Building NLP Evaluation Resources with LLMs and Community Engagement for Scale and Depth

Sunipa Dev
Google Research
sunipadev@google.com

Akshita Jha
Virginia Tech
akshitajha@vt.edu

Jaya Goyal
Circadian Connect
jaya@circadianconnect.com

Dinesh Tewari
Google Research
dineshtewari@google.com

Shachi Dave*
Google Research
shachi@google.com

Vinodkumar Prabhakaran*
Google Research
vinodkpg@google.com

Abstract

Measurements of fairness in NLP have been critiqued for lacking concrete definitions of biases or harms measured, and for perpetuating a singular, Western narrative of fairness globally. Current approaches to combat this issue through curation of resources face the significant challenge of achieving coverage over global cultures and perspectives at scale. In this paper, we demonstrate the utility and importance of complementary approaches that leverage both large generative models as well as community engagement in these curation strategies. We specifically target the harm of stereotyping and demonstrate a pathway to build a benchmark that covers stereotypes about diverse, and intersectional identities. We discuss the two approaches, their advantages and constraints, and the characteristics of the data they produce. We further discuss their potential to be used complementarily for better evaluation of stereotyping harms, in particular, for the African context.
WARNING: This paper contains examples of stereotypes that may be offensive.

1 Introduction

With the immense progress large language model capabilities [5, 24, 25, 6], the need for assessing their potential risks and harms to be contextually situated across global socio-cultural settings they are deployed has also been pointed out [22, 20]. This need in turn highlights the gaps in current evaluation paradigms, within which a vast majority of resources are in English, and/or is limited to a Western perspective of fairness and harms [15, 2]. This is especially troubling for evaluations that require socially situated benchmarks, for instance, to assess *stereotyping harms* that vary across cultures. Addressing this growing need for evaluation strategies to be more globally relevant has its own challenges. First, the scale of operation becomes massive, given how diverse different languages and cultures, and their associated axes of disparities are. Second, stereotypes can be locally situated; some stereotypes are prevalent only within a region and can be about people residing in it or outside it. Hence, a lack of involvement of some communities can result in major gaps in evaluations.

In this paper, we discuss the challenges in current evaluation paradigm, and then demonstrate how complementary approaches towards collecting stereotypes that target scale and depth can achieve greater coverage and address aforementioned challenges: our first approach involves generation of candidate stereotypes using large language models (LLMs), followed by human annotations to verify which associations are stereotypical; the second approach involves engaging with the communities with lived experiences of specific cultural contexts to collect the stereotypes known to them.

5th Deep Learning Indaba Conference (DLI 2023).

2 Complementary Approaches to Build Stereotype Resources

Stereotypes are generalizations about groups of people defined by their identity such as their gender, race, sexuality, age, etc. Stereotyping when propagated though language technologies can lead to many harmful outcomes including misrepresentation, targeted hateful speech generation, disparate access to resources, and opportunities [4, 9, 23]. There have been several efforts to build resources which document stereotypes in society [12], how they percolate into language technologies [17, 18, 2], and cause unfair model behavior [9, 14]. While existing stereotype resources are rich and enable model evaluations, most of them were collected by employing methods that rely on human annotations about statements describing a potential stereotype. However, stereotypes are not absolute, in that they vary by societies, communities, and individual experiences of people.

Large language models (LLMs) can be imagined as a lens on the society, since they are trained over copious amounts of naturally occurring, human-generated text that reflect the underlying societal context including social stereotypes. Their generations attempt to mimic human knowledge and predispositions, and has been shown to reproduce stereotypes [26, 9, 14]. Consequently, they can, inexpensively create generalizations that are diverse and representative of a wide range of identities across the globe [13, 15]. So we can tap into the generalizing capabilities of LLMs to create a broad-coverage candidate set for stereotypes. However, LLM generations are not always grounded factually, and reflect spurious correlations, and noise [1]. Hence, for usage as a stereotype resource, associations generated by LLMs about groups of people need to be validated for social presence of such stereotypes by human raters familiar with the corresponding socio-cultural contexts.

On the other hand, LLMs may not capture all social stereotypes globally. While they are trained on large amounts of data, there are still gaps in global representativeness in such data [6], which will also carry over to stereotype resources built using LLMs. Furthermore, since most state-of-the-art LLMs are trained on online data that has a Western lens [10], the stereotypes we get through LLMs may also reflect this Western gaze, and miss the nuances of stereotypes in local cultural contexts [15, 2]. Hence, it is important to complement the LLM-based approach with community engagements to build richer resources. Methods that rely on community engagement are expensive and time consuming but when used in targeted ways to understand one specific culture or society, they can provide depth and nuance to the collected stereotype resource.

In order to build a comprehensive stereotype benchmark, different strategies are warranted. Figure 1 imagines this juxtaposition of challenges and complementarity of community engaged and LLM generation based approaches. While LLMs provide a large set of generalizations they have learned, only a subset of them represent stereotypes that are widely held in societies. While human validation could help select those stereotypes, there is another large set of stereotypes that the LLMs may not have captured any associations about at all. In order to capture them, we will need to perform culturally situated community engagements with a diverse pool of participants.

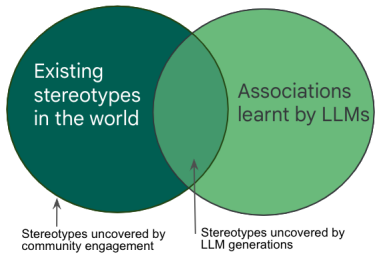

Figure 1: *Projected coverage of stereotypes uncovered by the approaches.* (Proportions of sets in image not to scale.)

3 Case Study

In this section, we summarize insights from two separate studies that take these complimentary approaches towards stereotype resource building, and outline their strengths and limitations. These studies have recently been published separately; in this paper, we discuss the methods only briefly, and focus on the insights that highlight the need for such complementary approaches. One approach crowd-sources stereotypes by engaging with communities [7], and the other uses generative models in conjunction with human annotations to scale coverage [11]. These complementary approaches can potentially be extended globally, to evaluate different types of sociotechnical harms such as hateful speech, toxic language and so on, which are also geo-culturally and socially situated.

3.1 LLM-based Stereotype Repository

Generative language models are powerful in learning from naturally occurring text and responding to prompts with text that is contextually meaningful. We prompt state-of-the-art language models PaLM [6] and GPT-3 [5] with stereotypes from existing datasets of stereotypes from NLP and social psychology literature [18, 17, 2, 21, 12]. The stereotypes selected for prompting were about global nationalities, and states in the United states and India. The prompts result in the models producing other such generalizations about geographical identities of persons, which are filtered and processed to obtain a candidate set. We then validated whether the associations in this candidate set are commonly known social stereotypes, for which we recruited annotators with diverse backgrounds (across gender) and geographic location that matches the associations.

Dataset: The resulting dataset contains about 7000 tuples, each with at least 3 human ratings whether the terms in the tuple represent a stereotype. Each tuple is consists of an *identity term* and an *attribute*. An *identity term* refers to a word or phrase that denotes a social group a person belongs to. An *attribute* refers to word(s)/phrase that describes a person or a group of people, such as adjectives or verbal predicates. Example stereotype tuples obtained using this approach include: (*Italian*, *gangsters*), (*Nigerian*, *scammers*), (*German*, *efficient*), to name a few. See [11] for more details about the dataset and the process followed.

3.2 Community Engagement based Stereotype Repository

Identities of persons can be intersectional, fine-grained, and also be more fluid than absolute categories. Additionally, each of these identities, associated generalizations and sentiments about them, and the potential harms they face from unfair technology is socially situated and differs by regions of the globe. Capturing these nuances require approaches that understand identities and stereotypes deeply for a given socio-cultural context, that may not be captured by the LLMs. We focus on India which yielded a large number of stereotypes in the LLM based approach. India is a country with 22 official languages, over 461 languages in use with many more dialects, 6 major religions, and many more such nuances which define individuals, their communities, and faced stereotypes. We employ an exploratory study design using surveys, distributed across 8 urban and suburban regions in India, which introduce the concept of stereotypes with examples of locally present stereotypes, followed by open-ended questions about what stereotypes the participant is aware of in their society. The stereotypes can be about any identity, or any combination of identities. For example, it can be about ethnic origin and caste such as 'Rajput', but also intersect with gender such as 'Rajput women'.

Dataset: The dataset created consists of about 2000 unique social stereotypes. In addition, it contains meta-data about how many persons with various identities (e.g., by gender, caste, and regional belonging) contributed the tuple as a stereotype. See [7] for more details about the dataset.

3.3 Complementary coverage and insights

The two approaches together yielded approximately 11,000 associations, with varying degrees of prevalence as social stereotypes. In this section we compare and contrast various aspects of tuples produced by both approaches.

Coverage of Identities: The LLM-based approach render the ability to scale up dataset creation many fold. In particular, the approach when restricted to generate for only region associated stereotypes, resulted in generation of candidate stereotype tuples for over 170 countries. This is 5 times the coverage of existing datasets such as StereoSet [17] and CrowS-Pairs [18]. In addition, it also contain stereotypes about states within India. Each identity term in this case is a demonym, restricted to countries and states. So, while the scale has been improved, the depth and granularity of identities understood is restricted. By engaging with communities in India, a larger number of identities, around 1000, are covered. These span demonyms, races, ethnicities, castes, religion, gender, sexuality, age, and more, including intersectional identities.

Coverage of Attributes: The LLM-based approach produced stereotype tuples, with over 10,000 different attributes. On the other hand, stereotypes collected by surveying communities contained about 2,000 distinct attributes. For both datasets, there is a substantial number of attribute terms that

are synonymous or alternate phrases for each other. While the absolute number of attributes produced does not directly imply richer stereotype data, diversity in attribute terms covered reflects indirectly on the diversity in the types of stereotypes about an identity that were uncovered.

Coverage of Stereotypes: Both approaches uncovered unique stereotypes with minimal overlap ($\leq$ 10 stereotypes). The LLM-based approach largely covered broad categories of demonyms, and yielded broad-strokes stereotypes such as 'Indian, vegetarian', while engagement with communities broke this stereotype down into smaller, more nuanced associations, such as 'Jain, vegetarian', where the identity is a religion category, 'Brahmin, vegetarians', where the identity term is an intersectional religion and caste category, and 'Punjabi, non-vegetarians', where the identity term is a state demonym. Furthermore, the generative approach hinges on the abilities of LLMs which in turn rely on their training data that is mostly in English and West-centric. Thus, stereotypes uncovered can sometimes have a Western perspective such as 'Indian, smelly', which was not present in the data produced through community engagement.

4 Discussion

In the paper, we presented two approaches to expand the coverage of stereotype resources used to evaluate language technologies. While we demonstrated the advantages of each individual method, it is also important to note how the complementary usage of the methods can lead to broad, and granular coverage of stereotype harms globally. Each method uncovered different kinds of stereotypes that were not found using the other. Additionally, the output of one method can serve as the seed for the other; the stereotypes recovered from engaging with communities can be used as prompts in subsequent usage of the generative approach using LLMs. Meanwhile, the generative approach highlights prevalence of associations and can help understand which communities to engage with for uncovering finer-grained stereotypes. Further, the collection of non-overlapping, complementary sets of stereotypes enhances coverage both in terms of global communities covered as well as fine-grained identities present in different regions. Measurements of harm in language tasks like question answering [14] and natural language inference [8] which are built on preferential associations with identities can leverage this more comprehensive list to make more holistic estimations.

Relevance to African NLP: Our work is inspired by the recent calls for a decolonial perspective towards AI [16, 3, 22], and in line with groundbreaking efforts within the African NLP community through participatory efforts such as the Masakhane project [19] that are aimed to improve existing NLP system capabilities in African langauges. However, as [2] points out, effective responsible AI interventions in these systems will require cross-cultural work that goes beyond cross-lingual capability building. Our work puts forth a set of complementary approaches towards building comprehensive fairness interventions in NLP systems equipped to operate in local contexts globally, including the African context. Crucially, this work aims to bridge the gap in societal understanding, in terms of stereotypes to test for and mitigate, that reflect the African societal context.

For instance, [11] points out that the stereotypes the LLMs detected about regional identities from Sub Saharan Africa had the highest likelihood for being offensive. If these associations were to be left unchecked, they risk being propagated and amplified through downstream applications of these generative models, such as search and conversational AI. While [7] is conducted within the Indian context, that study points out the importance of community insights in building more comprehensive resources. On that front, the successful community centered initiatives such as the Masakhane project provides promising avenues for building such effective and community-engaged benchmarks to put guardrails around African NLP systems.

Limitations: Stereotypes are subjective and socially situated. The absence of a stereotype in the lists collected by either approach does not imply that the stereotype does not exist in society or cannot be harmful to people. Any measurements built with these lists can still only make limited estimations, and more precautions should always be taken when deploying a model or tool with the specific use case at hand. Further, even with both approaches, we may not cover all possible regional identities and finer-grained examinations of stereotypes are possible. We also only work with English language text, and stereotypes written in English, and multi-lingual efforts are required to reflect some stereotypes present only within specific cultures.

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

[2] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Recontextualizing fairness in nlp: The case of india. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 727–740, 2022.

[3] Abeba Birhane. Algorithmic colonization of africa. *SCRIPTed*, 17:389, 2020.

[4] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[7] Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement, 2023.

[8] Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666, 2020.

[9] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only, November 2022. Association for Computational Linguistics.

[10] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Alex Koch, Nicolas Kervyn, Matthieu Kervyn, and Roland Imhoff. Studying the cognitive map of the u.s. states: Ideology and prosperity stereotypes predict interstate prejudice. *Social Psychological and Personality Science*, 9(5):530–538, 2018.

[13] Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. Araweat: Multidimensional analysis of biases in arabic word embeddings, 2020.

[14] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online, November 2020. Association for Computational Linguistics.

[15] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for hindi language representations. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), short*, 2022.

[16] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684, 2020.

[17] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, 2021.

[18] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 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)*, 2020.

[19] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, 2020.

[20] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence, 2022.

[21] Katherine H. Rogers and Dustin Wood. Accuracy of united states regional personality stereotypes. *Journal of Research in Personality*, 44(6):704–713, 2010.

[22] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Identifying sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2022.

[24] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

[25] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[26] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.