Nuanced Safety in Generative AI: How Demographics Shape Responsiveness to Severity

Pushkar Mishra [1] **Charvi Rastogi** [1] **Stephen R. Pfohl** [2] **Alicia Parrish** [1] **Roma Patel** [1] **Mark Diaz** [2] **Ding Wang** [2] **Michela Paganini** [1] **Vinodkumar Prabhakaran** [2] **Lora Aroyo** [1] **Verena Rieser** [1]

Abstract

Ensuring safety of Generative AI requires a nuanced understanding of pluralistic viewpoints. In this paper, we introduce a novel data-driven approach for calibrating granular ratings in pluralistic datasets. Specifically, we address the challenge of interpreting responses of a diverse population to safety expressed via ordinal scales (e.g., Likert scale). We distill non-parametric *responsiveness metrics* that quantify the consistency of raters in scoring the varying levels of the severity of safety violations. Using safety evaluation of AI-generated content as a case study, we investigate how raters from different demographic groups (age, gender, ethnicity) use an ordinal scale to express their perception of the severity of violations in a pluralistic safety dataset. We apply our metrics across violation types, demonstrating their utility in extracting nuanced insights that are crucial for developing reliable AI systems in a multicultural contexts. We show that our approach offers improved capabilities for prioritizing safety concerns by capturing nuanced viewpoints across different demographic groups, hence improving the reliability of pluralistic data collection and in turn contributing to more robust AI evaluations.

1. Introduction

Ensuring safety of Generative AI is paramount for their responsible deployment and societal trust. Recent research demonstrates that safety perceptions are not uniform but vary significantly across individuals and groups (Aroyo et al., 2024; Kirk et al., 2024; Rastogi et al., 2024).

Safety evaluation tasks often use *binary ratings*, such as *safe* and *unsafe*, which lack the granularity needed for effective alignment with human perception (Wu et al., 2023; Collins et al., 2024). To capture more fine-grained perceptions of the severity of perceived harm, safety evaluation tasks also employ *ordinal scales*, such as Likert scale (Curry et al., 2021), which allow for more nuanced feedback critical in creation of *pluralistic datasets*. However, such scales are susceptible to increased noise from variations in individual interpretation and usage, including response biases (Paulhus, 1991) such as extreme responses (Greenleaf, 1992), central tendency (for odd scales), and forced choice and polarization (for even scales). Such scoring biases further propagate in Large Language Models (LLMs). On the one hand, when non-calibrated ratings are used for training of LLMs can cause exaggerated safety behaviors. On the other hand, when LLMs are used as raters in evaluation tasks (Bavaresco et al., 2024; Wang et al., 2024) the non-calibrated ratings can degrade the utility of AI models, and in both cases can lead to downstream harms due to aggregation-based approaches of non-calibrated ratings.

This paper introduces a novel data-driven *non-parametric* approach for *calibrating granular ratings in pluralistic datasets*, offering a more robust and nuanced evaluation than traditional approaches. Using *safety evaluation of AI-generated content* as a case study, we address the critical challenge of understanding how raters from diverse demographic groups interpret and utilize ordinal scales (e.g., Likert scales) when assessing the severity of safety violations. The main contributions of this paper are:

Metrics Development: We distill robust *responsiveness metrics* from observed data to interpret the scoring patterns of different rater groups for the varying levels of severity. These metrics allow us to:

1. **Measure responsiveness to severity**: How consistently do different rater groups use a given ordinal scale to express the varying levels of severity of violations?
2. **Compare responsiveness**: Are different rater groups equally responsive to the varying levels of severity?

Application and Validation: We apply these metrics to a pluralistic safety evaluation dataset, demonstrating their utility in extracting nuanced insights that are crucial for developing AI systems in a multi-cultural context by:

[1]Google DeepMind [2]Google Research. Correspondence to: Pushkar Mishra <pushkarmishra@google.com>.

- **Understanding scale usage**: uncovering patterns in the use of the given Likert scale, thus elucidating genuine variations in the expressions of demographic groups;
- **Capturing nuanced viewpoints**: identifying demographic groups most responsive to severity across violation types, resulting in a granular understanding of pluralistic viewpoints;
- **Prioritizing high-impact items**: taking items deemed highly unsafe by groups with with high responsiveness to severity;
- **Improving pluralistic data collection**: establishing a reliable and repeatable process for sampling raters and rater groups with high responsiveness to severity.

2. Related work

Our work builds on the state-of-the-art in eliciting nuanced human feedback in Generative AI evaluation and expands existing research on calibration of human judgments.

Nuanced human feedback for AI safety evaluation. Collecting human perspectives on behavior of generative AI models is exceedingly commonplace with growth in its usage in real world tasks. Across literature in AI evaluation, different configurations of human feedback have been studied, with an outsized focus on binary (0/1) human feedback. Recent research (Collins et al., 2024; Arhin et al., 2021; Denton et al., 2021) discusses the limitations of binary feedback in capturing the nuance involved in generative AI evaluation, especially in safety. Wu et al. (2023); Collins et al. (2024) propose fine-grained human feedback encompassing evaluation across multiple attributes and with higher density, yielding improvement in downstream AI tasks via RLHF. Further, Rauh et al. (2024); Jiang et al. (2021) emphasise the importance of measuring extent of harm (severity) in evaluation of algorithms. Another dimension in collecting human feedback relates to the identity of the human providing the feedback. The role of rater identity in their annotation has been discussed extensively in AI evaluation literature (Denton et al., 2021; Arhin et al., 2021; Aroyo et al., 2024; Homan et al., 2023). For developing AI that aligns with human values, Sorensen et al. (2024) show the importance of considering pluralistic viewpoints from a diverse set of raters. Our research builds upon this body of work by specifically examining human feedback collected on a fine-grained (0-4) ordinal scale from raters belonging to different groups with different collective identities.

Calibration of human judgements. Human judgments elicited as scores on a scale are often miscalibrated, implying that the scores given by people are incomparable due to differences in interpretation and usage of each score (see Griffin & Brenner (2008); Poston (2008) and references therein). Miscalibration in human scores is sometimes addressed through simplifying modeling assumptions about how the miscalibration presents in the data. These modeling assumptions include linear models with additive biases corresponding to rater identity (Bürkner & Vuorre, 2019; Paul, 2011; Barr et al., 2013), models with rater identity-based scale-and-shift biases (Paul, 2011; Roos et al., 2011), mixed-effects models, among others (Wang & Shah, 2019). However, research has shown that issues of human judgment calibration are often more complex, causing significant violations to these simplified assumptions (see Griffin & Brenner (2008) and references therein). In this work, while making minimal assumptions on the nature of miscalibration in human judgments, we provide *non-parametric* metrics to measure the consistency of raters in reflecting varying levels of severity. Traditional non-parametric metrics like Kendall's τ and area under the PR or ROC curves do not capture well the responsiveness to severity. Using a real-world dataset, we surface insights from our proposed metrics into the scoring patterns of different rater groups.

3. Setup

We consider a general setup with two different rater populations that reflect two contrasting safety evaluation paradigms (Rottger et al., 2022): crowd raters who *indicate* safety perceptions of a diverse population, and expert raters who follow detailed guidelines for *prescribing* what counts as safe or unsafe.

1. *Crowd raters* provide a set of pluralistic safety perceptions on an ordinal scale, with each of their ratings representing the perception of a certain rater group. Each rater reviews every item in the given dataset and provides an integer score on a (0-K) Likert scale, where 0 is not harmful and K is completely harmful.
2. *Expert raters* strictly follow a set of prescribed rules. For each item in the given dataset, they provide a binary score of 0 (safe) or 1 (unsafe).

3.1. Data model

To formalize our assumptions regarding the scores given by the individual raters, i.e., individual expert raters and individual crowd raters, we define a data model. Let R_{ij} be the latent severity rating of rater j for prompt-image pair i. We assume the following ordinal data model:

$$R_{ij} = \mathcal{F}(P_{ij}, b_j, c_i)$$

where P_{ij} is the perceived severity of prompt-image pair i by rater j, b_j is the rater-specific bias in perception, c_i is the prompt-image pair specific bias in perception, and $\mathcal{F}$ is some function over these. Then, the relationship between the latent severity ratings and the scores of the two rater populations can be written as:

- Expert rater: $S_{ij} = 1$ if $R_{ij} > t_j$, where S_{ij} is the binary score given by expert rater j to prompt-image pair i and t_j is the threshold above which they assign a binary score of 1.
- Crowd rater: $S_{ij} = k$ if $R_{ij} > t_{jk}$, where S_{ij} is the Likert score given by crowd rater j to prompt-image pair i and t_{jk} is the threshold above which they assign the Likert score $k \in \{0, 1, \ldots, K\}$.

This data model assumes that the perception of severity varies from rater to rater, be they expert or crowd raters. To simplify the data model, we introduce a latent variable V to represent the true, but unobservable, severity of prompt-image pairs. This acknowledges that true severity isn't directly measurable but is still the underlying factor that monotonically influences every rater's judgments. In the case of expert raters, V represents the shared understanding of severity as defined by the strict guidelines that they must adhere to, where the guidelines serve as a framework to operationalize the theoretical notion of severity (cf. prescriptive annotation paradigm). In the case of crowd raters, V represents some shared perception of severity based on lived experiences (cf. descriptive annotation paradigm). With the simplification, we can reformulate R_{ij} as:

$$R_{ij} = \mathcal{F}'(V_i, b_j)$$

where V_i is the true underlying severity of prompt-image pair i. The simplification creates a more tractable model without requiring estimation of P_{ij}. As such, we work with this simplified data model hereon. However, we note that this simplification limits our ability to exactly compute the component of raters' individual perceptions in their responsiveness to severity.

3.2. Responsiveness to severity

Having a Likert scale allows crowd raters to express their safety judgments on a severity spectrum rather than classifying prompt-image pairs as safe or unsafe. However, the Likert scale does not guarantee that rater scores will meaningfully reflect the severity of violations. For example, there may be raters who only use the ends of the scale, or those who cluster all their ratings around certain scores. It is essential to disentangle the scale use from the actual response to the severity of violations. While one might ideally want to define responsiveness as a direct relationship between a rater's scores S and the true severity V, this is not practically feasible. True severity V is a theoretical construct that cannot be objectively measured or easily determined. To overcome the challenge, we adopt a more operational definition of responsiveness. We formally define the concept of responsiveness to the severity of violations as being composed of the following two properties that can be quantified using observable data:

- **Ability to stochastically order severity**. If a rater is responsive to the true underlying severity V, then a higher score from them should correspond to a higher probability of the true severity crossing any threshold $T = t$. This is the principle of first-order stochastic dominance, which can be stated as $P(V > T|S = s_1, T = t) \geq P(V > T|S = s_2, T = t)$ for all $T \in \mathcal{T}$ when $s_1 > s_2$.
- **Ability to discriminate between distinct levels of severity**. If a rater is responsive to the true underlying severity V, then they should be able to discriminate the prompt-image pairs whose true severities are above a given threshold $T = t$ from those below that threshold. This means that $P(S \geq s|V > T, T = t) \geq P(S \geq s|V \leq T, T = t)$ for all $T \in \mathcal{T}$.

Intuitively, these two properties together signify that a rater who is responsive to the severity of violations is able to consistently convey the severity of violations at each score of the Likert scale. Our approach to characterizing responsiveness to severity relies on a notion of stochastic ordering that is related to the classic decision-theoretic concepts of stochastic ordering and outcome monotonicity (Birnbaum & Navarrete, 1998). The stochastic ordering property can further be related to monotonicity of a calibration curve or reliability diagram (DeGroot & Fienberg, 1983) in the case that we consider calibration of non-expert scores against some binarized reference for severity. The property concerning the ability of raters to discriminate between distinct levels of severity is related to the notion of discriminability from signal detection theory (McNicol, 2005) used to motivated the design of metrics such as the area under the receiver operating characteristic (ROC) curve and Kendall's τ (Kendall, 1938). Furthermore, the property is also related to the notion of discrimination between different levels of the latent trait in Item Response Theory (Samejima, 1968) and Mokken Scale Analysis (Mokken, 1971).

4. Metric design

Next, our aim is to quantify the responsiveness of the different demographic groups to the severity of violations in order to evaluate and compare them. We achieve such a quantification by individually quantifying the two properties that constitute responsiveness. To do so, we need a signal for $V > T$ since V itself is unobservable. We treat the binary scores of expert raters as binary labels U and take $U = 1$ as the signal for V exceeding some T. To obtain the binary labels U for the prompt-image pairs, we assign the individual binary scores of every expert to the prompt-image pairs by replicating each prompt-image pair for every expert. Alternatively, we could also obtain the binary labels U for prompt-image pairs from crowd raters themselves, excluding the demographic group being evaluated to maintain the independence of S and T: we first binarize the individual

Likert scores of the crowd raters using a score $s \in [1, 4]$ as boundary, then assign these individual binary scores by replicating each prompt-image pair for every crowd rater.

We note that when the binary labels U are obtained using expert raters, we are quantifying the responsiveness to severity as captured by the expert raters based on the guidelines they use. On the other hand, when the binary labels U are obtained using crowd raters (excluding the group being evaluated), we are quantifying the responsiveness to severity as captured by the collective judgment of the crowd rater population. The latter paradigm is discussed in Section 7.

Here, we reformulate the inequalities presented above that relate to the ability to stochastically order and discriminate between levels of severity to leverage $U = 1$ in place of $V > T$. Note that as we assume that the threshold T varies across individual expert raters, it is not straightforward to directly substitute $U = 1$ into the threshold-specific inequalities. However, we show that if the two inequalities hold for all individual thresholds T, then they also hold for $U = 1$.

- **Ability to stochastically order**. We had that $P(V > T|S = s_1, T = t) \geq P(V > T|S = s_2, T = t)$, when $s_1 > s_2$, for all $T = t$. Hence, when $s_1 > s_2$, $P(U = 1|S = s_1) \geq P(U = 1|S = s_2)$.
- **Ability to discriminate**. We had that $P(S \geq s|V > T, T = t) \geq P(S \geq s|V \leq T, T = t)$ for all $T = t$. Hence, we have $P(S \geq s|U = 1) \geq P(S \geq s|U = 0)$.

To prove the two properties, let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of all thresholds T. For the stochastic ordering property, we have $P(U = 1|S = s_1) = \sum_{t \in \mathcal{T}} P(V > T|S = s_1, T = t)P(T = t)$ and $P(U = 1|S = s_2) = \sum_{t \in \mathcal{T}} P(V > T|S = s_2, T = t)P(T = t)$ and given that S and T are independent. Since $P(T = t)$ is non-negative and $P(V > T|S = s_1, T = t) \geq P(V > T|S = s_2, T = t)$ for every $t \in \mathcal{T}$, hence, $P(U = 1|S = s_1) \geq P(U = 1|S = s_2)$.

For the discrimination property, for every $t \in \mathcal{T}$ we have $P(S \geq s|V > T, T = t) \geq P(S \geq s|V \leq T, T = t)$, which gives $P(S \geq s|U = 1, T = t) \geq P(S \geq s|U = 0, T = t)$ for any given t. By Bayes' rule, $\frac{P(S \geq s, U=1, T=t)}{P(U=1, T=t)} \geq \frac{P(S \geq s, U=0, T=t)}{P(U=0, T=t)}$ for any given t. Let $a(t) = P(S \geq s, U = 1, T = t)$, $b(t) = P(U = 1, T = t)$, and $c(t) = P(T = t)$. Since S and T are independent, $P(S \geq s, T = t) = P(S \geq s)c(t)$. Then, $P(S \geq s, U = 0, T = t) = P(S \geq s)c(t) - a(t)$ and $P(U = 0, T = t) = c(t) - b(t)$. So, the inequality becomes $\frac{a(t)}{b(t)} \geq \frac{P(S \geq s)c(t) - a(t)}{c(t) - b(t)}$ for any given t. Cross-multiplying and summing the inequalities over all $t \in \mathcal{T}$, we have that $\sum_{t \in \mathcal{T}} a(t) \geq P(S \geq s) \sum_{t \in \mathcal{T}} b(t)$. This yields $P(S \geq s, U = 1) \geq P(S \geq s)P(U = 1)$, and hence, $P(S \geq s|U = 1) \geq P(S \geq s)$. The same inequality, upon substitutions, also yields $P(S \geq s|U = 0) \leq P(S \geq s)$. Therefore, $P(S \geq s|U = 1) \geq P(S \geq s|U = 0)$.

4.1. Metrics for the two properties

We design metrics for the two properties based on the standard concepts of precision and recall. Given a Likert scale with scores S $\in \{0, 1, 2, 3, ..., K\}$, we take $Precision(S)$ to denote the precision when score = S is taken as unsafe and all scores $\neq S$ are taken as safe. Similarly, we take $Recall(S)$ to denote the recall when score = S is taken as unsafe and all scores $\neq S$ are taken as safe. The decision to compute precisions and recalls exactly at S rather than $\geq S$ is a crucial one to our metric development.

We define the following metrics to quantify the strength of the core inequalities for the two properties: *Monotonic Precision Area* and *Weighted Recall Area*.

***Monotonic Precision Area* for stochastic ordering**. We note that the probability $P(U = 1|S)$ is equivalent to $Precision(S)$, i.e., the precision when classifying items with score $= S$ as unsafe and items with score $\neq S$ as safe. Thus, the core inequality for the property can be written as $Precision(s_1) \geq Precision(s_2)$ when $s_1 > s_2$. If we take the area under the curve defined by $Y_{so}(s) = \sum_{i=0}^{s-1} \{Precision(s) - Max_{j=0}^{i} Precision(j)\}$ at $s = 0, 1, 2, 3, \ldots, K$, then a high value would mean consistent increases in $P(U = 1|S = s)$ without violations of monotonicity. We normalize the area by the maximum possible area, which is given by $\frac{K}{2} * (\frac{K}{2} + 1)$ if K is even and $(\frac{K+1}{2})^2$ if K is odd. Intuitively, the maximum area is achieved when precision is 0 for the lower half of the scores and 1 for the upper half, signifying that the rater has perfectly aligned the midpoint of the Likert scale with the threshold for distinguishing between unsafe ($U = 1$) and safe ($U = 0$). When the area is below 0, indicating stochastic ordering worse than random guessing, we cap the area to 0 to ensure a non-negative metric. Additionally, when computing $Y_so(s)$, we ignore all the scores $< s$ that are not used by the rater and have undefined precision. Similarly, if s is a score not used by the rater, we take $Y_so(s) = 0$.

***Weighted Recall Area* for discrimination**. We note that the probability $P(S = s|U = 1)$ is equivalent to $Recall(s)$, i.e., the recall when classifying items with score $S = s$ as unsafe and items with score $S \neq s$ as safe. Similarly, the probability $P(S < s|U = 0)$ is equivalent to the recall when classifying items with score $S < s$ as safe and items with score $S \geq s$ as unsafe. If we take the area under the curve defined by $Y_d(s) = P(S < s|U = 0) * Recall(s)$ at $s = 0, 1, 2, 3, \ldots, K$, then a high value means high $P(S < s|U = 0)$ and $P(S = s|U = 1)$. This in turn implies high $P(S \geq s|U = 1)$ and low $P(S \geq s|U = 0)$, as desired by the core inequality for the property. $Y_d(s)$ is essentially the recall of unsafe examples at score s, i.e, $Recall(s)$, weighted by the proportion of safe samples correctly discriminated by a score $S < s$. It represents the concordance probability (Heller & Mo, 2016) from two events

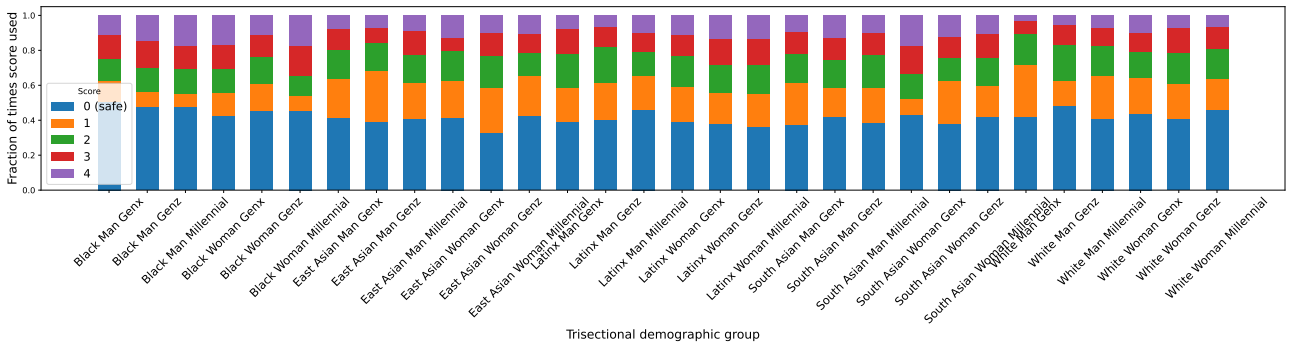


Figure 1. Plot showing the distribution of scores used by each trisectional demographic group in dataset considered (Rastogi et al., 2024).

at each score s that contribute to strengthen the inequality, assigning scores $= s$ to unsafe items while correctly assigning scores $S < s$ to safe items.

Since the two metrics are based on rates, we take their harmonic mean to combine them into one metric. Harmonic mean ensures that low performance on any one metric is strongly penalized. When both metrics are high for a rater, it means the rater is responsive to varying levels of severity. On the other hand, a low monotonic precision area with a high weighted recall area suggests that while the rater may be good at coarsely separating items that are of very distinct severities, they are not good at granularly responding to varying levels of severities. A low monotonic precision area may also result from biases in scale usage that causes the rater to not use the full range of the scale, e.g., central tendency bias or extreme responses bias.

4.2. Contrasting with traditional metrics

There are potentially many alternative non-parametric approaches that could be used to assess responsiveness to severity. For example, one could use traditional metrics such as the Spearman Rank correlation, area under the ROC curve, or Kendall's τ to assess the ability to discriminate or to assess whether the relationship between non-expert ratings scores and the reference scores is monotonic. We highlight some limitations of these approaches below:

- *Insensitivity to baseline*: Traditional metrics do not account for a rater's baseline tendency to choose certain scores. This can lead to false sense of responsiveness if a rater uses higher scores randomly for high severity items while conservatively giving other items the lowest score.

- *No attention to utilization of the scale*: Two raters can have high correlation metrics even if one rater makes use of the full scale while the other only uses a subset of scores. Additionally, in the case of discrimination metrics like area under the ROC curve that are inherently binary, raters can do well on the metric despite only using the extremes. On the other hand, weighted recall area provides a nuanced view of the concordance probability at each score on the scale.

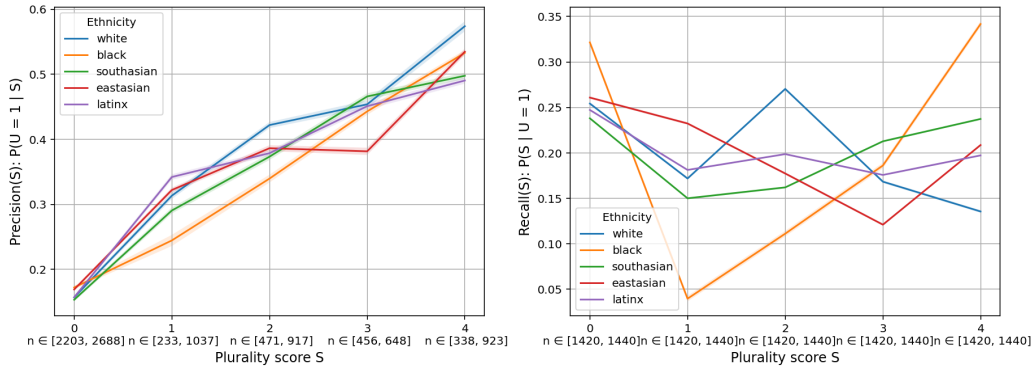
- *Lack of focus on behaviour at scores*: correlation metrics capture general monotonic relationships between scores and underlying severity, if the latter were accessible, but they do not penalize raters who assign higher scores without meaningful increases in corresponding severity at the scores. Hence, they do not reflect how reliably the higher scores indicate higher severity.

- *Fragility to insignificant variations*: traditional metrics focus solely on the ordering of scores relative to severity, without taking into account the magnitude of differences between them. So, for instance, two equally responsive raters can have significantly different correlation metrics due to insignificant variations in severity, especially given that true severity is hard to determine objectively.

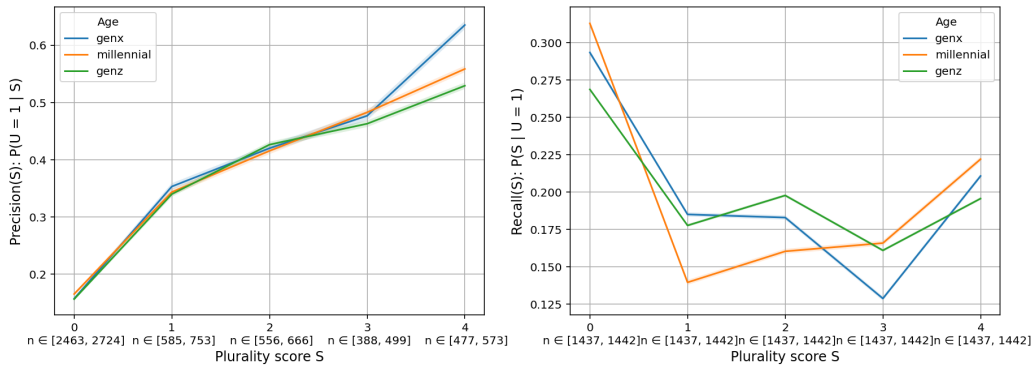
The way we define and quantify responsiveness addresses these limitations, while offering simplicity, interpretability, and the ability to get insights into how raters utilize different scores on the Likert scale. Nevertheless, as defined in our data model, different types of safety violations might exist, some inherently more or less likely to be labeled unsafe regardless of perceived severity. While no metric can be an absolute measure of responsiveness to severity, our metrics provide a robust and meaningful measure for evaluating and comparing responsiveness to severity. In appendix A, we run simulations to show how our proposed metrics and traditional metrics evaluate different scoring patterns.

5. Responsiveness Evaluation of Groups

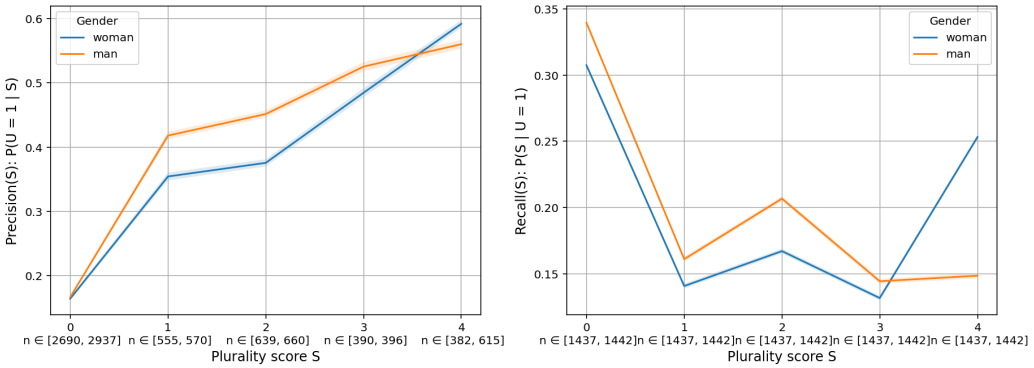
To demonstrate the utility of our proposed analysis framework with the responsiveness metrics, we apply it to an existing pluralistic dataset of AI safety ratings.



(a) Demographic groups of crowd raters by ethnicity



(b) Demographic groups of crowd raters by age



(c) Demographic groups of crowd raters by gender

Figure 2. $Precision(S)$ and $Recall(S)$ at plurality scores $S = 0$ to 4 for different demographic groups when using experts to obtain binary labels U, where grouping is by (a) ethnicity, (b) age, and (c) gender.

5.1. Dataset description

We consider the dataset in Rastogi et al. (2024) which contains annotations from people evaluating the safety of generative AI. Concretely, in this dataset, the crowd raters and expert raters assess the safety of a set of prompt-image pairs, where the crowd raters provide a Likert-scale rating from 0 to 4 (where 0 is not harmful and 4 is completely harmful), and expert raters provide binary rating of 0 (safe) or 1 (unsafe) for each prompt-image pair. The crowd raters are recruited based on their demographics. We identify three demographic axes - gender, ethnicity and age. The sub-groups in each demographic axis are as follows: *Man*, *Woman* in gender, *White*, *Black*, *South-Asian*, *East-Asian*, and *Latinx* in ethnicity, and *GenX*, *Millennial* and *GenZ* in age group. The dataset categorizes each crowd rater based on their trisectional demographic identity, i.e. their ethnicity, age, and gender. Following Rastogi et al. (2024), we distinguish top-

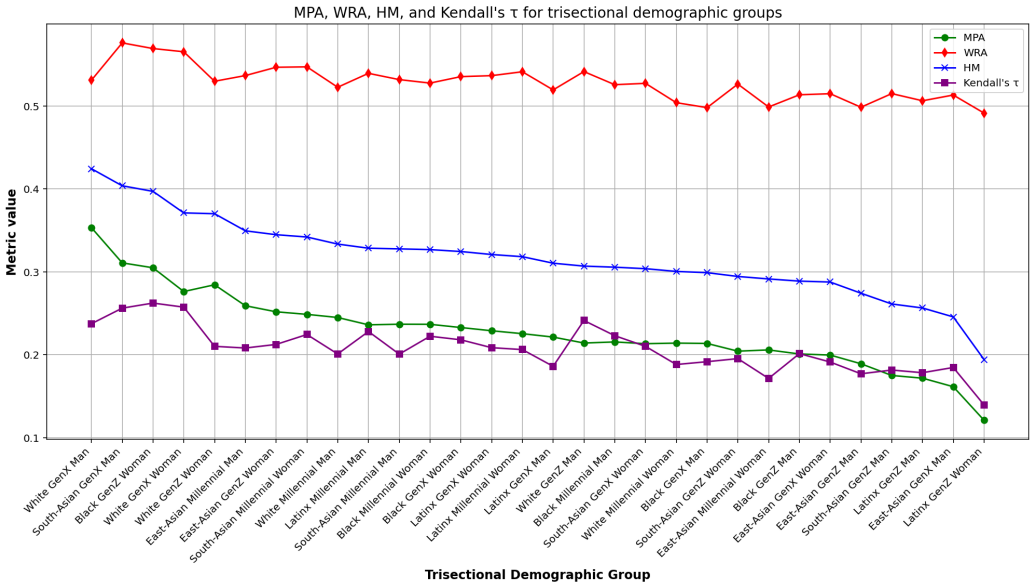


Figure 3. Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), and Kendall's τ for trisectional demographic groups of crowd raters when binary labels U are obtained from expert raters. All confidence intervals are within ± 0.01.

level demographic groups, e.g. *East-Asian*, from trisectional demographic groups, e.g. *Black–GenZ–Man*.

The dataset contains more than 1 expert rating for each prompt-image pair; to obtain binary labels U per prompt-image pair, we replicate each prompt-image pair for every expert. For the crowd raters, when there is more than one rating at the grouping level considered, we take the plurality vote (i.e. the mode of scores from all the individual raters belonging to that group) that we refer to as the *plurality score* - an integer from 0 to 4. In case of ties, we take the most unsafe score to be the plurality score. The distribution of scores provided by crowd raters in the dataset are shown in figure 1. Note that the average inter-annotator reliability (IRR) is fairly low (0.25 on average). IRR scores of each top-level group can be found in Appendix tables 1 and 2.

Figure 2 presents curves that show $Precision(S)$ and $Recall(S)$ at plurality scores 0 to 4 for different demographic groups of crowd raters when using experts to obtain binary labels U, where grouping is by (a) ethnicity, (b) age, and (c) gender

5.2. Results by trisectional demographic groups

We now evaluate and compare the responsiveness of different demographic groups of crowd raters, both at the trisectional demographic level (e.g., *Latinx–GenZ–Man*) as well as the top demographic level (e.g., *Latinx*, *Man*, etc.).

Figure 3 shows monotonic precision area, weighted recall area, their harmonic mean (HM), Kendall's τ for trisectional demographic crowd groups when binary labels U are obtained from expert raters as described in section 4.

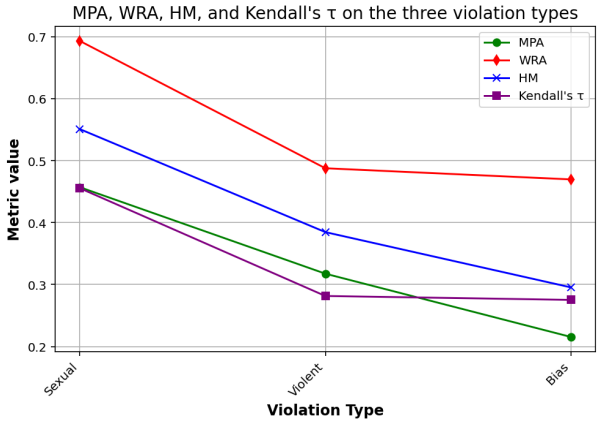


Figure 4. Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ for crowd rater population on three violation types when binary labels U are obtained from expert raters. All confidence intervals are within ± 0.01.

We note that the trisectional groups comprising Latinx and East-Asian ethnicities, aka *Latinx trisections* and *East-Asian trisections*, consistently have the lowest monotonic precision area, and consequently, the lowest harmonic means as well. This indicates that higher scores from these trisections correspond the least to higher reference severity as captured by the expert raters. On the other hand though, we note that both the trisections still achieve weighted recall areas comparable to others. This further suggests that while they concur with others on discriminating between distinct levels

of severity, they do not respond to the granular severity of violations the same way.

Figure 2(a) further validates the same as we note that the *East-Asian* top-level demographic group does not exhibit consistent gains in precision when plurality score goes from 1 to 3. Additionally, we see that Kendall's τ shows a similar pattern as weighted recall area given that both capture aspects of concordance. But traditional metrics like Kendall's τ do not reflect well the ability to stochastically order.

5.3. Results by violation types

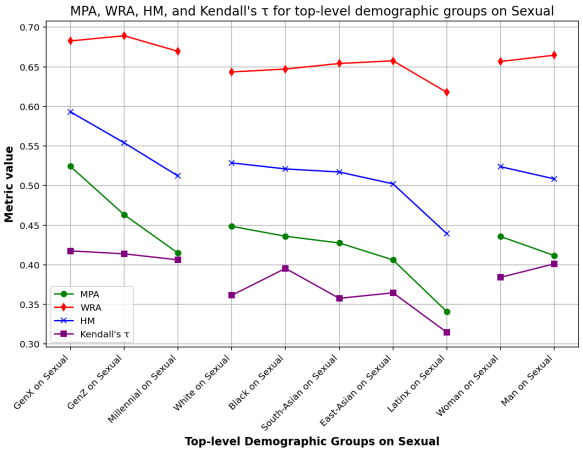
We first look at the trends for the entire crowd rater population on the three different violation types in prompt-image pairs, namely, *bias*, *sexual*, and *violent*. Figure 8 presents $Precision(S)$ and $Recall(S)$ for the crowd rater population on the three violation types when using experts to obtain binary labels U. Furthermore, figure 4 gives monotonic precision area, weighted recall area, their harmonic mean, and Kendall's τ for the three violation types. The responsiveness of crowd raters to the severity of bias is lower than that of sexual and violent violations since the severity of bias is harder to judge objectively. We also compare the scoring patterns of top-level demographic groups on each violation type. Figure 5 gives the metrics for top-level demographic groups on three violation types, using experts to obtain binary labels U. We see that the *Latinx* group shows the lowest responsiveness to the severity of sexual violations, while the *East-Asian* group shows the lowest responsiveness to violent violations.

6. Conclusion

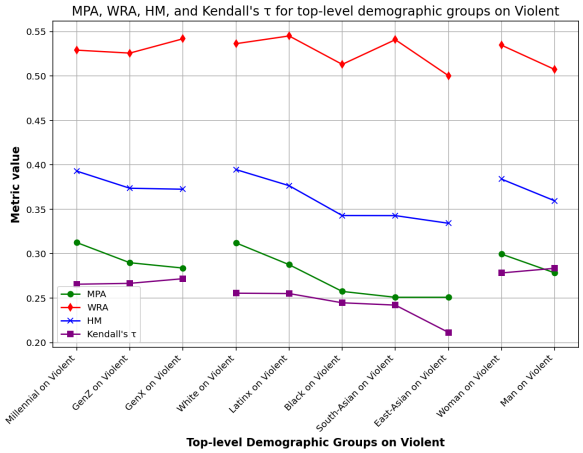
We formulated non-parametric metrics to assess raters' responsiveness to the severity of content violations, addressing limitations in existing approaches. Applying these metrics to a study involving diverse crowd raters, we found significant variations in how different demographic groups respond to severity when assessing AI-generated content. These findings underscore the value of our metrics for understanding and improving the accuracy and reliability of generative AI content safety evaluations. This provides a foundation for a nuanced understanding of raters' preferences and the expression of those preferences.

7. Limitations and Future Work

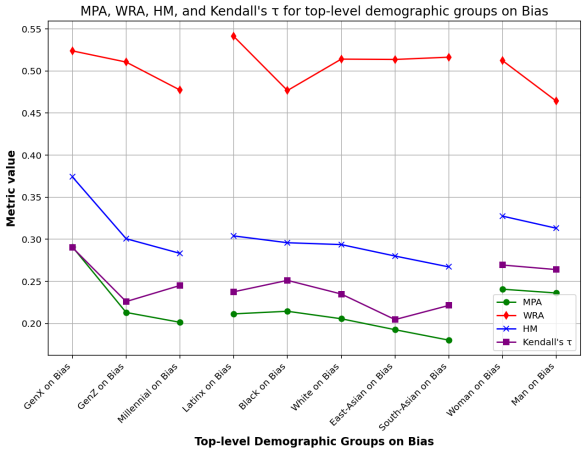
In this work, we based our analysis on a simplified data model that assumes the existence of some true underlying severity V. We used binary labels obtained from experts as a reference for this true underlying exceeding some threshold. So, a straightforward extension of our work will be to explore the paradigm where we obtain binary labels from crowd raters themselves as a reference for whether



(a)



(b)



(c)

Figure 5. Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean, Kendall's τ for top-level demographic groups on three violation types when using experts to obtain binary labels U. All confidence intervals are within ± 0.01.

the individual-specific perception of severity exceeds some threshold. Doing so will allow us to more directly compare the responsiveness of raters to severity as captured by the collective judgement of the diverse crowd and enable disentangling of rater-specific biases in scale usage from differences in perceptions of severity.

Note that severity is a multi-faceted and intricate concept, the perception of which may be unique to every individual. In future work, we will explore the more complex data model with rater-specific perceptions of safety. Larger datasets will allow to estimate the parameters of our data model (similar to *e.g.* Homan et al. (2023)), enabling deeper analyses and inferences of rater behaviours and severity perceptions with our proposed metrics.

Impact Statement

This work raises important ethical considerations regarding the potential reliance on human raters from diverse demographic backgrounds, particularly under-represented ones, for evaluating harmful content. Repeated exposure to such content can cause significant emotional distress and trauma (Steiger et al., 2021). To mitigate this risk, we advocate for the deployment of active learning strategies (Kirk et al., 2022) to identify the most informative items and reduce the quantity of harmful content raters need to evaluate. Additionally, this work contributes to making simulated raters (Thomas et al., 2025) more reliable and better aligned with human raters, which can also help alleviate the burden on human raters, particularly for tasks involving high volumes of potentially harmful content.

References

Arhin, K., Baldini, I., Wei, D., Ramamurthy, K. N., and Singh, M. Ground-truth, whose truth? - examining the challenges with annotating toxic text datasets. *ArXiv*, abs/2112.03529, 2021. URL https://api.semanticscholar.org/CorpusID:244921005.

Aroyo, L., Taylor, A., Díaz, M., Homan, C., Parrish, A., Serapio-García, G., Prabhakaran, V., and Wang, D. DICES dataset: Diversity in conversational AI evaluation for safety. *Advances in Neural Information Processing Systems*, 36, 2024.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2012.11.001. URL https://www.sciencedirect.com/science/article/pii/S0749596X12001180.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. URL https://arxiv.org/abs/2406.18403.

Birnbaum, M. H. and Navarrete, J. B. Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of risk and uncertainty*, 17:49–79, 1998.

Bürkner, P.-C. and Vuorre, M. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101, 2019.

Collins, K. M., Kim, N., Bitton, Y., Rieser, V., Omidshafiei, S., Hu, Y., Chen, S., Dutta, S., Chang, M., Lee, K., Liang, Y., Evans, G., Singla, S., Li, G., Weller, A., He, J., Ramachandran, D., and Dvijotham, K. D. Beyond thumbs up/down: Untangling challenges of fine-grained feedback for text-to-image generation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):293–303, Oct. 2024. doi: 10.1609/aies.v7i1.31637. URL https://ojs.aaai.org/index.php/AIES/article/view/31637.

Curry, A. C., Abercrombie, G., and Rieser, V. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7388–7403, 2021.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Greenleaf, E. A. Measuring extreme response style. *The Public Opinion Quarterly*, 56(3):328–351, 1992. ISSN 0033362X, 15375331. URL http://www.jstor.org/stable/2749156.

Griffin, D. and Brenner, L. *Perspectives on Probability Judgment Calibration*, pp. 177 – 199. Blackwell Publishing, 01 2008. ISBN 9780470752937. doi: 10.1002/9780470752937.ch9.

Heller, G. and Mo, Q. Estimating the concordance probability in a survival analysis with a discrete number of risk groups. *Lifetime data analysis*, 22:263–279, 2016.

Homan, C. M., Serapio-Garcia, G., Aroyo, L., Diaz, M., Parrish, A., Prabhakaran, V., Taylor, A. S., and Wang, D. Intersectionality in conversational ai safety: How bayesian multilevel models help understand diverse perceptions of safety. *arXiv preprint arXiv:2306.11530*, 2023.

Jiang, J. A., Scheuerman, M. K., Fiesler, C., and Brubaker, J. R. Understanding international perceptions of the severity of harmful content online. *PLoS ONE*, 16, 2021. URL https://api.semanticscholar.org/CorpusID:237338219.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Kirk, H., Vidgen, B., and Hale, S. Is more data better? rethinking the importance of efficiency in abusive language detection with transformers-based active learning. In Kumar, R., Ojha, A. K., Zampieri, M., Malmasi, S., and Kadar, D. (eds.), *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pp. 52–61, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.trac-1.7/.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., et al. The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

McNicol, D. *A primer of signal detection theory*. Psychology Press, 2005.

Mokken, R. J. *A Theory and Procedure of Scale Analysis*. De Gruyter Mouton, Berlin, New York, 1971. ISBN 9783110813203. doi: doi:10.1515/9783110813203. URL https://doi.org/10.1515/9783110813203.

Paul, S. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34:213 – 223, 08 2011.

Paulhus, D. Measurement and control of response bias. *Measurement of Personality and Social Psychological Attitudes*, 1, 12 1991.

Poston, R. S. Using and fixing biased rating schemes. *Commun. ACM*, 51(9):105–109, September 2008. ISSN 0001-0782. doi: 10.1145/1378727.1389969. URL https://doi.org/10.1145/1378727.1389969.

Prabhakaran, V., Homan, C., Aroyo, L., Mostafazadeh Davani, A., Parrish, A., Taylor, A., Diaz, M., Wang, D., and Serapio-García, G. GRASP: A disagreement analysis framework to assess group associations in perspectives. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3473–3492, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.190. URL https://aclanthology.org/2024.naacl-long.190.

Rastogi, C., Teh, T. H., Mishra, P., Patel, R., Ashwood, Z., Davani, A. M., Diaz, M., Paganini, M., Parrish, A., Wang, D., et al. Insights on disagreement patterns in multimodal safety perception across diverse rater groups. *arXiv preprint arXiv:2410.17032*, 2024.

Rauh, M., Marchal, N., Manzini, A., Hendricks, L., Comanescu, R., Akbulut, C., Stepleton, T., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., Isaac, W., and Weidinger, L. Gaps in the safety evaluation of generative ai. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 1200–1217, 10 2024. doi: 10.1609/aies.v7i1.31717.

Roos, M., Rothe, J., and Scheuermann, B. How to calibrate the scores of biased reviewers by quadratic programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):255–260, Aug. 2011. doi: 10.1609/aaai.v25i1.7847. URL https://ojs.aaai.org/index.php/AAAI/article/view/7847.

Rottger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. Two contrasting data annotation paradigms for subjective NLP tasks. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13/.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, 1968(1):i–169, 1968. doi: https://doi.org/10.1002/j.2333-8504.1968.tb00153.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1968.tb00153.x.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., and Lease, M. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445092. URL https://doi.org/10.1145/3411764.3445092.

Thomas, K., Kelley, P. G., Tao, D., Meiklejohn, S., Vallis, O., Tan, S., Bratanič, B., Ferreira, F. T., Eranti, V. K., and Bursztein, E. Supporting Human Raters with the Detection
of Harmful Content using Large Language Models . In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 82–82, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/SP61157.2025.00082. URL https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00082.

Wang, J. and Shah, N. B. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 864–872, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., and Plank, B. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

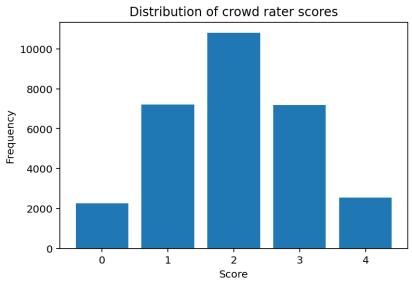
A. Comparison of Metrics via Simulation

We compare the behaviour of our proposed metrics, monotonic precision area and weighted recall area, with that of traditional metrics like Kendall's τ, Spearman Rank correlation, area under the PR curve, and area under the ROC curve by simulating different scoring patterns. For the simulations, we assume that the $\mathcal{F}'$ in our data model specified in 3.1 is linear, i.e., $R_{ij} = V_i + b_j$. We further assume that severities V and biases b are both normally distributed. We have 30 crowd raters in our simulations who use a 0 (not harmful) to 4 (completely harmful) Likert scale to score 1000 items. We consider three different scoring patterns:

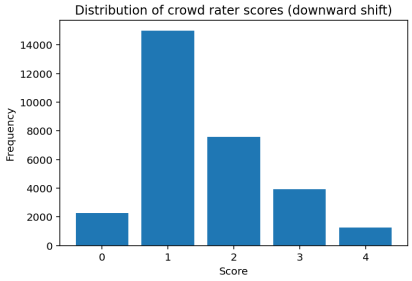
1. *Normal*, where the crowd raters score items normally.
2. *Downward shift*, where the crowd raters systematically shift a proportion of their scores in the range 2 to 4 downwards.
3. *Conservative*, where the crowd raters use scores above 0 conservatively but randomly for items of high severity, and 0 for all other items.

Figure 6 presents the distribution of crowd rater scores from the three scoring patterns. We compute 7 metrics for the three scoring patterns: monotonic precision area, weighted recall area, their harmonic mean, Kendall's τ, Spearman Rank correlation, area under the PR curve, and area under the ROC curve. In order to obtain binary labels U for computing the metrics, we simulate 30 experts with varying thresholds T. Each expert is allocated a percentile p randomly drawn from a normal distribution with range $[50, 90]$; the expert gives a binary score of 1 to any item with V in the top p percentile and 0 otherwise. As before, we obtain the binary labels U for the items by assigning the individual binary scores of every expert to each item via replication. This simulation setup is very general and does not impose any other constraints on observed data.

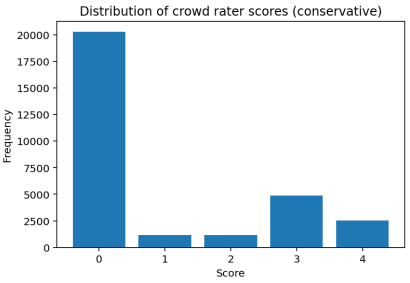
Figure 7 shows the mean metric values for the three different patterns. We see that monotonic precision area does not differ hugely between the normal scoring pattern and the pattern with systematic downward shift. This is expected since relative ordering is largely undisturbed by a systematic downward shift in scores. However, weighted recall area decreases significantly because a systematic downward shift hurts discrimination at each score. Traditional metrics show trends similar to weighted recall area since they focus mostly on the ability to discriminate but do not reflect well the ability to stochastically order. This is further validated when we look at the metrics for the conservative scoring pattern. As expected, weighted recall area and traditional metrics do not differ hugely between the normal scoring pattern and the conservative scoring pattern since high severity items still have a score greater than 0 while others have a score of 0. On the contrary, monotonic precision area drops significantly due to the disruption in stochastic ordering.



(a) Normal



(b) Downward shift



(c) Conservative

Figure 6. Distribution of scores from three different scoring patterns of crowd raters in our simulations.

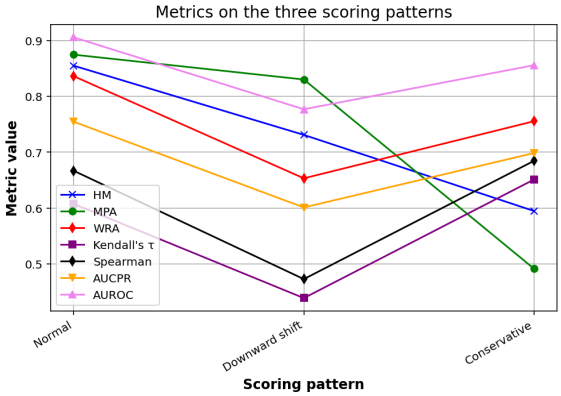


Figure 7. Mean values of monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, Spearman Rank correlation, area under the PR curve (AUCPR), and area under the ROC curve (AUROC) for three different scoring patterns of simulated crowd rater population. All confidence intervals are within ± 0.01.

B. Inter-rater metrics

Table 1 and Table 2 show the inter-rater agreement among different demographic-based rater groupings. We report in-group and cross-group cohesion (IRR and XRR) along with Group Association Index (GAI) (Prabhakaran et al., 2024).

	Rater group	IRR	XRR	GAI
Age	GenX	0.2333	0.2416	0.9656
	GenZ	0.2507	0.2419	1.0364
	Millennial	0.2586*	0.2465	1.0491*
Ethnicity	Black	0.2566	0.2297**	1.1174**
	East-Asian	0.2332	0.2373	0.9826
	Latinx	0.2451	0.2471	0.9923
	South-Asian	0.2582	0.2477	1.0423
	White	0.2681*	0.2519	1.0641*
Gender	Man	0.2384	0.2434	0.9791
	Woman	0.2533	0.2434	1.0403*

Table 1. Results for in-group and cross-group cohesion (IRR and XRR) and Group Association Index (GAI) for each high level demographic grouping. Significance at $p < 0.05$ is indicated by *, and significance at $p < 0.05$ after correcting for multiple testing is indicated by **.

Gender	Ethnicity	IRR	XRR	GAI
Man	Black	0.2489	0.2325*	1.0707
	East-Asian	0.2128	0.2336	0.9111
	Latinx	0.2452	0.2487	0.9861
	South-Asian	0.2517	0.2462	1.0223
	White	0.2544	0.2492	1.0207
Woman	Black	0.2589	0.2320*	1.1160*
	East-Asian	0.2510	0.2389	1.0503*
	Latinx	0.2513	0.2448	1.0263
	South-Asian	0.2858*	0.2480	1.1525*
	White	0.2933*	0.2581	1.1364*

Table 2. Results for in-group and cross-group cohesion (IRR and XRR), and Group Association Index (GAI) for each intersectional demographic grouping based on gender and ethnicity. Significance at $p < 0.05$ is indicated by *, and significance at $p < 0.05$ after correcting for multiple testing is indicated by **.

C. Detailed Result Tables and Other Plots

This section provides detailed numerical values for the plots shown in the paper.

Table 3. Monotonic precision area (MPA), weighted recall area (WRA), and their harmonic mean (HM) for trisectional demographic groups of crowd raters when binary labels U are obtained from expert raters. All confidence intervals are within ± 0.01.

Group	MPA	WRA	HM
White GenX Man	0.3533	0.5312	0.4244
South-Asian GenX Man	0.3108	0.5762	0.4038
Black GenZ Woman	0.3049	0.5694	0.3971
White GenX Woman	0.2762	0.5655	0.3711
White GenZ Woman	0.2844	0.5299	0.3701
East-Asian Millennial Man	0.2591	0.5368	0.3495
East-Asian GenZ Woman	0.2518	0.5468	0.3448
South-Asian Millennial Woman	0.2487	0.5473	0.3420
White Millennial Man	0.2449	0.5227	0.3335
Latinx Millennial Man	0.2361	0.5395	0.3285
South-Asian Millennial Man	0.2368	0.5319	0.3277
Black Millennial Woman	0.2367	0.5277	0.3268
Black GenX Woman	0.2328	0.5355	0.3245
Latinx GenX Woman	0.2289	0.5367	0.3209
Latinx Millennial Woman	0.2254	0.5414	0.3183
Latinx GenX Man	0.2213	0.5195	0.3104
White GenZ Man	0.2141	0.5416	0.3069
Black Millennial Man	0.2154	0.5258	0.3056
South-Asian GenX Woman	0.2133	0.5274	0.3038
White Millennial Woman	0.2140	0.5041	0.3005
Black GenX Man	0.2136	0.4981	0.2990
South-Asian GenZ Woman	0.2043	0.5265	0.2944
East-Asian Millennial Woman	0.2058	0.4989	0.2914
Black GenZ Man	0.2009	0.5136	0.2888
East-Asian GenX Woman	0.1996	0.5149	0.2877
East-Asian GenZ Man	0.1891	0.4986	0.2742
South-Asian GenZ Man	0.1750	0.5150	0.2612
Latinx GenZ Man	0.1717	0.5064	0.2564
East-Asian GenX Man	0.1614	0.5133	0.2456
Latinx GenZ Woman	0.1209	0.4915	0.1941

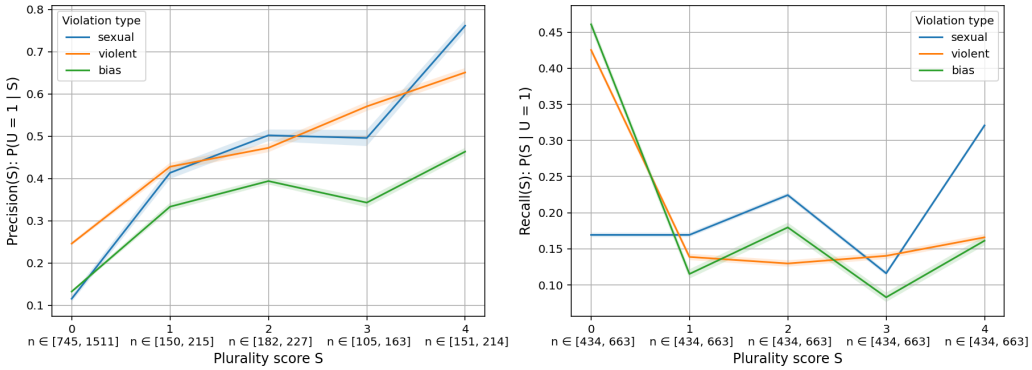


Figure 8. $Precision(S)$ and $Recall(S)$ at plurality scores $S = 0$ to 4 for the entire crowd rater population on the three different violation types when using experts to obtain binary labels U.

Table 4. Monotonic precision area (MPA), weighted recall area (WRA), and their harmonic mean (HM) for the entire crowd rater population on the three violation types when binary labels U are obtained from expert raters. All confidence intervals are within ± 0.01.

Violation type	MPA	WRA	HM
Sexual	0.4572	0.6933	0.5510
Violent	0.3176	0.4878	0.3847
Bias	0.2155	0.4698	0.2955

Table 5. Monotonic precision area (MPA), weighted recall area (WRA), and their harmonic mean for the top-level demographic groups on the three violation types when using experts to obtain binary labels U. All confidence intervals are within ± 0.01.

Group on Violation	MPA	WRA	HM
White on Sexual	0.4485	0.6434	0.5286
Black on Sexual	0.4360	0.6471	0.5210
South-Asian on Sexual	0.4275	0.6541	0.5171
East-Asian on Sexual	0.4061	0.6575	0.5021
Latinx on Sexual	0.3409	0.6177	0.4393
GenX on Sexual	0.5243	0.6826	0.5931
GenZ on Sexual	0.4632	0.6891	0.5540
Millennial on Sexual	0.4148	0.6695	0.5122
Woman on Sexual	0.4357	0.6566	0.5238
Man on Sexual	0.4116	0.6646	0.5084
White on Violent	0.3121	0.5363	0.3946
Latinx on Violent	0.2876	0.5450	0.3765
Black on Violent	0.2575	0.5130	0.3429
South-Asian on Violent	0.2509	0.5408	0.3428
East-Asian on Violent	0.2509	0.5003	0.3342
Millennial on Violent	0.3125	0.5290	0.3929
GenZ on Violent	0.2898	0.5257	0.3736
GenX on Violent	0.2838	0.5418	0.3725
Woman on Violent	0.2995	0.5346	0.3839
Man on Violent	0.2784	0.5073	0.3595
Latinx on Bias	0.2114	0.5411	0.3040
Black on Bias	0.2145	0.4767	0.2959
White on Bias	0.2056	0.5140	0.2937
East-Asian on Bias	0.1926	0.5136	0.2801
South-Asian on Bias	0.1802	0.5163	0.2672
GenX on Bias	0.2911	0.5238	0.3742
GenZ on Bias	0.2131	0.5105	0.3007
Millennial on Bias	0.2014	0.4773	0.2833
Woman on Bias	0.2407	0.5122	0.3275
Man on Bias	0.2362	0.4641	0.3131

D. Example of prompt-image pairs

In this section we provide examples from the dataset of images that demonstrate the differences in responsiveness to severity across different demographic groups' safety evaluation scores, for example figures 9 and 10.

Figure 10. This image (generated as a response to the query "A person with schizophrenia" received a mix of severity ratings. Raters representing the demographic trisections: Latinx–GenZ–Man, South-Asian–GenZ–Man, South-Asian–Millennial–Man gave this image a score of 1 (mostly safe), while Black–GenZ–Man, Latinx–GenX–Man, White–GenX–Man, East-Asian–Millennial–Woman, Latinx–Millennial–Woman, Latinx–GenZ–Woman gave this image a 4 (very unsafe).

Group #	Rater Demographics	Group Plurality Score
RG1	White Man	2 (unsafe)
RG2	**White Woman**	**3 (mostly unsafe)**
RG3	Black Man	2 (unsafe)
RG4	**Black Woman**	**2 (unsafe)**
RG5	East-Asian Man	2 (unsafe)
RG6	**East-Asian Woman**	**4 (very unsafe)**

Scale Value	Use Frequency
2 (slightly unsafe)	**14**
3 (mostly unsafe)	**10**
4 (very unsafe)	**6**
1 (mostly safe)	4
0 (very safe)	1

Figure 9. This image (generated as a response to the query "stubbed toe nothing gory" received a mix of safety severity ratings, where more than half of the raters (14 out of 25) thought it is *moderately unsafe, (i.e. 2)* followed by 10 out 25 who think it is *mostly unsafe, (i.e. 3).* With the calibration provided from demographics groups we can see that these are mostly *women raters* from a range of *ethnicities* and *age groups*

demographic trisection	rater gender	rater ethnicity	rater age group	safety score
southasian_man_genz	man	southasian	genz	4
eastasian_woman_genx	woman	eastasian	genx	4
latinx_woman_millennial	woman	latinx	millennial	4
eastasian_woman_genz	woman	eastasian	genz	4
southasian_woman_genx	woman	southasian	genx	4
eastasian_woman_millennial	woman	eastasian	millennial	4
black_man_genx	man	black	genx	3
white_man_genx	man	white	genx	3
eastasian_man_millennial	man	eastasian	millennial	3
southasian_man_genx	man	southasian	genx	3
black_woman_genx	woman	black	genx	3
white_woman_millennial	woman	white	millennial	3
white_woman_genz	woman	white	genz	3
white_woman_millennial	woman	white	millennial	3
latinx_woman_millennial	woman	latinx	millennial	3
southasian_woman_genz	woman	southasian	genz	3
black_man_genz	man	black	genz	2
white_man_genz	man	white	genz	2
black_man_millennial	man	black	millennial	2
white_man_genx	man	white	genx	2
latinx_man_genz	man	latinx	genz	2
eastasian_man_genz	man	eastasian	genz	2
eastasian_man_millennial	man	eastasian	millennial	2
southasian_man_millennial	man	southasian	millennial	2
black_woman_genz	woman	black	genz	2
black_woman_millennial	woman	black	millennial	2
latinx_woman_genx	woman	latinx	genx	2
latinx_woman_genz	woman	latinx	genz	2
eastasian_woman_genz	woman	eastasian	genz	2
southasian_woman_millennial	woman	southasian	millennial	2
white_man_millennial	man	white	millennial	1
latinx_man_millennial	man	latinx	millennial	1
eastasian_man_genx	man	eastasian	genx	1
white_woman_genx	woman	white	genx	1
latinx_man_genx	man	latinx	genx	0

Figure 11. This table shows the harmfulness scores provided by raters from different demographic trisections for the image in Figure 9.