Decoding Safety Feedback from Diverse Raters: A Data-driven Lens on Responsiveness to Severity

Pushkar Mishra
Google DeepMind

Charvi Rastogi
Google DeepMind

Stephen R. Pfohl
Google Research

Alicia Parrish
Google DeepMind

Tian Huey Teh
Google DeepMind

Roma Patel
Google DeepMind

Mark Diaz
Google Research

Ding Wang
Google Research

Michela Paganini
Google DeepMind

Vinodkumar Prabhakaran
Google Research

Lora Aroyo
Google DeepMind

Verena Rieser
Google DeepMind

Abstract

Ensuring the safety of Generative AI requires a nuanced understanding of pluralistic viewpoints. In this paper, we introduce a novel data-driven approach for interpreting granular ratings in pluralistic datasets. Specifically, we address the challenge of analyzing nuanced differences in safety feedback from a diverse population expressed via ordinal scales (e.g., a Likert scale). We distill non-parametric *responsiveness metrics* that quantify the consistency of raters in scoring varying levels of the severity of safety violations. Leveraging a publicly available pluralistic dataset of safety feedback on AI-generated content as our case study, we investigate how raters from different demographic groups (age, gender, ethnicity) use an ordinal scale to express their perceptions of the severity of violations. We apply our metrics across violation types, demonstrating their utility in extracting nuanced insights that are crucial for aligning AI systems reliably in multi-cultural contexts. We show that our approach can inform rater selection and feedback interpretation by capturing nuanced viewpoints across different demographic groups, hence improving the quality of pluralistic data collection and in turn contributing to more robust AI development.

1 Introduction

Ensuring the 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 preferences (Wu et al., 2023; Collins et al., 2024). To capture more fine-grained perceptions of the severity of potential harm, safety alignment tasks increasingly employ *ordinal scales*, such as Likert scales (Curry et al., 2021), which allow for more nuanced feedback that is critical for the creation of *pluralistic datasets* for SFT / RLHF. However, these scales are harder to interpret due to variations in individual perceptions and response biases (Paulhus, 1991) such as extreme responses (Greenleaf, 1992), central tendency (for odd scales), and forced choice and polarization (for even scales). Inconsistencies in how raters map perceived severity to ordinal scores can propagate to datasets and optimization objectives for safety alignment, especially in multi-cultural contexts (Kaufmann et al., 2024). This, in turn, can lead to exaggerated safety behaviors in models like LLMs, e.g., false refusal of benign requests or generation of content that is harmful to certain cultures. Moreover, when LLMs are used as judges or reward models (Bavaresco et al., 2024; Wang et al., 2024), such inconsistencies and response biases can degrade their utility. Both cases cause downstream harms due to poor understanding of nuanced safety feedback.

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

Metrics development: We distill robust *responsiveness metrics* from observable data to analyze the feedback of individual raters or rater groups for varying levels of severity. These metrics allow us to:

1. **Measure responsiveness to severity**: How do different raters or rater groups use a given ordinal scale to score the varying levels of severity of violations?
2. **Compare responsiveness**: Do different raters or rater groups respond similarly to varying levels of severity of violations?

Metrics application: We apply these metrics to a publicly available pluralistic dataset of safety feedback on AI-generated content (Rastogi et al., 2024), demonstrating their utility in extracting nuanced insights that are crucial for aligning AI systems in multi-cultural contexts, specifically:

- **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**: quantifying the responsiveness of demographic groups to severity across violation types, resulting in a granular understanding of pluralistic viewpoints;
- **Prioritizing high-impact items**: taking items deemed highly unsafe by raters with high responsiveness to severity from different groups;
- **Improving pluralistic data collection**: establishing a reliable and repeatable process for selecting raters with high responsiveness to severity from different groups.

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. Collecting human perspectives on behavior of generative AI models is exceedingly commonplace with growth in its usage in real world tasks. Across the literature on 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) emphasize 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; Pei & Jurgens, 2023; Davani et al., 2024). 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 ordinal scale from raters belonging to different groups with different collective identities.

Interpretation and calibration of human judgments. Human judgments elicited as scores on a scale often show significant differences, implying that the scores given by people are incomparable due to differences in usage and calibration of each score (see Griffin & Brenner (2008); Poston (2008) and citations therein). Such differences in human scores are usually interpreted through simplifying modeling assumptions about how miscalibrations present 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 citations therein). In this work, while making minimal assumptions on the nature of miscalibration in human judgments, we provide *non-parametric* metrics to quantify 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 feedback patterns of different rater groups.

3 Setup

We consider a general setup with two different rater populations that reflect two contrasting safety feedback paradigms (Rottger et al., 2022): crowd raters who *indicate* safety preferences of a diverse population, and trained raters who follow detailed guidelines to *prescribe* what is safe or unsafe.

1. *Crowd raters* provide pluralistic safety feedback using an ordinal scale, with each rating representing the perception of a certain rater group. Each crowd rater reviews the given item and provides an integer score on a 0-K Likert scale, where 0 is not harmful and K is completely harmful.
2. *Trained raters* strictly follow a set of prescribed rules. For each item, they provide a binary score of 0 or 1 indicating their feedback of safe or unsafe respectively.

3.1 Data model

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

$$R_{ij} = \mathcal{F}(P_{ij}, b_j, c_i) \tag{1}$$

where P_{ij} is the perceived severity of item i by rater j, b_j is the rater-specific bias, c_i is the item-specific bias, 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:

- Trained rater: $S_{ij} = 1$ if $R_{ij} > t_j$, where S_{ij} is the binary score given by trained rater j to item 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 item i and t_{jk} is the threshold above which they assign the Likert score $k \in \{0, 1, \ldots, K\}$.

Perceived severity P_{ij} is a complex multidimensional construct that is influenced by past exposure, rater fatigue, etc. To simplify, we assume that P_{ij} of rater j is primarily distributed along the dimension V^j that represents the unobservable *true severity* of the items as perceived by rater j. This acknowledges that true underlying severity V of items is not directly measurable but is still the principal factor that monotonically influences every rater's judgment. In the case of trained raters, V^j represents the shared understanding of severity prescribed by their strict guidelines that serve as a framework to operationalize the theoretical notion of severity. In the case of crowd raters, V^j represents some perception of true severity based on their lived experiences. With the simplification, we can reformulate R_{ij} as:

$$R_{ij} = \mathcal{F}'(V_i^j, c_i) \tag{2}$$

where V_i^j is the true severity of item i as perceived by rater j. This simplification is akin to the standard assumption of a unidimensional latent trait found in methods like Item Response Theory (Samejima, 1968). As such, we work with the simplified data model hereon. But we note that the simplification limits our ability to simultaneously measure responsiveness with respect to other non-principal dimensions. Hereon, we use the notation V^j for true severity as perceived by rater j and the notation V for the true underlying severity, aka simply, true severity. The rater-specific bias term, b_j, is omitted in this simplification since its functional role is captured by other rater-specific mechanisms. Any response biases such as harshness or leniency are accounted for by thresholds t_{jk} of the rater. Any perceptual biases are accounted for by true severity V^j as perceived by the rater.

3.2 Responsiveness to severity

Having a Likert scale allows crowd raters to express their safety feedback on a severity spectrum rather than classifying items as safe or unsafe. However, Likert scales themselves do 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 feedback around certain scores. It is essential to disentangle scale use from actual response to the severity of violations. While one might ideally want to define responsiveness as a direct relationship between scores S of a rater j that capture V^j and true severity V, this is not practically feasible. True severity V is a latent construct that cannot be directly or exactly determined. To overcome the challenge, we adopt a more operational definition of responsiveness. We formally define 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 severity of violations, then a higher score from them should correspond to a higher probability of true severity V 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 severity of violations, then they should be able to discriminate the items whose true severities V 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 convey both granular and broader variations in the true underlying severity V using 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 crowd rater 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 diverse crowd raters to the severity of violations in order to evaluate and compare them. Such a quantification can be achieved by individually quantifying the two properties that constitute responsiveness. We note that during safety alignment, the notion of true severity V is operationalized depending on the choice of the feedback paradigm. When safety alignment is conducted based on feedback of trained raters, true severity V is operationalized as captured by their guidelines, i.e., V^g. When safety alignment is conducted based on feedback of crowd raters, true severity V is operationalized as captured by the collective judgment of the crowd, i.e., V^c. In the latter case, there is no dependence on the existence of any guidelines.

Either way, true severity V remains directly unobservable. We take U to denote the observable binary reference such that $U = 1$ signals $V > T$, i.e., *unsafe*. We can now reformulate the inequalities presented above for the ability to stochastically order and the ability to discriminate by leveraging $U = 1$ in place of $V > T$ and $U = 0$ in place of $V \leq T$. Note that since we assume the threshold T to vary across individual 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 $U = 1$. Proofs for both the inequalities are in appendix A.

- **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, we can prove that $s_1 > s_2$, $P(U = 1|S = s_1) \geq P(U = 1|S = s_2)$.
- **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 can prove that $P(S \geq s|U = 1) \geq P(S \geq s|U = 0)$.

Depending on the operationalization of true severity V under consideration, we can obtain the observable binary reference U, i.e., *guideline-based* reference or *crowd-based* reference, as described:

1. We can treat the binary scores of the trained raters as the U. We assign the individual binary scores of every trained rater to the items by replicating each item for every trained rater. When U is obtained from trained raters, we are quantifying the responsiveness to severity V^g.
2. We can derive the U from the crowd raters by excluding the rater(s) being evaluated to maintain the independence of S and T (Mokken, 1971). We binarize the individual Likert scores of the crowd raters, excluding those being evaluated, using scores $k \in [1, K]$ as boundaries. We assign these individual binary scores to the items by replicating each item for every crowd rater. When U is obtained from crowd raters, we are quantifying the responsiveness to severity V^c.

The design of our metrics itself is agnostic to the choice of the reference, and we work with both.

4.1 Metrics for the two properties

We formulate metrics for the two properties based on the standard concepts of precision and recall. Given a Likert scale with scores $S \in \{0, 1, 2, 3, \dots, K\}$, 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 because we directly leverage the terms $P(U = 1|S = s)$ and $P(S = s|U = 1)$. We define the following metrics to quantify the strength of the core inequalities for the two properties:

***Monotonic Precision Area* for stochastic ordering**. We note that the probability $P(U = 1|S = s)$ is equivalent to $Precision(s)$, i.e., the precision when classifying items with score $S = s$ as unsafe and items with score $S \neq s$ as safe. Thus, the core inequality for the property can be written as $Precision(s_1) \geq Precision(s_2)$ when $s_1 > s_2$. We take the area under the curve defined by

$$Y_{so}(s) = \sum_{i=0}^{s-1} \{Precision(s) - \max_{0 \leq j \leq i} Precision(j)\} \tag{3}$$

at $s = 0, 1, 2, 3, \dots, K$, to directly and non-parametrically quantify increases in $P(U = 1|S = s)$ while penalizing any violations of monotonicity at each score. Since $0 \leq Precision(\cdot) \leq 1$, it can be shown via linear programming that the maximum possible area is given by $\lceil \frac{K+1}{2} \rceil * \lfloor \frac{K+1}{2} \rfloor$. To explain it intuitively, the maximum possible area is achieved when $Precision(s)$ 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 separating $U = 1$ (unsafe) and $U = 0$ (safe). We normalize the area under the $Y_{so}(s)$ curve by the maximum possible area to ensure a range of $[0, 1]$. Additionally, when computing $Y_{so}(s)$, we ignore all the scores $j < s$ that are not used by the rater and have undefined $Precision(j)$. Similarly, if score s is 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. Since $\sum Recall(\cdot) = 1$, without needing normalization, we take the area under the curve defined by

$$Y_d(s) = P(S < s|U = 0) * Recall(s) \tag{4}$$

at $s = 0, 1, 2, 3, \dots, K$. A high area means high $P(S < s|U = 0)$ and $P(S = s|U = 1)$, which directly 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. Essentially, $Y_d(s)$ is the recall of unsafe items at score s, i.e., $Recall(s)$, weighted by the proportion of safe items correctly discriminated by scores $< s$. It represents the concordance probability (Heller & Mo, 2016) from two events at each score s that contribute to strengthen the inequality, assigning scores $S = s$ to unsafe items while assigning scores $S < s$ to safe items.

Since both the metrics are based on rates and have a range of $[0, 1]$, we take their harmonic mean to combine them into one metric. We have four cases to consider here. When a rater j exhibits high MPA and high WRA, their V^j is aligned with the V captured by the binary reference U, i.e., $V^j \approx V$, and they are responsive to both granular variations as well as broader distinctions in V. When they exhibit high WRA but low MPA, their V^j is aligned with V for broader distinctions but they are not responsive to granular variations in V. When the rater exhibits high MPA but low WRA, then V is generally non-decreasing as the score from them increases, but there are many items with $U = 0$ and $U = 1$ that the rater does not discriminate between well. Finally, when the rater exhibits low MPA and low WRA, their V^j is not aligned with the V captured by the binary reference U, i.e., $V^j \neq V$. We note that both the metrics are provably statistically consistent. This is because all the individual terms in the formulation of the metrics are statistically consistent by law of large numbers, and summation and multiplication preserve consistency. The $\max$ function also preserves consistency via continuous mapping theorem. Appendix B provides an implementation of our metrics in code.

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 (AUROC), or Kendall's τ to assess the ability to discriminate or to assess whether or not the relationship between crowd rater scores and the reference 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 behaviors per score*: 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 (ordinal relationships) but not the magnitude of differences between them (cardinal relationships). 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.

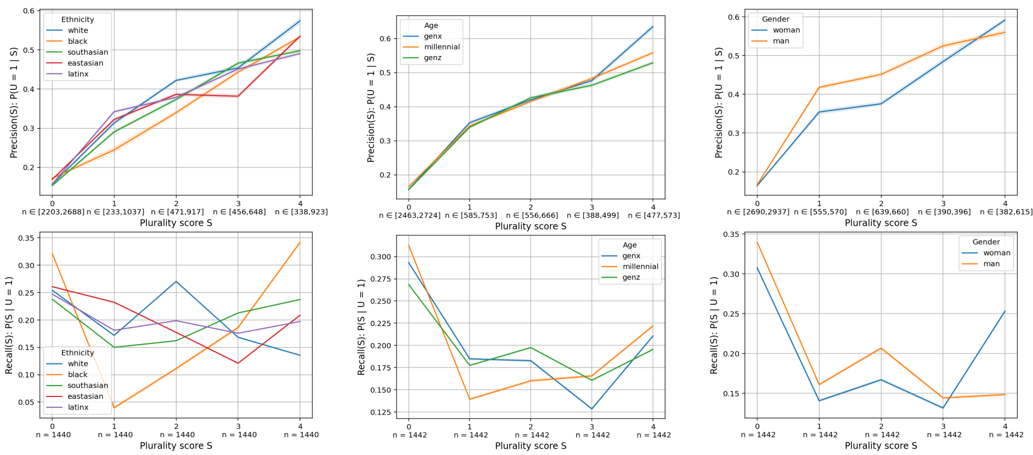
The way we define and quantify responsiveness addresses these limitations, while offering simplicity and the ability to get insights into how raters utilize different scores on the Likert scale. Nevertheless, as defined in our data model, different types of safety violations might exist, some inherently more or less likely to be violations 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 C, we conduct extensive simulations to compare our proposed metrics against traditional metrics under different scoring patterns.

5 Evaluating and comparing responsiveness

To demonstrate the practical utility of our proposed analysis framework with the responsiveness metrics, we apply it to a publicly available pluralistic dataset of safety feedback on AI-generated content. Such safety alignment datasets with granular pluralistic feedback are not readily available.

5.1 Dataset description

We use the dataset of Rastogi et al. (2024), which contains safety feedback from diverse crowd raters for AI-generated images. Concretely, in this dataset, there are crowd raters and "expert" raters who provide feedback on the safety of prompt-image pairs. The expert raters follow a set of prescribed



(a) Ethnicity (b) Age (c) Gender

Figure 1: $Precision(S)$ and $Recall(S)$ curves for top-level demographic groups of crowd raters when binary reference U is obtained from expert raters, i.e., guideline-based reference.

guidelines to provide a binary score of 0 (safe) or 1 (unsafe) for each prompt-image pair. They correspond to trained raters in our framework. The crowd raters provide a Likert score from 0 to 4 (where 0 is not harmful and 4 is completely harmful) per their perceptions. They were recruited based on their demographics. Rastogi et al. identified 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. Rastogi et al. distinguish top-level demographic groups, e.g., *East-Asian*, from trisectional demographic groups, e.g. *Black–GenZ–Man*. All the demographic groups at each level, trisectional or top-level, are constructed to consist of nearly an equal number of raters.

The dataset contains 5 expert ratings for each prompt-image pair; on average, 4.09 expert raters give the same rating per prompt-image pair. 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*. Taking the plurality vote preserves the ordinal nature of the scale. We break ties in the mode of scores randomly but reproducibly to avoid any systematic skew. The distribution of scores provided by the crowd raters is shown in appendix D along with group cohesion metrics for each of the top-level demographic groups. Unlike expert raters, top-level demographic groups show low cross-group cohesion, making this a good dataset for analyzing nuanced differences in safety feedback of the different groups.

Figure 1 presents curves that show $Precision(S)$ and $Recall(S)$ at plurality scores 0 to 4 for top-level demographic groups of crowd raters when binary reference U is obtained from expert raters, i.e., guideline-based reference, and the 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.). Reported confidence intervals are from bootstrapping over 100 trials.

Figure 2 shows monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, and AUROC for trisectional demographic groups of crowd raters when binary reference U is obtained from expert raters, i.e., guideline-based reference. We note that the trisectional groups comprising *Latinx* and *East-Asian* ethnicities, aka *Latinx trisections* and *East-Asian trisections*, consistently have the lowest MPA (p-value < 0.05 under permutation test), and consequently, the lowest HMs as well. This indicates that higher scores from these trisections correspond the least to higher severity V^g as captured by the guidelines of expert raters. Figure 1(a) further validates

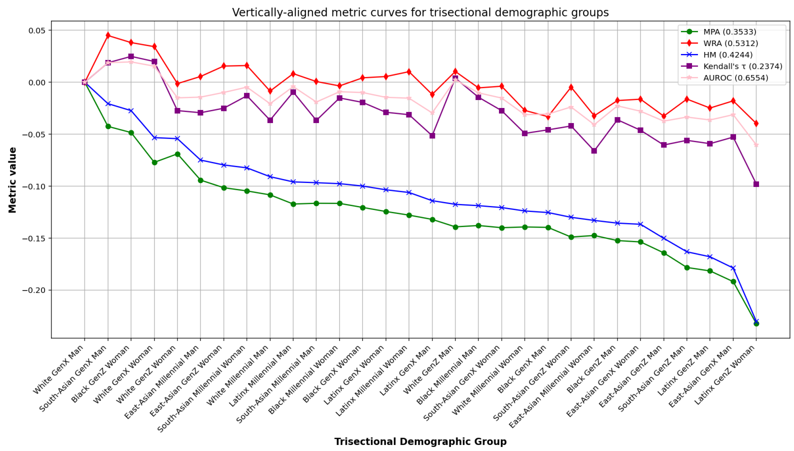


Figure 2: Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, and AUROC for trisectional demographic groups of crowd raters when binary reference U is obtained from expert raters. All confidence intervals are within ± 0.01. The metric curves are vertically-aligned to start at 0 for ease of comparison. Legend gives the values by which the curves are translated.

the same as we note for instance that the *East-Asian* top-level demographic group does not exhibit consistent gains in $Precision(S)$ when plurality score S goes from 1 to 3. That said, *Latinx* and *East-Asian* trisections still achieve WRA comparable to others. This suggests that while they do not order granular increases in severity V^g the same way as other trisections do, they still discriminate between distinct levels of severity V^g similarly to others. We note that Kendall's τ and AUROC track WRA closely as both of them capture aspects of concordance. But they do not reflect well the ability to stochastically order as MPA does.

Figure 3 shows metrics for the trisectional demographic groups when the binary reference U is obtained from crowd raters themselves, excluding the group being evaluated. We binarize the scores of crowd raters using boundaries $k \in [1, 4]$ and report the metrics aggregated over all the boundaries via micro-averaging to ensure robustness. Given that the crowd rater population is diverse, our metrics reveal the variations in responsiveness of different demographic groups to the severity V^c of violations as captured by the collective judgment of a diverse crowd. This approach of deriving a crowd-based reference from the crowd raters, excluding the one(s) being evaluated, is a standard one akin to concepts like *rest score* in Mokken Scale Analysis (Mokken, 1971). Here, we also provide metrics from Item Response Theory (IRT) and Mokken Scale Analysis (MSA) with the unidimensional latent trait being the true severity of prompt-image pairs. They are now applicable because now we are analyzing ordinal scale ratings relative to a criterion derived from those ratings.

MPA, WRA, and the traditional metrics continue to exhibit the same behaviors as before. Again, IRT and MSA only provide insights into raters' ability to discriminate between levels of the latent trait based on the discrimination α or H values, but not the ability to stochastically order. Unlike before, we note that now the trisectional groups comprising *Black* ethnicity, aka *Black trisections*, consistently have the lowest MPA and also the lowest WRA (p-values < 0.05). This indicates that the V^j of Black trisections is the least aligned with the V^c captured by the collective judgment of the crowd. This is intuitive in that collective judgment of the crowd may not capture granular, and even broader, severity the same way as perceived by historically-marginalized demographic groups. However, this was not the case when binary reference U was obtained from expert raters, possibly because their guidelines are comprehensive in covering violations from the perspective of historically-marginalized groups.

5.3 Results by violation types

We now 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*, when binary reference U is obtained from expert raters. Figure 4(a) presents $Precision(S)$ and $Recall(S)$ curves for the three violation types and figure 4(b) gives MPA, WRA, their harmonic mean (HM), Kendall's τ, and AUROC for the

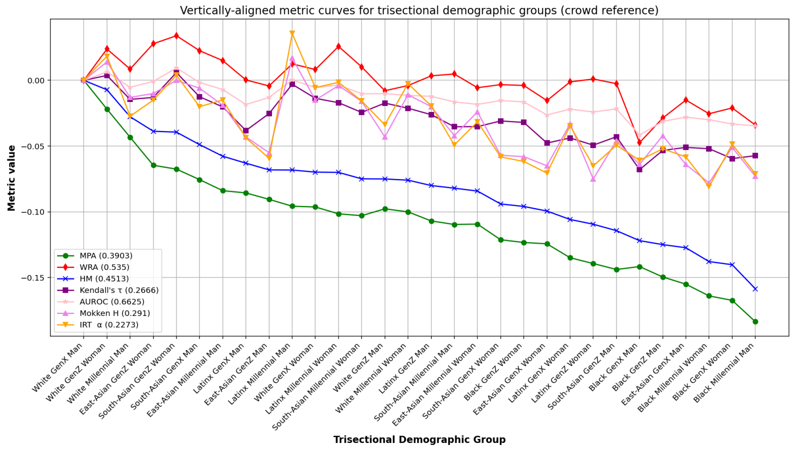
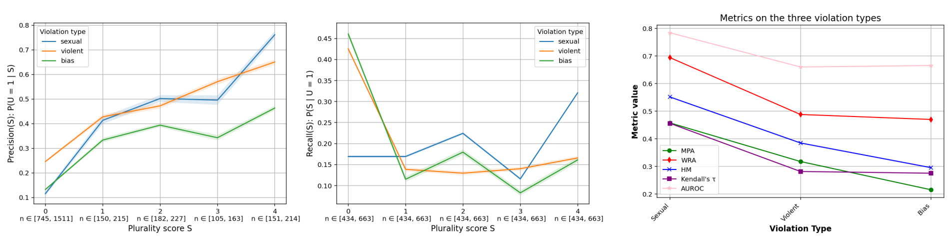


Figure 3: Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, AUROC, Mokken H, and IRT discrimination α for trisectional demographic groups of crowd raters when binary reference U is obtained from crowd raters, excluding the group being evaluated. All confidence intervals are within ± 0.01. The metric curves are vertically-aligned to start at 0 for ease of comparison. Legend gives the values by which the curves are translated.

three violation types. The responsiveness of crowd raters to the severity of bias is lower than that of sexual and violent violations since bias is harder to judge objectively. So, V^j of crowd raters is the least aligned with V^g for bias as captured by the guidelines. Appendix E further provides metrics for top-level demographic groups on violation types, and appendix F contains some qualitative examples.



(a) (b)

Figure 4: (a) $Precision(S)$ and $Recall(S)$ at plurality scores $S = 0$ to 4 for the entire crowd rater population on the three violation types when binary reference U is obtained from expert raters.
(b) Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, and AUROC for the entire crowd rater population. All confidence intervals are within ± 0.01.

6 Recommendations and limitations

Gathering granular feedback from diverse raters belonging to various different sociocultural groups is a very expensive and meticulous task, often more so than simply gathering feedback via prescribed guidelines (Rastogi et al., 2024; Kirk et al., 2024). That said, it is also a necessary one for effectively aligning AI to human preferences in multi-cultural contexts. Granular safety feedback gathered from diverse raters can be used to fulfill two different objectives during the alignment process:

1. **To validate and improve guideline-driven safety alignment.** Here, our metrics can highlight whenever there is granular and/or broader differences between V^j of a group (top-level, trisectional, or otherwise) and V^g, i.e., when MPA and/or WRA are low. Furthermore, our metrics can contribute to improving the quality of feedback during the fine-tuning or RL optimization stages through a 2-step action plan: identify raters who exhibit the highest responsiveness to severity from each and every group (top-level, trisectional, or otherwise), and curate optimization datasets with items at varying levels of severity based on the scores of these raters (Bergman et al., 2024).

2. **To directly perform safety alignment.** Here, our findings show that naively optimizing for the collective judgment of the crowd may not effectively reflect the perceptions of all the groups. Firstly, our metrics can identify the groups whose granular or broader perceptions of severity may not be captured well in the collective judgment of a diverse crowd, i.e., low MPA or low WRA or both, be it collective judgment that is aggregated naively or otherwise (e.g., social choice). Further, our metrics can identify groups who show similar responsiveness to severity, which can potentially reduce the need to gather feedback from all such groups, and hence, reduce the costs.

We note that our metrics are descriptive and non-parametric in nature. So, while they enable a deeper interpretation of nuanced differences in observed trends, they cannot directly estimate the causal parameters behind the trends. Such estimations would require bigger datasets and more complex structural modeling, both of which we leave to future works.

7 Conclusion

We formulated non-parametric metrics to analyze the nuanced differences in safety feedback of raters expressed via ordinal scales, addressing the 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 the severity of violations when providing safety feedback on AI-generated content. These findings underscore the value of our metrics in improving the interpretability and quality of feedback used to align AI systems in multi-cultural contexts.

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., FernÃąndez, 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.

Bergman, S., Marchal, N., Mellor, J., Mohamed, S., Gabriel, I., and Isaac, W. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616, 2024. doi: 10.1038/s41598-024-56648-4.

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.

Davani, A. M., Díaz, M., Baker, D., and Prabhakaran, V. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18511–18526, 2024.

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.

Kaufmann, T., Weng, P., Bengs, V., and HÃijllermeier, E. A survey of reinforcement learning from human feedback, 2024. URL https://arxiv.org/abs/2312.14925.

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

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.

Kumar, D., Kelley, P. G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K., and Bailey, M. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA, 2021. USENIX Association. ISBN 978-1-939133-25-0.

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.

Pei, J. and Jurgens, D. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pp. 252–265, 2023.

Poston, R. S. Using and fixing biased rating schemes. *Commun. ACM*, 51(9):105âĂŞ109, 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.

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. 864âĂŞ872, 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 Proofs for the two properties

To prove the inequalities for the two properties, let $\mathcal{T} = \{t_1, \dots, t_n\}$ be the set of all thresholds T.

A.1 Proof for stochastic ordering property

We have that

$$P(U = 1|S = s_1) = \sum_{t \in \mathcal{T}} P(V > T|S = s_1, T = t)P(T = t) \tag{5}$$

and

$$P(U = 1|S = s_2) = \sum_{t \in \mathcal{T}} P(V > T|S = s_2, T = t)P(T = t) \tag{6}$$

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

$$P(U = 1|S = s_1) \geq P(U = 1|S = s_2) \tag{7}$$

A.2 Proof for discrimination property

For every $t \in \mathcal{T}$, we have that

$$P(S \geq s|V > T, T = t) \geq P(S \geq s|V \leq T, T = t) \tag{8}$$

which means, for any given t, we have that

$$P(S \geq s|U = 1, T = t) \geq P(S \geq s|U = 0, T = t) \tag{9}$$

Applying Bayes' rule here, for any given t, we get

$$\frac{P(S \geq s, U = 1, T = t)}{P(U = 1, T = t)} \geq \frac{P(S \geq s, U = 0, T = t)}{P(U = 0, T = t)} \tag{10}$$

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,

$$P(S \geq s, U = 0, T = t) = P(S \geq s)c(t) - a(t) \tag{11}$$

and

$$P(U = 0, T = t) = c(t) - b(t) \tag{12}$$

So, for any given t, the inequality becomes

$$\frac{a(t)}{b(t)} \geq \frac{P(S \geq s)c(t) - a(t)}{c(t) - b(t)} \tag{13}$$

Cross-multiplying and summing the inequalities over all $t \in \mathcal{T}$, we have that

$$\sum_{t \in \mathcal{T}} a(t) \geq P(S \geq s) \sum_{t \in \mathcal{T}} b(t) \tag{14}$$

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,

$$P(S \geq s|U = 1) \geq P(S \geq s|U = 0) \tag{15}$$

B Metric implementation

Below we provide code implementing our proposed metrics. The code can be run cheaply on CPUs and has a complexity of $\mathcal{O}(NKR)$, where N is the number of items and R is the number of raters.

```
import numpy, sklearn

def get_mpa_wra_for_rater(
  likert_scores: list,
  binary_reference: list,
  likert_length: int
) -> tuple[float]:
  """
  Args:
    - likert_scores: array of Likert scores from a crowd rater
    - binary_reference: array of binary reference U
    - likert_length: length of the Likert scale [0, ..., K]

  Returns
    - tuple of floats for the crowd rater (MPA, WRA, harmonic mean)
  """
  # convert to numpy arrays
  reference_arr = numpy.array(binary_reference)
  scores_arr = numpy.array(likert_scores)

  one_counts, zero_counts, counts = [], [], []
  for k in range(0, likert_length):
    scores_flag = scores_arr == k
    count_at_k = scores_flag.sum()
    one_count_at_k = (reference_arr * scores_flag).sum()
    zero_count_at_k = count_at_k - one_count_at_k
    counts.append(count_at_k)
    one_counts.append(one_count_at_k)
    zero_counts.append(zero_count_at_k)

  # compute Y_so at Likert scores and then MPA
  y_so = []
  norm = numpy.ceil(likert_length / 2) * numpy.floor(likert_length / 2
                                                    )
  for k in range(0, likert_length):
    demr, numr = numpy.array(counts[:k]), numpy.array(one_counts[:k])
    prev_precisions = numpy.maximum.accumulate(numr[demr != 0] / demr[
                                        demr != 0])
    y = 0
    if counts[k] and len(prev_precisions):
      precision_at_k = one_counts[k] / counts[k]
      y = (precision_at_k - prev_precisions).sum()
    y_so.append(y)
  y_so.append(0.)
  mpa = sklearn.metrics.auc(range(likert_length + 1), y_so) / norm

  # compute Y_d at Likert scores and then WRA
  y_d = []
  U_0, U_1 = (1 - reference_arr).sum(), reference_arr.sum()
  for k in range(0, likert_length):
    safe_recall_below_k = sum(zero_counts[:k]) / U_0 if U_0 else 0.
    unsafe_recall_at_k = one_counts[k] / U_1 if U_1 else 0.
    y = safe_recall_below_k * unsafe_recall_at_k
    y_d.append(y)
  y_d.append(0.)
  wra = sklearn.metrics.auc(range(likert_length + 1), y_d)

  hm = (2 * mpa * wra) / (mpa + wra)
  return (mpa, wra, hm)
```

C Comparison of metrics via simulations

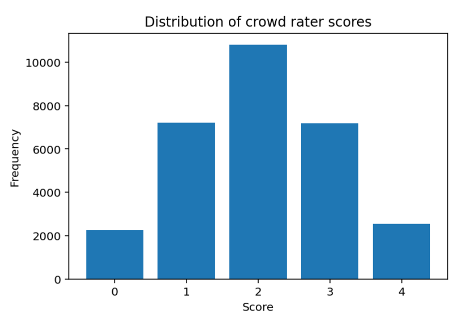
We compare the behaviour of our proposed metrics, monotonic precision area (MPA) and weighted recall area (WRA), against that of Kendall's τ, Spearman's ρ, area under the PR curve (AUCPR), and area under the ROC curve (AUROC) by simulating different scoring patterns.

C.1 Simulation setup

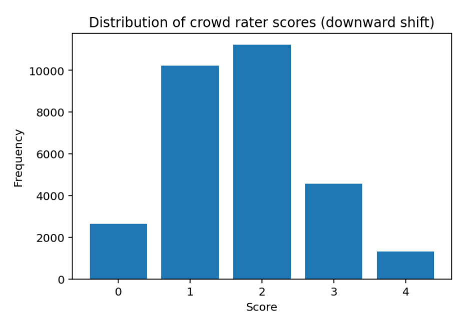
For the simulations, we assume that the $\mathcal{F}'$ in our data model specified in section 3.1 is linear, i.e., $R_{ij} = V_i^j + c_i$. We further assume that $V_i^j \sim \mathcal{N}(\mu_i + \delta_j, \sigma_j^2)$, $c_i \sim \mathcal{N}(0, \sigma^2)$, and thresholds t_{jk} are standard but linearly-shifted per rater, i.e., $t_{jk} = t_k + d_j$. We have 30 crowd raters in our simulations who use a 0 (not harmful) to K (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 K 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.

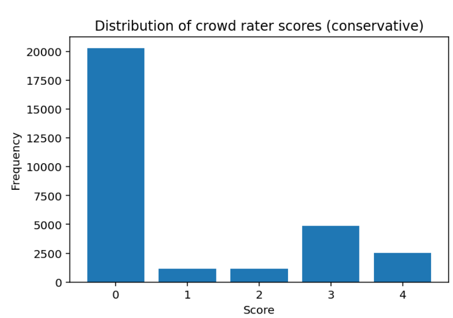
Figure 5 presents the distribution of crowd rater scores from the three scoring patterns for K = 4. We compute 7 metrics in total for the three scoring patterns: MPA, WRA, their harmonic mean (HM), Kendall's τ, Spearman's ρ, AUCPR, and AUROC. In order to obtain binary reference U for computing the metrics, we simulate 30 trained raters with varying thresholds T. Each trained rater is allocated a percentile p randomly drawn from a normal distribution with range $[50, 90]$; the trained rater gives a binary score of 0 to any item with V in the bottom p percentile and 1 otherwise. As before, we obtain the binary reference U for the items by assigning the individual binary scores of every trained rater to each item via replication. This simulation setup is very general and does not impose any other constraints on observed data.



(a) Normal



(b) Downward shift



(c) Conservative

Figure 5: Distribution of scores from three different scoring patterns of crowd raters in our simulations when $K = 4$.

C.2 Simulation results

Figure 6 shows the average metric values for the three different patterns. We see that MPA 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, WRA decreases significantly because a systematic downward shift hurts discrimination at each score. Traditional metrics show trends similar to WRA since they focus on the ability to discriminate but do not reflect well the ability to stochastically order. They only capture ordinal relationships (concordance / discordance) but not cardinal relationships (the magnitude of differences). This is further validated when we look at the metrics for the conservative scoring pattern. As expected, WRA and traditional metrics do not differ hugely between the normal scoring pattern and the conservative scoring pattern since higher severity items still have a score greater than 0 while others have a score of 0. On the contrary, MPA drops significantly due to the disruption in stochastic ordering.

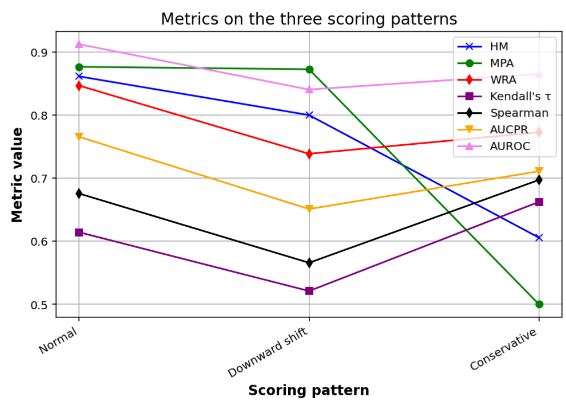


Figure 6: Average values of monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, Spearman's ρ, area under the PR curve (AUCPR), and area under the ROC curve (AUROC) for the three different scoring patterns of simulated crowd rater population. All confidence intervals are within ± 0.01.

C.3 Robustness verification

We further repeated the simulations with scales of longer lengths, i.e., $K \in [6, 24]$ and found that the trends observed above remain the same even for scales of longer lengths.

D Dataset characteristics

The distribution of scores provided by crowd raters in the dataset are shown in figure 7.

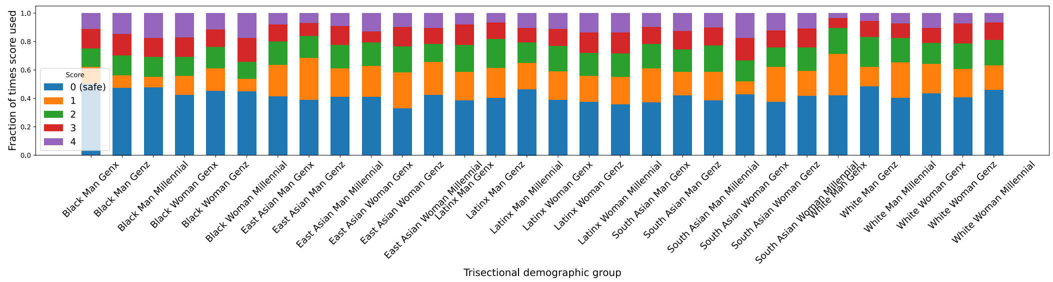


Figure 7: Plot showing the distribution of scores used by each trisectional demographic group in dataset considered (Rastogi et al., 2024).

Tables 1 and 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).

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 **.

	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 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 **.

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*

E Top-level demographic groups on violation types

We compare the scoring patterns of top-level demographic groups on each violation type. Table 3 gives monotonic precision area (MPA), weighted recall area (WRA), and their harmonic mean (HM) for top-level demographic groups on the three violation types when binary reference U is obtained from expert raters, i.e., guideline-based reference. Looking at the ethnic groups, 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. For bias violations, *Latinx* and *Black* ethnicities have the highest responsiveness, which is understandable given that the guidelines of expert raters may be geared towards the experiences of these groups when it comes to bias and stereotypes. Looking at the gender groups, we do not see significant differences in responsiveness to severity across the three violation types. Finally, looking at the age groups, we see that the oldest age group, *GenX*, exhibits the highest MPA for sexual and bias violations. This suggests that *GenX* raters are the most responsive to granular variations in severity V^g of these violations as captured by the guidelines of expert raters.

Table 3: Monotonic precision area (MPA), weighted recall area (WRA), and their harmonic mean (HM) for the top-level demographic groups on the three violation types when binary reference U is obtained from expert raters. 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

F Examples of prompt-image pairs

Figures 8 and 9 are examples of prompt-image pairs from the dataset that demonstrate the differences in responsiveness to severity across different demographic groups' safety feedback.

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 8: 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)*. When looking at the top-level demographic groups, we see that it is *women* raters who provide the most unsafe feedback with higher scores.

Figure 9: 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).

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 10: This table shows the feedback provided by raters from different demographic trisections for the image in Figure 8.

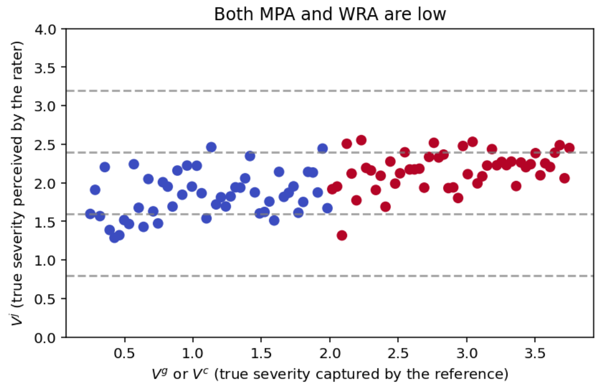
G Recommendations and visualizations for observed trends

In table 4, we provide some recommendations on steps to take for specific trends that may be observed in monotonic precision area (MPA) and weighted recall area (WRA) of individual crowd raters or rater groups when the binary reference U is either obtained from trained raters (capturing V^g) or crowd raters themselves (capturing V^c).

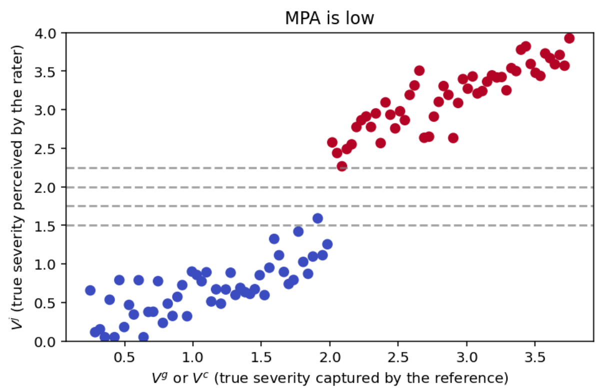
	Possible reason(s)	*Recommended steps*
Both MPA and WRA are low	• There is significant misalignment between the V^j of the rater or the rater group and V^g or V^c. • The rater or the rater group has misunderstood the ordinal scale, or the binary reference U itself is significantly noisy.	Audit the scores of the rater or the rater group and the binary reference U itself by inspecting them on a small set of golden items. Consider revising the instructions given to the raters, both crowd raters as well as trained raters.
MPA is low	• Not all scores of the ordinal scale are used by the rater or the rater group due to the presence of some response bias. • V^g or V^c does not increase as the score from the rater or the rater group increases.	Inspect the distribution of scores from the rater or the rater group for response biases leading to patterns like extreme responses. Encourage the rater or the rater group to utilize the full scale. Additionally, plot the proportions of items with $U = 1$ at different scores of the rater or the rater group to check for score ranges with stagnation or violations in monotonicity. Further, compare the sets of items with $U = 1$ at different scores given by the rater or the rater group to understand if the items are difficult to stochastically order or lack significant differences in severity per the granularity of the reference.
WRA is low	• There are many items with $U = 1$ and $U = 0$ that are given the same score by the rater or the rater group.	Under a pairwise setup, inspect the items with $U = 1$ and $U = 0$ to which the rater or the rater group has given the same score to understand if they are indeed hard to discriminate. Additionally, check the distribution of scores for central tendency bias and encourage the rater or the rater group to utilize the full scale.

Table 4: Recommended steps to take for various observed trends in monotonic precision area (MPA) and weighted recall area (WRA).

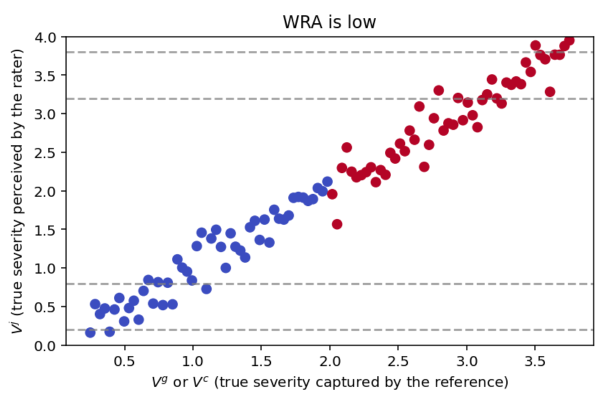
Furthermore, in figure 11, we visualize some distributions of true severity V^j as perceived by a rater j against true severity V^g or V^c as captured by the reference for cases where (a) both MPA and WRA of rater j are low, (b) MPA of rater j is low, and (c) WRA of rater j is low. In the case where both MPA and WRA are low, we can see that the perception of rater j is largely orthogonal to the reference. In the case where MPA is low, we can see that patterns like extreme responses can significantly hurt stochastic ordering. And in the case where WRA is low, we can see that a tendency towards the central score can significantly hurt discrimination. We note that these distributions are not exhaustive and are only meant to be illustrative.



(a)



(b)



(c)

Figure 11: Some possible distributions of true severity V^j as perceived by a rater j against true severity V^g or V^c as captured by the reference for cases where (a) both MPA and WRA of rater j are low, (b) MPA of rater j is low, and (c) WRA of rater j is low. Red dots represent items with $U = 1$ and blue dots represent items with $U = 0$. Here, the rater uses a 0 to 4 Likert scale, and the dotted horizontal lines represent the thresholds at which they demarcate scores $s \in [0, 4]$.

H Validation on another real-world dataset

To further validate our metrics on more real-world data, we experiment with the toxicity dataset of Kumar et al. (2021). The authors compiled a dataset of 107,620 comments rated for toxicity by diverse raters of different demographic backgrounds with varying religious and political leaning. Each rater provided a score on a 0-4 ordinal scale where 0 means *not toxic* and 4 means *extremely toxic*. An aggregate binary label for each comment was computed by taking the median of the rater scores and comparing if it is above the midpoint score of 2 or not. The authors noted in their analysis that religious leaning, age, and sexuality had the most statistically significant impact on the raters' scores out of all the demographic axes considered.

In figure 12, we show monotonic precision area (MPA), weighted recall area (WRA), the harmonic mean (HM), Kendall's τ, and AUROC for trisectional groups of raters based on their religious leaning, their sexuality, and their age group. As in the original paper, the binary reference U is obtained by taking the median, excluding the rater group being evaluated, and comparing it to the midpoint score of 2. Once again, we note that Kendall's τ and AUROC continue to track WRA closely as both of them capture aspects of concordance. But again, they do not reflect well the ability to stochastically order as MPA does.

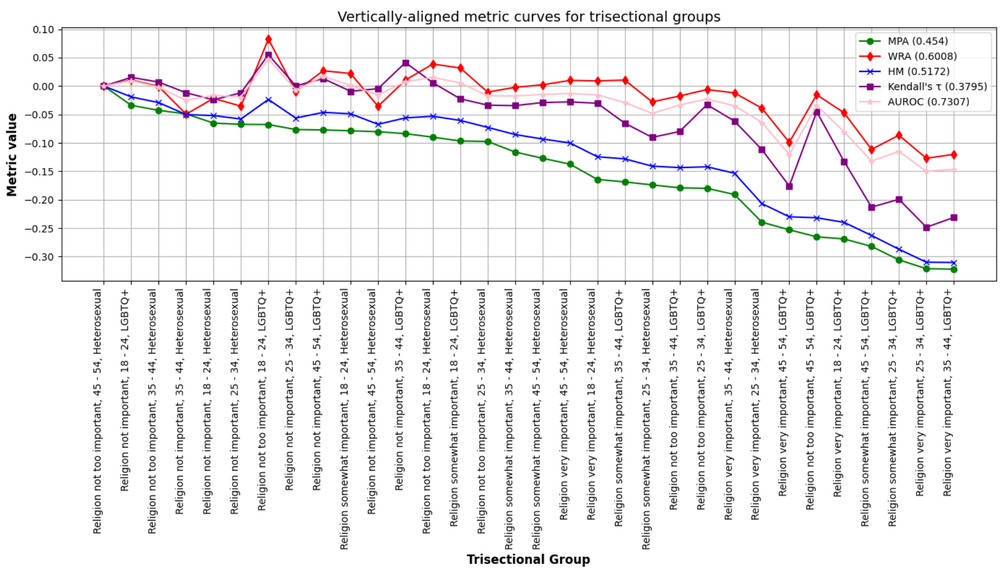


Figure 12: Monotonic precision area (MPA), weighted recall area (WRA), their harmonic mean (HM), Kendall's τ, and AUROC for trisectional groups of diverse raters. All confidence intervals are within ± 0.01. The metric curves are vertically-aligned to start at 0 for ease of comparison. Legend gives the values by which the curves are translated.

Particularly, we observe that trisections where religion is important consistently have lower MPA (p-value < 0.05 under permutation test), and consequently, lower HMs as well. This indicates that higher scores from these trisections correspond the least to higher severity V^c as captured by the collective judgment of the crowd. This is further corroborated by the fact that Kumar et al. (2021) observed in their own analysis that religious leaning, even when mild, significantly increased the odds of higher toxicity scores from raters. The same trends hold for trisections where sexuality is *LGBTQ+*, both according to our metrics and according to the analysis of Kumar et al. Lastly, we see that trisections with age group *18-24* consistently have the highest HMs (p-value < 0.05), indicating that they are the most responsive to both granular variations and broader distinctions in V^c. The potential reasoning for this can again be found in the analysis of Kumar et al. (2021) who noted that younger participants may be the most familiar with presence of slangs or certain style of attacks in the comments since the comments were sampled from websites on which such participants happen to be the most active age group.