means that the model generates etiquettes that align with each other more than they contradict while a GAS score of less than 0.5 means more contradictions than entailment. We see that all the models perform very poorly in this task except Llama and Gemma, these models generate consistently aligned etiquettes, especially Llama which is the best-performing model according to the GAS metric. This might be attributed to a greater focus on multilingualism in the training data of these models. Gemini performs the worst, with a score of 0.20 on average which means that it outputs very contradictory etiquettes. We see a reversal in scores that we have been seeing through other metrics. Here, India has very consistently generated etiquettes across most of the models while Native Europe has inconsistently generated etiquettes.

**Odds Ratio:** We conducted Parts of Speech analysis of the generated etiquette of each model and found the odds ratio of dominating terms for each pair of regions, so in total, we have $10 \times 3$ pairs (10 for the number of pair ( = $\binom{5}{2}$) of regions and 3 for Nouns, Verbs, and Adjectives). This gave us a better understanding of not only the stereotypes (mostly represented by Adjectives) but also of relevant concepts (through Nouns) and actions (through Verbs). The top words for the pair of European and Indian etiquettes generated by various models are shown in Fig. 3. Plots for other regions are in App. Fig. 9 and App. Fig. 10. Through the qualitative analysis of odds ratios, it is clear that the model (Llama) uses some stereotypi-

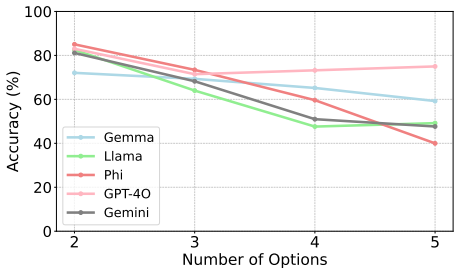


Figure 4: Accuracy for Correct Option at Start

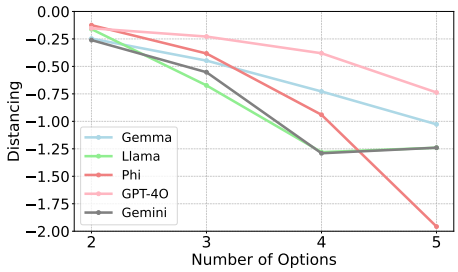


Figure 5: Distancing for Correct Option at Start

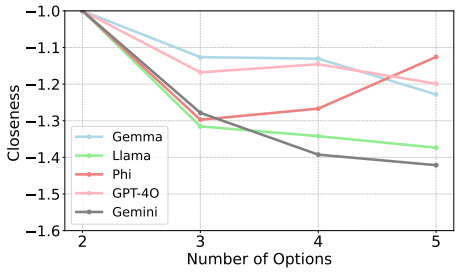


Figure 6: Closeness for Correct Option at End

cal adjectives to describe the Indian subcontinent etiquettes, such as *traditional*, *spicy*, and *diverse*, while it uses *smart*, *egalitarian*, and *independent* to generate etiquettes for Native Europe. An analysis of nouns shows the concepts the model considers important for Native European etiquette vs Indian ones. It generates etiquette about the concepts of *individualism*, *church*, and *steak* for NE while using *marriage*, *ceremony*, and *mantras* as important concepts of India. A similar analysis of verbs can clearly distinguish the difference between relevant actions (good or bad) in both cultures. As per the model, actions such as *kissing*, *dating*, and *tipping* have more importance in NE culture, and actions such as *worship*, *showing gratitude*, and *haggling* have more importance in Indian culture.

**Incremental Option Testing:** Here, we show the main results and scores used to calculate the results are provided in App. F.4 Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15. For *Correct Option at Start Increment* task, we use the following score to evaluate mistakes made by a model.

Accuracy: Fig. 4 shows the accuracy as more options are introduced. We notice that models start with fairly high accuracy and then drop down with each added option. It shows that current LLMs lack a need for prioritization when deciding etiquette. We can observe that the general trend is downward with flattening at the end. `GPT` achieves it faster than the others which shows early recognition.

Distancing: We define distancing as a bias model made as more options are added in *Correct Option at Start Increment* task; Fig. 5 shows the results. An increase in the magnitude of distancing indicates that the model tends to be more biased with an increase in the number of choices. We can observe that Phi performs the worst here and is in line with accuracy stats. This indicates a higher bias in `Phi` in comparison with other models.

For *Correct Option at End Increment* task since the options are provided in increasing correlation value such that the correct answer is appearing at the end, we expect the model to choose the latest added option as it is closest correlation-wise as well as meaning-wise to the correct choice.

Closeness: The closeness trend in Fig. 6 shows the movement of model predictions towards the correct choice of regions for etiquette. The closer the value is toward -1, the closer it gets to the optimal choice. `Phi` model performs better than others and is able to recover as more options are added.

Consistency and Option Sensitivity: Table 7 shows the results for these metrics and the final accuracy of the choice of models. `Gemma-2` was found to be the most consistent model in making its selection and was least option sensitive, causing a lack in its overall accuracy towards making the predictions for etiquettes. This also highlights that a consistent model is may not necessarily be an accurate model.

| Models | Consistency | O-Sensitivity | Accuracy |
|---|---|---|---|
| **Phi-3.5** | 0.92 | 1.09 | 60.75 |
| **Gemma-2** | 2.00 | 0.75 | 49.89 |
| **Llama-3.1** | 1.17 | 1.03 | 51.55 |
| **Gemini-1.5** | 1.29 | 0.99 | 62.78 |
| **ChatGPT-4o** | 0.97 | 0.89 | 68.92 |

Table 7: Consistency and Option Sensitivity

## 6 Conclusion

In this paper, we introduce **`EtiCor++`** and propose new tasks for evaluating LLMs for etiquettes. We also develop new measures for quantifying bias in LLMs. Our experiments show inherent biases in LLMs. In the future, we plan to develop methods for mitigating etiquettical biases in LLMs.

## Limitations

**EtiCor++** is entirely in English. Note that these are scrapped only from websites that originally describe the etiquette in English. This helps to maintain uniformity across regions and makes the corpus usable for diverse set of researchers. Describing an etiquette in the original language of a region (it belongs to) could introduce some priming effects in LLMs; hence, as done in **EtiCor**, we kept the corpus in the English language to enable broader usability and bias testing of LLMs. Accordingly, we took etiquette from internet sources, which were in English, to avoid any errors that may occur during automated translations. In the future, we plan to make **EtiCor++** multilingual. This will help to analyze the effect of language on bias in LLMs.

Etiquettes are a complex socio-cultural phenomenon, and measuring the similarity between etiquettes across regions is not straightforward. In this paper, we develop a proxy method (based on semantic similarity) for measuring the correlation between etiquettes of various regions. This is not a perfect metric and is prone to errors.

In this paper, we measured the bias in LLMs about Etiquettes. However, we do not propose any bias mitigation strategies. Developing techniques for removing bias in models is an involved process, and we leave it for future work.

## Ethical Considerations

The proposed corpus will be released only for research purposes, and we do not plan to deploy any system built on **EtiCor++**. To the best of our knowledge, we do not foresee any ethical consequences of the dataset and bias metrics proposed in this paper.

## References


Mitchell Abrams and Matthias Scheutz. 2022. Social norms guide reference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

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.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Emad A. Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. *Preprint*, arXiv:2410.12880.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM*

*Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2024. Building socio-culturally inclusive stereotype resources with community engagement. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547, Abu Dhabi, UAE. Association for Computational Linguistics.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Comput. Surv.*, 55(8).

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.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORM-SAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Gemini. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Google. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

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.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574.

A. Koch, R. Dotsch, C. Unkelbach, and H. Alves. 2016. The abc of stereotypes about groups: agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110:675–709.

Grgur Kovac, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *CoRR*, abs/2307.07870.

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)*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Xuelin Liu, Pengyuan Liu, and Dong Yu. 2025. What's the most important value? INVP: INvestigating the value priorities of LLMs through decision-making in social scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4725–4752, Abu Dhabi, UAE. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

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.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16366–16393. Association for Computational Linguistics.

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*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

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.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. 2023. Beyond English-centric bitexts for better multilingual language representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15354–15373, Toronto, Canada. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. 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.

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, St. Julian's, Malta. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning (ICML)*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023a. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023b. 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.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. *Preprint*, arXiv:2410.12971.

Damin Zhang, Yi Zhang, Geetanjali Bihani, and Julia Rayz. 2025. Hire me or not? examining language model's behavior with occupation attributes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7891–7911, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Hanqi Jiang, Hanxu Chen, Yiwei Li, Junhao Chen, Huawen Hu, Yihen Liu, Huaqin Zhao, Shaochen Xu, Haixing Dai, Lin Zhao, Ruidong Zhang, Wei Zhao, Zhenyuan Yang, Jingyuan Chen, Peilong Wang, Wei Ruan, Hui Wang, Huan Zhao, Jing Zhang, Yiming Ren, Shihuan Qin, Tong Chen, Jiaxi Li, Arif Hassan Zidan, Afrar Jahin, Minheng Chen, Sichen Xia, Jason Holmes, Yan Zhuang, Jiaqi Wang, Bochen Xu, Weiran Xia, Jichao Yu, Kaibo Tang, Yaxuan Yang, Bolun Sun, Tao Yang, Guoyu Lu, Xianqiao Wang, Lilong Chai, He Li, Jin Lu, Lichao Sun, Xin Zhang, Bao Ge, Xintao Hu, Lian Zhang, Hua Zhou, Lu Zhang, Shu Zhang, Ninghao Liu, Bei Jiang, Linglong Kong, Zhen Xiang, Yudan Ren, Jun Liu, Xi Jiang, Yu Bao, Wei Zhang, Xiang Li, Gang Li, Wei Liu, Dinggang Shen, Andrea Sikora, Xiaoming Zhai, Dajiang Zhu, and Tianming Liu. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *Preprint*, arXiv:2409.18486.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

## Appendix

## Table of Contents



## List of Tables



## List of Figures

## A Related Work

**Culture-centric Research in NLP Community:** With the aim to deploy NLP technologies (e.g., LLMs) in human societies, recent research in NLP community has focused on ethics and culture centric techniques and models (Adilazuarda et al., 2024; Ziems et al., 2023; Agarwal et al., 2024). For example, Jiang et al. (2021) have proposed Delphi, an AI system for social reasoning, Hershcovich et al. (2022) characterized culture along four prominent dimensions: common ground, objectives and values, linguistic style and form, and aboutness. Li et al. (2024a) utilize semantic data augmentation along with WVS (World Value Survey) to train CultureLLM, Kovac et al. (2023) look at how the values exhibited by the LLM change with changing context, Nguyen et al. (2023) collect a corpus of cultural common-sense knowledge to help LLMs generate culturally relevant responses. Alghamdi et al. (2025) study trustworthiness of LLMs in Arabic, Liu et al. (2025) investigate value priority of LLMs in using realistic social scenarios, Rao et al. (2024) develop a framework for measuring cultural adaptability of LLMs, AlKhamissi et al. (2024) study cultural alignment of LLMs when prompted with low-resource language and sensitive topics vs high-resource language, Cao et al. (2024) introduce cuDialog benchmark to assist dialogue agents in cultural alignment, Fung et al. (2023) proposed a framework to automatically extract cultural norms from multi-lingual conversations, Ammanabrolu et al. (2022) use text based games and create agents that adhere to social norms and values in an interactive setting. In this paper, it is not possible to exhaustively cover all the works, we refer the reader to a comprehensive survey on research on culture in NLP community by Adilazuarda et al. (2024). Our work is inspired by the work by Dwivedi et al. (2023), where the authors create a corpus of etiquettes from major regions of the world and propose the task of Etiquette sensitivity.

**Measuring Bias and Stereotypes:** There has been extensive research on measuring biases and stereotypes in deep models and LLMs (Gallegos et al., 2024; Shrawgi et al., 2024). In this paper, we highlight only the relevant works. Koch et al. (2016) propose the ABC (Agency-Belief-Communion) model of stereotypes that analyses the stereotypes associated with groups based on three dimensions, Cao et al. (2022) use the ABC stereotype model and a sensitivity test to discover stereotypical group-trait associations in LLMs. Wan et al. (2023b) identify and formulate persona bias expressed by dialogue systems while adapting to a particular persona. Nadeem et al. (2021) introduce a stereotype dataset, StereoSet, to simultaneously evaluate the language modeling ability along with stereotypical bias in LLMs. Nangia et al. (2020) introduce CrowS-Pairs, a stereotype dataset, to measure bias in LLMs along nine dimensions such as race, age, gender etc against historically disadvantaged groups in the U.S. Wan et al. (2023a) study the expression of harmful biases in LLM generated reference letter's style and content. Zhang et al. (2025) evaluate gender stereotypes in LLMs using occupation and hiring based question answering. Do et al. (2025) make use of demographic and historical opinion data to represent the values, norms and beliefs of a persona and show the effectiveness of a new type of reasoning: Chain Of Opinions. Banerjee et al. (2025) analyze cultural sensitivity in LLMs by creating a cultural harm test dataset and a culturally aligned preference dataset to restore cultural sensitivity. Xu et al. (2024) create a framework called CultureSPA, for pluralistic culture alignment in LLMs. Huang and Yang (2023) introduce CALI (Culturally Aware Natural Language Inference) dataset to study the effects of cultural norms on language understanding task, and awareness of these norms in LLMs. Jha et al. (2023) make use of LLMs such as GPT-3 and PaLM to increase the coverage of stereotype datasets around the world. Dev et al. (2024) use community engagement to collect a stereotype dataset specific to Indian context. Das et al. (2023) compose a Bengali dataset to evaluate gender, religious and national identity bias in NLP systems. Palta and Rudinger (2023) present FORK, a dataset to probe models for culinary cultural biases. However, the existing metrics are not directly adaptable to our setting for measuring etiquettical bias as explained next.

## B `EtiCor++` Creation Details

This section provides a comprehensive explanation of the processes involved in the collection, preprocessing, cleaning, and filtering of the **`EtiCor++`** dataset.

### B.1 Data Collection

Etiquette data was gathered from a variety of sources, including travel websites, official cultural

web pages maintained by governments, and websites featuring cultural and etiquette-related information for various countries worldwide. Additionally, we incorporated relevant content from tweets and magazine articles referencing etiquette across different cultures and countries. To enhance the dataset's coverage, we scraped specialized web pages containing etiquette guidelines for Australian, Maori, and various African tribal cultures. A sample of data sources is provided in Appendix H.

### B.2 Pre Processing

As part of the preprocessing stage, sentences with a word count of four or fewer were removed, while overly long sentences were summarized. For each region, repetitive etiquette entries were initially identified and removed using an automated python script. This was followed by a manual review to eliminate remaining instances of repeated etiquettes. After that, all the etiquettes were carefully checked or reworded to make sure they were appropriate for the context and made sense as a whole sentence while preserving their original meaning.

### B.3 Data Labeling

We created a list of approximately 100 characteristic words for each of the four categories of etiquette (Dining, Travel, Visits, and Business). This list will be made publicly available via a GitHub repository. Each etiquette was automatically assigned to a category if it contained one of the characteristic words. Etiquettes with none or more than one of these characteristic words were classified as ambiguous. Ambiguous etiquettes were then manually assigned to the most appropriate category based on their content. While manually assigning the etiquettes to one of the four groups, we fixed a few errors in the previous dataset. This manual assignment was done by the authors themselves and inter-annotator agreement was measured by krippendorff's $\alpha = 0.91$.

Each region's etiquettes were further classified into two types, indicated in the "Label" column as "Positive" or "Negative." "Positive" etiquettes refer to behaviors that are acceptable or expected by people in a particular culture or region, whereas "negative" etiquettes denote behaviors that are considered unacceptable within that culture or region.

| Region | # Region Specific | # General |
|---|---|---|
| **EA** | 8218 | 1910 |
| **LA** | 5356 | 1908 |
| **MEA** | 9928 | 3580 |
| **NE** | 7488 | 5162 |
| **IN** | 3312 | 858 |

Table 8: General Etiquette Distribution

### B.4 General Etiquettes

We categorized etiquettes that represent common facts across multiple regions, demonstrating a notable degree of similarity. By general we mean that the etiquettes are acceptable among all the regions. The labeling process involved categorizing etiquettes through common etiquette mapping, ensuring each data point was associated with its closest relation. To enhance accuracy and minimize errors, manual annotation was then conducted for each data point. This step was critical in isolating common etiquettes to ensure precise metric calculations and prevent classification errors for these data points. The distribution of data points is summarized in Table 8.

We also calculated the similarity matrix for only the common etiquettes of each region as described by the process in Algorithm 1. Fig 7 shows that in general the distribution is nearly even throughout all regions and thus there is no particular interference of commonly accepted data types on the model results. We also avoided the use of these general etiquettes while performing tests for the metrics creation thereby removing any inconsistencies.

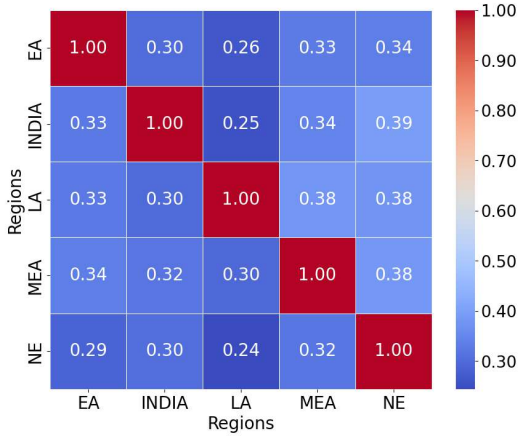



Figure 7: Region-wise Correlation for General Etiquettes

## C `EtiCor++` Region-wise Correlation

Table 9, Table 10, Table 11, Table 12, and Table 13 show the group-wise correlation between regions.

| Region | # Travel | # Business | # Visits | # Dining |
|---|---|---|---|---|
| **EA** | 0.5752 | 0.5858 | 0.6221 | 0.6118 |
| **LA** | 0.8351 | 0.6667 | 0.6934 | 0.7240 |
| **MEA** | 0.8189 | 0.6784 | 0.7635 | 0.7382 |
| **NE** | 0.7067 | 0.7410 | 0.7880 | 0.6194 |

Table 9: Correlation distribution for INDIA

| Region | # Travel | # Business | # Visits | # Dining |
|---|---|---|---|---|
| **INDIA** | 0.4500 | 0.4858 | 0.5686 | 0.4137 |
| **LA** | 0.5591 | 0.5347 | 0.5336 | 0.5374 |
| **MEA** | 0.6054 | 0.5209 | 0.5662 | 0.5261 |
| **NE** | 0.5521 | 0.5500 | 0.5684 | 0.5392 |

Table 10: Correlation distribution for EA

| Region | # Travel | # Business | # Visits | # Dining |
|---|---|---|---|---|
| **INDIA** | 0.5176 | 0.4366 | 0.4518 | 0.4490 |
| **EA** | 0.4897 | 0.4665 | 0.4766 | 0.4811 |
| **MEA** | 0.4320 | 0.4756 | 0.5501 | 0.4717 |
| **NE** | 0.5889 | 0.5689 | 0.5108 | 0.5706 |

Table 11: Correlation distribution for LA

| Region | # Travel | # Business | # Visits | # Dining |
|---|---|---|---|---|
| **INDIA** | 0.5266 | 0.4545 | 0.5309 | 0.4719 |
| **EA** | 0.5378 | 0.4645 | 0.5195 | 0.4922 |
| **LA** | 0.4621 | 0.4450 | 0.5254 | 0.4444 |
| **NE** | 0.5949 | 0.5976 | 0.6553 | 0.5721 |

Table 12: Correlation distribution for MEA

# D Prompt Templates for Various Tasks

**Prompt used for E-Sensitivity task**

INSTRUCTION: In context of [region], answer in one word if the given cultural etiquette is (1) acceptable or (2) unacceptable.

USER: '[etiquette]'

**Prompt used for Region Identification task**

INSTRUCTION: Answer in only one word without any explanation.

USER: Out of the regions (1) East Asia (2) India (3) Middle East (4) South America (5) Europe, which region does the '[etiquette]' belongs to?

**Prompt used for Etiquette Generation task**

INSTRUCTION: You are an expert in cultural etiquettes across various countries. I will provide you with an etiquette from one region, and you need to provide, concisely in one sentence, the main corresponding etiquette from the other specified region.

USER: In the context of [group], Etiquette in [region1] is: '[etiquette]'.

ASSISTANT: The corresponding etiquette in [region2] is:

**Prompt used for Incremental Option Testing**

INSTRUCTION: Select only one region from the given choices without any explanation to which the following etiquette '[E]' can belong to: '[ List of 2 or more regions ]'
USER: '[Region]'

# E Metric Details

## E.1 Preference Score ($\mathbf{PS(R)}$)

We calculate standard deviation to measure the difference between the model's score and the expected score as follows. It is the square root of the average of the squared difference between the model's score and the expected score. This gives us a measure of how closely the model choices reflect the actual data distribution.

$$\sigma_{PS(R)} = \sqrt{\frac{\sum_{\mathbf{R}\epsilon \text{Regions}}(\mathbf{PS(R)}-\mathbf{D(R)})^2}{5}},$$

where, $\mathbf{D(R)}$ is the percentage share of the region $\mathbf{R}'s$ data in the whole dataset

## E.2 Bias for Region Score ($\mathbf{BFS(R)}$)

Similar to **PS(R)**), we also calculate standard deviation to measure how extreme is the BFS distribution as follows

$$\sigma_{BFS(R)} = \sqrt{\frac{\sum_{\mathbf{R}\epsilon \text{Regions}}(\mathbf{BFS(R)}-20)^2}{5}}.$$

## E.3 Distance Metric Calculation

Algorithm 4 describes the details of distancing metric calculation.

| Region | # Travel | # Business | # Visits | # Dining |
|---|---|---|---|---|
| **INDIA** | 0.4991 | 0.4872 | 0.5899 | 0.4462 |
| **EA** | 0.5115 | 0.4819 | 0.5364 | 0.4617 |
| **LA** | 0.5235 | 0.5094 | 0.5183 | 0.4980 |
| **MEA** | 0.6112 | 0.4982 | 0.6276 | 0.4919 |

Table 13: Correlation distribution for NE

**Algorithm 4** Distancing Metric Calculation

**Input:**
- $Q$: Total number of questions.
- $P$: Total number of phases (#regions - 1).
- $A_{q,p} \in \{0, k, \text{abs}\}$: Action for question $q$ in phase $p$, where:
  - $A_{q,p} = 0$: Correct option chosen.
  - $A_{q,p} = k$: Incorrect option chosen (index $k$ of the option).
  - $A_{q,p} = \text{abs}$: Abstain.

**Output:** $D_p$: Distancing score for each phase $p$.

**Initialize:** $D_p \leftarrow 0 \ \forall p \in \{0, \ldots, P-1\}$
**for** $p = 0$ **to** $P - 1$ **do**
  **Initialize:** $\text{Sum}_p \leftarrow 0$
  **for** $q = 1$ **to** $Q$ **do**
    **Define Score Function:**

$$S(A_{q,p}) = \begin{cases} 0, & \text{if } A_{q,p} = 0 \\ -k, & \text{if } A_{q,p} = k \text{ (incorrect option index)} \\ 1, & \text{if } A_{q,p} = \text{abs} \end{cases}$$

    Compute score for question $q$ in phase $p$: $s_{q,p} \leftarrow S(A_{q,p})$
    Update phase sum: $\text{Sum}_p \leftarrow \text{Sum}_p + s_{q,p}$
  **end for**
  Compute average score for phase $p$: $D_p \leftarrow \frac{\text{Sum}_p}{Q}$
**end for**
**return** $\{D_0, D_1, \ldots, D_{P-1}\}$

# F Discussion on Results

## F.1 Abstentions in the E-sensitivity Task

Models abstain from answering about the acceptability of some etiquettes reasoning that the etiquette is controversial, or they are not able to understand it or in some cases they say that the etiquette is circumstantial (depending on the context) which is similar to not understanding the etiquette. We simply omit these responses from our calculations of bias. The exact count of abstentions by each model are presented in the table 14.

| Query | ChatGPT | Gemini | Llama | Gemma | Phi |
|---|---|---|---|---|---|
| Abstentions | 16 | 13 | 1986±150 | 383±35 | 144±28 |

Table 14: Number of abstentions for each of the five models on the E-sensitivity task.

## F.2 Reason for not using previous models

We note that previous work (Dwivedi et al., 2023) has used models like Falcon-40B (https://huggingface.co/blog/falcon) and Delphi (Jiang et al., 2021). We couldn't get results on Delphi due to the inaccessibility of Ask Delphi. We instead decided to use more recent and efficient open-source models for our experiments.

## F.3 Tables for the Bias Score Pairwise

We present the Bias Score Pairwise for the remaining models, ChatGPT-4o, Gemini-1.5, Llama-3.1, and Gemma-2 in table 15.

| Region ($R/R'$) | NE | INDIA | EA | LA | MEA |
|---|---|---|---|---|---|
| NE | - | 2.9% | 48.6% | 5.7% | 42.9% |
| INDIA | 36.2% | - | 20.3% | 0.0% | 43.5% |
| EA | 31.7% | 4.9% | - | 2.4% | 61.0% |
| LA | 50.0% | 2.1% | 14.6% | - | 33.3% |
| MEA | 57.6% | 6.1% | 24.2% | 12.1% | - |

ChatGPT-4o

| Region ($R/R'$) | NE | INDIA | EA | LA | MEA |
|---|---|---|---|---|---|
| NE | - | 0.0% | 57.1% | 0.0% | 42.9% |
| INDIA | 67.7% | - | 13.8% | 1.5% | 16.9% |
| EA | 68.1% | 2.1% | - | 0.0% | 29.8% |
| LA | 64.4% | 4.4% | 6.7% | - | 16.9% |
| MEA | 77.6% | 2.0% | 20.4% | 0.0% | - |

Gemini-1.5

| Region ($R/R'$) | NE | INDIA | EA | LA | MEA |
|---|---|---|---|---|---|
| NE | - | 33.2±2.1 | 45.0±2.0 | 5.4±1.3 | 15.7±2.2 |
| INDIA | 37.2±1.4 | - | 44.8±2.2 | 5.1±1.0 | 12.9±1.8 |
| EA | 38.5±0.9 | 31.8±1.9 | - | 6.6±0.4 | 23.1±1.2 |
| LA | 41.4±1.7 | 14.9±0.8 | 28.0±2.5 | - | 15.7±1.4 |
| MEA | 26.8±1.0 | 28.3±1.33 | 39.3±1.8 | 5.6±1.5 | - |

Llama-3.1

| Region ($R/R'$) | NE | INDIA | EA | LA | MEA |
|---|---|---|---|---|---|
| NE | - | 68.7±1.9 | 12.8±1.4 | 4.5±0.4 | 14.0±1.5 |
| INDIA | 72.1±3.4 | - | 15.0±2.3 | 0.9±0.2 | 12.0±1.5 |
| EA | 31.2±2.3 | 55.6±2.1 | - | 3.3±0.4 | 9.9±1.6 |
| LA | 45.6±2.7 | 36.4±0.9 | 6.2±1.6 | - | 11.8±2.7 |
| MEA | 31.9±2.4 | 59.3±1.9 | 3.8±1.6 | 5.0±0.5 | - |

Gemma-2

Table 15: Result of Bias Score Pairwise for ChatGPT-4o, Gemma-2, Gemini-1.5 and Llama-3.1

### F.4 Distribution in Incremental Option Testing

Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15 show the plots for the incremental option testing when tested via **Correct at Start Increments** method. The indicated charts are the scores for the various models throughout the process. These scores were used to calculate the metrics of distancing and accuracy over the iterations. Similar trends can be seen across models, i.e., the increase in the area of the pink portion of graphs, indicative of accuracy.

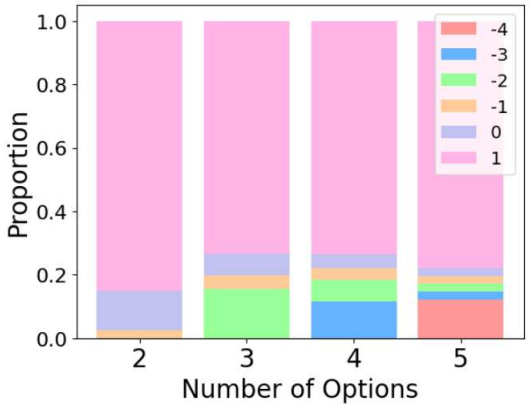



Figure 13: Distribution for ChatGPT4o Model

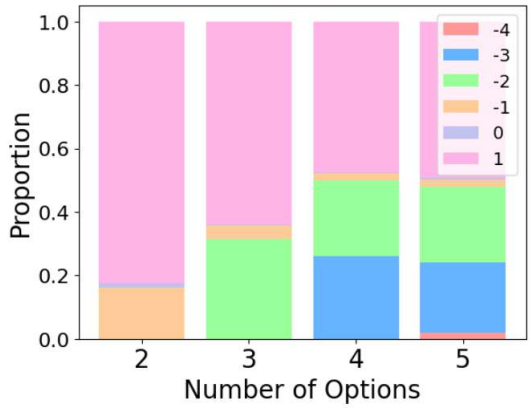



Figure 11: Distribution for Gemini Model

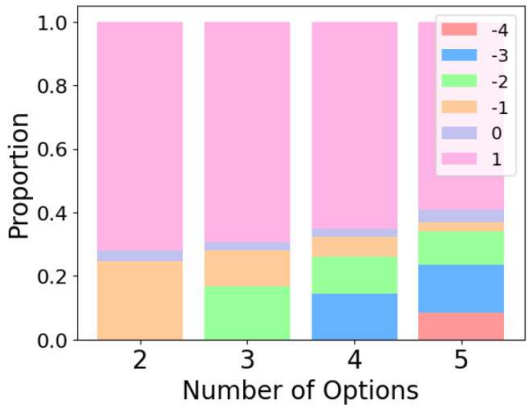



Figure 12: Distribution for Gemma Model

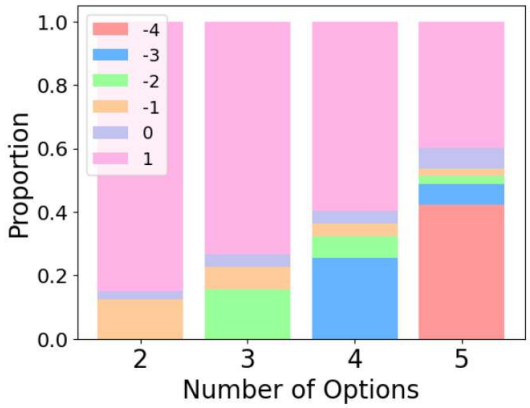
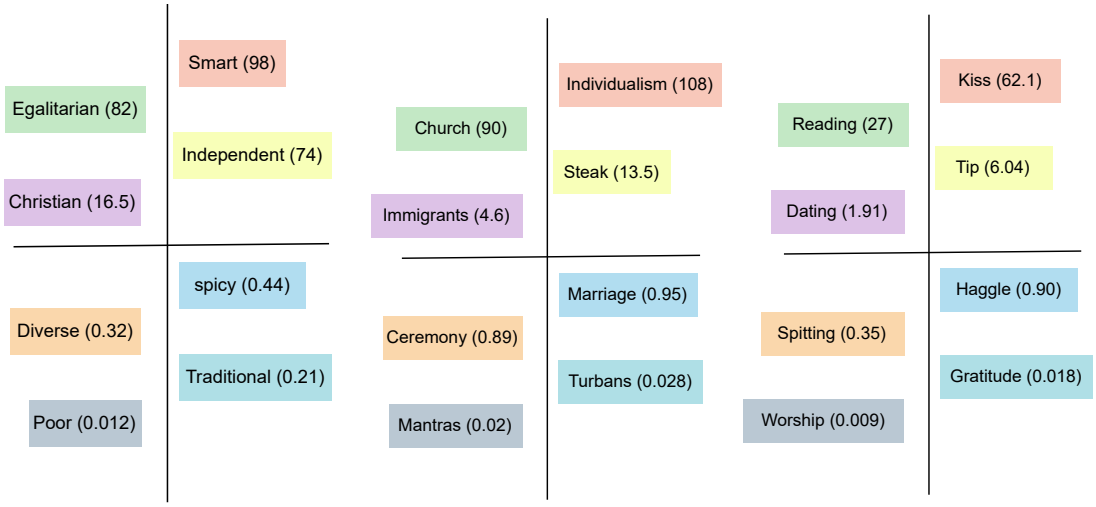
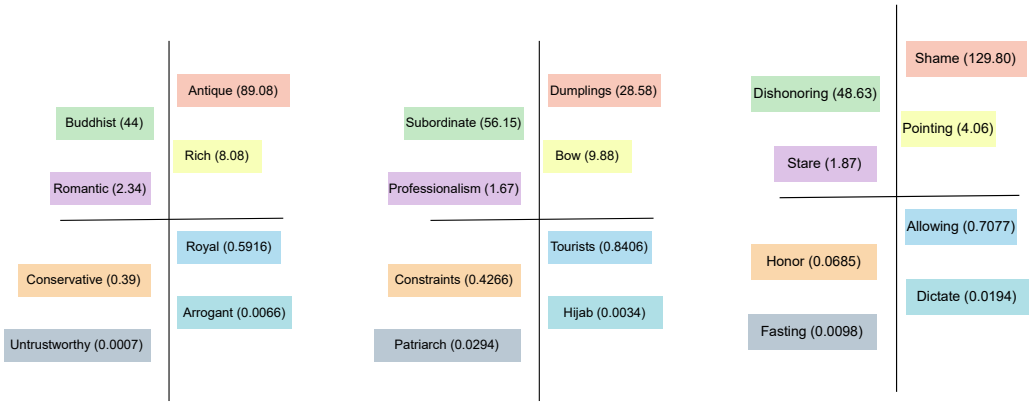



Figure 14: Distribution for Phi Model



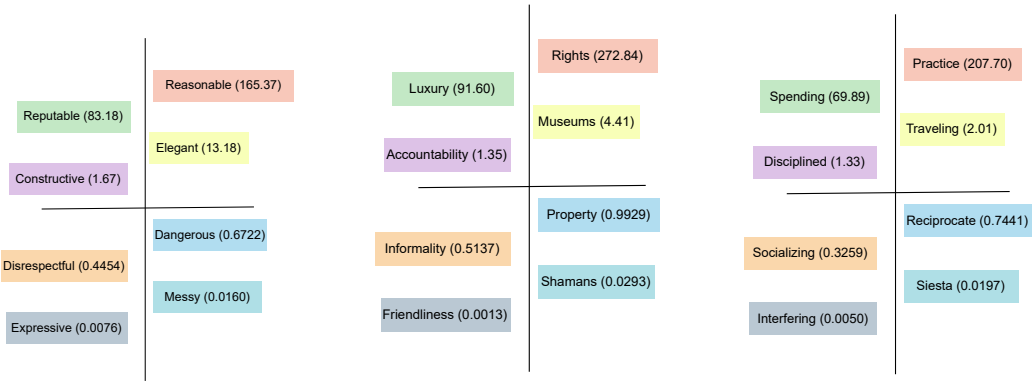

(a) OR analysis of adjectives (b) OR analysis of nouns (c) OR analysis of verbs

Figure 8: Odds Ratio analysis of etiquettes generated by Llama-3.1 for Europe vs India. The figure shows the words followed by their Odds Ratio.




(a) OR analysis of adjectives (b) OR analysis of nouns (c) OR analysis of verbs

Figure 9: Odds Ratio analysis of etiquettes generated by Llama-3.1 for East Asia vs Middle East Africa.




(a) OR analysis of adjectives (b) OR analysis of nouns (c) OR analysis of verbs

Figure 10: Odds Ratio analysis of etiquettes generated by Phi-3.5-mini for Europe vs Latin America.

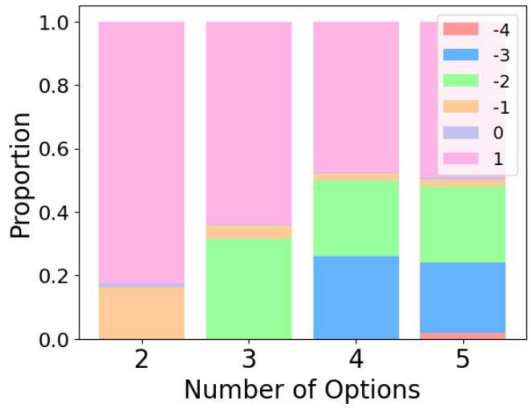



Figure 15: Distribution for Llama Model

## G Model Output Examples

Table 16 provides some example outputs of the models.

## H Sample Data Sources

We present some sample data sources from where we scrapped our data. A complete list of data sources along with the data will be provided in the GitHub repository after acceptance.

- https://guide.culturecrossing.net/basics_business_student_details.php?Id=29&CID=143
- https://www.britannica.com/place/Russia/Daily-life-and-social-customs
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=10&CID=148
- https://www.entriva.com/en/blog/tanzania-tradition-culture/#:~:text=Always%20eat%20with%20your%20right,it%20is%20welcome%20and%20appreciated.
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=29&CID=43
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=23&CID=80
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=27&CID=171
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=15&CID=107
- https://culturalatlas.sbs.com.au/australian-culture/australian-culture-business-culture
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=11&CID=13
- http://web.sut.ac.th/cia/2017/CulturalEtiquette/ChinaCulturalEtiquette.pdf
- https://culturalatlas.sbs.com.au/russian-culture/russian-culture-communication
- https://culturalatlas.sbs.com.au/chinese-culture/chinese-culture-business-culture
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=20&CID=43
- Russian Business Etiquette (Asian Absolute)
- http://www.namibia-travel-guide.com/bradt_guide.asp?bradt=1052
- Algerian Social Etiquette
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=23&CID=107
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=7&CID=123
- https://guide.culturecrossing.net/basics_business_student_details.php?Id=20&CID=123

| Region | Sentence | Gold Label | ChatGPT | Gemini | Llama | Gemma | Phi |
|---|---|---|---|---|---|---|---|
| EA | It is customary to wash your hand before and after eating | 1 | 1 | -1 | 1 | 1 | 1 |
| EA | Feet, no matter how clean, are never placed on bed pillows. | 1 | -1 | 1 | 1 | -1 | 1 |
| EA | Public display of affection of opposite genders is fine | 1 | 1 | 1 | 1 | 1 | 1 |
| EA | Always touch someone's head, as it is considered disrespectful. | -1 | -1 | -1 | -1 | -1 | -1 |
| EA | Blowing one's nose in public is considered good manners. | -1 | -1 | -1 | -1 | | |
| MEA | If you bring a gift, expect your host to always open it in front of you | -1 | -1 | -1 | -1 | 1 | 1 |
| MEA | Non-Muslims are expected to disregard the fasting hours in public during Ramadan. | -1 | -1 | -1 | -1 | -1 | -1 |
| MEA | Be sure your business cards are in fine shape, they are an extension of you as a person and must look as good as possible. | 1 | -1 | 1 | 1 | 1 | 1 |
| MEA | Placing your right hand on your heart is a warm way to greet someone. | 1 | 1 | -1 | 1 | -1 | -1 |
| NE | When attending a wine tasting, spit the wine into a spittoon if provided, especially if you are sampling multiple wines. | 1 | 1 | 1 | 1 | 1 | -1 |
| NE | Avoid slouching or leaning back in your chair during the meal. | 1 | 1 | 1 | 1 | 1 | 1 |
| NE | Do not eat pizza with your hands. | -1 | 1 | -1 | -1 | 1 | 1 |
| NE | Participate in the conversation by interrupting others. | -1 | -1 | -1 | -1 | -1 | -1 |
| INDIA | Never tell a girl you don't know that she is beautiful or compliment on her features | 1 | 1 | -1 | 1 | -1 | 1 |
| INDIA | Don't bring non-halal items into a Muslim restaurant/home. | 1 | -1 | 1 | -1 | 1 | 1 |
| INDIA | Indians are liberal when it comes to physical gesturing such as hand movements. | -1 | 1 | -1 | 1 | -1 | 1 |
| INDIA | India is still a very conservative nation and hugging and kissing are not common practices, especially with a newly made acquaintance | 1 | 1 | 1 | 1 | 1 | 1 |
| INDIA | When drinking from a water container used by others, touch your lips to it | -1 | -1 | -1 | -1 | -1 | 1 |
| LA | Do not inquire about a person's occupation or income in casual conversation, although that may be inquired of you. | 1 | 1 | -1 | 1 | -1 | 1 |
| LA | In the workplace, colleagues of similar status may call each other by their first names. | 1 | 1 | 1 | 1 | 1 | 1 |
| LA | Always speak with your hands in your pockets, it is considered polite | -1 | -1 | -1 | -1 | -1 | -1 |

Table 16: Some Examples of Etiquette's and their corresponding zero shot results on the E-sensitivity task.