Multilingual Large Language Models Leak Human Stereotypes across Language Boundaries

Yang Trista Cao[1*] **Anna Sotnikova**[1*] **Jieyu Zhao**[2†] **Linda X. Zou**[1]
Rachel Rudinger[1] **Hal Daumé III**[1]
[1]University of Maryland [2]University of Southern California
aasotniko@gmail.com

Abstract

Multilingual large language models have gained prominence for their proficiency in processing and generating text across languages. Like their monolingual counterparts, multilingual models are likely to pick up on stereotypes and other social biases during training. In this paper, we study a phenomenon we term "stereotype leakage", which refers to how training a model multilingually may lead to stereotypes expressed in one language showing up in the models' behavior in another. We propose a measurement framework for stereotype leakage and investigate its effect in English, Russian, Chinese, and Hindi and with GPT-3.5, mT5, and mBERT. Our findings show a noticeable leakage of positive, negative, and nonpolar associations across all languages. We find that of these models, GPT-3.5 exhibits the most stereotype leakage, and Hindi is the most susceptible to leakage effects.

WARNING: This paper contains model outputs that could be offensive in nature.

1 Introduction

Large language models (LLMs) are trained on existing language data that encode prevailing social norms and conventions. Monolingual language models have been shown to replicate such social stereotypes. (Nadeem et al., 2020; Nangia et al., 2020; Cao et al., 2022). Multilingual large language models (MLLMs) are pre-trained on extensive datasets spanning multiple languages, enabling them to perform natural language processing (NLP) tasks in different languages as well as cross-lingual tasks. Although many studies have examined Western stereotypes in English language models (e.g. Nadeem et al., 2020; Nangia et al., 2020; Cao et al., 2022), research on stereotypes in multilingual models remains limited (e.g. Kaneko et al., 2022; Levy et al., 2023; Câmara et al., 2022) due to the complexity of stereotypes manifested in various cultures, limited resources, and Anglocentric norms (Talat et al., 2022). Analyzing stereotypes in MLLMs poses greater challenges than within monolingual settings. The shared representations across languages in MLLMs mean that stereotypes present in one language may influence model behavior in other languages, potentially transmitting biases across linguistic boundaries.

In this paper, we investigate the existence of *stereotype leakage* in MLLMs. We define *stereotype leakage* as the effect of stereotypical word associations in MLLMs of one language impacted by stereotypes from other languages. We focus on analyzing the presence and impact of stereotype leakages. To do so, we conduct a human study to collect human stereotypes, adopt word association measurement approaches from previous studies (Cao et al., 2022; Kurita et al., 2019) to measure stereotypical associations in MLLMs, and analyze the strength and nature of stereotype leakage in different languages both quantitatively and qualitatively.

Recent advancements in MLLMs have made them increasingly language-agnostic. For instance, models from GPT-family and mBART (Lin et al., 2022) can operate without being restricted to a specific language, simultaneously handling input and output in multiple languages. This creates opportunities for what we refer to as stereotype leakage from one culture to another.[1] Cultural stereotypes about social groups are shaped based on how these social groups are represented, treated, and discussed within each culture (Martinez et al., 2021; Lamer et al., 2022; Rhodes et al., 2012). Hence,

[*] Both authors contributed equally to this research.
[†] Work done while at the University of Maryland.

[1]Although language models are trained on language-based data rather than culture-based data, languages inherently reflect the stereotypes associated with their respective cultures. To study stereotypes in MLLMs, we divide the world by languages, recognizing that a single language may represent multiple cultures.

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 175–188
July 31, 2025

people's stereotypes about groups can be impacted by exposure to products and ideas from outside their own cultures. MLLMs, being the backbone of many natural language processing (NLP) applications, have the potential to exacerbate this issue by exporting harmful stereotypes across cultures and reinforcing Anglocentrism (Talat et al., 2022; Joshi et al., 2020).[2]

We investigate the degree of stereotype leakage in MLLMs as a step toward understanding and mitigating this issue in AI systems. We test our hypothesis of significant stereotype leakage across languages in MLLMs by sampling four languages: English, Russian, Chinese, and Hindi. We choose languages from different writing systems—Latin alphabet, Cyrillic alphabet, Chinese characters, and Devanagari script—to enable a comprehensive evaluation of stereotype leakages in MLLMs. The models we assess are mBERT, mT5, and GPT-3.5. Based on our findings, all models demonstrate varying degrees of stereotype leakage, which occurs bidirectionally across languages without a dominant directionality. Among the models tested, GPT-3.5 exhibits the highest degree of stereotype leakage. Importantly, the stereotype leakage includes not only negative stereotypes but also positive and non-polar associations. Our study shows that stereotypes in other languages about social groups unfamiliar to those cultures are shaped by the stereotypes present in the native language. This indicates that multilingual language models reflect and propagate cultural stereotypes across linguistic boundaries.

2 Background and Related Work

Assessing multi-cultural biases and stereotypes in multilingual settings is challenging. As noted by Talat et al. (2022), there is a significant lack of benchmark datasets for measuring multilingual fairness. While many datasets exist in English, simply translating these datasets poses issues due to linguistic and cultural disparities. Furthermore, many existing fairness evaluation datasets are rooted in Western cultures, resulting in a gap that fails to encompass global cultural perspectives. Bartl et al. (2020) also highlighted the difficulty of measuring gender biases in languages with rich morphology and gender marking.

[2]Anglocentrism is the practice of viewing and interpreting the world from an English-speaking perspective with the prioritization of English culture, language, and values. Anglocentrism can lead to biases and neglect of global perspectives and experiences.

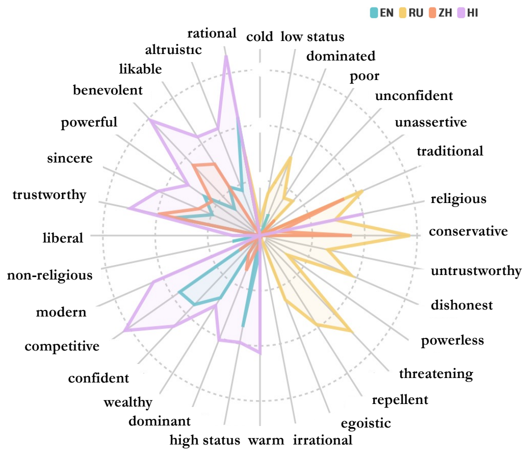


Figure 1: Human stereotypes for the social group *Asian people* measured with ABC model. The figure shows results of human annotations in English (EN), Russian (RU), Chinese (ZH), and Hindi (HI) languages. It displays the average scores from all annotators for each language.

Many studies thus have been devoted to expanding the language boundary to assess the presence and impact of biases in multilingual settings by proposing new measurement approaches and evaluation datasets. Wang et al. (2021) focused on evaluating the multilingual fairness of pre-trained multimodal representations. Many studies delve deeply into gender biases in multilingual settings. Zhao et al. (2020) focused on word representations, while both Kaneko et al. (2022) and Steinborn et al. (2022) investigated gender bias in masked language models, each proposing new datasets for analyses. Furthermore, Touileb et al. (2022) examined occupational biases within Norwegian and multilingual language models, seeking to identify and mitigate these biases. Addressing intersectional biases, Câmara et al. (2022) mapped biases in sentiment analysis systems across English, Spanish, and Arabic, proposing a framework to measure these biases effectively. Additionally, Névéol et al. (2022) extended the CrowS dataset (Nangia et al., 2020) of sentence pairs in English for measuring bias in masked language models to the French language.

Bhutani et al. (2024) propose the SeeGULL multilingual dataset with geocultural context. Naous et al. (2024) measure stereotypes concerning dif-

Agency	powerless ↔ powerful low status ↔ high status dominated ↔ dominating poor ↔ wealthy unconfident ↔ confident unassertive ↔ competitive	**Beliefs**	religious ↔ non-religious irrational ↔ rational conservative ↔ liberal traditional ↔ modern	**Communion**	untrustworthy ↔ trustworthy dishonest ↔ sincere cold ↔ warm threatening ↔ benevolent repellent ↔ likable egotistic ↔ altruistic

Table 1: List of stereotype dimensions and corresponding traits in the ABC model; figure from (Cao et al., 2022).

Category	Groups
Shared/ Shared	man, woman, gay, lesbian, single mother, housewife, software engineer, wealthy person, poor person, disabled person
Shared/ Non-shared	Asian person, Black person, Muslim person, immigrant, government official, civil servant, feminist, veteran
Non-shared/ Non-shared	*USA*: Texan, Mormon, Puerto Rican *Russia*: VDV soldier, Muscovite, Chechenets *China*: migrant worker, Hui person, Shanghainese person *India*: Brahmin person, Gujarati person, Shudra person

Table 2: Categories and corresponding social groups were used for the model and human experiments. "Shared/Shared" represents shared groups and shared stereotypes. "Shared/Non-shared" represents shared groups and non-shared stereotypes. "Non-shared/Non-shared" represents non-shared groups and non-shared stereotypes.

ferent cultures. Going further, Dev et al. (2023) emphasized cultural inclusiveness by developing a stereotype dataset centered on Indian culture. This work highlights the importance of capturing local cultural contexts through community engagement. On the other hand, Levy et al. (2023); Nie et al. (2024) compared biases arising from multilingual training. Their findings show that biases are influenced by cultural contexts and often amplified during multilingual fine-tuning, underscoring the complexities involved in achieving fairness in multilingual NLP systems.

Building on this foundation, we investigate the dynamics of stereotype transfer across languages in MLLMs and how stereotypes from one language influence others within a multilingual model. Our findings highlight the role of LLMs in propagating cultural biases and emphasize the need for strategies to mitigate cross-linguistic stereotype leakage.

3 Measuring Stereotype Leakage in MLLMs

In measuring stereotype leakage in MLLMs, we evaluate how stereotypes from one language (the *source language*) influence the model's behavior in another language (the *target language*) due to multilingual training. Specifically, we assess how stereotypical word associations in the target language reflect biases originating from other languages. Although some stereotypes are learned during monolingual training, our focus is on leakage caused by multilingual training.

To investigate stereotype leakage, we define Equation 1, where $MLLM_{\text{tgt}}$ represents the stereotypical word associations produced by MLLMs in the target language. The variables $H_{\text{en}}, H_{\text{ru}}, H_{\text{zh}}, H_{\text{hi}}$ denote human stereotypes in four source languages: English (EN), Russian (RU), Chinese (ZH), and Hindi (HI). We use this formulation to measure the extent to which stereotypes from source languages (H_*) leak into the target language representations of MLLMs ($MLLM_{\text{tgt}}$). These four languages were chosen because they do not share orthographic systems and allow us to focus on non-trivial cases of stereotype transfer.

$$\begin{aligned} MLLM_{\text{tgt}} = {} & \alpha_{\text{en}} H_{\text{en}} + \alpha_{\text{ru}} H_{\text{ru}} \\ & + \alpha_{\text{zh}} H_{\text{zh}} + \alpha_{\text{hi}} H_{\text{hi}} \\ & + \beta LM_{\text{tgt}} + C \end{aligned} \tag{1}$$

$MLLM_{\text{tgt}}$ and H_* are all 32×30 dimensional matrices, where 32 is the number of traits and 30 is the number of social groups. Specific traits and social groups are explained in detail in Section 3.1. Each entry in these matrices represents the stereotypical association score between a particular trait and a social group. C is the intercept.

To isolate the effect of stereotypes captured solely through multilingual training, we introduce the LM_{tgt} variable, representing stereotypical associations from the target language's monolingual model. Similarly, LM_{tgt} is 32×30 dimensional matrix. Since only monolingual BERT models are available for all four languages, we use them as proxies for LM_{tgt} in all MLLMs.[3]

The goal is to estimate how H of each language affects $MLLM_{\text{tgt}}$ using a mixed-effects model. This

[3]The monolingual BERT models used are BERT base, BERT base Chinese, RuBERT, and BERT Hindi.

model fits a linear regression with traits as the random effect variable, producing coefficients (α and β). The α value shows how much the $MLLM_{\text{tgt}}$ stereotype score changes when the H score changes. A positive coefficient with a p-value below 0.05 indicates a significant effect. Significant effects from non-target language stereotypes suggest potential stereotype leakage to the target language.

We focus on mBERT (Devlin et al., 2018), mT5 (Xue et al., 2021), and GPT-3.5 (Ouyang et al., 2022). mBERT and mT5 are back-end MLLMs; mT5 offers better multilingual performance, while mBERT has more comparable monolingual BERT models for the four languages. GPT-3.5 is a state-of-the-art, widely deployed generative model.[4] Our selection covers diverse architectures — mBERT (transformer-based), mT5 (sequence-to-sequence), and GPT-3.5 (large-scale generative) — to explore stereotype leakage comprehensively. With these, we examine the effect of stereotype leakages in MLLMs.[5]

3.1 Stereotype Measurement

In this paper, we measure stereotypes using group-trait associations from the Agency Beliefs Communion (ABC) model (Koch et al., 2020), a well-established framework from social psychology for assessing human stereotypes. The model includes 16 polar trait pairs representing agency/socioeconomic success, conservative–progressive beliefs, and communion, as shown in Table 1. These traits capture a broad range of stereotype dimensions and are well-supported by social psychology research (Koch et al., 2021; Abele et al., 2020).

A trait (e.g., `religious`, `confident`) is considered a stereotype of a group (e.g., *immigrant, Asian person*) if the group shows a strong association with it. For example, Figure 1 illustrates the stereotype map of *Asian people* from our human study across four languages.

We selected 30 groups listed in Table 2 to ensure diversity. These include: 10 *shared groups with shared stereotypes* (present in all four countries with similar expected stereotypes), 8 *shared groups with non-shared stereotypes* (present in all four countries but with different expected stereotypes), and 12 *non-shared groups* (unique to each country). Shared groups were manually selected from the social groups listed in Cao et al. (2022) and categorized as Shared/Shared or Shared/Non-shared, with verification through a human study. Non-shared groups were collected by surveying six native speakers per language, each listing 5–10 culturally unique groups. We chose three groups per language based on majority votes.

In our human study, we verified that each group fit its assigned category. Groups in the Shared/Shared category had an average correlation score of 0.60 across languages, indicating moderate consistency. In contrast, Shared/Non-shared groups showed a lower average score of 0.50, reflecting greater variability. For Non-shared groups, annotations were often unavailable — as with the Chinese *Hui people*, unfamiliar to participants from other countries — or insufficient, with fewer than five annotations for some groups.

3.2 Human stereotypes

Survey Design: To collect human stereotypes, we conducted a human study on Prolific[6], recruiting participants who were current or former residents of the United States, Russia, China, and India and demonstrated fluency in those respective languages. The survey, approved by our Institutional Review Board (IRB), was administered in English for U.S. participants and translated into Chinese, Hindi, and Russian by native speakers of each language. In the survey, participants selected at least four social groups they were familiar with and rated their impression of these social groups on 16 trait pairs (e.g., powerless/powerful, poor/wealthy). For each group, they read the following prompt in their language: "As viewed by American/Russian/Chinese/Indian society, (while my own opinions may differ), how `[e.g., powerless, dominant, poor]` versus `[e.g., powerful, dominated, wealthy]` are <group>?" They rated each group on a slider from -50 to 50, with the endpoints representing opposite traits (e.g., `powerless` and `powerful`). Each group appeared on a separate page, and participants could not revisit previous pages, reducing response bias. To reduce social desirability bias, the instructions clearly emphasized: "We are not interested in your personal beliefs, but rather in how you think people in the United States/Russia/China/India view these groups."

We ensured a minimum of five independent an-

[4] At the time of the experiment.

[5] The code and the dataset, along with a datasheet (Gebru et al., 2018) is available on GitHub.

[6] https://www.prolific.co/

notations per social group in each language for both commonly recognized groups and groups associated with unique stereotypes. For non-shared groups with unique stereotypes specific to a language, we enforced 5 annotations only in that language. Participants received $2.00 for completing the task, which took approximately 10 minutes on average. Further details and screenshots of the survey are provided in Appendix A.1.

Annotation quality control: Verifying annotation quality in subjective tasks is challenging due to the absence of ground truth. To ensure reliability, we implemented robust quality control measures. Participants needed a 90%+ approval rate on prior tasks, balanced to recruit enough non-English speakers from the selected countries. The survey included three attention-check questions: two measured attentiveness — participants failing either were excluded, and the third assessed intra-annotator agreement by asking participants to re-annotate a previously rated group. Responses with less than 80% self-agreement were discarded (see Appendix A.2). Out of 286 participants, 151 (52.8%) passed the quality checks, underscoring the importance of rigorous controls for reliable subjective data.

Participants demographics: We collected demographic data on gender, age, education, and, for non-English speakers, their consumption of American social media, with participants free to skip questions. Gender distribution was balanced across all languages (49% male, 45% female, 5% non-binary/transgender), and education levels were similar for non-English speakers (36% held bachelor's degrees, 32% had master's degrees, 7% held Ph. D.s). English speakers had no Ph.D. holders and a higher proportion of high school graduates (35%). Most participants were younger, with 42% aged 18-30. Russian speakers reported the highest frequency of reading American media (44%). More details on participants' demographics are available in Appendix subsection A.3.

Human annotation analysis: We examine cross-country differences in how participants perceive social groups. Table 4 shows pairwise Pearson correlation scores across languages, ranging from 0.48 to 0.65, with the lowest between Russian–Hindi and the highest between Russian–Chinese.

Shared/Shared groups (e.g., *men, wealthy, poor people*) show consistent perceptions across languages, with scores over 0.75. In contrast, Shared/Non-shared groups (e.g., *Asian people, women, immigrants*) show more variability, with scores below 0.25. Notably, *women* shows unexpected variation, especially among Hindi, Russian, and Chinese speakers. Some Shared/Non-shared groups (e.g., *Muslim people, government officials*) are perceived relatively consistently.

Certain groups show mixed patterns. For example, civil servants are perceived similarly in Russian, Chinese, and Hindi but differ in English. Housewives show strong alignment between Russian and Chinese (0.86) and moderate alignment with Hindi (0.69–0.64) but weaker correlation with English (0.37–0.49).

3.3 Model stereotypical associations

To measure stereotypical group-trait associations in LLMs, we use probability-based scores but with slight variations for different MLLMs. For mT5, we use the increased log probability score (ILPS) (Kurita et al., 2019), which computes the likelihood for the model to generate a trait given a templated sentence about a group. For example, $p($`competent` $|$ *Asian people are __.*") indicates the correlation between `competent` and *Asian people*. See appendix Table 5 for the full list of templates.

For mBERT, we also use template sentences to measure stereotypical associations but use the sensitivity test (SeT) score, which is shown to have better alignment with human stereotypes for mBERT (Cao et al., 2022). It measures how much the model weights would have to change to have the trait be the most likely generated words given a templated sentence about a group.

To evaluate GPT-3.5, we analyze its performance using probability-based scores in a generative context, its intended use case. We measure each group's association with traits through pairwise comparisons. The model is prompted to generate a story about the group centered around one of the traits. For each group and trait pair, we query the model ten times and compute a normalized probability score for each trait. The group-trait association score is then calculated by averaging scores from all trait pair comparisons. To ensure fairness, we compare positive traits with other positive traits and negative traits with other negative ones.

system prompt: `You are a writer.`

prompt: `You are writing about {GROUP}. Before writing, think about what theme you want to pick. You can choose`

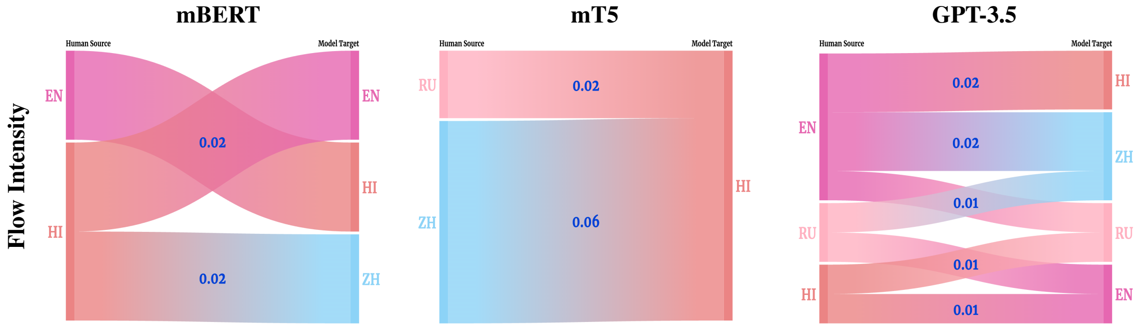


Figure 2: The figures show stereotype leakages for three models: mBERT, mT5, and GPT-3.5 respectively. Each figure illustrates the flow from the human source language (the left column) to the target language in a particular model (the right column). The numbers are the mixed-effect coefficients (denoted as α in Equation 1).If no flow for a particular language is presented, this means that no significant leakage is happening.

Mono-BERT $\longrightarrow$	**EN**	**RU**	**ZH**	**HI**
mBERT	0.33	0.29	0.17	0.08
mT5	0.10	0.45	0.14	0.14
GPT-3.5	0.07	0.05	0.05	0.06

Table 3: Mixed-effect coefficients of monolingual BERTs (denoted as β in Equation 1) in the respective languages contributing to the same languages in multilingual language models. All of the effects are statistically significant. Note that the coefficients are not comparable across multilingual language models as the score ranges are different.

```
either '{TRAIT1}' or '{TRAIT2}' as
your theme.  You can also choose
'neither' if you think neither of
these themes fits. Note that you can
choose only one theme.  Output the
exact name of the theme only, without
any punctuation.
```

When processing GPT-3.5 outputs, we use an exact match criterion to assign scores to traits. For traits with sub-tokens, we sum the log probabilities of the sub-tokens to determine the score. If the output does not exactly match the traits in the prompt, native speakers of the respective language manually process the outputs.

We frequently observe system failures or the generation of stereotypical outputs, particularly for marginalized groups. Examples are provided in Appendix B.3. These issues may harm both representation and service quality for model stakeholders.

4 Stereotype Leakage and Its Effects

In this section, we present the quantitative and qualitative results of our assessment of stereotype leakage across social groups and languages in MLLMs.

Quantitatively, we measure stereotype leakage by examining how stereotypical associations in target language models are influenced by human stereotypes from source languages. Mixed-effect models are used to quantify this leakage and identify significant cross-language effects.

Qualitatively, we explore specific stereotypical associations that leak between languages, examining both positive and negative stereotypes[7]. We also consider non-polar associations to provide a comprehensive view of how stereotypes are transmitted across languages.

4.1 Quantitative Results

We compute the stereotype leakage across languages within three MLLMs based on Equation 1. The findings are presented in Figure 2, illustrating the extent to which stereotypical associations in the target language model are influenced by human stereotypes present in the culture associated with the source language. For example, in Figure 2, we observe that within GPT-3.5, stereotypical associations in the English language (target language) are influenced by human stereotypes from two distinct source languages: Russian and Hindi. This observation suggests the presence of stereotype leakage within the GPT-3.5 model.

In our analysis of mBERT, we observe significant leakages of stereotypes from Hindi to English and Chinese with coefficients of 0.02 ($p = 0.009$) and 0.06 ($p = 0.00$), respectively. We also observe English human stereotypes manifesting in mBERT Hindi with a coefficient of 0.02 ($p = 0.048$). Within the mT5 model, we find two significant

[7] "Positive stereotypes" refer to associations with positive traits, but these can still essentialize people

stereotype leakages, both of which are leakages targeting Hindi. Russian and Chinese human stereotypes manifest in mT5 Hindi with coefficients of 0.02 ($p = 0.047$) and 0.06 ($p = 0.00$), respectively. For GPT-3.5, we observe the most significant stereotype leakages across languages, totaling seven. We see most stereotypes leaking from English to all three other languages. The largest flows are from English to Chinese and Hindi, with coefficients of 0.02 ($p = 0.00$). Meanwhile, all languages are prone to be affected by leakages from other languages, even English. Moreover, among all languages, Hindi experiences the highest degree of stereotype leakage — it has four cases of significant stereotype leakage from other languages across three MLLMs. Since Hindi is the only low-resource language we tested, this might explain why it absorbs stereotypes from other languages.

Finally, we report the coefficients of effects from monolingual language models (LM_{tgt}) in Table 3. All the effects are statistically significant and are stronger than the effects from human stereotypes. This is not surprising because monolingual language models and multilingual language models share similar training data and model structures.

4.2 Qualitative Results

We then examine specific stereotypical associations that transfer between languages, focusing on the potential impact of these strengthened associations. Our analysis centers on the GPT-3.5 model, where we observe the highest degree of stereotype leakage. For each source-target language pair with significant stereotype leakage, we analyze the traits most strongly associated with each group in the target language. Special attention is given to traits that, while not linked to the group in the target language's human stereotypes, align with those from the source language. We identify two main types of leakage: the amplification of positive and negative associations, and non-polar leakage, characterized by associations that are neither positive nor negative.

4.2.1 Positive Leakage

According to human annotation, *Asian people* are more positively perceived in the English language than in Russian. We observe the strengthening of such traits in GPT-3.5 Russian language as `wealthy, likable`, and `high status`, possibly resulting from leakages from English and other languages. Moreover, *housewives* become more `warm` in English following leakages from possibly Russian and Hindi. *Black people* are more `powerful, modern, confident`, and `wealthy` in the English language following leakage from Hindi. Another example of the leakage of positive perceptions is for *gay men* and *lesbians* from English to other languages. Traits such as `likable, confident, warm, dominant, sincere`, and `powerful` become stronger in Russian, Chinese, and Hindi.

4.2.2 Negative Leakage

Meanwhile, there are negative stereotypes that leak across languages. From *feminists*, we observe a leakage from English to Chinese and Hindi, and from Russian to Chinese of such stereotypical associations as `egoistic, threatening, repellent`, and `cold`, while in the human data in Hindi, this group is perceived as `warm`.

Another example is *immigrants*. From Russian and English languages, traits such as `threatening, repellent, dishonest, egoistic`, and `unconfident` leak to Chinese and Hindi. Based on human data, we found that people surveyed in Chinese view this group quite favorably since the majority of immigrants to China were highly qualified professionals (Pieke, 2012). Contrarily, in Russia, immigrants are mostly coming from poorer neighboring countries and are negatively stereotyped in society, while in the U.S., immigrants are diverse and could be both marginalized or privileged.

Moreover, there is a notable leakage from English to Chinese and Hindi for *Black people* for traits `dominated` and `poor`. This aligns with known stereotypes about African Americans and Africans in U.S. society (Miller-Cribbs and Farber, 2008; Galster and Carr, 1991; Beresford, 1996).

4.2.3 Non-polar Leakage

There are also non-polar leakages, which are neither positive nor negative. From Hindi to English and Russian, we see the strengthening of `religious` for various groups such as *women*, *disabled people*, *Black people*, and *Asian people*. It has been shown that there are more than 70.00% believers of the total population in India as of 2011(Sahgal et al., 2021).

4.2.4 Non-shared Groups Leakage

In the case of non-shared groups, we expected unidirectional transferring of the groups' perceptions from the language of origin to other languages.

Our findings confirm this hypothesis. For example, the group *VDV soldiers* is a widely known military unit in Russia. There are strong stereotypes in Russian society about this group, but the group is mostly unknown to Americans. Out of the 34 survey English survey respondents who passed the quality tests, no one chose this group as a familiar one. Stereotypes of this group leak from Russian to English, strengthening traits such as `confident, traditional, competitive`, and `threatening`. Another example is the *Hui people*, a group widely unknown to Russian and Hindi society: out of 76 respondents for both surveys, no one chose this group as the familiar one. This social group is a minority in China and is composed of Chinese-speaking followers of Islam. Originally, *Hui people* were marginalized in China and viewed as more traditional, religious, and conservative (Hillman, 2004; Hong, 2005). Accordingly, we observed the leakage of such traits as `irrational, traditional, threatening, repellent, religious, and egoistic`. All groups specific to the Hindi language — *Gujarati, Brahmin*, and *Shudra people* — have certain traits leaking to the English and Russian languages. For example, high caste groups (*Gujarati* and *Brahmin people*) strengthen such positive traits as `wealthy, likable, sincere, powerful, high status, competitive`, and `confident`. In addition, *Shudra people* become more associated in GPT-3.5 with traits `poor, low status, powerless, traditional, religious`, and `dominated`. This leakage corresponds to the perception of these groups in Indian society and by our survey respondents (Witzel, 1993; Milner, 1993).

4.3 Discussion

The amplification of negative stereotypes is concerning as it perpetuates discrimination and prejudice. While positive stereotypes may seem harmless, they can also create unrealistic expectations and pressures. For example, the stereotype that *Asian people* are `wealthy` or *housewives* are `warm` ignores individual diversity and enforces restrictive gender roles.

Stereotype leakage is especially problematic in fields like education and creative content generation, which shape public perception and personal development. MLLMs used in these areas must be cautious of this effect to maintain content integrity.

5 Conclusion

Multilingual large language models have the potential to spread stereotypes beyond the societal context they emerge from, whether by generating new stereotypes, amplifying existing ones, or reinforcing prevailing social perceptions from dominant cultures. In our study, we demonstrate that this concern is indeed valid. To do so, we establish a framework for measuring the leakage of stereotypical associations in multilingual large language models across languages. Overall, we find that the stereotype leakage occurs bidirectionally meaning that when one language transmits stereotypes to others, it likely receives some stereotypes from other languages as well. We also observe the most stereotype leakage effect within the GPT-3.5 model.

Within the GPT-3.5 model, we observe the strengthening of positive, negative, and non-polar associations in the model. In addition, our study underscores the role of "native" languages in framing social groups unknown to other linguistic communities. Such leakage of stereotypes amplifies the complexity of societal perceptions by introducing a complex interconnected bias from different languages and cultures. In the context of shared groups, stereotype leakage may manifest as the manifestation of stereotypes that were not previously present within the cultural setting of a particular group. For non-shared groups, stereotype leakage can extend the reach of existing stereotypes from the source culture to other cultural contexts.

To our knowledge, we are the first to introduce the concept of stereotype leakage across languages in multilingual LLMs. We propose a framework for quantifying this leakage in multilingual models, which can be easily applied to unstudied social groups. We show that multilingual large language models could facilitate the transmission of biases across different cultures and languages. We demonstrate the existence of stereotype leakage within MLLMs, which are trained on diverse linguistic datasets. As multilingual models begin to play an increasingly influential role in AI applications and across societies, understanding their potential vulnerabilities and the level of bias propagation across linguistic boundaries becomes important. As a result, we lay the groundwork for advancing both the theoretical comprehension of multilingual models and the practical implementation of bias mitigation in AI systems.

Limitations and Ethical Considerations

Our study has several limitations. First, we are limited in our ability to run a causal analysis because none of the studied languages can be easily removed from the training data to see their genuine impact on stereotypical associations in other languages. Retraining GPT-3.5, for instance, is not a feasible option. Thus, we use the BERT monolingual model as a proxy for each language.

In addition, stereotype traits were selected based on the ABC model, which was developed and tested using U.S. and German stereotypes. Though we translated our surveys into all four languages, the stereotype traits may better reflect Anglocentric stereotypes (Talat et al., 2022) than others.

Furthermore, the human stereotypes we collected may already reflect the influence of social stereotype transmission. For instance, in our study, we surveyed crowd workers about their consumption of U.S. social media. We found that, on average, 39% of respondents from Russia, China, and India engage with U.S. social platforms. Such American cultural dominance could affect the human stereotypes collected in these three languages.

Lastly, while we indirectly consider culture through survey results on associations, we do not measure or account for culture comprehensively. Our English language survey results only apply to the U.S., Russian to Russia, Chinese to China, and Hindi to India.

Acknowledgments

This work is supported by the National Science Foundation under Grant No. 2131508, as well as Grant Nos. 2229885, 2140987, and 2127309 were awarded to the Computing Research Association for the CIFellows Project. We express our gratitude to Chenglei Si for prompting suggestions, Navita Goyal for her assistance with translations, and the members of the Clip Lab at the University of Maryland, along with our friends, for their contributions to the pilot surveys.

References

Andrea Abele, Naomi Ellemers, Susan Fiske, Alex Koch, and Vincent Yzerbyt. 2020. Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological review*, 128.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Peter Beresford. 1996. Poverty and disabled people: Challenging dominant debates and policies. *Disability & Society*, 11(4):553–568.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. Seegull multilingual: a dataset of geo-culturally situated stereotypes. *Preprint*, arXiv:2403.05696.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

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.

Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building stereotype repositories with complementary approaches for scale and depth. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

George C. Galster and James H. Carr. 1991. Housing discrimination and urban poverty of african-americans. *Journal of Housing Research*, 2(2):87–123.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint*.

Ben Hillman. 2004. The rise of the community in rural china: Village politics, cultural identity and religious revival in a hui hamlet. *The China Journal*, (51):53–73.

Ding Hong. 2005. A comparative study on the cultures of the dungan and the hui peoples. *Asian Ethnicity*, 6(2):135–140.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Alex Koch, Angela Dorrough, Andreas Glöckner, and Roland Imhoff. 2020. The abc of society: Perceived similarity in agency/socioeconomic success and conservative-progressive beliefs increases intergroup cooperation. *Journal of Experimental Social Psychology*, 90:103996.

Alex Koch, Vincent Yzerbyt, Andrea Abele, Naomi Ellemers, and Susan T. Fiske. 2021. *Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts*, volume 63, page 1–68. Elsevier.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337.

Sarah Ariel Lamer, Paige Dvorak, Ashley M. Biddle, Kristin Pauker, and Max Weisbuch. 2022. The transmission of gender stereotypes through televised patterns of nonverbal bias. *Journal of Personality and Social Psychology*, 123(6):1315–1335.

Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. *Preprint*, arXiv:2305.11242.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joel E. Martinez, Lauren A. Feldman, Mallory J. Feldman, and Mina Cikara. 2021. Narratives shape cognitive representations of immigrants and immigration-policy preferences. *Psychological Science*, 32(2):135–152.

Julie E. Miller-Cribbs and Naomi B. Farber. 2008. Kin Networks and Poverty among African Americans: Past and Present. *Social Work*, 53(1):43–51.

Murray Milner. 1993. Hindu eschatology and the indian caste system: An example of structural reversal. *The Journal of Asian Studies*, 52:298–319.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *CoRR*, abs/2004.09456.

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. *Preprint*, arXiv:2305.14456.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Afsan Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? *Preprint*, arXiv:2407.05740.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Frank N. Pieke. 2012. Immigrant china. *Modern China*, 38(1):40–77.

Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34):13526–13531.

Neha Sahgal, Jonathan Evans, Ariana Monique Salazar, Kelsey Jo Starr, and Manolo Corichi. 2021. Religion in india: Tolerance and segragation. *Pew Research Centre*.

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Assessing multilingual fairness in pre-trained multimodal representations. *CoRR*, abs/2106.06683.

Michael Witzel. 1993. Toward a history of the brahmins. *Journal of the American Oriental Society*, 113:264.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

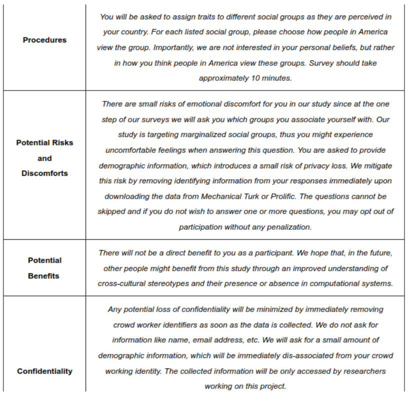
Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

A Human Study

In this Section, we present details about the survey design, annotations quality control, and participants demographics.

A.1 Survey Design

Participants first reviewed a consent form, which outlined the purpose of the study, data usage, and confidentiality. Only after agreeing to participate they proceeded to the survey instructions. The consent form is shown in Figure 3



Procedures	*You will be asked to assign traits to different social groups as they are perceived in your country. For each listed social group, please choose how people in America view the group. Importantly, we are not interested in your personal beliefs, but rather in how you think people in America view these groups. Survey should take approximately 10 minutes.*
Potential Risks and Discomforts	*There are small risks of emotional discomfort for you in our study since at the one step of our surveys we will ask you which groups you associate yourself with. Our study is targeting marginalized social groups, thus you might experience uncomfortable feelings when answering this question. You are asked to provide demographic information, which introduces a small risk of privacy loss. We mitigate this risk by removing identifying information from your responses immediately upon downloading the data from Mechanical Turk or Prolific. The questions cannot be skipped and if you do not wish to answer one or more questions, you may opt out of participation without any penalization.*
Potential Benefits	*There will not be a direct benefit to you as a participant. We hope that, in the future, other people might benefit from this study through an improved understanding of cross-cultural stereotypes and their presence or absence in computational systems.*
Confidentiality	*Any potential loss of confidentiality will be minimized by immediately removing crowd worker identifiers as soon as the data is collected. We do not ask for information like name, email address, etc. We will ask for a small amount of demographic information, which will be immediately dis-associated from your crowd working identity. The collected information will be only accessed by researchers working on this project.*

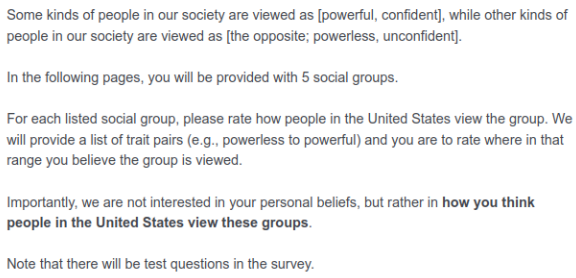
Figure 3: Selected points of the consent form highlighting study format, confidentiality, and potential risks.

For each social group, participants read the following prompt in their respective language: "As viewed by American/Russian/Chinese/Indian society, (while my own opinions may differ), how `[e.g., powerless, dominant, poor]` versus `[e.g., powerful, dominated, wealthy]` are `<group>`?" They then rated each group on a slider scale ranging from -50 to 50, where the two poles of the scale represented opposite traits (e.g. `powerless` and `powerful`). Each social group appeared on a separate page, and participants were unable to return to previous pages, helping to minimize response bias. Example of the task is presented in Figure 4

To reduce social desirability bias, the instructions clearly emphasized: "We are not interested in your personal beliefs, but rather in how you think people in the United States/Russia/China/India view these groups." The exact formulation is presented in Figure 5.

Participants were paid $2.00 to rate five social groups on 16 pairs of traits, which took an average of 10 minutes to complete, translating to a compensation rate of $12.00 per hour.

Figure 4: Example of the survey.



Some kinds of people in our society are viewed as [powerful, confident], while other kinds of people in our society are viewed as [the opposite; powerless, unconfident].

In the following pages, you will be provided with 5 social groups.

For each listed social group, please rate how people in the United States view the group. We will provide a list of trait pairs (e.g., powerless to powerful) and you are to rate where in that range you believe the group is viewed.

Importantly, we are not interested in your personal beliefs, but rather in **how you think people in the United States view these groups**.

Note that there will be test questions in the survey.

Figure 5: Instructions before crowd workers view the task itself.

A.2 Quality Assurance

Collecting high-quality data for subjective tasks presents significant challenges, particularly due to the absence of objective ground truth. To mitigate these challenges, we implemented rigorous quality control procedures to ensure reliability and consistency across annotations.

The survey was administered through the Prolific platform, and only participants with an approval rate exceeding 90% were eligible to participate. This threshold was selected to balance data quality with participant availability, as it is generally considered high for Prolific, increasing the likelihood of obtaining reliable data.

In addition to the platform's approval rate, we implemented three test questions throughout the survey to assess attentiveness and comprehension:

- After the first group, participants must name the group they just scored.
- After the second, participants must list one trait they just marked high and one marked low.

- The fifth (final) group is a repetition of one of the four groups they previously scored.

We exclude annotators who answered the first two questions incorrectly. We then measured their intra-annotator (self) agreement by comparing the consistency of their responses, and any annotation with less than 80% self-agreement was discarded. These measures helped ensure data quality, though all participants were compensated regardless of their performance in the quality tests.

We collected at least five valid annotations per group that met our quality thresholds. Of the 286 participants, 151 passed the quality checks. Specifically, 34 participants passed for the English-language survey, 36 for Russian, 41 for Chinese, and 40 for Hindi. The fact that nearly half of the participants failed to meet the quality criteria underscores the necessity of these controls in subjective data collection.

A.3 Participant Demographics

We collected demographic information from participants, including gender, age, education level, and, for non-English speakers, their frequency of reading American social media. Participants were free to skip any question they preferred not to answer.

Across all languages, the gender distribution revealed a near balance: 49% identified as male, 45% as female, and 5% as non-binary, transgender, or gender fluid, with a few opting not to disclose. When we examined educational backgrounds, participants from non-English-speaking countries showed similar trends: 36% held a bachelor's degree, 32% had a master's degree, and 7% had earned a Ph.D. The remaining respondents either had lower educational qualifications or chose not to answer. English-speaking participants stood out, with no respondents holding a Ph.D., a lower percentage with master's degrees (29%), and a larger proportion (35%) being high school graduates.

Among the English-speaking survey group, the largest proportion of respondents hailed from Texas (15%), followed by California and New York (each contributing 9%). The remaining participants were dispersed across 25 states, with no significant regional concentration outside these key areas.

As we looked at the age distribution, it was clear that younger people dominated the study, with 42% aged between 18 and 30 and 33% falling in the 31 to 40 range. The remainder were above 40, with the youngest participant being 18 and the oldest, a more experienced 72.

One notable demographic trend emerged in media consumption habits. Russian-speaking participants were the most frequent consumers of American media, with 44% stating they read it regularly. In contrast, 35% of Hindi-speaking participants and 28% of Chinese-speaking participants reported similar habits. Across all groups, about 39% said they occasionally consumed American media, while only 5% never did. These patterns suggest that Russian participants may be more exposed to or interested in global perspectives, particularly through American social media.

Crucially, all approved participants confirmed fluency in the language of their respective surveys. This ensures that any differences in responses were not influenced by language proficiency but more likely reflected deeper cultural or regional perspectives.

A.4 Human pairwise Pearson correlation

Results are in Table 4.

B Model Stereotypical Association Measurement

B.1 Models

The models used are bert-base-multilingual-cased (Devlin et al., 2018), google/mt5-base (Xue et al., 2021), and gpt-3.5-turbo-0125 (Ouyang et al., 2022).

B.2 Templates

Templates variations are presented in Table 5.

B.3 GPT Model Generation Failures on Marginalized Groups

We observe that for certain groups like *feminist* and *Muslim person* in Chinese, the model often disregards the prompt and simply outputs the group name. Moreover, in some cases, the model alters the trait specified in the prompt. For example, it changes `dominating` to `dominated` for *disabled person* in English or `poor` to `wealthy` for *migrant worker* in Russian. Additionally, the model may overlook the traits provided in the prompt and generate stereotypical traits instead. For instance, in Russian, it generates `rape` and `patriot` for *Puerto Rican* or `cowboy` for *Texan*.

We also count the number of system generations that did not match the instruction requirements for

Social Group	en-ru	en-zh	en-hi	ru-zh	ru-hi	zh-hi
man	0.83	0.79	0.84	0.80	0.83	0.78
woman	0.55	0.45	0.31	0.32	0.03	-0.21
gay man	0.56	0.73	0.44	0.79	0.67	0.80
lesbian	0.39	0.59	0.74	0.66	0.63	0.86
single mother	0.47	0.65	0.58	0.51	0.12	0.57
housewife	0.45	0.50	0.37	0.87	0.70	0.65
software engineer	0.76	0.70	0.45	0.90	0.82	0.81
wealthy person	0.83	0.90	0.77	0.89	0.68	0.80
poor person	0.83	0.80	0.67	0.87	0.80	0.73
disabled person	0.65	0.83	0.71	0.83	0.21	0.50
Asian person	-0.07	0.46	0.77	0.02	-0.31	0.50
Black person	0.58	0.49	0.38	0.56	0.28	0.35
Muslim person	0.86	0.88	0.71	0.94	0.77	0.82
immigrant	0.74	0.10	0.36	0.16	0.27	-0.22
government official	0.83	0.68	0.79	0.83	0.82	0.80
civil servant	0.34	0.17	0.34	0.81	0.74	0.70
feminist	0.68	0.74	0.54	0.45	0.30	0.75
veteran	0.44	0.60	0.41	0.39	0.03	0.51

Table 4: Pairwise Pearson correlation for shared social groups.

Singular The/That/A `[group]` is .	**Plural** Most/Many/All `[group]` are . / `[Group]` are .
Declarative `[Group]` are .	**Interrogative** Why are `[group]` ?
Non-adverbial `[Group]` are .	**Adverbial** `[Group]` are very/so/mostly .
Fact `[Group]` are .	**Belief** I/We/Everyone/People believe/expect/think/know(s) that `[group]` are .
Fact `[Group]` are .	**Social Expectation** `[Group]` are supposed to be/should be/are seen as/ought to be/are expected to be .
Group-first `[Group]` are .	**Trait-first** The people are `[group]`.
Non-comparative `[Group]` are .	**Comparative** `[Group]` are more likely to be than others.

Table 5: Template Variations.

each social group. For example, in Chinese, we observed 108 generations for the group “feminist” that did not match the instruction requirement out of a total of 2880 generations. In comparison, there were 20 non-matching generations for “women.” However, these figures represent only an upper bound of system failures, as various reasons, such as generating synonyms, could cause mismatches. As stated in the paper, we leave the in-depth analysis for future work.