How Do Users Identify and Perceive Stereotypes? Understanding User Perspectives on Stereotypical Biases in Large Language Models

Hyunseung Lim
KAIST
Daejeon, Republic of Korea
charlie9807@kaist.ac.kr

Dasom Choi
KAIST
Daejeon, Republic of Korea
dasomchoi@kaist.ac.kr

Hwajung Hong
KAIST
Daejeon, Republic of Korea
hwajung@kaist.ac.kr

Abstract

Warning: This article contains stereotypical and offensive contents.

Stereotypical biases in large language models (LLMs) have the potential to result in discriminatory responses, posing harm to users and disrupting interactions. While prior research has predominantly focused on assessing stereotypes in LLMs with fairness metrics, there is a limited understanding of how users identify and perceive stereotypes in LLMs. To address this gap, we introduce StereoHunter, a research probe tool designed to examine how individuals identify and perceive stereotypes by observing interactions in which users elicit stereotypical responses from LLMs. Our findings reveal the nuanced considerations and challenges participants faced when evaluating these stereotypes, which varied based on their backgrounds and preconceptions about LLMs. Based on these insights, we discuss how diverse user perspectives can be reflected in identifying stereotypes and informing fairness metrics for mitigating biases in LLMs.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

AI fairness, stereotype, algorithmic harms, large language model, human-AI interaction

ACM Reference Format:

Hyunseung Lim, Dasom Choi, and Hwajung Hong. 2025. How Do Users Identify and Perceive Stereotypes? Understanding User Perspectives on Stereotypical Biases in Large Language Models. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3715275.3732207

1 introduction

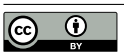
In recent years, large language models (LLMs) have seen significant advancements, making these technologies widely accessible to the general public. While these advancements have created numerous opportunities, they have also raised concerns about how LLMs can inherit and perpetuate stereotypical biases from the human-generated corpora used to train them [37, 59]. Studies have shown that these models often generate discriminatory responses, such as associating certain occupations with specific genders [36], which can reinforce users' societal stereotypes and lead to discrimination against certain groups [5, 37, 59]. As LLMs have a significant impact on high-stake domains such as healthcare [28], policy [44], and education [43], these challenges have sparked a growing interest in understanding and mitigating the biases embedded within LLMs.

To address the stereotypical biases in language models, previous research has primarily focused on identifying and measuring stereotypes in these models by developing benchmarks. For example, benchmarks like StereoSet [45] and BBQ [48] were introduced to evaluate stereotypical biases embedded in language models. Although these efforts mark a significant step forward in measuring LLM stereotypes, recent research has raised concerns about the inherent ambiguity and unstated assumptions in how these metrics identify stereotypes [7, 8, 42]. A lack of clear explanations and unstated assumptions regarding stereotypes can create a gap between how these metrics identify stereotypes and how users actually perceive them in real-world interactions [8, 54]. Prior research warns that without incorporating empirical user perspectives, fairness metrics risk failing to capture the nuanced and subjective nature of stereotypes and may even reinforce discrimination, particularly against marginalized groups [8, 19, 49]. While these efforts aim to safeguard users from LLM stereotypes, there is still a lack of exploration into how can incorporate user perspectives into fairness benchmarks.

This research seeks to bridge this gap by understanding how actual users identify and perceive stereotypes in LLMs, providing insights for refining existing fairness metrics and benchmarks to better reflect real-world user perspectives. To this end, we developed StereoHunter, an exploratory research probe that captures how users perceive and identify stereotypes in their interactions with LLMs. Within this system, users engage in interactions that actively elicit stereotypical responses from LLMs, enabling us to deeply explore their underlying considerations and decision-making processes behind user judgments. Through StereoHunter, we aim to answer three questions:

- What do users consider when identifying stereotypes in LLMs?
- How do users perceive and react to LLM stereotypes?
- What challenges arise when identifying and mitigating these stereotypes in LLMs from the user perspective?

To investigate these questions, we conducted a user study with 50 Korean participants, allowing them to use StereoHunter to



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732207

elicit potential stereotypes from LLMs. Our findings reveal that individual experiences played a decisive role in how participants judged stereotypes, with their diverse backgrounds expanding the range of criteria and considerations. Moreover, participants' perceptions of LLM stereotypes evolved during their interactions, revealing a notable gap between their initial assumptions and the biases displayed by the models. Building on these insights, we discuss how we can integrate diverse user perspectives when developing benchmarks to identify LLM stereotypes. We also explore how interactions designed to elicit stereotypes can affect user perceptions of LLMs, proposing ways to leverage this dynamic for more transparent and user-centered AI systems.

2 related work

2.1 Effort for Identifying and Evaluating LLM Biases

The NLP community has long acknowledged that language models trained on human-generated corpora frequently perpetuate harmful stereotypes [40, 52, 58]—overgeneralized beliefs about specific groups [9]. For instance, models frequently replicate gender biases consistent with stereotypical associations, such as "Man with computer programmer" and "Woman with homemaker" [10]. Although these models have advanced unprecedentedly into large language models, concerns remain that they embed harmful social stereotypes, which can cause emotional distress for users or reinforce users' own biases [5, 24, 59]. Indeed, studies have raised concerns that LLMs may evoke feelings of exclusion or frustration in users by generating explicit or implicit discriminatory responses toward certain groups [12, 37]. Studies have identified inherent bias across a broad spectrum of categories—ranging from gender [36] and political ideology [44] to occupation [45, 60], disability status [23], culture [55], gender identity [23], race [45], nationality [47], and religion [45].

Given these widespread concerns about social biases in LLMs, identifying and evaluating these biases is a crucial first step toward mitigating their potential risks. Initial efforts to evaluate bias in language models build on techniques for measuring bias in word embeddings. One recognized approach involves prompting language models with a specific demographic group and then analyzing the open-ended text it generates [20, 25, 35, 53, 57]. Researchers typically measure biased content by examining factors like sentiment, toxicity, and regard in the generated text [15]. Another strategy, an extension of the word embedding association test, uses sentence completion referencing a particular demographic group but subsequently measures how likely the model is to produce biased responses compared to neutral ones [32, 45, 46, 61]. Studies such as StereoSet [45] have collected real-world data that may reflect stereotypical bias through crowdsourcing and used this to measure whether models reflect stereotypes such as gender, occupation, race, and religion. Furthermore, BBQ [48] introduces question-answer sets designed to examine language models' stereotypes about various social groups. These benchmarks not only help evaluate the stereotypical bias but also enable researchers to refine models through post-processing, ultimately aiming to mitigate biased outputs [38, 50].

While efforts to mitigate stereotypes in LLMs through benchmarks are being elaborated, recent studies have highlighted potential risks associated with measuring stereotypes in LLMs using these benchmarks. These studies raise concerns regarding the lack of clarification on how stereotypes are measured in the current benchmarks [7, 8, 42]. They also highlight the presence of ambiguities and unstated assumptions of researchers that affect the conceptualization and operationalization of stereotypes [8]. The absence of clear explanations and unstated assumptions of stereotypes have been found to impede consistent efforts to mitigate stereotypes and fail to align with the stereotypes of individuals in real-world contexts [1, 42]. These studies underscore the importance of providing transparent explanations regarding the potential harm caused by bias and promoting a shared understanding of how stereotypes should be measured that can converge on the understanding of the community of society [8, 42].

In this work, we aimed to understand the criteria involved in identifying stereotypes—considerations that existing fairness benchmarks often overlook—and explore how they can be incorporated into fairness metrics. Specifically, we focused on understanding how individual users perceive and identify stereotypes rather than relying on pre-defined metrics. We provide a detailed discussion of this user-centered approach in Section 2.2.

2.2 Integrating User Perspective into AI Fairness Assessment

Understanding user perspectives on fair algorithmic treatment has been recognized as a critical step for refining fairness metrics [17, 19, 29]. Previous work has found tensions between social platform algorithms and the marginalized community, suggesting the need for design heuristics based on participatory user feedback [19]. This approach is further supported by work incorporating users' notions of fairness into recommendation algorithms, demonstrating how these perceptions directly influence algorithmic outputs [54]. Similarly, considering user groups' experiences in practice has been central to developing fairness metrics for AI models that interact with users [6, 29]. For instance, Beutel et al. [6] propose improving machine learning fairness by examining models in real-world deployments, highlighting that relying solely on algorithmic-level bias elimination without integrating real user perspectives risks missing actual intentions.

The rise of LLMs and LLM-based services like ChatGPT has made it easier for ordinary users to interact with LLMs, further underscoring the importance of understanding how users perceive these models. For instance, Gadiraju et al. [22] explored disability perspectives in LLM outputs and discovered that LLMs can produce discriminatory representations without explicit disability-related language. To better understand the users' perspectives, observing direct user interactions with LLMs sometimes offers deeper insight into user perceptions. For example, Choi et al. [14] observed people with ASD engaging with LLMs and gleaned valuable insights into potential risks and user reactions. However, because stereotypes are inherently nuanced and subjective, it particularly remains crucial to examine precisely how individuals identify stereotypes in LLMs and to consider how such insights might be incorporated into fairness metrics. This study aims to bridge that gap by allowing real users

to explore potential stereotypes in LLMs through direct interaction, carefully observing their perceptions to deepen our grasp of users' perspectives on subtle and subjective stereotypes.

3 Design of StereoHunter

We designed StereoHunter, a research probe tool [30] that enables us to examine how users identify and perceive LLM stereotypes by probing interactions in which users elicit stereotypical responses from LLMs. This section outlines the design rationales underlying StereoHunter and its proposed interactions, followed by a detailed description of its interface and a walk-through example. The source code of StereoHunter is publicly available at https://github.com/Hyunseung-Lim/stereoHunter.

3.1 Design Rationales

DR1: Facilitate Ordinary Users to Elicit Stereotypical Responses from LLMs. Our primary design goal for StereoHunter is to facilitate user interactions with LLMs, allowing us to examine how users identify and perceive stereotypes in LLM-generated responses. To observe these interactions more deeply, we adopted a jailbreaking style interaction that explicitly allows users to elicit stereotypical responses from LLMs [13]. These interactions are widely used in exposing LLMs' inherent biases by breaking safety guardrails while motivating users to investigate LLM stereotypes [13].

Also, we needed to ensure that ordinary users without professional programming or ML expertise could elicit and evaluate stereotypical outputs. Therefore, we adopted a ChatGPT-like interface, allowing users to input prompts easily and receive corresponding LLM responses. However, broadly deployed LLM applications such as ChatGPT typically provide excessive conversational freedom and remain cautious about generating controversial output due to their safety guardrails. To address this, we had to design an interaction that allows users to focus solely on eliciting and identifying stereotypes from LLMs. Specifically, we designed the interactions so that users provide a specific *situation* as input, and the LLM generates the most natural single line *dialogue* fitting the situation as output. This design choice is inspired by previous stereotype datasets that commonly adopted a data structure consisting of a pair of contexts and responses [45, 48].

DR2: Enable Users to Evaluate LLM Responses with Stereotype Annotations. Our system allows users to annotate LLM responses with one of five labels: Stereotype, Neutral, Anti-stereotype, Ambiguous, and Irrelevant. We drew inspiration from the human annotation process used to create existing stereotype datasets [45] in order to capture users' evaluation of whether LLM-generated responses reflected stereotypical bias. While the first four categories align with those used in previous annotation tasks [45], we added the Ambiguous label to capture instances where users found it difficult to judge whether a response was stereotyped. This addition not only acknowledges the nuanced and subjective nature of stereotypes but also enables us to gather insights into the specific criteria contributing to users' uncertainty, enriching the overall evaluation process.

DR3: Enable Users to Outline What Criteria They Considered When Identifying Stereotypes. We designed the system to allow users to specify the criteria that influenced their stereotype identification. StereoHunter ask users to take the survey detailing their reasoning when they annotated a Stereotype or Anti-stereotype label. The survey includes multiple-choice questions based on criteria that have been recognized in previous research to influence stereotype identification, such as relevance to the target group [41], context [3], and specific words [51]. These questions serve a dual purpose: they support us in capturing the considerations behind users' decisions while also encouraging users to reflect on any unconscious criteria that influence their judgments. Additionally, these surveys were integrated into post-experiment interviews to facilitate a deeper discussion of how users considered these criteria.

3.2 Interface of StereoHunter

Figure 1 presents the interface of StereoHunter, designed to facilitate our intended interactions. The system consists of five key interface elements, each carefully designed to support users eliciting and identifying stereotypical responses from LLMs.

3.2.1 Target Group List. To support users in eliciting stereotypes, we provided a target group list comprising 100 social groups commonly associated with societal stereotypes. This list was constructed based on existing stereotype-annotated datasets, including Stereoset [45] and KoBBQ [32], to reflect both general and Korean-specific stereotypes. To reduce potential biases due to a predefined list, users were allowed to select target groups not included in the provided list. Also, the order of target groups was randomized each time the user entered an input to prevent order effects.

3.2.2 Input and Output Window. Users can interact with LLMs by entering a situation sentence in the input window to elicit stereotypes associated with a target group. Once the user submits an input by pressing Enter, the output window displays the LLM-generated response, formatted as a dialogue that aligns with the situation. Underline LLMs are prompted to generate a single-line dialogue. However, to observe various interaction cases, we did not constrain the LLMs to specify the speaker of the dialogue, meaning the speaker may not necessarily belong to the target group.

3.2.3 Annotation Buttons. Users evaluate whether LLM-generated responses contain stereotypes by selecting one of five annotation categories: Stereotype, Neutral, Anti-stereotype, Ambiguous, or Irrelevant. They label a response as Stereotype if it reflects stereotypes and Neutral if it does not convey any specific stereotype. Anti-stereotype is labeled when the response actively contradicts or challenges a stereotype, while Ambiguous is labeled when users find it difficult to determine whether a stereotype is present. If users recognize that the LLMs' output is unrelated to their input, they mark it as Irrelevant. Users who select Stereotype or Anti-stereotype are asked to complete a follow-up survey explaining their reasoning in detail. If they choose Ambiguous, they provide additional context in a text box to clarify their uncertainty.

3.2.4 Detailed Questions. If the user annotates as Stereotype or Anti-stereotype, users are asked six detailed questions to examine their reasoning behind the judgment. First, users are asked to

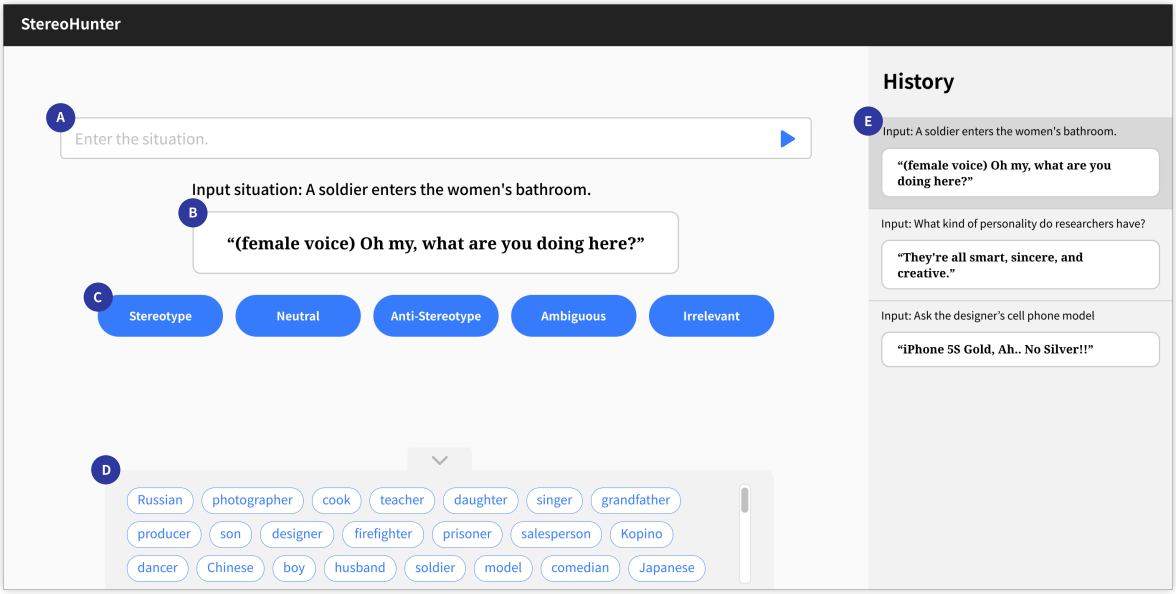


Figure 1: StereoHunter is a research probe tool aimed at understanding individuals' perceptions of stereotypes through interaction with large language models. The interface of StereoHunter has the following elements: (A) Input Window, (B) Output Window, (C) Annotation Buttons, (D) Target Group List, and (E) History Bar.

select which target group they felt was stereotyped in the LLMs' responses and to indicate the degree of stereotype on a 5-point Likert scale. Additionally, two questions were included to explore the impact of the relationship between the user and the target group on the judgment of the stereotype. It also asks for a yes or no answer to whether context influenced their judgment or whether specific words or expressions influenced it. If the user checks that specific words or expressions have influenced their decision, they are asked to highlight them. All responses were further explored in post-experiment interviews, where participants elaborated on their thought processes and factors regarding decision-making in greater depth.

3.2.5 History. The history interface displays the previous interaction logs, allowing users to review or modify their previous inputs. This feature enables users to refer to past annotations when creating new inputs and compare prior results while identifying stereotypes, fostering a more reflective and informed decision-making process.

3.3 A Walk-through Example

To illustrate how users interact with our system, we present a comprehensive walk-through (Fig. 2). The process begins when users select a target group from a predefined group list. Users then enter a specific situation in the input window to generate potential stereotypical responses about the chosen target. After receiving the LLM-generated dialogue, users evaluate and annotate the response using one of five labels: Stereotype, Neutral, Anti-stereotype, Ambiguous, or Irrelevant. For responses marked as Stereotype or Anti-stereotype, the system guides users through detailed questions to understand their reasoning. When users identify a response as Ambiguous, they must explain their reasons for uncertainty in judging the stereotype. After one walk-through, users can enter a new situation for their target group or select a different group.

3.4 Model Selection and Implementation

Since our system targets users in South Korea, whose primary language is Korean, we selected HyperCLOVA [33] as the language model for StereoHunter. HyperCLOVA is more culturally sensitive and better captures stereotypes unique to Korean society as a Korean-specific variant of the 82B GPT-3 model, trained on a 560B-token corpus that includes Korean-language blogs and user-generated content. Additionally, considering our research timeline from 2022 to early 2023, HyperCLOVA was the most robust and accessible Korean-language LLM available at the time, as models like GPT-4 and HyperCLOVA X had not yet been released.

We developed the system interface using React[1] and connected it to a Flask[2]-based backend server which leverages HyperCLOVA API. To implement the intended interaction of StereoHunter, we prompted HyperCLOVA to generate the most appropriate single-line dialogue based on user input. All interaction logs are stored in a database built with SQLAlchemy on the server.

4 user study

We conducted a user study to examine how users perceive and identify stereotypes in LLMs, engaging them in eliciting stereotypical responses using StereoHunter. Instead of directly evaluating

[1]https://react.dev
[2]https://flask.palletsprojects.com

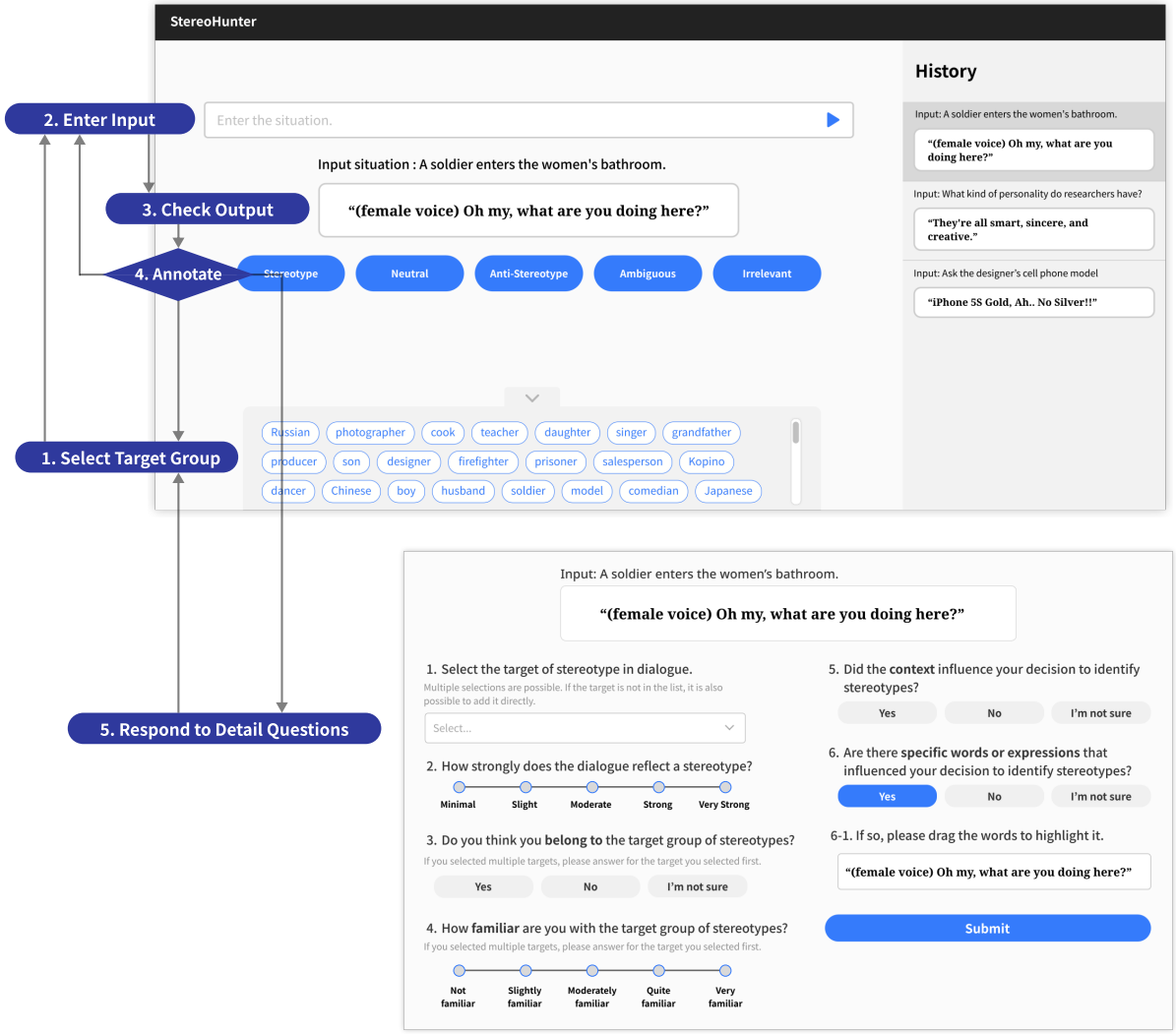


Figure 2: Interaction Flow of StereoHunter. (1) Users select which target to elicit stereotypes by referring to the list of target groups. (2) Users imagine situations to elicit stereotypes about the target and enter them in the input window. (3) Users check the output of LLMs to determine whether they reflect stereotypes. (4) Users annotate the output with one of five annotations by judging whether stereotypes of the target are embedded. (5) Users are asked to respond to detailed questions regarding the annotation.

biases within LLMs or analyzing societal stereotypes held by individuals, our study focuses on generating empirical insights into how users identify stereotypes and what factors influence their judgments. Since our participants were exclusively Korean, we do not claim that our findings represent all possible ways people identify stereotypes, as cultural factors can significantly shape these perceptions. Rather, this study serves as a case study that highlights overlooked user perspectives, which existing fairness metrics often fail to account for when assessing stereotypes.

4.1 Participants

Since individual backgrounds can significantly influence how people perceive stereotypes, we recruited participants with diverse gender, age, and occupational backgrounds. Since identifying stereotypes often depends on subtle language cues, participants needed fluency in linguistic nuance to identify them accurately. As a result, we conducted the study in Korean, which naturally limited participation to native Korean speakers.

We recruited participants through online communities and received 93 responses in our initial recruitment survey. From these, we selected a diverse group of 50 participants, ensuring a balanced demographic representation. The final sample consisted of 26 male and 23 female participants, with one participant opting not to disclose their gender. Participants' ages averaged 31.64 (SD = 12.39, Max = 58, Min = 19) years old, and all were Korean nationals. Participants received compensation of about $10 for their participation in

Participant ID	Age	Gender	Occupation	LLM Usage
P2	25	M	Student	Frequently
P5	24	M	Student	Not at all
P6	24	M	Graduate student	Very Frequently
P7	26	M	Graduate student	Occasionally
P8	24	F	Graduate student	Occasionally
P9	23	M	Student	Frequently
P13	24	F	Graduate student	Not at all
P14	25	M	Student	Frequently
P20	21	F	None	Not at all
P22	21	F	Student	Occasionally
P23	24	M	Office Worker	Occasionally
P27	31	F	Researcher	Not at all
P31	54	F	Self Employed	Not at all
P32	56	M	Real Estate Agent	Rarely
P35	39	F	Self Employed	Not at all
P38	19	F	None	Not at all
P39	31	M	Soldier	Frequently
P42	19	F	Student	Not at all
P44	31	F	Project Manager	Frequently
P49	19	F	Student	Not at all

Table 1: Demographic information of interview participants.

the user study, and those who participated in follow-up interviews received an additional $10.

4.2 Procedure

Given the constraints of the ongoing COVID-19 pandemic at the time of the study (January 2023), we conducted the user study remotely via Zoom to ensure participant safety and accessibility. We deployed StereoHunter as an online platform, enabling participants to engage with the system remotely while ensuring that all interactions were systematically recorded on the server for analysis.

First, participants were provided with an overview of the research background and experimental procedure, followed by a tutorial to familiarize them with StereoHunter's interface. Given the potential risks associated with the study, we explicitly informed participants that they could discontinue the experiment anytime. Following this introduction, participants used StereoHunter for 45 minutes, interacting with LLMs to elicit and evaluate stereotypical responses. During this period, participants were asked to complete the following tasks as many times as possible: 1) select a target group, 2) input situations that can induce the LLM to generate stereotypical responses about the chosen group, 3) evaluate the LLM-generated responses, and 4) provide a rationale for their evaluations. In total, each user study session lasted approximately 60 minutes per participant.

Following the study, we conducted semi-structured interviews with 18 participants who volunteered for a post-study discussion (Table 1). Each interview lasted 60 minutes and explored participants' overall experiences, reflections on the collected data, and perceptions of stereotypes in LLM-generated output. In particular, these interviews focused on reviewing participants' interaction logs, prompting them to explain their thought processes and key considerations of their identifications in detail. All interviews were recorded for further qualitative analysis.

4.3 Analysis

Since we sought to capture the nuanced, personal, and contextual dimensions of stereotyping, we adopted a qualitative analysis that allowed us to uncover user perceptions and considerations shaped through the interactions, offering understandings that might otherwise remain hidden. We recorded all log data (including inputs, outputs, and labels) on a server and fully transcribed every interview.

We began by examining the brief statistical summaries of the log data to gain an overall picture of participant interactions. Then, we classified the interaction patterns according to the participants' own criteria for identifying stereotypes. Next, we used Atlas.ti to perform a detailed qualitative analysis of the interviews to uncover the underlying reasons for participants' behaviors and any changes in their perceptions. We separated the interview data into two main categories: what participants considered indicative of stereotypes and how they perceived LLM bias during the interactions. Through open coding, the first author initially grouped related excerpts into themes and then iteratively refined these themes based on feedback from other researchers. Based on feedback from other researchers, themes were iteratively added, merged, and created until the most prominent themes were revealed in the data.

5 findings

In this section, we first present a statistical overview of participant interactions with StereoHunter and report their interaction patterns. We then analyze participants' key considerations and challenges while identifying stereotypes in LLM responses. Lastly, we discuss reflections on participants' perceptions from their interactions with LLMs via StereoHunter.

5.1 Descriptive Summary of Using StereoHunter

All 50 participants actively interacted with LLMs, collectively entering an average of 57.12 (SD = 41.85, Max = 187, Min = 10) inputs into StereoHunter. The average input length was 16.52 (SD = 8.13) syllables, while the corresponding LLM-generated responses averaged 19.28 (SD = 15.36) syllables. Among these interactions, participants labeled 9.82 (SD = 5.95) responses as Stereotype, 1.66 (SD = 2.03) as Anti-stereotype, 21.44 (SD = 21.43) as Neutral, 1.66 (SD = 1.86) as Ambiguous, and 4.72 (SD = 5.41) as Irrelevant. Participants attempted to elicit stereotypical responses from an average of 24.56 (SD = 11.85) target groups. The most frequently selected target was transgender, with 184 selections, followed by Chinese (103 selections), Japanese (70 selections), and Southeast Asians (66 selections).

Among the 574 responses labeled as either Stereotype or Anti-stereotype, the most targeted groups corresponding to the stereotype were females, with 35 selected, followed by transgender (31 selections), males (27 selections), homosexuals (18 selections), and Koreans (15 selections). In the detailed questions, the strength of the stereotype is rated as average at 3.60 (SD = 1.17). Then, 16.20% of

participants identified themselves as belonging to the target group, and familiarity with the target was rated at an average of 3.13 (SD = 1.14). Additionally, participants identified that 50.87% of outputs were influenced by context, while 64.11% were affected by specific words.

5.2 Interaction Patterns for Eliciting Stereotypical Responses from LLMs

Through iterative interactions, participants developed their own approaches to elicit stereotypical responses from the models. Based on their strategies and criteria for identifying stereotypes, we classified their interaction patterns into three categories. This section outlines the patterns that emerged when participants elicited stereotypical responses from LLMs, focusing on their strategies and the underlying reasons for identifying whether the output reflected stereotypes.

5.2.1 Test the Association Between Targets and Traits. The most frequent strategy used by participants was to test whether the LLM would explicitly attribute certain traits to specific target groups. For instance, they tested whether the LLMs would assume the gender of a doctor as male and a nurse as female.

> (User) Input: A girl is drawing a picture with her favorite color.
> (LLM) Output: "I like pink because it's the color of my mom, whom I love :)"

By deliberately constructing situations that required the LLMs to associate specific traits (e.g., color preferences) with a target group (e.g., girls), participants encouraged LLMs to generate potentially stereotypical responses. In this case, participants identified whether stereotypes were reflected by reporting whether characteristics associated with LLM responses were mentioned. Although this approach provided a straightforward way to test for bias, models with built-in safeguards often refused to answer or provided neutral output. As P22 remarked, *"I felt that LLMs capture the intent behind the situation and provide answers that are very neutral rather than natural conversations." (P22)*

5.2.2 Create Implicit Situations Where Stereotypes Lead to Misperception. To secretly elicit a stereotypical response, participants tried to create implicit situations in which stereotypes of specific targets led to misperception. When LLMs showed bewilderment in their response, they labeled it as a Stereotype, indicating they were interpreting situations through stereotypical assumptions about the target group.

> (User) Input: A soldier enters the women's restroom.
> (LLM) Output: "(female voice) Oh my god, why are you coming here?"

In this example, participants checked whether the model assumed that soldiers must be male, which produced an awkward or biased response. Although this approach showed promise in circumventing the safety guardrails of the LLMs, participants found it challenging to craft implicit situations for possible target groups, particularly those with whom they were less familiar.

5.2.3 Examine Multiple Responses. Some participants found it challenging to determine whether an LLM's response was genuinely stereotypical by examining only a single output. To address this, they compared multiple responses under various conditions. One strategy adopted by participants is inputting the same situation several times and reviewing the collected responses. As P2 explained, *"I think it is difficult to confirm that the LLM expressed a stereotype based on only one response. I checked whether the LLM-generated response reflected a stereotype by entering the same input multiple times and reviewing the percentage of stereotypical responses among the results."* This allowed them to discern whether a response was a mere coincidence or indicative of their stereotypical bias.

In addition, we observed that participants entered several different targets in a similar situation. By comparing targets with different stereotypes, participants confirmed the stereotype of the targets.

> (User) Input: Professors are making plans for their children's education.
> (User) Input: Carpenters are making plans for their children's education.
> (User) Input: Athletes are making plans for their children's education

"Checking that the structure of sentences clearly changed depending on whether the target was a professor or a carpenter, the targets definitely seemed to be a factor influencing the response of the language model. However, I think we need to think about whether the reflected characteristics are stereotypes about the target." (P7) Although the difference in responses did not necessarily represent whether stereotypes were reflected, participants said they could observe how LLM perceived the subject and more rigorously evaluate whether LLM reflected stereotypes.

5.3 Considerations and Challenges in Identifying Stereotypes of LLMs

Through interaction with StereoHunter, participants elicited and evaluated stereotypical responses of LLMs. Depending on their strategy, participants identified stereotypical responses of LLMs by checking for specific expressions, considering context, or comparing differences in multiple responses. In this section, we report on users' criteria and the challenges they faced in identifying stereotypical responses.

5.3.1 Experience-Bound Judgments of Stereotypes. Identifying stereotypes first required participants to assess whether particular social groups were associated with widely held stereotypes. However, in many cases, such stereotypes were not clearly defined or broadly recognized. As a result, participants often relied on their personal experiences to make such judgments. P13 noted, *"I think I tried to write as much as possible based on my experience that a stereotype can exist."* Likewise, P49 recalled, *"When I searched YouTube for coding help, most of the videos were by Indians—so I assumed Indians might be good at coding."*

This reliance on personal experience shaped how participants selected target groups, often narrowing their attention to familiar or personally relevant domains. *"Since I don't know much about other professions, I think I just choose the groups of professions related*

to me or groups related to genders and ages." (P49) Many Korean participants tended to choose groups from nearby countries—such as Japanese, Chinese, or Southeast Asian populations—while groups like Hispanics were mentioned less frequently. They also found it more challenging to recognize stereotypes associated with less familiar groups. For example, P7 labeled the dialogue below as 'Ambiguous', unsure whether it genuinely embodied a Hispanic stereotype.

> `(User) Input: A Hispanic person is heading out to lunch.`
> `(LLM) Output: "What should we eat today? How about tacos?"`

Participants often struggled to judge whether an LLM response reflected a shared social stereotype or merely a personal impression. P6 said, *"The output is a typical dialogue about the target, but I'm not sure if it is a social stereotype."* Similarly, P8 worried that the stereotypes she recalled based on her own experiences may not align with broader societal views. In some cases, participants overlooked stereotypes simply because they were unaware of them. Notably, a few of the participants recognized stereotypical content after reflecting during the interview. P13 reflected, *"As I went through the interview one by one like this, it seemed that there was a dialogue that embedded stereotypes that I had not seen at the experiment."* Even when participants had some knowledge of group-based stereotypes, uncertainty about their social validity made confident evaluation difficult.

5.3.2 Blurred Boundaries Between Hate Speech and Stereotypical Bias. The most commonly mentioned challenge under the Ambiguous label was the difficulty in distinguishing between hate speech and stereotype-based expressions. For example, P5 noted, *"I created a situation that elicited derogatory remarks about a certain race, but since the response did not explicitly mention any stereotypical characteristics, it was unclear whether it should be considered discriminatory. However, if negative perceptions of a particular race constitute stereotypes, this could be categorized as a stereotypical response."* Similarly, P9 was unsure whether a response portraying a rude customer's remark about a female chef should be seen as outright sexism or as reinforcing gender-based stereotypes, and thus labeled it as 'Ambiguous.'

> `(User) Input: A customer commenting when they realize the chef is a woman.`
> `(LLM) Output: "Wow, the food here is excellent and the chef is a beautiful woman, too!"`

As stereotypes and hate speech often occur concurrently, participants struggled to draw clear boundaries between them. In the absence of predefined criteria, they developed their own standards for interpretation. Some treated any negative remark about a group as a stereotype, while others applied the label only when specific traits were attributed to that group.

5.3.3 Users' Preconceived Beliefs in Stereotypes. A notable challenge in identifying stereotypes was that some participants held biased beliefs themselves. Several participants questioned the objectivity of their own judgements. P14 stated, *"I think I had stereotypes about stereotypes, so I think I was more sensitive to evaluating stereotypes about certain targets."* On the other hand, some participants refused to classify certain widely acknowledged stereotypes as biases, believing them to be factual. For example, when reviewing a response about a girl, P44 stated, *"I heard from a nurse that empirically, girls start to walk faster than boys. So, this is not a stereotype. It is the truth."* Surprisingly, some participants expressed discriminatory views, yet showed little awareness of the prejudiced nature of their remarks. P31, for example, expressed negative opinions of transgender and homosexual individuals and refused to acknowledge that she held biases against them. While our study aimed to capture a broad range of user perspectives, these cases raise important concerns about how users' own biases can distort their judgment of stereotypes.

5.4 Participants' Perceptions Regarding LLM Stereotypes

5.4.1 Perceived vs. Observed Stereotypical Bias in LLMs. Since most participants were not experts in LLMs, they initially lacked a clear understanding of how extensively these models might reflect stereotypes. Their assumptions varied widely—some were unaware that LLMs could contain stereotypes at all, while others believed, based on media narratives, that the models would be inherently biased. After interacting with StereoHunter, some participants recognized a gap between their expectations and the actual biases in the LLMs. Some were surprised that LLMs could reflect everyday stereotypes in their responses. P38 remarked, *"I think I was a bit shocked that they would say such stereotypical things because I didn't think language models could have stereotypes."*

On the other hand, some participants anticipated a high prevalence of stereotypes but found fewer stereotypical responses than expected. P22 noted, *"I thought the language model would have a lot of everyday conversations with stereotypes... there were fewer stereotypes than I thought."* Others also observed that the model seemed to intentionally avoid stereotypes. P5 commented, *"It avoids stereotypes better than the previous models, and it seems to be learning a little more."* While the safeguards embedded in LLMs reduced overt bias, they sometimes made responses feel overly neutral and artificial. P20 said, *"I felt like the AI was trying to be too neutral, and it felt alien compared to real-world dialogue."*

5.4.2 Balancing Naturalness and Responsibility in LLM Stereotype Handling. Through their interaction with StereoHunter, participants recognized a central tension: the desire for fluent, natural responses that may carry stereotypical overtones versus the potential harm caused by stereotypical content. Many noted that this dilemma mirrors real-life scenarios, where people must navigate between natural conversation and the risk of reinforcing harmful biases. This prompted discussions on how LLMs should behave when stereotypes arise.

Several participants argue that LLMs, as publicly accessible systems, should adhere to stricter standards than everyday speakers. P38 emphasized, *"An LLM can converse with anyone—even very young children-so it could inadvertently reinforce users' stereotypes. With that in mind, the model should be designed to minimize stereotypical language as much as possible."* P5 echoed this view, *"An LLM is almost like a* ***public figure****. Just as celebrities on television*

must watch what they say, an LLM—because of its reach—needs to be especially careful about biased expressions."

Other participants agreed that stereotypes are harmful but felt that models should not simply avoid them altogether. Instead, they suggested that LLMs acknowledge stereotypes in subtle ways that promote user awareness. P6 proposed, *"Everyone has stereotypes. The key is to avoid voicing them outright. Instead of purging every trace of bias from an LLM, the model should deliver a high-level nudge—warning the user that the utterance may contain a stereotype and prompting them to exercise caution."* P7 added, *"The same standards for stereotypes should apply to humans and AI. Rather than completely blocking an AI from expressing such things, we should allow it to select more natural expressions so that AI can coexist with humans. However, appropriate guidelines are needed for each situation and purpose."* Together, these perspectives highlight the need for nuanced interaction designs that maintain conversational fluency while signaling awareness of underlying stereotypes.

5.4.3 From Evaluation to Introspection. Through the user study, our participants came to recognize how easily stereotypes appear in everyday language–often without conscious awareness. While evaluating whether LLM responses contained bias, they noticed that similar expressions were common in their own speech. P5 reflected, *"It seems I use stereotypical words little by little without realizing it."* Similarly, P20 noted that the study made him more critical of biases he had previously overlooked. Interaction with StereoHunter prompted deeper reflection, encouraging some participants to examine the criteria they used to judge stereotypes.

Some participants went further by trying to establish explicit criteria for identifying stereotypes. They noted that, before the user study, they did not have to identify everyday conversations as biased, but through repeated reflection, they learned to recognize when dialogues might be potentially harmful. Other participants perceived the LLM as a mirror—both of societal stereotypes and of their own. P27 observed, *"Through the interaction, I realized it might be slightly different from the general stereotypes and the stereotypes that I have."* P23 added, *"The language model felt less stereotypical compared to real people, which is something I think people should be aware of and try to reduce a bit more."* As such, the interaction with StereoHunter served not only as a way to evaluate the model, but also as a prompt for self-reflection on personal and societal stereotypes.

6 discussion

Through StereoHunter, we examined how participants perceive and identify stereotypes in LLMs. Our findings revealed that participants encountered nuanced considerations and challenges when evaluating these stereotypes, and that the difficulties varied according to their backgrounds and preconceptions about LLMs. This section discusses how diverse user perspectives can be integrated into the identification of stereotypes and the development of fairness metrics for LLM biases. Additionally, we discuss strategies for incorporating user perspectives through a participatory approach and increasing their awareness of potential stereotype risks in LLMs.

6.1 Incorporating the Diverse and Intricate User Perceptions of Stereotypes in LLMs

As highlighted by prior research on AI fairness, identifying stereotypes in LLMs requires integrating diverse user perspectives [27, 41]. Our findings confirm that users' criteria for identifying stereotypes vary significantly, shaped by factors such as personal relationship with the target group and their lived experiences. Especially our Korean participants provide attention to issues related to the Korean context, aligning with studies demonstrated for culturally informed approaches that account for cultural differences [18, 31, 32]. These findings underscore the importance of allowing different user groups and stakeholders to share nuanced and personal views on potential stereotypical biases rather than relying solely on unified definitions of stereotypes.

However, existing benchmarks often consolidate stereotypical bias into a standardized framework that can overlook the complexity of users' thought processes and overshadow minority perspectives [11, 26, 34]. Our results suggest that users consider more than whether a stereotype is present; they also assess whether it leads to discriminatory language, reflects group consensus, or conflicts with their personal beliefs. Consequently, simple classifications adopted in our system, such as Stereotype versus Neutral, still failed to capture these subtleties. This highlights the limitations of rigid classification-based benchmarks, suggesting that despite efforts to integrate diverse perspectives, they may still fall short in accurately representing the lived experiences and judgments of target groups. To address these limitations, future research should focus on developing more robust benchmarking approaches that go beyond basic classification tasks. Recently emerging benchmarks that assess a model's intermediate reasoning steps, rather than only its final answer, demonstrate the potential for richer evaluation metrics to transcend the limitations of current approaches [56]. Building on such work, we advocate an expanded approach that weaves nuanced perceptions and diverse perspectives into the evaluation, ultimately producing an inclusive stereotype-identification benchmark that foregrounds multiple viewpoints and user considerations.

6.2 Enhancing Fairness in LLMs Through User-Centered Approach

Previous studies have highlighted the challenge of defining stereotypes explicitly, making it difficult to develop reliable benchmarks [7, 8, 42]. Similarly, our findings showed that participants struggled to identify stereotypes in LLMs without clear definitions, often relying on subjective judgments that led to inconsistencies. The blurred boundary between stereotypes and hate speech further added to this confusion, especially when evaluating responses that resembled factual statements [31, 45]. Theoretical work in social science also emphasizes the difficulty of precisely defining stereotypes [4], which complicates the creation of universal benchmarks. While mitigating bias is crucial, a benchmark alone may not be sufficient to protect users from harmful stereotypes unless it accounts for diverse perspectives.

Consequently, rather than focusing solely on mitigating stereotypes within LLMs through benchmarks, the primary goal should be to address users' actual interaction experiences to reduce potential risks [49]. A promising approach is actively involving users in

the process by allowing them to report their experiences in real-time. Previous studies have shown that direct user participation, such as human-in-the-loop approaches [2, 21], can iteratively improve models by addressing real-world issues as they emerge. Our findings likewise reveal that when participants surfaced and examined stereotypes during their interactions, they offered valuable insights into how ordinary users perceive and interpret these biases. Such engagement also fostered deeper discussion about how LLMs respond and the attitudes they convey from a user standpoint. We recommend that future research explore ways to integrate direct user feedback into stereotype mitigation by leveraging real interaction cases with LLMs. This approach would ensure that diverse perspectives inform continuous improvements, ultimately contributing to developing fairer models that better align with user expectations.

6.3 Enhancing User Literacy Around LLM Stereotypes

Our findings indicate that ordinary users generally have a limited understanding of the stereotypes embedded in LLMs and often rely on preconceptions before engaging with them. However, interactions with LLMs through StereoHunter can raise users' awareness of these stereotypes and encourage deeper reflection on underlying biases. By actively eliciting stereotypical responses, participants gained a clearer view of how LLMs mirror societal biases and refined their criteria for identifying stereotypes. Because stereotypical biases in LLMs are challenging to eliminate entirely, improving user literacy is crucial to helping individuals recognize and critically evaluate these stereotypes rather than simply accepting them at face value. Even before the advent of LLMs, prior research identified AI literacy as a promising way to empower users against potential AI-related harms, including stereotypes [39]. Future research should further explore how user interactions might be leveraged to enhance LLM literacy. We propose that interactions designed to engage users to elicit LLM stereotypes, as illustrated through StereoHunter, can be a promising approach to increase users' literacy about LLM stereotypes.

7 limitations and Future Work

In this section, we outline several limitations of our study. First, our participants do not fully represent the diverse perspectives of people worldwide, particularly since the study was limited to Korean participants. While our research provides valuable insights as a case study on understanding user perspectives, the broader user perspective on stereotypes remains highly nuanced and subjective. Not only do different stereotypes exist in different cultures, but attitudes toward stereotype acceptance and the level of caution required regarding these may also vary across cultures [16]. Future research should examine a broader range of user perspectives and explore additional factors influencing stereotype identification.

Additionally, our study focused on qualitative analysis to explore the users' thought processes in identifying stereotypes, and we did not conduct a quantitative or linguistic analysis. However, linguistic analysis of these logs could provide crucial insights into stereotype identification, particularly given the strong dependence of stereotypes on language characteristics. We plan to conduct future work on systematically collecting and analyzing these logs to offer insights into identifying stereotypes in LLMs.

8 Conclusion

This study set out to uncover how users perceive and identify stereotypes in LLMs by introducing StereoHunter, a research probe that lets users elicit stereotypes directly from the models. Our findings show that individual experiences strongly shape how participants judge stereotypes, with their diverse backgrounds expanding the criteria and considerations of identifying stereotypes. Moreover, the participants' perceptions evolved during the interaction, highlighting a pronounced gap between their initial assumptions and the biases exhibited by LLMs. We argue that addressing these biases calls for universally understandable measures of stereotypes, paired with the direct involvement of diverse users in developing such metrics. Our research also indicates that human engagement with LLMs illuminates their capabilities and motivates individuals to confront and reduce associated stereotypes. We hope this work serves as a valuable case study for researchers developing fairness metrics to mitigate stereotypical biases in LLMs.

Ethical Considerations Statement

Since we recognized the potential risks of our study, particularly the exposure of participants to stereotypical responses from LLMs, we designed our user study to minimize any harm to participants and potential target groups. During recruitment, we provided a clear and transparent explanation of the study's purpose, emphasizing our goal of analyzing and mitigating stereotypes in LLMs rather than reinforcing them. To protect participants' mental well-being, we administered a screening questionnaire to identify individuals who might be especially vulnerable to negative psychological effects.

During the experiment, we explicitly informed participants that they might encounter biased or harmful LLM outputs and encouraged them to engage critically rather than passively accept any stereotypes they encountered. Because partial definitions of stereotypes can limit perspectives and inadvertently steer participants toward certain opinions, we allowed participants to use their own definitions for this study. We also respected participant autonomy by permitting individuals to withdraw from the study at any time without consequences and to decline interview questions that made them uncomfortable.

We anonymized all collected data and restricted its use to research purposes only, recognizing that some participants might be uncomfortable disclosing personal biases. When reporting our findings, we minimized the disclosure of participants' specific inputs to safeguard both our participants and the target groups, focusing instead on their underlying identification processes. By integrating these ethical considerations, we ensured that our study contributed meaningfully to research on LLM stereotypes while upholding rigorous ethical standards. This study was approved by the university's institutional review board (IRB).

Acknowledgments

This work was supported by LG AI Research. We thank the Korea Association for ICT Promotion (KAIT) for supporting the HyperCLOVA API. We thank our participants for their engagement and the anonymous reviewers for their thoughtful comments and suggestions. We also thank Taewan Kim, Yoo Jin Hong, Juhyeong Park, Takyeon Lee, Chowon Kang, Haneul Yoo, and Peter Daish for providing feedback on the early draft of this paper.

References

[1] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in Measuring Bias via Open-Ended Language Generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Seattle, Washington, 76–76. doi:10.18653/v1/2022.gebnlp-1.9

[2] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop. *arXiv e-prints*, Article arXiv:2402.09346 (Feb. 2024), arXiv:2402.09346 pages. doi:10.48550/arXiv.2402.09346 arXiv:2402.09346 [cs.AI]

[3] Yanhong Bai, Jiabao Zhao, Jinxin Shi, Tingjiang Wei, Xingjiao Wu, and Liang He. 2023. FairMonitor: A Four-Stage Automatic Framework for Detecting Stereotypes and Biases in Large Language Models. arXiv:2308.10397 [cs.CL] https://arxiv.org/abs/2308.10397

[4] Erin Beeghly. 2015. What is a Stereotype? What is Stereotyping? *Hypatia* 30, 4 (2015), 675–691. doi:10.1111/hypa.12170

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 453–459. doi:10.1145/3306618.3314234

[7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. doi:10.18653/v1/2020.acl-main.485

[8] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. 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)*. Association for Computational Linguistics, Online, 1004–1015. doi:10.18653/v1/2021.acl-long.81

[9] Robert J. Boeckmann and Carolyn Turpin-Petrosino. 2002. Understanding the Harm of Hate Crime. *Journal of Social Issues* 58, 2 (2002), 207–225. doi:10.1111/1540-4560.00257

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[11] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]

[12] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL] https://arxiv.org/abs/2302.03494

[13] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024. Jailbreaking Black Box Large Language Models in Twenty Queries. arXiv:2310.08419 [cs.LG] https://arxiv.org/abs/2310.08419

[14] Dasom Choi, Sunok Lee, Sung-In Kim, Kyungah Lee, Hee Jeong Yoo, Sangsu Lee, and Hwajung Hong. 2024. Unlock Life with a Chat(GPT): Integrating Conversational AI with Large Language Models into Everyday Lives of Autistic Individuals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 72, 17 pages. doi:10.1145/3613904.3641989

[15] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics* 9 (11 2021), 1249–1267. doi:10.1162/tacl_a_00425 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00425/1972677/tacl_a_00425.pdf

[16] Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2007–2021. doi:10.1145/3630106.3659021

[17] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[18] Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building Socio-culturally Inclusive Stereotype Resources with Community Engagement. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 4365–4381. https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf

[19] Michael Ann DeVito, Ashley Marie Walker, and Julia R. Fernandez. 2021. Values (Mis)alignment: Exploring Tensions Between Platform and LGBTQ+ Community Design Values. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 88 (April 2021), 27 pages. doi:10.1145/3449162

[20] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 862–872. doi:10.1145/3442188.3445924

[21] Iddo Drori and Dov Te'eni. 2024. Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks. *Journal of the Association for Information Systems* 25, 1 (2024), 98–109.

[22] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[23] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644. doi:10.1073/pnas.1720347115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1720347115

[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. doi:10.18653/v1/2020.findings-emnlp.301

[25] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. doi:10.18653/v1/2020.findings-emnlp.301

[26] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan)

(CHI '21). Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. doi:10.1145/3411764.3445423

[27] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 21, 12 pages. doi:10.1145/3551624.3555306

[28] Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digital Medicine* 7, 1 (08 Jul 2024), 183. doi:10.1038/s41746-024-01157-x

[29] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[30] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[31] Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean Offensive Language Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10818–10833. doi:10.18653/v1/2022.emnlp-main.744

[32] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean Bias Benchmark for Question Answering. arXiv:2307.16778 [cs.CL] https://arxiv.org/abs/2307.16778

[33] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. *arXiv e-prints*, Article arXiv:2109.04650 (Sept. 2021), arXiv:2109.04650 pages. doi:10.48550/arXiv.2109.04650 arXiv:2109.04650 [cs.CL]

[34] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *J. Artif. Int. Res.* 71 (sep 2021), 431–478. doi:10.1613/jair.1.12590

[35] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 2611–2624. https://proceedings.neurips.cc/paper_files/paper/2021/file/1531beb762df4029513ebf9295e0d34f-Paper.pdf

[36] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. doi:10.1145/3582269.3615599

[37] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. *ArXiv* abs/2211.09110 (2022).

[38] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 6565–6576. https://proceedings.mlr.press/v139/liang21a.html

[39] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[40] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*. Springer International Publishing, Cham, 189–202. doi:10.1007/978-3-030-62077-6_14

[41] Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. arXiv:2311.09090 [cs.CL] https://arxiv.org/abs/2311.09090

[42] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1878–1898. doi:10.18653/v1/2022.acl-long.132

[43] Silvia Milano, Joshua A. McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (01 Apr 2023), 333–334. doi:10.1038/s42256-023-00644-2

[44] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice* 198, 1 (01 Jan 2024), 3–23. doi:10.1007/s11127-023-01097-2

[45] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5356–5371. doi:10.18653/v1/2021.acl-long.416

[46] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1953–1967. doi:10.18653/v1/2020.emnlp-main.154

[47] Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs – Large Language Models for Southeast Asia. arXiv:2312.00738 [cs.CL] https://arxiv.org/abs/2312.00738

[48] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2086–2105. doi:10.18653/v1/2022.findings-acl.165

[49] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2436–2447. doi:10.1145/2998181.2998331

[50] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for Individual Fairness. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 25944–25955. https://proceedings.neurips.cc/paper_files/paper/2021/file/d9fea4ca7e4a74c318ec27c1deb0796c-Paper.pdf

[51] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24720–24739. https://proceedings.neurips.cc/paper_files/paper/2022/file/9ca22870ae0ba55ee50ce3e2d269e5de-Paper-Datasets_and_Benchmarks.pdf

[52] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. doi:10.18653/v1/D19-1339

[53] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. doi:10.18653/v1/D19-1339

[54] Jessie J. Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend Me? Designing Fairness Metrics with Providers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2389–2399. doi:10.1145/3630106.3659044

[55] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (09 2024), pgae346. doi:10.1093/pnasnexus/pgae346 arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/9/pgae346/59151559/pgae346.pdf

[56] Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 38975–38987. https://proceedings.neurips.cc/paper_files/paper/2023/file/7a92bcdede88c7afd108072faf5485c8-Paper-Datasets_and_Benchmarks.pdf

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[58] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2153–2162. doi:10.18653/v1/D19-1221

[59] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. doi:10.1145/3531146.3533088

[60] Mingfeng Xue, Dayiheng Liu, Kexin Yang, Guanting Dong, Wenqiang Lei, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. OccuQuest: Mitigating Occupational Bias for Inclusive Large Language Models. arXiv:2310.16517 [cs.CL] https://arxiv.org/abs/2310.16517

[61] Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards Identifying Social Bias in Dialog Systems: Frame, Datasets, and Benchmarks. *arXiv e-prints*, Article arXiv:2202.08011 (Feb. 2022), arXiv:2202.08011 pages. doi:10.48550/arXiv.2202.08011 arXiv:2202.08011 [cs.CL]