# Beyond Demographics: Uncovering Latent User Groups for Fair Toxic Content Detection on LLMs

**Anonymous ACL submission**

## Abstract

Most fairness evaluations in Large Language Models (LLMs) focus on the disparities between groups with different demographic features, but these evaluations often ignore the impact of non-demographic features, resulting in overlooking groups with common dispositional tendencies. In response, we seek to understand the significance, reliability, and effect of non-demographic features in evaluating LLMs' fairness. We use three prior datasets with diverse non-demographic features to investigate these aspects. We cluster groups based on features' internal similarities and measure their difference in tagging positive class labels. Then, we ensure reliability by eliminating the impact of messages and validating groups' tendencies against prior studies. After that, we investigate whether LLMs can benefit these groups equally, which evaluates the LLMs' representative bias toward different groups. Our results suggest that non-demographic features can effectively cluster groups with notable different tendencies in tagging positive class labels, and LLMs do not provide equal benefits to these groups. Notably, despite the inequality, the downstream effects can significantly vary based on message types and group dispositional tendencies. Our findings call for consideration of both factors above in evaluating fairness, which is currently lacking in mainstream studies.

## 1 Introduction

LLMs can learn, perpetuate, and amplify harmful social biases. Previous studies have proven that LLMs can generate harmful content when no proper intervention is undergone (Gallegos et al., 2024). To be more specific, the bias stems from imbalanced power relationships during the constructing phase and under-representative demographic groups (Fleisig et al., 2023; Ferrara, 2023).

As a result, prior studies have investigated the impact of demographic features, such as gender, race, age, political tendencies and education, on identifying toxic language in LLMs (Beck et al., 2024; Haller et al., 2024). Furthermore, several techniques, data augmentation, loss function modification, and weight redistribution, have been introduced to mitigate bias (Qian et al., 2022; Woo et al., 2023; Orgad and Belinkov, 2023).

However, a few studies have proven that demographic-based solutions can inadvertently introduce bias (Cheng et al., 2023a; Deshpande et al., 2023). That is because these solutions usually represent stereotypical associations instead of being grounded in well-supported theories (Nangia et al., 2020; Nadeem et al., 2021).

At the same time, current studies suggested that individual subjectivity can be another critical factor in understanding the perception of harmful and toxic content (Plank, 2022; Sandri et al., 2023; Wan et al., 2023; Cabitza et al., 2023). This individual subjectivity is usually measured by non-demographic measurement scales, such as psychological measurement scales generally associated with psychological theories (Yao et al., 2024; Sap et al., 2022; Balakrishnan et al., 2020). Nevertheless, the effectiveness of these non-demographic features on the bias/fairness of LLMs has yet to be thoroughly investigated.

In response, this paper explores the complexity of imbalanced power dynamics by examining the role of non-demographic features in clustering social groups and identifying toxicity in LLMs. We are particularly interested in uncovering latent groups that meet any of these two conditions.

- Condition 1: Certain groups, clustered by non-demographic features, tag toxic language more or less frequently than other groups, or
- Condition 2: these groups receive different levels of benefits from LLMs

That is because the condition above may indicate that (1) these specific non-demographic features are crucial in tagging toxic language, or (2) people with these features are under-representative in LLMs.

Our main contributions are three-fold.

- First, our findings show that non-demographic features can uncover latent groups for fair detection of toxic content in LLMs. This revelation is substantiated by statistical significance in the Ratio of Positive Class (RPC) and the performance of LLMs across these groups (Section 4).
- Second, our experiments suggest that the imbalanced power relationship does not fully encapsulate fairness, as dispositional tendency also plays a vital role in perceiving harm. More specifically, a group's under or fair representation does not always correlate with the benefits received from the LLMs (Section 5).
- Third, we highlight a few critical gaps in current fairness assessments. By focusing solely on demographic attributes, current studies risk perpetuating harm toward groups that are seemingly demographically privileged but, in fact, vulnerable due to latent mental or contextual disadvantages (Section 6.4). Additionally, we emphasise the importance of carefully selecting datasets for fairness evaluation because these choices can significantly influence assessment results (Section 6.2).

## 2 Background

Studies have demonstrated that LLMs are not consistently effective in identifying toxic language (Kolla et al., 2024; Kruschwitz and Schmidhuber, 2024). For instance, Park et al. (2024) found that LLMs can generate near-zero response variation when dealing with diversity between individuals. Overlooking individuals' diversity in identifying toxic language on LLMs can lead to severe consequences (Cheng et al., 2023b; Gallegos et al., 2024).

Regarding dealing with individuals' diversity, some studies proposed demographic-based solutions for more fine-grained detection (Kocoń et al., 2021; Mishra et al., 2018); however, Hung et al. (2023) suggested that the improvement of the downstream performance gains from demographic features does not necessarily stem from demographic knowledge. As a result, these solutions raise concerns about introducing bias and stereotypes through demographic attributes (Cheng et al., 2023a; Deshpande et al., 2023).

A similar concern was raised in LLMs. Beck et al. (2024) evaluated the effectiveness of incorporating users' demographic attributes into LLMs-based detection systems and concluded that while there are potential benefits to using these attributes, they must be applied cautiously, as outcomes can significantly vary depending on the settings.

Lastly, recent studies call for a reevaluation of social group definitions, emphasizing that individuals possess intersectional identities that blend privileged and marginalised demographic attributes (Gallegos et al., 2024; Devinney et al., 2022). This perspective underscores the critical need to address intersectional biases in identifying toxic language (Ovalle et al., 2023; Lalor et al., 2022).

## 3 Our Proposal - Explore Latent Social Groups by Non-demographic Features

This paper concentrates on the impact of non-demographic features in forming social groups and identifying toxicity in LLMs. By employing a reverse engineering approach, we aim to uncover latent social groups that show distinct behaviours in tagging toxic languages or in the benefits received from LLMs. In other words, if groups with attribute A show different metrics than those with attribute B, then A and B represent two distinct latent groups regardless of demographic similarities/differences.

### 3.1 Difference Makes Groups

Most current studies concentrate on demographic groups, usually clustered by gender, race, education, or political tendencies. The underpinning assumption is that because these groups generally differ in social, historical, and political aspects, LLMs can introduce bias and stereotypes toward them. However, as discussed in the prior section, the social groups based on demographic features are not always reliable (Nangia et al., 2020; Nadeem et al., 2021; Beck et al., 2024).

In addition, we observed that the primary purpose of clustering social groups is to evaluate whether (1) certain groups are more or less reactive to messages than others or (2) LLMs provide equal benefits across groups, usually minor ones (Fleisig et al., 2023; Ferrara, 2023; Gallegos et al., 2024; Chu et al., 2024). In other words, the primary reason for clustering social groups is to evaluate performance disparities between them.

Based on the observation above, we suggest that if a group with specific non-demographic features meets any condition above, this group can also seem a latent group worth further investigation. That is because these conditions can strongly indicate (1) crucial non-demographic features or (2) a sign of an under-representation group in LLMs.

### 3.2 Research Question

The reverse engineering approach discovers latent social groups by identifying abnormalities in tagging toxic languages or in receiving benefits from LLMs. These results help clarify two questions.

**Question 1: Can non-demographic features be used to identify latent social groups?** This question directly stems from observing the significant impact of individual subjectivity in perceiving toxic language. In our experiment setting, this question will be measured by whether groups with specific non-demographic features find statistically more or less toxic language than other groups (see section 4).

**Question 2: Would LLMs provide equal benefits to these groups?** This research question mainly focuses on opportunities and representative bias that will be measured by True Positive Rate (TPR), Fairness Violation (FV), and Remaining Harm (RH) (see Section 5).

### 3.3 Research Method

Figure 1 outlines the overall method, which comprises two primary studies - creating groups with non-demographic features and evaluating fairness. Regarding creating groups, it creates groups with non-demographic features and manages the possible impacts of messages. First, we select datasets comprising non-demographic features, and each dataset is clustered into k groups using the K-means method based on the internal similarities of selected features (Section 4.1). After that, these groups are evaluated from three aspects : (i) Impact from messages—ensuring that differences between groups are due to selected features rather than the messages themselves; (ii) Significance—assessing group's differences in RPC through a statistic lens; and (iii) Reliability—validating analysis results against prior studies (Section 4.2).

Regarding elevating fairness, an experiment is conducted to investigate the difference in performance between a baseline and specific groups (Section 5). First, each selected dataset is evaluated by the LLMs to establish a baseline, representing the average benefit users can receive from the LLMs. From each data set, we select at least two groups that are statistically different from others in terms of their RPC. Then, we compare the difference in performance and benefits received by these selected groups against the baseline. The models are considered fair if LLMs provide equal benefits to baseline and selected groups. If there are discrepancies in the benefits, indicating unequal treatment, the LLMs are considered to exhibit fairness issues.

## 4 Study One: Can non-demographic features be used to identify latent social groups?

In this section, we created groups with non-demographic features and evaluated their significance and reliability.

### 4.1 Creation of Groups

This research uses three datasets, each concentrating on different non-demographic features, to help identify latent groups from various perspectives. The datasets and clustering processes are described below.

#### 4.1.1 Clustering Processes

All datasets undergo the same clustering processes. First, the K-means approach divides each dataset into K groups based on the internal similarities of selected features. We determined the final K using the Elbow method and Calinski-Harabasz indexes to ensure all groups are purely driven by internal similarity rather than subjective human input. Notably, all groups are distinct, without overlap between the data.

In addition, to assess the potential randomness in group differences, we generated two shuffled sets for each dataset. These shuffled sets maintain the same groups' numbers and sample size as the original but lack internal similarity within groups. This approach helps determine whether these observed differences among groups could be arbitrary.

Lastly, considering the varying features of selected datasets, such as the difference in the number of annotators and scoring schemes, specific pre-processing steps were applied to each dataset before clustering. The details of the pre-processing steps are further elaborated on in the respective sections of the study.

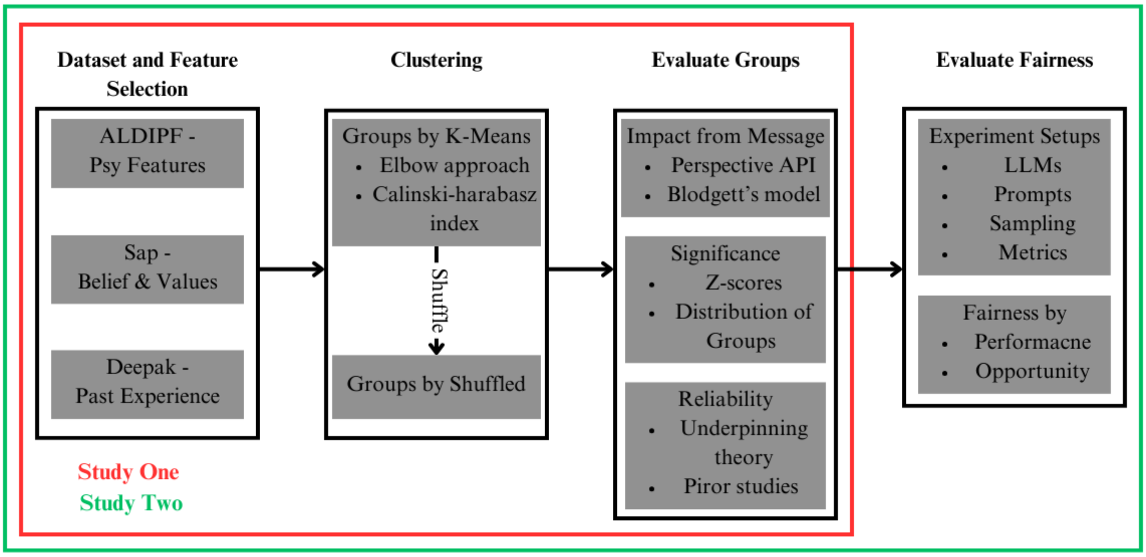


Figure 1: Research Method

#### 4.1.2 Selection of Datasets

We surveyed currently available datasets that (1) focus on toxic language detection, (2) contain non-demographic features [1], and (3) focus on readers' perceptions or opinions rather than imposing definitions. Furthermore, to address the varied nature of toxic language, we excluded datasets that were solely focused on racism, sexism, counter language, and sarcasm. After evaluation, three datasets met our criteria and were selected, as shown in Table 1. The description of each dataset is detailed in the following sections.

#### 4.1.3 Clustering Details

**Kumar Dataset.** The Kumar dataset comprises 500k annotated messages along with a wide range of demographic and non-demographic features collected from 17k annotators. We chose the "toxic score" column to define class labels, which reflect the users' feelings toward a message. The original "toxic score" column is presented on a 5-point scale (0 to 4), in which higher scores indicate greater perceived toxicity. For better comparability, we transferred 0 and 1 to Negative (not toxic) and the rest to Positive, as in prior studies.

Regarding clustering, we were particularly interested in features that may represent users' online exposure and cyberbullying experience. **Notably, to simplify the features, three features- using social media, video, and forums, were integrated into one synthetic column: exposure to social media**. Any positive answer in these features is marked as a positive in the synthetic column. As a result, four features were selected as follows: (1) considering toxic comments as a problem, (2) personally seen toxic content, (3) having personally been targeted, and (4) social media exposure.

The Kumar dataset was split into 19 groups (see Table 2), and the difference in the RPC among groups is significant on K-means sets. In contrast, no notable difference is observed in the shuffled sets (see Appendix 9).

**Sap Dataset.** The original Sap dataset contains 3.5k lines of data comprising a wide range of demographic and non-demographic features. Importantly, recognising the impact of demographic dialectal variation and anti-Balck meaning, the original dataset allocated toxic messages into three categories. This research only selected ONI messages, which were exclusively for vulgar messages. As a result, only 1k lines of data were selected for clustering. Additionally, the class labels were defined by the "to you" column, representing whether the user perceives a message as toxic. The original "to you" column was presented on a 5-point scale (1 to 5), in which the higher the number, the higher the toxicity. For better comparability, we transferred 1 and 2 to Negative (not toxic) and the rest to Positive, as in prior studies.

Regarding clustering, we were particularly interested in the individual's attitudes measured by a few different attitude scales (see Table 1) from prior established social science studies (Steg et al., 2014; Pulos et al., 2004; Bouchard Jr. and McGue,

[1]We selected only non-demographic features that intuitively impact individuals' perception for clustering.

| Dataset | Selected Non-demographic Features | Value Range | Selected N |
|---|---|---|---|
| Kumar et al. (2021) | 1. Toxic Comments Problem 2. Personally Seen Toxic Content 3. Personally Been Target 4. Exposure to Social Media (using social media, video, or forums) | 1. 5-Point Likert 2. Yes/No 3. Yes/No 4. Yes/No | 500k |
| Sap et al. (2022) | 1. Free of Speech 2. Harm of Hate Speech 3. Racist Beliefs 4. Traditionalism 5. Linguistic Purism 6. Empathy 7. Altruism | 1-7. 5-Point Likert | 1K |
| Yao et al. (2024) | 1. Other Down 2. Need for Achievement 3. Rationality 4. Need for Comfort 5. Self-Down 6. Need for Approval 7. Demand for Fairness 8. Irrationality | 1-2. 3 to 15 3-7. 4 to 20 8. 22 to 110 | 108k |

Table 1: Dataset Description

| Dataset | Group | Sample | RPC % | Z-score |
|---|---|---|---|---|
| Kumar | k1 | 9780 | 18.55 | -2.33 |
| | k2 | 98700 | 20.71 | -1.85 |
| | k18 | 10800 | 55.86 | 5.89 |
| | k19 | 2420 | 60.33 | 6.87 |
| | k20 | 960 | 62.08 | 7.259 |
| Sap | s1 | 168 | 59.52 | -0.53 |
| | s5 | 144 | 77.08 | 3.08 |
| Yao | y1 | 4527 | 15.63 | -3.85 |
| | y6 | 27480 | 37.33 | 0.74 |

Table 2: Clustering Results. Only the groups selected for further investigation are presented here, with the complete table available in Appendix B.

2003; McConahay, 1986; Cowan et al., 2002). Sap et al. (2022) observed a strong association between such attitudes and annotators' behaviour of tagging toxic messages.

Considering the limited number of 128 annotators, we split the Sap dataset into five groups (see Table 2). This is a trade-off between distinguishing internal similarity and ensuring a reasonable number of annotators for each group. Apart from an abnormal group (s5), no significant difference between k-means and shuffled sets (see Appendix 9).

**Yao Dataset.** The Yao dataset was created based on the ABC model (Ellis, 1991; Ellis and MacLaren, 1998; DiGiuseppe et al., 2018), which suggests that consequences (class labels) are co-created by triggers (messages) and individuals' psychological features. It contains 100k lines of data consisting of three eight psychological features from 505 annotators (see Table 2).

The eight psychological features were selected for clustering groups. These features originate from the Shortened General Attitude and Belief Scale (SGABS), which can measure one's attitudes and beliefs. They have been widely used in clinical settings to anticipate one's general well-being or differentiate particular groups of individuals from others (Ciarrochi and Bailey, 2009; DiGiuseppe et al., 2018; David et al., 2019; Owings et al., 2013).

The Yao dataset was split into six groups (see Table 2), and the difference in the RPC among groups is significant on the K-means set, while this difference is not observed in the shuffled sets (see Appendix 9).

### 4.2 Evaluation of Groups

As shown in Table 2, some groups tagged significantly more/fewer messages as a positive class; nevertheless, the same effect was not always observed in the shuffled groups. In this instance, we would like to clarify further (i) whether this disparity between groups stems from the difference in messages and (2) whether the disparity is significant and reliable.

#### 4.2.1 Impact From Messages

We want to ensure that group disparity derives from selected non-demographic features rather than messages. To evaluate the differences and impact of messages, we use (1) Perspective API to evaluate the overall toxicity of messages and (2) Blodgett et al. (2016) model to evaluate the demographic dialectal variation.

Perspective API has been widely used as a benchmark to evaluate the toxicity of messages. 100 samples were randomly selected from each group. Then, the Perspective API elevated these samples' toxicity, and each group's overall toxicity was presented in mean and standard deviation. Notably, prior studies have shown that Perspective API struggles with multilingual code-switching and demographic dialectal variations (Badjatiya et al., 2019; Lees et al., 2022). As a result, direct comparisons across different datasets may lead to misleading conclusions.

Figure 2 shows no notable difference in toxicity between groups for the Sap and the Yao datasets. Despite a more erratic curve for the Kumar dataset, the difference between groups was still negligible compared to the difference in the ratio of toxic language.

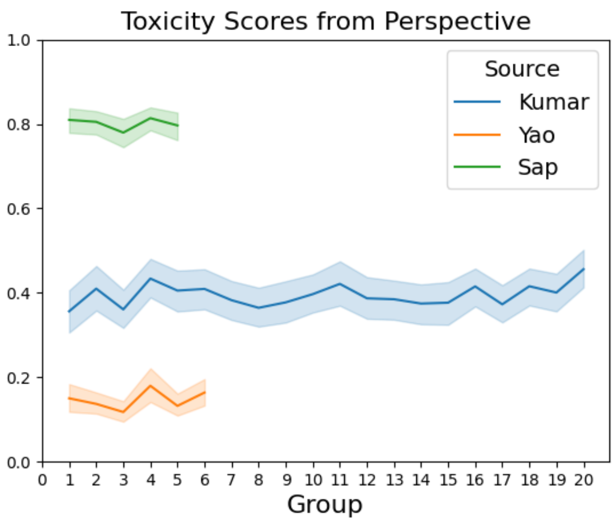


Figure 2: Toxicity Scores from Perspective API.

Blodgett's model specialises in evaluating demographic dialectal variation. Similarly, 100 samples were randomly selected from each group and fed to the Blodgett model. We focused on AAE and SAE due to their significant influence on toxicity evaluation (Davidson et al., 2019; Sap et al., 2019), and the results were presented in mean and standard deviation by dialectal types. Results suggested no notable difference between groups for all datasets.

Since no notable difference exists between groups' messages for the selected dataset, the disparity between groups can derive from members' subjectivity.

#### 4.2.2 Significance & Reliability

Significance focuses on whether these new groups created by the K-means are statistically different from others. The difference between groups is evaluated using the Z-score that shows the probability of a group with a ratio of harmful class occurring within a normal distribution. Additionally, the Z-score can evaluate how well the proposed method differentiates particular groups from others.

Z-score = $\frac{X-\mu}{\sigma}$, where $X$ is the RPC of a selected group, $\mu$ is the mean of RPC of the corresponding dataset, and $\sigma$ is the standard deviation of the RPC of the corresponding dataset. Results are shown in figure 3. Notably, the y-aix is normalised by $\frac{\text{Group Sample}}{\text{Total Sample}}$, which indicates the portion of a group.

Reliability concentrates on whether the differences between groups are aligned with prior studies' (1) observation and (2) explanation. We are particularly interested in factors that make groups tag significantly more/less toxic language than others.

**Kumar Groups.** The distribution of groups' Z-scores is erratic, and little groups sit between ± 1 standard deviation. The overall distribution is right-skewed with a long tail. Most users identify less toxic language than the average; nevertheless, most groups identify more toxic language than the average. Importantly, a few abnormal groups (see Table 2) identified significantly more/less toxic language, ranging from -2.33 to 6.89 Z-score values.

The analysis results (see Table 4 and 5) are generally aligned with the prior studies that suggest (1) previous experience with being targeted increases the RPC and (2) prior experience with witnessing toxic content decreases the RPC. Additionally, groups that have seen content and have never been personally targeted always tag fewer messages as a positive class than average. By contrast, the rest of the groups tag more messages as a positive class than average. Lastly, we noticed that when groups share the same experience, the feature *considering toxic comments as a problem* shows a weak negative correlation with RPC, which was not discussed in prior studies.

**Sap Groups.** Regarding the Z-scores, most of the groups are located between 0 and—0.5 standard deviation. The difference between most groups is negligible. In addition, the overall distribution is right-skewed. Most groups identify slightly less toxic language than the average. However, group 5 identified significantly more toxic language and impacted the average.

Prior studies are not well supported by analysis results regarding reliability (see Figure 6). First, despite a positive correlation between Empathy, Altruism, and labelling toxic language, this correlation seems weak in our experiment since there is no notable difference in these two aspects between

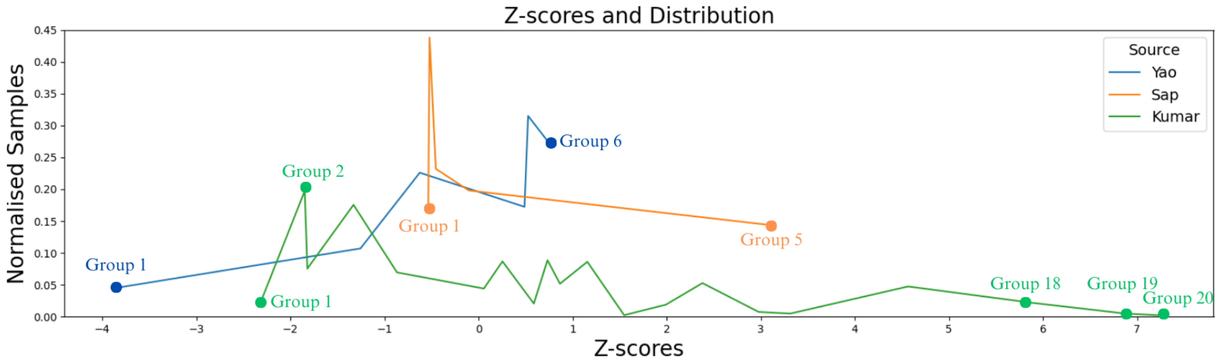


Figure 3: Z-scores and Distribution

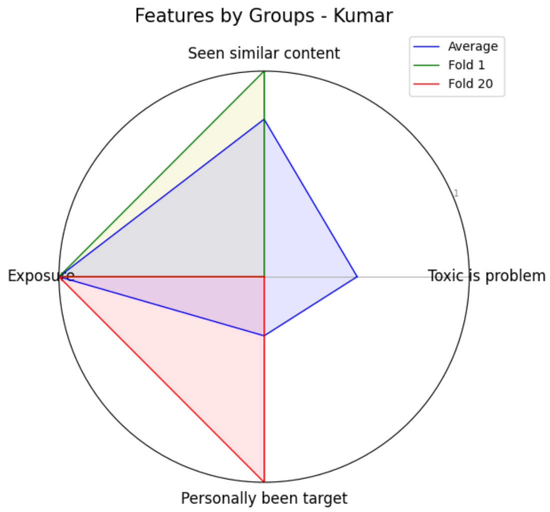


Figure 4: Features by Folds - Kumar. Note: the values are normalised.


Figure 5: Z-scores and Past Experience - Kumar. The numbers refer to the mean of the Z-scores of groups share the same experience.

group 5 and the average. Second, Free Speech and Racism are believed to have negatively correlated with labelling toxic language; however, there is no notable difference in these two aspects between group 5 and group 1.

**Yao Groups.** Regarding the Z-scores, most groups are located between ± 1 standard deviation. In addition, the overall distribution is left-skewed. Most groups identify more toxic language than the average. However, group 4 identified significantly less toxic language, almost reaching a -4 standard deviation.

Regarding reliability, the analysis results are generally supported by prior studies that suggest a positive correlation between Irrationality and labelling toxic languages (see Figure 7). In other words, people with lower Irrationality scores can be less reactive to toxic language and vice versa. Group 4 identified the less toxic language as having the lowest irrationality scores. By contrast, group 6 identified the more toxic language as having higher irrationality scores than average.

## 5 Study Two: Would LLMs provide equal benefits to these groups?

As discussed in the previous section, the disparity between some groups is statistically significant. Additionally, these disparities can derive from groups' non-demographic features that cause group members to possess certain tendencies because the difference in messages is negligible. As a result, this section would like to clarify further whether the LLMs can equally benefit these groups with these certain tendencies. In other words, we evaluate whether LLM's exhibit representational bias toward various groups.

### 5.1 Experiment Setups

**LLMs and Prompts**: The experiments were conducted on two out-of-the-box LLMs - GPT-3.5 Turbo and Llama3.1-70B with a temperature setting of 0. Additionally, our prompts follow the framework proposed by Eager and Brunton (2023), which divides prompts into a few essential components. Additionally, to make the most of the LLMs,

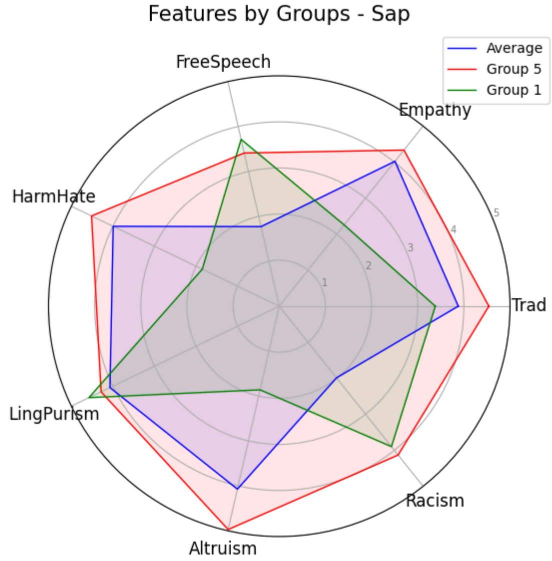


Figure 6: Features by Groups - Sap

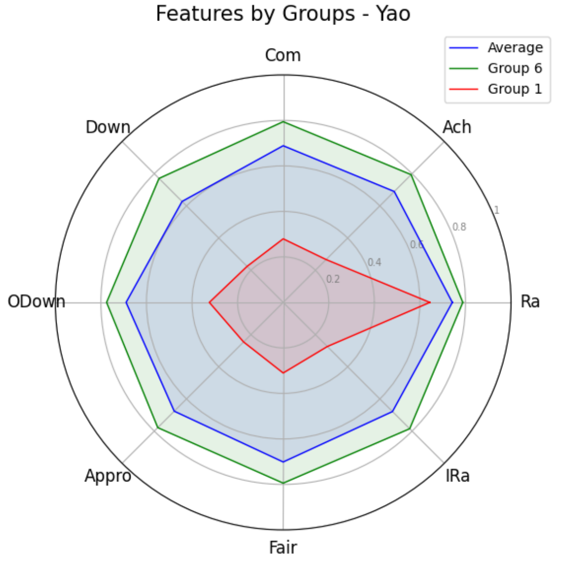


Figure 7: Features by Groups - Yao. Note: the values are normalised.

we applied the attribute prompt technique to co-create prompts with them (Yu et al., 2023). The complete prompt is provided in Appendix X, and the design of the prompts is as follows:

- The task with details: the LLMs are asked to determine whether a message is toxic based on a given definition. Importantly,
- Output: A score from 0 to 1, where 0 means absolutely not harmful, and 1 means definitely detrimental.

**Sampling Strategy**: Only the groups significantly different from others were selected for this fairness evaluation. Additionally, considering the difference in group size, we follow a rule of thumb that selects 10% of the data for sampling if this is more than 100 and does not exceed 1000.

**Evaluation Metrics**: Five metrics were selected due to their significance in measuring fairness between groups: Weighted F1 scores, Macro F1 scores, Accuracy, True Positive Rate (TPR), Fairness Violation (FV), and Remaining Harm (RH). The former three metrics are commonly used to measure a model's performance. The latter three represent the distinct benefits and drawbacks of implementing toxic language detection (Lalor et al., 2022; Liao and Naghizadeh, 2023).

FV = $\max_{g \in G_f} |\text{TPR}_g - \text{TPR}_D|$, where $G_\text{f}$ is the set of non-demographic groups for analysis (Yang et al., 2020). $TPR_\text{g}$ refers to the TPR of an LLM on the instance in $_g$, while $TPR_\text{D}$ indicates the overall TPR of an LLM on a dataset.

**Definition** (*Remaining Harm*). RH is a new metric that is introduced in this paper. It indicates the number of messages that can possibly harm users after being filtered by an LLM. The lower the RH, the more benefits an LLM can provide. In practice, RH refers to $\frac{\text{Number of False Negative}}{\text{Number of Total Samples}}$.

### 5.2 Results

Results are presented in Table 3 and 4. In most cases, there is no significant difference in performance and fairness between ChatGPT and Llamma. The following sections concentrate on the differences between groups within the same LLM.

#### 5.2.1 Fairness Evaluation for Kumar

LLMs do not benefit groups equally, and this dataset has the highest disparity between groups, reflected by its highest FV values. Regarding group differences, k2, which has the highest sample and tagged fewer messages as a positive class, benefited the most on every metric. By contrast, k20, which tagged the most messages as a positive class, benefited the least on every metric. k20 is nearly seven times more likely to be harmed than k2.

#### 5.2.2 Fairness Evaluation for Sap

LLMs do not benefit groups equally; nevertheless, the impact of this unfairness could be negligible because of the very few remaining harms ranging from 1 to 7%. Additionally, this dataset has the lowest FV among all. Regarding group differences, LLMs performed better on s5, which tagged the most messages as a positive class. However, s5 still has a higher RH than Group 1 due to their significant RPC.

#### 5.2.3 Fairness Evaluation for Yao

LLMs do not provide equal benefits to groups. Regarding group differences, y1, which tagged the lowest messages as a positive class, benefited more from the LLMs on every metric. Compared to y6, y1 has almost double the TPR and is nearly three times less likely to be harmed.

## 6 Discussion & Limitation

Based on our findings, we discuss the effectiveness of the proposed approach, its practical implications, and limitations in improving toxic language detection to serve groups with diverse non-demographic features better.

### 6.1 Can non-demographic features be used to identify latent social groups?

Our experiment provided empirical evidence suggesting that non-demographic features can uncover latent user groups that merit additional attention. These groups have significantly different RPCs and receive fewer benefits from LLMs that are not arbitrary. Additionally, the reliability of clustering groups is supported by the fact that it can generally explain the observed disparities between groups.

Nevertheless, the effectiveness of using non-demographic features for clustering can be hindered by a few factors. For instance, the choice of ideal K values and the size of the datasets. Despite using the Elbow method and Calinski-Harabasz indexes, the selection of ideal K can remain highly subjective, especially when the intersection of two indexes does not exist. Additionally, the dataset's size can significantly impact the choice of K, as discussed in Sap's dataset.

### 6.2 Would LLMs provide equal benefits to these groups?

LLMs do not provide equal benefits to groups with different non-demographic features. In other words, LLMs exhibit representational bias toward various groups.; nevertheless, the downstream effects of this inequality vary greatly across datasets. For instance, the disparity between Sap's groups is negligible after LLMs' filtering, whereas the gap between Yao's groups remains significant and almost unchanged after filtering.

Additionally, we noticed that the nature of the collected messages can significantly influence fairness evaluations. For instance, Sap's messages were pre-selected using Zhou et al. (2021) vulgarity list, which means most messages contain vulgar terms. This section contributes to its high TRP in the filtering phase. As a result, despite notable disparities in RPC, Sap's dataset exhibited the lowest RH between groups. In contrast, Yao's dataset, which mainly comprised contested messages, showed a lower TPR and the disparity between groups persisted after filtering. Notably, this finding does not align with Wiegand et al. (2019), who suggest that datasets with a higher proportion of explicit toxic messages could be more sensitive to bias.

These findings highlight an important consideration when selecting datasets for fairness evaluation. Like Sap's dataset, a crafted dataset can better focus on toxic language attributes, but using such a dataset for evaluation may overlook the diversity of real-world messages. On the other hand, a dataset without pre-selection may better reflect real-world scenarios but can contain noise that dilutes its effectiveness.

### 6.3 Fairness & Imbalanced Power Relationship

An imbalanced power relationship in participation does not entirely capture the nuances of fairness. The majority commonly receive more benefits than minorities, but this is not consistently observed. For instance, although Y1 is much smaller than Y6, Y1 received significantly more benefits. Similarly, despite K1 and k18 being similar in size, their benefits differed substantially. This result aligns with Cabello et al. (2023), who argue that bias and fairness are not always correlated.

We suggest that a group's dispositional tendency is another crucial yet underexplored factor in fairness evaluation. Dispositional tendency refers to stable traits influencing an individual's behaviour across various situations. Such tendencies may cause a small group to be more or less reactive to toxic language. Consequently, groups like y1 may still receive greater benefits despite their underrepresentation, whereas groups like k18 may receive fewer benefits, even with fair representation.

### 6.4 Latent Minority Group Within a Demographically Privileged Group

We noticed that some demographically privileged groups may actually be latent minority groups due to their mental or contextual disadvantages. For

| LLMs | Groups | RPC | Acc | WTD F1 | Macro F1 | RH* | TPR | FV* |
|---|---|---|---|---|---|---|---|---|
| ChatGPT | Overall | 0.3130 | 0.6570 | 0.6689 | 0.6370 | 0.1020 | 0.6741 | |
| | k1 | 0.1769 | 0.6217 | 0.6664 | 0.5569 | 0.0573 | 0.6763 | |
| | k2 | 0.2070 | 0.6760 | 0.7063 | 0.6284 | 0.0480 | 0.7681 | |
| | k18 | 0.5330 | 0.5440 | 0.5442 | 0.5440 | 0.2580 | 0.5159 | |
| | k19 | 0.5828 | 0.5148 | 0.5173 | 0.5140 | 0.3047 | 0.4772 | |
| | k20 | 0.6970 | 0.4737 | 0.4868 | 0.4708 | 0.4316 | 0.3881 | 0.2860 |
| Llamma | Overall | 0.3130 | 0.6420 | 0.6550 | 0.6268 | 0.0930 | 0.7029 | |
| | k1 | 0.1769 | 0.6227 | 0.6674 | 0.5616 | 0.0521 | 0.7052 | |
| | k2 | 0.2070 | 0.6630 | 0.6949 | 0.6204 | 0.0430 | 0.7923 | |
| | k18 | 0.5330 | 0.5320 | 0.5323 | 0.5319 | 0.2610 | 0.5103 | |
| | k19 | 0.5828 | 0.5148 | 0.5181 | 0.5114 | 0.2840 | 0.5127 | |
| | k20 | 0.6970 | 0.4949 | 0.5148 | 0.4796 | 0.3636 | 0.4783 | 0.2246 |

Table 3: Results for Fairness Evaluation - Kumar. Note 1: * indicates that lower values correspond to better performance. Note 2: This table presents the sample's RPC rather than the group's original PRC.

instance, for k18, k19 and k20 groups, most users display privileged demographic traits: male, white, non-transgender, hold an undergraduate degree, and are between 25-34 years old. Additionally, 7.1 % of users in these groups possess all these privileged traits, higher than the average of 5.9%. However, despite these demographic advantages, members of this group exhibit heightened sensitivity to toxic language and fewer benefits from LLMs, suggesting that factors beyond demographics, such as experience with cyberbullying, play a significant role in shaping their perceptions.

This finding highlights a critical gap in fairness assessments based solely on demographic features, as these features do not always capture a group's mental or contextual disadvantages. Consequently, some demographically privileged groups, which are, in fact, minority groups in other respects, may be overlooked and excluded from fairness evaluations.

### 6.5 Limitations

Although this paper aims to uncover latent groups, the k-means method may overlook groups that comprise a smaller portion of the population. Since k-means clusters are based on the internal similarity of majority patterns, smaller or less represented groups may be treated as noise and ignored. Additionally, the k-means method cannot explain the interactions between selected features, as it only identifies general tendencies among groups. Further steps are required to assess the reliability and significance of these compound factors.

## 7 Conclusion

This paper demonstrates that non-demographic features can effectively uncover groups that may confront inequality when interacting with LLMs. The proposed "difference makes groups" approach can help identify and support at-risk users more accurately. Importantly, this paper challenges the prevailing notion of fairness that 'with demographic features in hand, everything is sexism and racism.' However, we are not downplaying the importance of demographic features; rather, we aim to highlight the often-overlooked majority demographic groups who may be mentally or contextually disadvantaged. Lastly, we suggest that fairness evaluations incorporate more nuanced psychological and contextual factors for more LLMs to capture the full spectrum of disadvantages.

## References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, WWW '19, page 49–59, New York, NY, USA. Association for Computing Machinery.

Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia. 2020. Improving cyberbullying detection using twitter users' psychological features and machine learning. *Computers & Security*, 90:101710.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Compu-*

| LLMs | Groups | RPC | Acc | WTD F1 | Macro F1 | RH* | TPR | FV* |
|---|---|---|---|---|---|---|---|---|
| ChatGPT | Overall | 0.6333 | 0.7750 | 0.7500 | 0.7152 | 0.0167 | 0.9737 | |
| | s1 | 0.5200 | 0.6900 | 0.6613 | 0.6570 | 0.0200 | 0.9615 | |
| | s5 | 0.7800 | 0.8300 | 0.8176 | 0.7181 | 0.0500 | 0.9359 | 0.0378 |
| Llamma | Overall | 0.6333 | 0.7500 | 0.7202 | 0.6804 | 0.0250 | 0.9605 | |
| | s1 | 0.5200 | 0.7100 | 0.6831 | 0.6792 | 0.0100 | 0.9808 | |
| | s5 | 0.7800 | 0.7900 | 0.7746 | 0.6518 | 0.0700 | 0.9103 | 0.0502 |
| ChatGPT | Overall | 0.3440 | 0.6310 | 0.5948 | 0.5247 | 0.2650 | 0.2297 | |
| | y1 | 0.1881 | 0.7257 | 0.7380 | 0.5931 | 0.1106 | 0.4118 | 0.1821 |
| | y6 | 0.3700 | 0.5910 | 0.5526 | 0.4955 | 0.2920 | 0.2108 | |
| Llamma | Overall | 0.3440 | 0.6520 | 0.5887 | 0.5042 | 0.2910 | 0.1541 | |
| | y1 | 0.1881 | 0.7633 | 0.7555 | 0.5876 | 0.1327 | 0.2941 | 0.1400 |
| | y6 | 0.3700 | 0.5800 | 0.5087 | 0.4339 | 0.3340 | 0.0973 | |

Table 4: Results for Fairness Evaluation - Sap & Yao. (Cont).


*tational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Thomas J. Bouchard Jr. and Matt McGue. 2003. Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, 54(1):4–45.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 370–378, New York, NY, USA. Association for Computing Machinery.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1504–1532. Association for Computational Linguistics, Stanford University.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023b. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *SIGKDD Explor. Newsl.*, 26(1):34–48.

Joseph Ciarrochi and A. Bailey. 2009. *A CBT-practitioner's Guide to ACT: How to Bridge the Gap between Cognitive Behavioral Therapy and Acceptance and Commitment Therapy*, volume 50.

Gloria Cowan, Miriam Resendez, Elizabeth Marshall, and Ryan Quist. 2002. Hate speech and constitutional protection: Priming values of equality and freedom. *Journal of Social Issues*, 58(2):247–263.

Daniel O. David, Raymond DiGiuseppe, Anca Dobrean, Costina Ruxandra Păsărelu, and Robert Balazsi. 2019. *The Measurement of Irrationality and Rationality*, pages 79–100. Springer International Publishing, Cham.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Raymond DiGiuseppe, Russell Leaf, Bernard Gorman, and Mitchell W. Robin. 2018. The development of a measure of irrational/rational beliefs. *Journal* *of Rational-Emotive & Cognitive-Behavior Therapy*, 36(1):47–79.

B. Eager and R. Brunton. 2023. Prompting higher education towards ai-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, 20(5).

Albert Ellis. 1991. The revised abc's of rational-emotive therapy (ret). *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, 9(3):139–172.

Albert Ellis and Catharine MacLaren. 1998. *Rational emotive behavior therapy: A therapist's guide*. The practical therapist series. Impact Publishers, Atascadero, CA, US.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. Available at SSRN: `https://ssrn.com/abstract=4614228` or `http://dx.doi.org/10.2139/ssrn.4614228`.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pages 1–83.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2024. OpinionGPT: Modelling explicit biases in instruction-tuned LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86, Mexico City, Mexico. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Udo Kruschwitz and Maximilian Schmidhuber. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA. USENIX Association.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Yiqiao Liao and Parinaz Naghizadeh. 2023. Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8764–8772.

John B. McConahay. 1986. Modern racism, ambivalence, and the modern racism scale. In John F. Dovidio and Samuel L. Gaertner, editors, *Prejudice, discrimination, and racism*, pages 91–125. Academic Press.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 496–511, New York, NY, USA. Association for Computing Machinery.

Larry R. Owings, Gregory L. Thorpe, Evan S. McMillan, Ronald D. Burrows, Scott T. Sigmon, and Dawn C. Alley. 2013. Scaling irrational beliefs in the general attitude and belief scale: An analysis using item response theory methodology. *SAGE Open*, 3(2).

P.S. Park, P. Schoenegger, and C. Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Steven Pulos, Jeff Elison, and Randy Lennon. 2004. The hierarchical structure of the interpersonal reactivity index. *Social Behavior and Personality: An International Journal*, 32(4):355–359.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Linda Steg, Goda Perlaviciute, Ellen van der Werff, and Judith Lurvink. 2014. The significance of hedonic values for environmentally relevant attitudes, preferences, and actions. *Environment and Behavior*, 46(2):163–192.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc.

Tsungcheng Yao, Sebastian Binnewies, Ernest Foo, and Masoumeh Alavi. 2024. See the words through my eyes: The role of personality traits in abusive language detection. *SSRN Electronic Journal*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European*

*Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

| Group | Messages | RPC | Z-score |
|---|---|---|---|
| 1 | 9780 | 18.55 | -2.33 |
| 2 | 98700 | 20.71 | -1.85 |
| 3 | 37660 | 20.84 | -1.82 |
| 4 | 87760 | 23.07 | -1.33 |
| 5 | 34700 | 25.17 | -0.87 |
| 6 | 22020 | 29.36 | 0.056 |
| 7 | 43360 | 30.26 | 0.254 |
| 8 | 10240 | 31.78 | 0.587 |
| 9 | 44280 | 32.44 | 0.732 |
| 10 | 25820 | 33.04 | 0.865 |
| 11 | 43020 | 34.36 | 1.156 |
| 12 | 1140 | 36.14 | 1.548 |
| 13 | 9380 | 38.16 | 1.991 |
| 14 | 26320 | 39.92 | 2.38 |
| 15 | 3680 | 42.64 | 2.978 |
| 16 | 2400 | 44.17 | 3.315 |
| 17 | 23660 | 49.85 | 4.566 |
| 18 | 10800 | 55.86 | 5.889 |
| 19 | 2420 | 60.33 | 6.873 |
| 20 | 960 | 62.08 | 7.259 |

Table 5: Cluster Results by Group - Kumar

## A Details for Clustering Process

The final K was determined using the Elbow method and Calinski-Harabasz indexes (see Figure 8).To be more specific, the elbow method was done using *kmeans.fit*, which focuses on the Sum of squared distances. The calinski-Harabasz index was done using *calinski_harabasz_score*, concentrating on the sum of between-cluster dispersion and within-cluster dispersion. Additionally, features' normalisation was done *StandardScaler*. Lastly, the K-means was done by using *kmeans.fit_predict*,

## B Cluster Results by Group

Table 5, 6, and 7 present the cluster results for each dataset by group. Notably, due to the smaller number of contributors in Sap's dataset, which may impact the generalisability of the results, an additional column is included to highlight this concern.

## C RPC by Data Set

Figure 9 presents the difference in the ratio of positive class (RPC) between k-means and shuffled sets. This plot treats each group with the same weight.

| Group | Messages | RPC | Z-score | Ctr n=132 |
|---|---|---|---|---|
| 1 | 168 | 59.5238 | -0.5352 | 19 |
| 2 | 438 | 59.589 | -0.5218 | 51 |
| 3 | 232 | 59.9138 | -0.4548 | 27 |
| 4 | 198 | 61.6162 | -0.1039 | 22 |
| 5 | 144 | 77.0833 | 3.0847 | 13 |

Table 6: Cluster Results by Group - Sap

| Group | Messages | RPC | Z-score |
|---|---|---|---|
| 1 | 4527 | 15.6395 | -3.8465 |
| 2 | 10700 | 27.8972 | -1.256 |
| 3 | 22590 | 30.8809 | -0.6254 |
| 4 | 17234 | 36.1495 | 0.4881 |
| 5 | 31496 | 36.3316 | 0.5266 |
| 6 | 27480 | 37.329 | 0.7374 |

Table 7: Cluster Results by Group - Yao

## D Demographic Dialectal Variation

Figure 10 shows the results of demographic dialectal variation based on Blodgett et al. (2016) model.

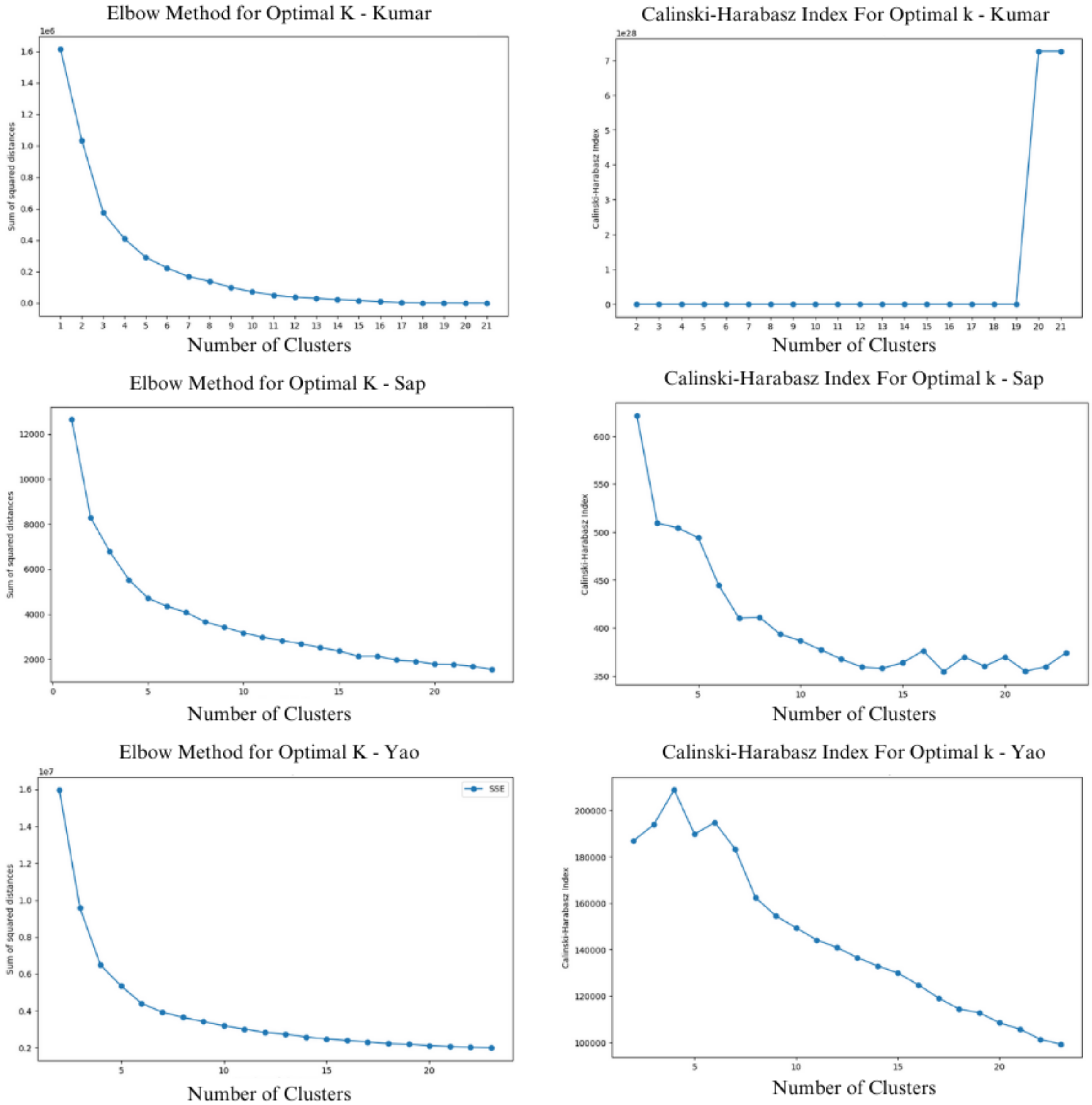


Figure 8: Elbow method and Calinski-Harabasz indexes for Clustering Process.

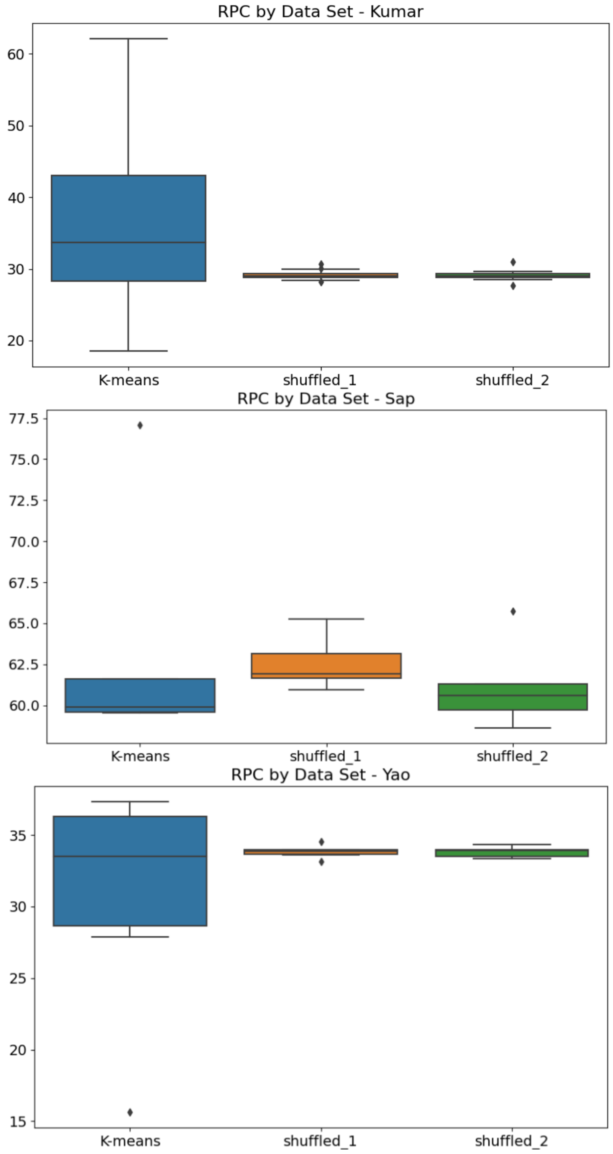


Figure 9: RPC by Data Set

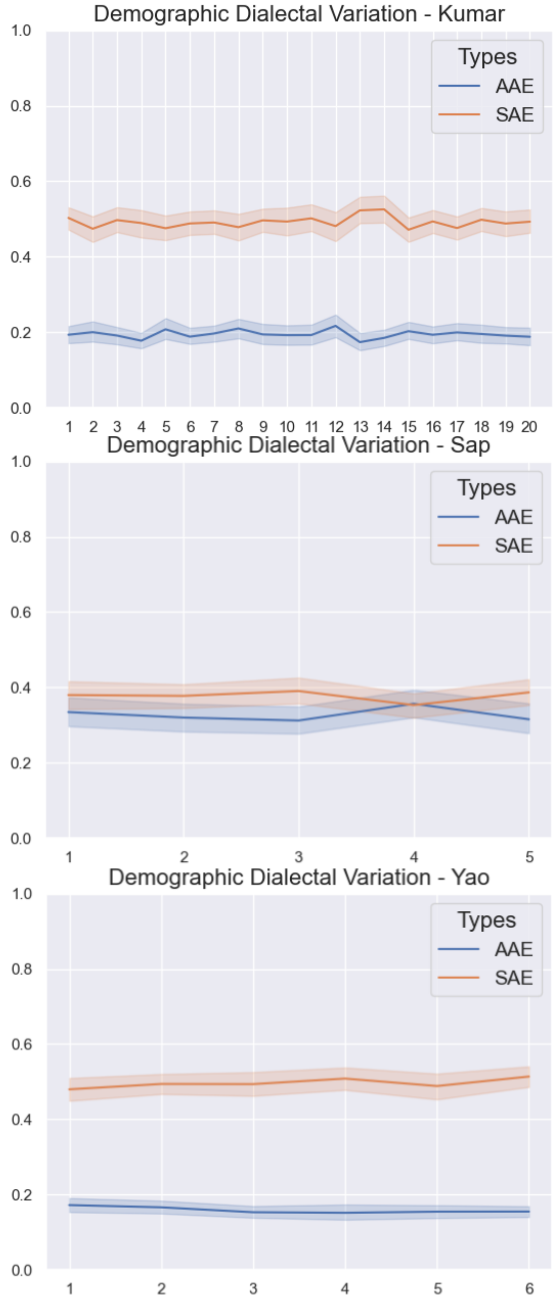


Figure 10: Demographic Dialectal Variation