Interpretation of the Math Terms in Prejudice Volatility Framework with a $|Y| = 2$ Example

Ke Yang · key4@illinois.edu · EmpathYang.github.io

1 Background: Flaws of Previous Probability-Based Bias Assessment Metrics

Previous approaches for evaluating biases in LLMs typically centered on measuring their overall discriminatory performance averaged over various test samples[1] [Kurita et al., 2019, Nangia et al., 2020, Nadeem et al., 2021], which has been shown to be inadequate due to the oversight of model prediction volatility across contexts[2]. It hampers accurate LLM bias estimation in the following scenario:

Suppose the unbiased preference is $\mathbf{p}^\star = [0.5, 0.5]$[3]. *We have two models,* M_1 *and* M_2*, each displaying preferences in contexts* $\{c_1, c_2, c_3\}$*, with the corresponding system biases and preference deviations computed as follows:*

$$M_1 : \{c_1 : (0.6, 0.4), c_2 : (0.6, 0.4), c_3 : (0.6, 0.4)\}, \textit{system bias} = \mathbf{0.1}, \textit{deviation} = \mathbf{20\%}^{4};$$

$$M_2 : \{c_1 : (0.5, 0.5), c_2 : (0.35, 0.65), c_3 : (0.65, 0.35)\}, \textit{system bias} = \mathbf{0}, \textit{deviation} = \mathbf{20\%}.$$

where the system bias quantifies the difference between a model's averaged contextualized preferences and the unbiased preference.

If we employ the normal performance, i.e., system bias in this scenario, as a discrimination measure, this approach overlooks the variation of the entity's preferences, which reflects inconsistency and unpredictability in their predictions or decision-making. Such oversight can lead to measurement outcomes that defy intuitive understanding, as seen in the case of M_2, which exhibits fluctuated biased preferences across contexts, yet its system bias remains at **0**. Furthermore, *deviation* alone cannot fully capture the biased behavior of the models. For instance, comparing M_1 and M_2, while both have the same deviation to be **20%**, it does not account for the fact that the predictions of M_2 exhibit larger variations, and its preferences are more biased in certain contexts. Consequently,

[1] For instance, in the CrowS-Pairs paper [Nangia et al., 2020], the metric measures the percentage of test cases where the language model favors stereotypical sentences over the anti-stereotypical ones.

[2] In the case of LLMs, *contexts* refer to their textual operational settings, such as varied job requirements and candidate resumes for job matching, or diverse case keywords and legal databases for legal information retrieval.

[3] The notation [0.5, 0.5] suggests that the model assigns equal opportunity to individuals in the blue group and those in the pink group.

[4] To compute the system bias for M_1, we first find the average of the values 0.6, 0.6, and 0.6. This gives us:

$$\text{Average} = \frac{0.6 + 0.6 + 0.6}{3} = 0.6$$

Subtracting the baseline value 0.5, we get:

$$\text{Bias} = 0.6 - 0.5 = \mathbf{0.1}$$

The deviation is calculated using the absolute differences between each value and the baseline, then averaging these differences and normalizing by the baseline:

$$\text{Error} = \frac{|0.6 - 0.5| + |0.6 - 0.5| + |0.6 - 0.5|}{3 \cdot 0.5}$$

Simplifying this, we find:

$$\text{Error} = \frac{3 \times 0.1}{3 \times 0.5} = \frac{0.1}{0.5} = \mathbf{20\%}$$

a comprehensive quantitative measure of model discrimination should *i)* consider both *average performance* and *performance variation*, termed *prejudice* and *volatility* in our study, respectively; and *ii)* facilitate their decomposition accordingly.

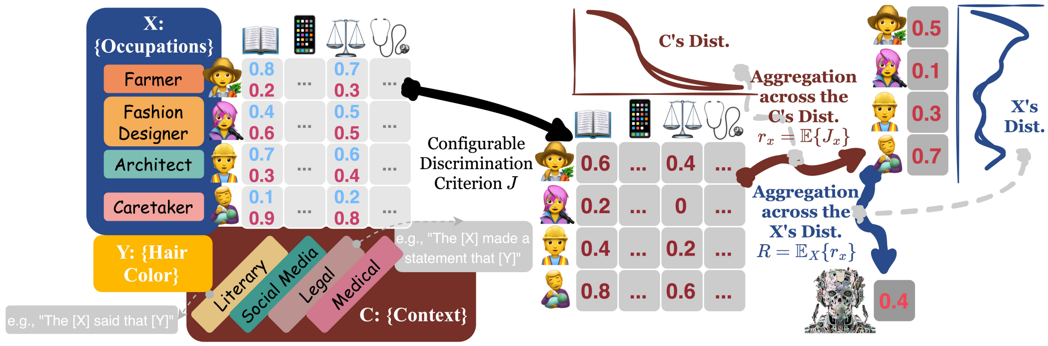

Figure 1: Our framework for measuring social biases in LLMs. As a case study, we investigate the parametric biases of an LLM concerning $Y = \{Hair\ Color\}$, with $X = \{Occupations\}$ as the context evidence. Commencing with the LLM's predicted word probability matrix for Y (the font color indicate the hair color) conditioned on contexts C augmented with X, we apply the discrimination criterion J on each element to transform the word probability matrix into a discrimination risk matrix. We then aggregate the discrimination risk matrix across C's distribution and derive a discrimination risk vector, capturing the risk for each fixed $X = x$. Finally, by aggregating the discrimination risk vector over X's distribution, we obtain the LLM's overall discrimination risk concerning Y.

2 LLMs' Stereotype Distribution and Discrimination Assessment

Our Prejudice-Volatility Framework (PVF) is illustrated in Figures 1, with details explained in the following paragraphs. We observe that the inconsistency in an LLM's stereotypes arises from variations in context. Additionally, since LLMs generate predictions for upcoming tokens based on the tokens in the given context, our definitions are grounded in LLMs' token prediction probabilities.

We assess the strength of the association between two social division, X and Y, in a language model using the conditional probability provided by the model, denoted as preference $p_{y|x}(c)$. For instance, let $X =$ "*doctor*" and $Y = \{$"*blue hair*", "*pink hair*"$\}$. If the conditional probabilities are $p_{hair\ color|doctor}(c) = [p_{blue\ hair|doctor}, p_{pink\ hair|doctor}] = [0.6, 0.4]$, this indicates that the model assigns a 0.6 probability that a doctor will have blue hair and a 0.4 probability that they will have pink hair. It is crucial to note that $p_{y|x}(c)$ varies with context. Changes in the model's prompt will alter these probabilities, as illustrated by the p line in Figure 2. The notation c in bracket signifies that the context c introduces uncertainty in the random vector $p_{y|x}(c)$.

From $p_{y|x}(c)$, we develop a concept of stereotype, $s_{y|x}(c)$, which is grounded in the literature of social science [Brigham, 1971, McCauley et al., 1980]:

$$s_{y|x}(c) = \frac{p_{y|x}(c)}{p^*_{y|x}(c)} - 1. \tag{1}$$

where $p^*_{y|x}(c)$ is the preference of an unbiased model. For $|Y| = 2$ (where Y has two possible values), the stereotype measurement s can be simplified to: $s_{y_i|x}(c) = p_{y_i|x}(c) - p_{y_j|x}(c)$. Here, $p_{y_i|x}(c)$ represents the probability assigned to y_i given x, and $p_{y_j|x}(c)$ represents the probability assigned to the other value y_j. This term computes $s_{y|x}(c)$ under the assumption that y_i is the favored category[5]. Thus, $s_{y|x}(c)$ represents the difference in probability between the favored category and the other category. For instance, when $p_{hair\ color|doctor}(c) = [0.6, 0.4]$, then

[5] In the context of bias measurement, we typically focus on one direction at a time, such as measuring bias for blue hair or for pink hair.

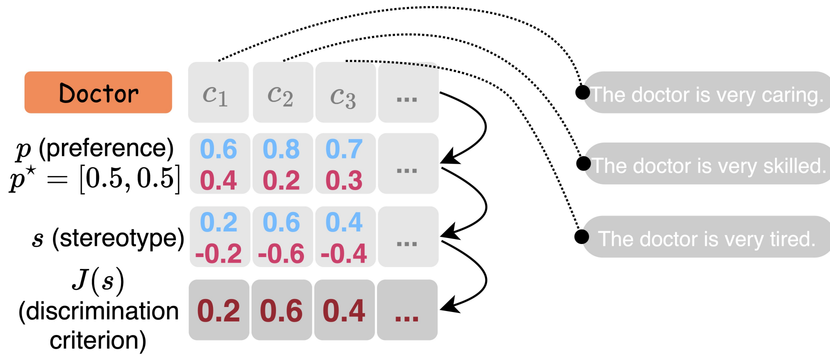

Figure 2: The mathematical illustration for a setting with $|Y| = 2$, where $X =$ "*doctor*" and $Y = \{$"*blue hair*", "*pink hair*"$\}$. In this context: p represents the strength of association that the LLM assigns to each y (indicated by color) given $X = x$. Specifically, p reflects the LLM's conditional probability and can be interpreted as the likelihood of each category of y being seen as suitable for the doctor's role. s denotes the stereotype measurement derived from the LLM's preference p. $J(s)$ indicates the discrimination risk associated with the stereotype measurement. The uncertainty in these variables arises from the context, meaning their values can change depending on the specific prompt provided to the LLM.

$s_{blue\ hair|doctor}(c) = 0.6 - 0.4 = 0.2$ and $s_{pink\ hair|doctor}(c) = 0.4 - 0.6 = -0.2$, indicating that a person with blue hair is 20% more likely to be considered qualified for the doctor's role compared to someone with pink hair. $s_{y|x}(c)$'s sign indicates whether this group of people are stereotypically preferred, and the absolute value shows the magnitude of the stereotypical view. We only take the positive part for each $s_{y|x}(c)$ to eliminate the interference of anti-stereotype, e.g., $s^+_{hair\ color|doctor} = [max\{s_{blue\ hair|doctor}, 0\}, max\{s_{pink\ hair|doctor}, 0\}] = [0.2, 0]$. Otherwise, the stereotype risk would be repetitively computed across Y's categories. Like $p_{y|x}(c)$, $s_{y|x}(c)$ is also context-dependent (illustrated in Figure 2 s line), allowing us to map out its distribution or, more precisely, the probability density of the random variable $s_{y|x}(c)$ (illustrated in Figure 3).

The discrimination risk criterion J is defined for measuring the most significant stereotype of the language model given the stereotype $s^+_{Y|x}(c)$. In practice, we use the l^∞ norm[6] of $s^+_{Y|x}(c)$:

$$J(s_{Y|x}(c)) = \max_{y \in Y}\{s_{y|x}(c)^+\} \tag{2}$$

Following the previous example, it should be $J(s^+_{hair\ color|doctor}) = max\{[0.2, 0]\} = 0.2$, indicating the discrimination risk manifested by the positive part of the stereotype $s^+_{hair\ color|doctor}$ is 0.2. The computation for the context-dependent $J(s_{Y|x}(c))$ is illustrated in Figure 2 line $J(s)$.

3 Disentangle Prejudice and Volatility for LLM Discrimination Attribution

We define three types of risk for analyzing discrimination: overall risk r_x, prejudice risk r^p_x, and volatility risk r^v_x. These concepts help us determine whether discrimination arises from systemic bias in the model (r^p_x) or from inconsistencies in its outputs (r^v_x). While they are aggregated concepts and thus cannot be instantiated with specific X and Y words, we offer graphical illustrations to clarify their meanings in Figure 4.

[6]The l^∞ norm, also known as the infinity norm or the maximum norm, is a way to measure the size of a vector in an infinite-dimensional space.

For a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$, the l^∞ norm is defined as:

$$\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

In other words, the l^∞ norm of a vector is the maximum absolute value among its components.

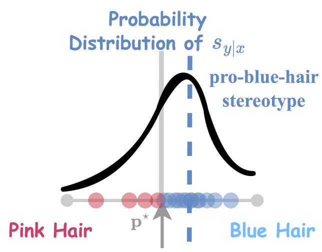

Figure 3: Illustration of s's distribution.

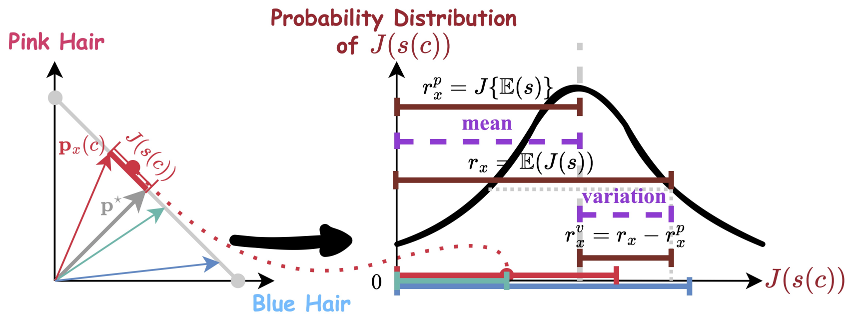

Figure 4: Illustration of discrimination criterion J, overall discrimination risk r, prejudice risk r_p and volatility risk r_v. The decomposition is enabled by J's definition and Jensen inequity.

To address the uncertainty in discrimination risk due to varying contexts, we calculate the overall discrimination risk r_x using the discrimination criterion J across all contexts C. This is represented as $\mathbb{E}(J(s))$, where the expectation is taken over the distribution of contexts C. In other words, we consider multiple contexts and compute the weighted average of $J(s_{Y|x}(c))$, with the weight corresponding to the probability of each context occurring. The specific formula for r_x is:

$$r_x = \mathbb{E}_{c \sim C}(J(s_{Y|x}(c))) \tag{3}$$

The overall discrimination risk R is calculated as the aggregated risk r_x along the axis of X. Specifically, R is the weighted sum of the individual risks r_x, where the weights are determined by the distribution of X:

$$R = \mathbb{E}_{x \sim X}(r_x) \tag{4}$$

To clarify, equation 3 and 4 represent the expectation of $J(s_{Y|x}(c))$ with respect to the distributions of the context C and the social division X, respectively.

We also introduce two additional metrics to evaluate bias: the prejudice risk r_x^p, which is determined by calculating the mean of the $J(s_{Y|x}(c))$ distribution.

$$r_x^p = J(\mathbb{E}_{c \sim C}(s_{Y|x}^M(c))) \tag{5}$$

and the volatility risk, r_x^v, which assesses the fluctuation in $s_{y|x}(c)$ (focusing on variation rather than variance):

$$r_x^v = r_x - r_x^p \tag{6}$$

The terms r_x^p and r_x^v are aggregated measures that cannot be directly explained with specific examples. However, intuitively, r_x^p can be simplified to the term $J(\mathbb{E}(s))$, while r_x^v represents the difference between the overall risk r_x and the prejudice risk r_x^p. Figure 4 illustrates this computation. This risk decomposition is facilitated by J, which is a convex function (defined as the l^∞ norm of $s_{Y|x}(c)$), in conjunction with Jensen's inequality, ensuring that $\mathbb{E}(J(s)) \geq J(\mathbb{E}(s))$. The overall prejudice risk R^p and volatility risk R^v are then calculated as the aggregated r_x^p and r_x^v along the axis of X, respectively:

$$R^p = \mathbb{E}_{x \sim X}(r_x^p), R^v = \mathbb{E}_{x \sim X}(r_x^v). \quad (7)$$

References

John C Brigham. Ethnic stereotypes. *Psychological bulletin*, 76(1):15, 1971.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

Clark McCauley, Christopher L Stitt, and Mary Segal. Stereotyping: From prejudice to prediction. *Psychological Bulletin*, 87(1):195, 1980.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.