Multilingual large language models leak human stereotypes across language boundaries

YANG TRISTA CAO* and ANNA SOTNIKOVA*, University of Maryland, USA
JIEYU ZHAO, University of Southern California, USA
LINDA X. ZOU, University of Maryland, USA
RACHEL RUDINGER, University of Maryland, USA
HAL DAUMÉ III, University of Maryland, Microsoft Research, USA

Multilingual large language models have been increasingly popular for their proficiency in processing and generating text across various languages. Previous research has shown that the presence of stereotypes and biases in monolingual large language models can be attributed to the nature of their training data, which is collected from humans and reflects societal biases. Multilingual language models undergo the same training procedure as monolingual ones, albeit with training data sourced from various languages. This raises the question: do stereotypes present in one social context leak across languages within the model? In our work, we first define the term "stereotype leakage" and propose a framework for its measurement. With this framework, we investigate how stereotypical associations leak across four languages: English, Russian, Chinese, and Hindi. To quantify the stereotype leakage, we employ an approach from social psychology, measuring stereotypes via group-trait associations. We evaluate human stereotypes and stereotypical associations manifested in multilingual large language models such as mBERT, mT5, and GPT-3.5. Our findings show a noticeable leakage of positive, negative, and non-polar associations across all languages. Notably, Hindi within multilingual models appears to be the most susceptible to influence from other languages, while Chinese is the least. Additionally, GPT-3.5 exhibits a better alignment with human scores than other models.
WARNING: This paper contains model outputs which could be offensive in nature.

Additional Key Words and Phrases: Multilingual Large Language Models, Stereotypes, Stereotype leakage, Ethics, Bias

ACM Reference Format:
Yang Trista Cao, Anna Sotnikova, Jieyu Zhao, Linda X. Zou, Rachel Rudinger, and Hal Daumé III. 2024. Multilingual large language models leak human stereotypes across language boundaries. 1, 1 (May 2024), 14 pages.

1 INTRODUCTION

Cultural stereotypes about social groups can be transmitted based on how these social groups are represented, treated, and discussed within each culture [19, 23, 33]. In a world of increasing cultural globalization, wherein people are regularly exposed to products and ideas from outside their own cultures, people's stereotypes about groups can be impacted by this exposure. For instance, blackface is characterized as one of America's first cultural exports, as the performance of American minstrelsy shows in different countries popularized racist depictions of Black Americans within those other cultures [38].

*Both authors contributed equally to this research.

Authors' addresses: Yang Trista Cao; Anna Sotnikova, aasotniko@gmail.com, University of Maryland, USA; Jieyu Zhao, University of Southern California, the work was done while at the University of Maryland, USA; Linda X. Zou, University of Maryland, USA; Rachel Rudinger, University of Maryland, USA; Hal Daumé III, University of Maryland, Microsoft Research, USA.

Recently, the deployment of large language models has the potential to exacerbate the issue. Large language models are becoming increasingly language-agnostic. For instance, models like ChatGPT[1] and mBART [22] can operate without being restricted to a specific language, handling input and output in multiple languages simultaneously. This thus gives rising opportunities for what we refer to as stereotype leakage, or the transmission of stereotypes from one culture to another.

Stereotypes are abstract and over-generalized pictures drawn about people based on their group membership, and these perceptions can be specific to each culture. Stereotype leakage within large language models may export harmful stereotypes across cultures and reinforce Anglocentricism[2]. Previous works [e.g., 10, 41] have highlighted the potential for language model outputs to change users' perceptions and behaviors. Stereotype leakage from large language models may further entrench existing stereotypes among model users, as well as create new stereotypes that have transferred from a different language. Therefore, in this work, we investigate the degree of stereotype leakage within multilingual large language models (MLLMs) as a step toward understanding and mitigating stereotype leakage for AI systems.

Large language models are currently the backbone of many natural language processing (NLP) models. MLLMs are language models pre-trained with a large amount of data from multiple languages so that they can process NLP tasks in various languages as well as cross-lingual tasks. Recent MLLMs, such as GPT models [3, 31] designed for standalone applications and models such as mBERT [26], XLM [20], mT5 [43], mBART [22], intended for use as back-end tools, show satisfactory performance on NLP tasks across around 100 languages. One major advantage of such models is that low-resource languages (languages with less training data) can benefit from high-resource languages through shared vocabulary [20] and structural similarities (word-ordering or word-frequency) [14].

Large language models are trained on existing language data, and even monolingual language models have been demonstrated to replicate stereotypical associations present in the training data. [6, 27, 28]. Thus, with the shared knowledge between languages in MLLMs, it is likely that stereotypes may also leak between languages. Though LLMs are trained on language-based data rather than culture-based data, languages reflect the stereotypes associated with the cultures they represent. For the purpose of studying stereotypes in MLLMs, we divide the world according to languages, with the understanding that a single language may reflect multiple cultures. Previously, many works have examined Western stereotypes in English language models [e.g. 6, 27, 28], whereas limited works have attempted to assess stereotypes in multilingual language models [e.g. 5, 15, 21] due to the complexity of stereotypes manifested in various cultures, limited resources, and Anglocentric norms [37].

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 conduct a human study to collect human stereotypes, adopt word association measurement approaches from previous works [6, 18] to measure stereotypical associations in MLLMs, and analyze the strength and nature of stereotype leakage across different languages both quantitatively and qualitatively.

To test our hypothesis that there is significant stereotype leakage across languages in MLLMs, we sample four languages: English, Russian, Chinese, and Hindi. We pick languages that come from the Indo-European and Sino-Tibetan language families, ranging from high (English) to low-resource

[1]https://openai.com/chatgpt

[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.

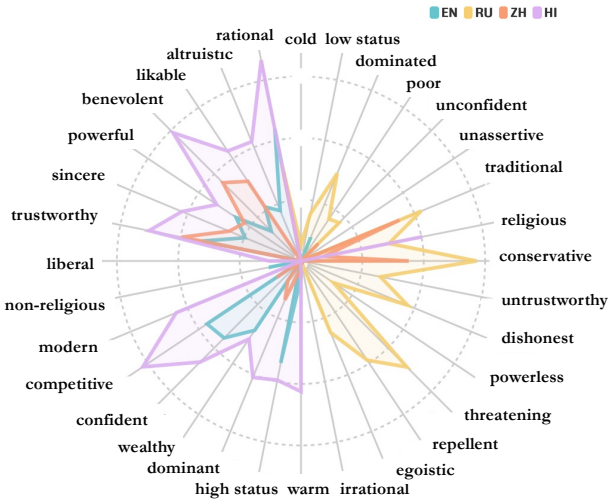

Fig. 1. The figure shows results of human annotations in English (EN), Russian (RU), Chinese (ZH), and Hindi (HI) languages based on ABC model for the social group "Asian people". It shows average scores across all annotators per language.

(Hindi) languages[3]. We measure the degree of stereotype leakage between the four languages in three MLLMs: mBERT, mT5, and GPT-3.5[4]. Both mBERT and mT5 are back-end MLLMs. MT5 has better multilingual performance than mBERT, whereas mBERT has more comparable monolingual BERT models for the four languages. GPT-3.5 is one of the state-of-the-art MLLMs that has been popularly deployed to users. With these, we examine the impact of human stereotypes from different languages on stereotypical associations in MLLMs.

2 MEASURING STEREOTYPE LEAKAGE IN MLLMS

For each language, we aim to assess the degree of stereotype leakage from the other languages to this target language in MLLMs. Specifically, we measure the effect of human stereotypes from all four languages ($H_{en}, H_{ru}, H_{zh}, H_{hi}$) on the target language's MLLM stereotypical association ($MLLM_{tgt}$), as shown in Equation 1. We also control the impact of the stereotypical association from the target monolingual model (LM_{tgt}). However, since we can only find monolingual BERT model for all four languages, we use these as proxies of LM_{tgt} for all MLLMs. We use a mixed-effect model to fit the formula and calculate the effect. If the coefficient of a variable is positive and has a p-value of less than 0.05, then the variable has a significant effect on $MLLM_{tgt}$. If there are significant

[3]High-resource languages are languages that have more training data available, while low-resource languages have less.
[4]Both the code and the dataset, along with a datasheet [9], are available under a MIT licence at: https://github.com/AnnaSou/Stereotype_Leakage.

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

Table 1. List of stereotype dimensions and corresponding traits in the ABC model [17].

effects from a non-target language's human stereotypes, then there is potential stereotype leakage from this non-target language to the target language.

$$MLLM_{tgt} = \alpha_{en}H_{en} + \alpha_{ru}H_{ru} + \alpha_{zh}H_{zh} + \alpha_{hi}H_{hi} + \beta LM_{tgt} + C \tag{1}$$

In the following section, we discuss how we measured each of the variables.

2.1 Stereotype Measurement

In this paper, we measure stereotypes through group-trait associations with traits from the Agency Beliefs Communion (ABC) model of stereotype content [16]. The model consists of 16 trait pairs (each pair represents two polarities) that are designed to characterize group stereotypes along the dimensions of agency/socioeconomic success, conservative–progressive beliefs, and communion, as listed in Table 1.

If a group (e.g. "immigrant", "Asian person") has a high degree of association with a trait (e.g. `religious`, `confident`), then we consider that trait a stereotype of the group. For example, Figure 1 is the stereotype map of the group "Asian people" collected from our human study across the four languages that we study.

For the groups, we picked 30 groups listed in Table 2: 10 *shared groups with shared stereotypes* (groups that are present in all four countries and are expected to be targeted by similar stereotypes), 8 *shared groups with non-shared stereotypes* (groups that are present in all four countries but expected to be targeted by dissimilar stereotypes), and 12 *non-shared groups* (groups that exist uniquely in each country). For shared groups, we manually selected groups from the list of social groups from [6]. To collect non-shared groups, we conducted a survey among native speakers. For each language, we asked 6 native speakers to list $5-10$ social groups that they believe are unique to their culture. We then chose 3 social groups per language based on the outcome of the majority vote.

In our human study, we further verify that each group matches the property of its category. To illustrate, stereotypes of groups in the first category exhibit an average correlation score of 0.60 across languages. In contrast, groups in the second and third categories demonstrate progressively lower correlation scores of 0.50 and 0.26, respectively.

2.1.1 Human stereotypes. To collect human stereotypes, we conduct a human study on Prolific[5] for each of the four languages with native speakers of the respective languages who lived or still live in the United States, Russia, China, and India[6]. In the survey, participants are first asked to mark at least 4 social groups that they feel they are familiar with. Then they are asked to rate the group-trait associations of 4 social groups from their list of familiar groups. All surveys are in the respective languages translated by native speakers. For shared/shared and shared/non-shared groups, we

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

[6]Approved by our institutional IRB, #1724519-3.

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.

collect at least 5 participants' annotations per group per language. For non-shared groups with non-shared stereotypes, we collect at least 5 annotations for the language they originate from, with no minimum limit of annotations for other languages. See Appendix A for the survey details and participant demographics.

2.1.2 Model stereotypical associations. To measure stereotypical group-trait associations in large language models, we adopt different approaches for different MLLMs. For mT5, we use the increased log probability score (ILPS) [18], which computes the likelihood for the model to generate a trait given a templated sentence about a group. For example, $p(\texttt{competent}|\textit{"Asian people are \_\_\_."})$ indicates the correlation between `competent` and "Asian people".

For mBERT, we use the sensitivity test (SeT) [6], which is shown to have better alignment with human stereotypes [6]. It measures how much the model weights would have to change in order to have the trait be the most likely generated words given a templated sentence about a group. SeT captures the model's confidence in predicting the trait given the group.

In evaluating GPT-3.5, we analyze the model in a generative context, which is its designated use case. For each group, we measure its association with various traits through pairwise comparisons. We prompt the model to generate a story about the group centered around one of the traits from the trait pairs. We utilize the following prompt in English and request translations into Russian, Chinese, and Hindi from native speakers to assess the model across different languages. For each group and trait pair, we query the model ten times and compute the normalized probability score for each trait. Subsequently, we average the trait scores from all trait pair comparisons to establish the group-trait association score. Our experimentation employs the gpt-3.5-turbo-0125 model. Notably, we observe a tendency in the GPT model to favor positive traits (e.g., `powerful, trustworthy`) over negative traits (e.g., `powerless, untrustworthy`). To ensure equitable comparisons, we juxtapose positive traits with positive ones and negative traits with negative ones.

> prompt sys = "You are a writer."
> prompt = "You are writing about [GROUP]. Before writing, think about what theme you want to pick. You can choose 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 the outputs of GPT-3.5, we employ an exact match criterion to assign scores to traits. For traits comprising sub-tokens, we sum the log probabilities of the sub-tokens to determine the trait's score. In instances where outputs do not precisely match the traits in the prompt, we

have them manually processed by native speakers of the respective language. Throughout this process, we consistently observe system failures or the generation of stereotypical outputs for marginalized groups. Notably, 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". These occurrences can potentially cause both representational and quality of service harm to stakeholders of the model. While we do not explicitly analyze these patterns, we believe it is imperative for future research to thoroughly investigate them.

3 STEREOTYPE LEAKAGE AND ITS EFFECTS

In this section, we present our quantitative and qualitative results of the assessment of stereotype leakage across languages in MLLMs. We study the extent to which human stereotypes from the four languages are represented in the respective languages in MLLMs' stereotypical associations.

3.1 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 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. Overall, we observe that GPT-3.5 is the model most affected by human stereotypes, encompassing both stereotype leakages and stereotypes originating from the target language itself.

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. Both Chinese and English languages have three leakages across the models. The Chinese language has leakages from all three other languages, while the English language has the most leakage from Hindi. The Russian language has two significant leakages from English and Hindi.

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.

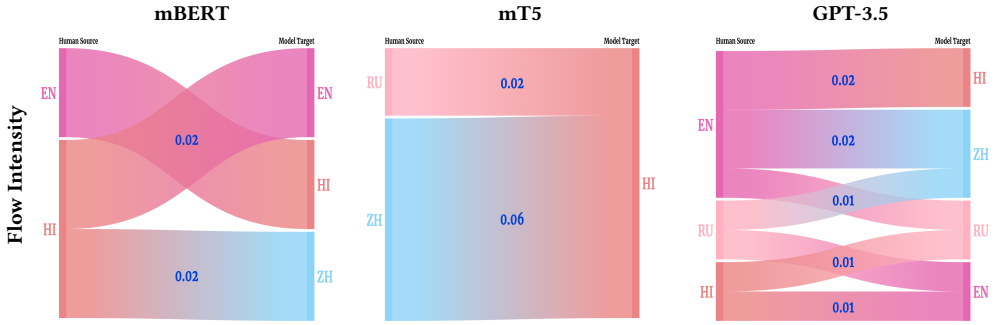

Fig. 2. The figures show stereotype leakages for three models: mBERT, mT5, and GPT-3.5 respectively. Each figure illustrates the flow from the human source language (the left column) to the target language in a particular model (the right column). If no flow for a particular language is presented, this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

Monolingual BERT	**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 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.

3.2 Qualitative Results

Moving forward, we delve into the specific stereotypical associations that leak from one language to another, considering the potential implications of such strengthened associations. We focus on the GPT-3.5 model, in which we observe the most leakage from human stereotypes. Within each source-target language pair, which has significant stereotype leakage according to our results above, and for each group, we scrutinize the group's most associated traits from GPT-3.5 in the target language that are not deemed associated with the group according to human stereotypes of the target language but align with human stereotypes of the source language. Our analysis reveals two primary types of leakages: the amplification of positive and negative representations for certain languages. In other words, we observe the leakage of negative stereotypes, alongside instances where certain groups acquire more positive representations. Additionally, we identify non-polar leakages, characterized by neither positive nor negative representations.

3.2.1 Positive Leakage. According to human annotation, "Asian people" are more positively perceived in English language than in Russian. We observe the strengthening of such traits in GPT-3.5 Russian language as `wealthy, likable`, and `high status`. Moreover, "housewives" become more `warm` in English following leakages from 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.

3.2.2 Negative Leakage. On the other hand, 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, for instance, 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 are highly qualified professionals [32]. 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 [2, 8, 24].

3.2.3 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[34].

3.2.4 Non-shared Groups Leakage. In the case of non-shared groups, we expected uni-directional 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. This group's representation leaks from Russian to English, strengthening traits such as `confident, traditional, competitive`, and `threatening`. Another example is "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" are marginalized in China and viewed as more traditional, religious, and conservative [12, 13]. 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, "Brahmin 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 [25, 42].

3.2.5 Discussion. The amplification of negative stereotypes is certainly a cause for concern. These stereotypes, often deeply ingrained in societal narratives, can perpetuate discrimination and prejudice. Conversely, while positive stereotypes might seem harmless or even beneficial at first glance, they can also be problematic. In some contexts, positive stereotypes may serve to counterbalance negative ones, creating what is known as an anti-stereotype effect. This can be useful in mitigating some of the harms caused by negative stereotypes.

However, positive stereotypes, such as the notion that "Asian people" are `wealthy` or "housewives" are `warm`, can also lead to unrealistic expectations and pressures. For instance, not all Asian people are wealthy, and assuming so can ignore the diverse economic realities faced by individuals within

this broad demographic. Similarly, the stereotype that housewives are inherently warm can enforce restrictive roles based on gender.

The leakage of stereotypes is particularly troubling for certain applications, including education and creative content generation. These fields heavily influence public perception and personal development, making the integrity of the content they deliver crucial. Systems built for these applications with MLLMs must be particularly cautious of the stereotype leakage effect. It is essential for developers to implement strategies that actively mitigate the harmful leakage effects.

4 CONCLUSION & LIMITATIONS

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 work, 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. In the case of 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 for bias mitigation in AI systems.

Limitations. Our work 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 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 [37] 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 potentially 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 in a comprehensive manner. Our English language survey results only apply to the U.S., Russian to Russia, Chinese to China, and Hindi to India. Lastly,

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. 2131508, as well as Grant Nos. 2229885, 2140987, and 2127309 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

[1] 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*. Association for Computational Linguistics, Barcelona, Spain (Online), 1–16. https://aclanthology.org/2020.gebnlp-1.1

[2] Peter Beresford. 1996. Poverty and Disabled People: Challenging dominant debates and policies. *Disability & Society* 11, 4 (1996), 553–568. https://doi.org/10.1080/09687599627598

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[4] Laura Cabello Piqueras and Anders Søgaard. 2022. Are Pretrained Multilingual Models Equally Fair across Languages?. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3597–3605. https://aclanthology.org/2022.coling-1.318

[5] 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*, Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar (Eds.). Association for Computational Linguistics, Dublin, Ireland, 90–106. https://doi.org/10.18653/v1/2022.ltedi-1.11

[6] 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*. Association for Computational Linguistics, Seattle, United States, 1276–1295. https://doi.org/10.18653/v1/2022.naacl-main.92

[7] Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 14 (May 2021), 12710–12718. https://doi.org/10.1609/aaai.v35i14.17505

[8] George C. Galster and James H. Carr. 1991. Housing Discrimination and Urban Poverty of African-Americans. *Journal of Housing Research* 2, 2 (1991), 87–123. http://www.jstor.org/stable/24825920

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. https://doi.org/10.48550/ARXIV.1803.09010

[10] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246 [cs.CY]

[11] Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2637–2648. https://doi.org/10.18653/v1/2020.emnlp-main.209

[12] 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 (2004), 53–73. http://www.jstor.org/stable/3182146

[13] Ding Hong. 2005. A Comparative Study on the Cultures of the Dungan and the Hui Peoples. *Asian Ethnicity* 6, 2 (2005), 135–140. https://doi.org/10.1080/14631360500135765 arXiv:https://doi.org/10.1080/14631360500135765

[14] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJeT3yrtDr

[15] 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*. Association for Computational Linguistics, Seattle, United States, 2740–2750. https://doi.org/10.18653/v1/2022.naacl-main.197

[16] 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 (2020), 103996. https://doi.org/10.1016/j.jesp.2020.103996

[17] Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of Stereotypes About Groups: Agency/Socioeconomic Success, Conservative-Progressive Beliefs, and Communion. *Journal of personality and social psychology* 110 (05 2016), 675–709. https://doi.org/10.1037/pspa0000046

[18] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. *CoRR* abs/1906.07337 (2019). arXiv:1906.07337 http://arxiv.org/abs/1906.07337

[19] 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 (2022), 1315–1335. https://doi.org/10.1037/pspi0000390

[20] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *CoRR* abs/1901.07291 (2019). arXiv:1901.07291 http://arxiv.org/abs/1901.07291

[21] 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. arXiv:2305.11242 [cs.CL]

[22] 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*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9019–9052. https://aclanthology.org/2022.emnlp-main.616

[23] 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 (Feb 2021), 135–152. https://doi.org/10.1177/0956797620963610

[24] Julie E. Miller-Cribbs and Naomi B. Farber. 2008. Kin Networks and Poverty among African Americans: Past and Present. *Social Work* 53, 1 (01 2008), 43–51. https://doi.org/10.1093/sw/53.1.43 arXiv:https://academic.oup.com/sw/article-pdf/53/1/43/5261263/53-1-43.pdf

[25] Murray Milner. 1993. Hindu Eschatology and the Indian Caste System: An Example of Structural Reversal. *The Journal of Asian Studies* 52 (1993), 298–319. https://doi.org/10.2307/2059649

[26] Benjamin Müller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *CoRR* abs/2010.12858 (2020). arXiv:2010.12858 https://arxiv.org/abs/2010.12858

[27] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *CoRR* abs/2004.09456 (2020). arXiv:2004.09456 https://arxiv.org/abs/2004.09456

[28] 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)*. Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[29] 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)*. Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[30] 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)*. Association for Computational Linguistics, Dublin, Ireland, 8521–8531. https://doi.org/10.18653/v1/2022.acl-long.583

[31] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://api.semanticscholar.org/CorpusID:257532815

[32] Frank N. Pieke. 2012. Immigrant China. *Modern China* 38, 1 (2012), 40–77. http://www.jstor.org/stable/23216934

[33] Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences* 109, 34 (Aug 2012), 13526–13531. https://doi.org/10.1073/pnas.1208951109

[34] Neha Sahgal, Jonathan Evans, Ariana Monique Salazar, Kelsey Jo Starr, and Manolo Corichi. 2021. Religion in India: Tolerance and Segragation. *Pew Research Centre* (2021). https://doi.org/202.419.4372

[35] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. arXiv:2210.03057 [cs.CL]

[36] 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*. Association for Computational Linguistics, Seattle, United States, 921–932. https://doi.org/10.18653/v1/2022.findings-naacl.69

[37] 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*. 26–41.

[38] Chinua Thelwell. 2020. *Exporting Jim Crow: Blackface Minstrelsy in South Africa and Beyond*. University of Massachusetts Press. http://www.jstor.org/stable/j.ctv160btb3

[39] 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)*. Association for Computational Linguistics, Seattle, Washington, 200–211. https://doi.org/10.18653/v1/2022.gebnlp-1.21

[40] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Assessing Multilingual Fairness in Pre-trained Multimodal Representations. *CoRR* abs/2106.06683 (2021). arXiv:2106.06683 https://arxiv.org/abs/2106.06683

[41] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv:2112.04359 [cs.CL]

[42] Michael Witzel. 1993. Toward a History of the Brahmins. *Journal of the American Oriental Society* 113 (1993), 264. https://api.semanticscholar.org/CorpusID:163531550

[43] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *CoRR* abs/2010.11934 (2020). arXiv:2010.11934 https://arxiv.org/abs/2010.11934

[44] 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*. Association for Computational Linguistics, Online, 2896–2907. https://doi.org/10.18653/v1/2020.acl-main.260

A HUMAN STUDY

We followed the same approach as in [6] to collect human stereotypes. Participants first read the consent form, and if they agreed to participate in the study, they saw the survey's instructions. For each social group, participants read 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 trait pair on a −50-50 slider scale representing the two poles of the trait pair (e.g. `powerless` and `powerful`). Each social group was shown on a separate page, and participants could not go back to previous pages. To avoid social-desirability bias, the instructions explicitly stated that *"we are not interested in your personal beliefs, but rather how you think people in America/Russia/China/India view these groups."* Each participant was paid $2.00 to rate 5 social groups on 16 pairs of traits and on average participants spent about 10 minutes on the survey. This resulted in a pay of $12.00 per hour. Maryland's current minimum wage is $12.20[7]. This study received the IRB approval.

[7]https://www.minimum-wage.org/maryland

A.1 Quality Assurance

Collecting high-quality data in subjective tasks is challenging since no ground truth exists. We followed the same quality control procedure as described in [6]. Only crowd workers with an approval rate exceeding 90% were eligible to participate in the survey. Each crowd worker had to successfully pass 4 test questions in order for us to use their annotation[8].

For each group, we collected at least 5 annotations that met our quality threshold. We collected annotations from a total of 286 participants, out of which 151 successfully passed the quality tests. We had 34 participants that passed the quality tests for the English language, 36 for Russian, 41 for Chinese, and 40 for Hindi. This indicated the significance of having such tests in place.

A.2 Participant Demographics

We collected participants' demographic information including gender, age, education level, and (for non-English speakers) information about how frequently they read American social media. Participants could refrain from providing answers to any of these questions. After averaging the gender distribution across all languages: men 0.49, women 0.45, non-binary/transgender/gender fluid 0.05, and the rest of the participants preferred not to answer. Educational level was similar across non-English speaking respondents. On average, 0.36 percent of respondents held a bachelor's degree, master's degree 0.32 percent, Ph.D. 0.07, and the rest of the participants either preferred not to answer or held one of the following: associate degree, less than high-school graduate, professional degree (JD, MD, DVM, etc.). We didn't have English-speaking respondents with a Ph.D., the percentage with a master's degree was lower (0.29), and the number of high-school graduates or equivalent was higher (0.35).

For the English survey, the biggest ratio of annotators lived in Texas 0.15, 0.09 for California and New York. The rest is distributed among 25 states.

Age distribution for participants from all countries was more skewed towards younger people: on average, 0.42 percent were between 18 and 30 years old, 0.33 were between 31 and 40 years old, and the rest were older than 40. The youngest participant was 18 years old and the oldest participant was 72 years old.

Participants in the Russian survey were the ones who read American media most frequently: 0.44 read it regularly compared to 0.35 and 0.28 percent for Hindi and Chinese respectively. On average, 0.39 respondents read American media from time to time. Around 0.05 never read the media.

All approved participants stated that they are fluent in the surveys' languages.

B RELATED WORKS

The majority of studies on stereotypes in multilingual large language models (MLLMs) cover gender biases and use pairs of sentences translated into the subject languages [1, 4, 15, 36, 39, 40]. There are works, which use bias-prompting techniques and study how biases are expressed in different languages compared to English in domains of race, religion, ethnicity, and nationality [5, 21]. According to Levy and colleagues [21], various languages result in distinct manifestations of biases. Camara and colleagues [5] propose a framework to measure uni-sectional and intersectional biases across models trained on sentiment analysis tasks. There is work that compares how linguistically fair across different languages are multilingual models [7]. Zhao and colleagues [44] analyze bias in multilingual word embeddings and create a dataset in four languages. Numerous studies have put forth multilingual datasets for a wide range of tasks. Another work introduces a template-based anti-reflexive bias challenge dataset for Danish, Swedish, Chinese, and Russian languages that all

[8]All participants were paid regardless of the quality check results.

have anti-reflexive gendered pronouns [11]. Shi and colleagues developed a benchmark dataset for arithmetic reasoning in 10 languages and showed that large pre-trained language models such as GPT3 are capable of performing multi-step reasoning across multiple languages [35]. There is the CrowS dataset of sentence pairs in English for measuring bias in masked language models [29] and its extension to French language [30].