Introduction:

The paper addresses fairness challenges in NLP evaluations, proposing a dual methodology with Large Language Models (LLMs) and community engagement.

Methodology:

LLM-based Approach:
- Utilizes language models to generate and validate stereotypes.
- Emphasizes extensive coverage across 170+ countries.

Community Engagement:
- Directly collects stereotypes from diverse communities.
- Leverages surveys for nuanced, local perspectives.
- Enriches data with meta-information.

Challenges:

Global Scale:
- Recognizes the need for comprehensive coverage across diverse global languages and cultures.

Differential Prevalence:
- Acknowledges the variation of stereotypes across different regions.

Gaps in Evaluation:
- Highlights risks of excluding communities, leading to increased disparities in harms.

Western Perspective:
- Expresses concerns about benchmarks reflecting a Western gaze.

Insights and Results:

LLM-based Approach:
- Successfully probes language models, achieving extensive coverage.
- Validates stereotypes through human annotators.

Community Engagement:
- Utilizes surveys across 8 regions in India, collecting ~2000 unique stereotypes.
- Enriches the dataset with meta-information.

Future Work:

- Advocates for complementary use of LLMs and community engagement.

- Explores scalable community engagement methods, maintaining local relevance.
- Consider extension to other harms, such as hateful speech.
- Recognizes the importance of multilingual efforts for region-specific stereotypes.

Conclusion:

The paper underscores the significance of a dual approach for globally relevant stereotype evaluation, providing a foundation for addressing biases in language technologies and encouraging ongoing efforts toward inclusive NLP evaluations.