Bias and Reliability in AI Safety Assessment: Multi-Facet Rasch Analysis of Human Moderators

Chunling Niu[1], Kelly Bradley[2], Biao Ma[1], Brian Waltman[1], Loren Cossette[1], & Rui Jin[3]

[1]University of the Incarnate Word
[2]University of Kentucky
[3]Shenzhen University, P. R. China

Abstract

Using Multi-Facet Rasch Modeling on 36,400 safety ratings of AI-generated conversations, we reveal significant racial disparities (Asian: 39.1%, White: 28.7% detection rates) and content-specific bias patterns. Simulations show that diverse teams of 8-10 members achieve over 70% reliability versus 62% for smaller homogeneous teams, providing preliminary evidence-based guidelines for AI-generated content moderation.

1 Background

As conversational AI systems proliferate, ensuring reliable human evaluation of AI-generated content safety becomes critical. Modern generative AI systems like LaMDA rely heavily on human judgment to assess response safety, particularly for nuanced content requiring contextual understanding. However, the demographic composition of evaluation teams and its impact on AI safety assessment remains understudied.

Recent work documents bias in content moderation (Aroyo et al., 2023; Goyal et al., 2022), but few studies examine how rater demographics affect evaluation of AI-generated conversational content specifically. This distinction matters because AI-generated conversations present unique challenges: subtle harmful content, contextual nuances, and adversarial prompting designed to elicit unsafe responses.

The Diversity in Conversational AI Evaluation for Safety (DICES) dataset (Aroyo et al., 2023) established foundations for understanding demographic effects in AI safety evaluation but lacks detailed bias analysis or reliability optimization guidelines. Prior research shows significant demographic disparities in toxicity ratings, particularly affecting African American and LGBTQ populations (Goyal et al., 2022), yet the interaction between rater demographics and AI conversation characteristics remains unexplored.

2 Aims

This study addresses three critical research questions:

1. **Quantify human rater disparities**: Do significant demographic differences exist in safety detection rates for AI-generated conversations, and what is their magnitude?
2. **Identify content-specific patterns**: How do demographic bias patterns vary across different AI conversation topics (health, political, legal, racial content, etc.)?
3. **Optimize rater team composition**: What rater team configurations achieve optimal reliability while maintaining demographic diversity for AI safety evaluation?

We employ Multi-Facet Rasch Modeling (MFRM) to simultaneously model rater, conversation types, and demographic effects while providing actionable guidelines for assembling effective AI-generated content safety evaluation teams.

3 Sample(s)

We analyze the DICES-350 dataset (Aroyo et al., 2023) containing safety evaluations of 350

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 393–397
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

adversarial human-AI conversations generated using Google's LaMDA. It is a relatively new dataset from NeurIPS that is gaining traction in the field. The dataset includes:

3.1 Conversations

Three hundred and fifty (n=350) multi-turn interactions were created as the corpus data by human agents instructed to generate adversarial prompts designed to elicit unsafe responses. Conversations span health (8%), political (18%), racial (25%), gender/sexual (14%), legal (3%), violence (1%), and miscellaneous (30%) topics. Expert annotations indicate 40% benign, 20% debatable, 20% moderate, and 20% extreme harm levels.

3.2 Raters

One hundred and four (n=104) demographically diverse raters provided 36,400 total safety judgments. Demographics were consolidated for statistical power based on initial exploratory analysis results:

- **Race**: Asian (n=21), White (n=25), Other races (n=58, including Black/African American, Latin X, Latino, Hispanic or Spanish Origin, and Multiracial)
- **Age**: GenZ 18-24 (n=49), Millennial 25-34 (n=28), GenX+ 35+ (n=27)
- **Gender**: Male (n=47), Female (n=57)

3.3 Ratings

Granular safety assessments were collected from the raters across conversation legibility, harmful content (8 sub-questions), unfair bias (4 sub-questions), misinformation, political affiliation, and policy violations using three-point scales (*No/Unsure/Yes*).

4 Methods

4.1 Multi-Facet Rasch Analysis

We implemented MFRM using generalized linear mixed models with logistic regression:

$$logit\left(P(unsafe_{rating} = 1)\right) = \beta^0 + \beta^1(race) + \beta^2(content) + \beta^3(gender) + \beta^4(race \times content) + \beta^5(gender \times content) + (1|rater) + (1|conversation) \quad (1)$$

This simultaneously estimates conversation difficulty (random effect), rater severity (random effect), racial bias (fixed effects), content-specific bias (interaction terms), and gender effects. Model fitted using maximum likelihood estimation in R (lme4 package).

4.2 Empirical Reliability Simulation

We developed bootstrap simulation using real rating patterns:

1. *Sample teams from actual demographic distributions (3-10 members)*
2. *Calculate pairwise reliability for multiply rated conversations*
3. *Estimate consensus via majority vote aggregation*
4. *Bootstrap replicates across 500 iterations for stable estimates*

We tested 12 team configurations across four content types, simulating realistic AI safety evaluation scenarios.

5 Results

5.1 Demographic Effects

Our MFRM analysis reveals differential patterns across demographic groups in safety detection of AI-generated conversational content. The model achieved excellent fit (AIC: 9,716.9, BIC: 9,933.2) with successful convergence across all parameters.

Primary Demographic Effects: While we analyzed race, age, and gender effects simultaneously, racial differences emerged as the most substantial and consistent predictor of safety detection patterns. Age effects were modest (GenZ vs Millennial: $\beta = -0.12$, $p = 0.67$; GenX+ vs Millennial: $\beta = +0.08$, $p = 0.78$), and gender effects were non-significant (Male vs Female: $\beta = -0.03$, $p = 0.85$). Based on these preliminary findings and space constraints, we focus our detailed analysis on racial bias patterns, which showed the strongest effects and clearest interaction patterns with content types.

Figure 1 shows the box plots of rater severity (in logits) across 28 demographic subgroups defined by combinations of race/ethnicity, gender, and age, revealing systematic differences in how different demo groups evaluate AI-generated content.

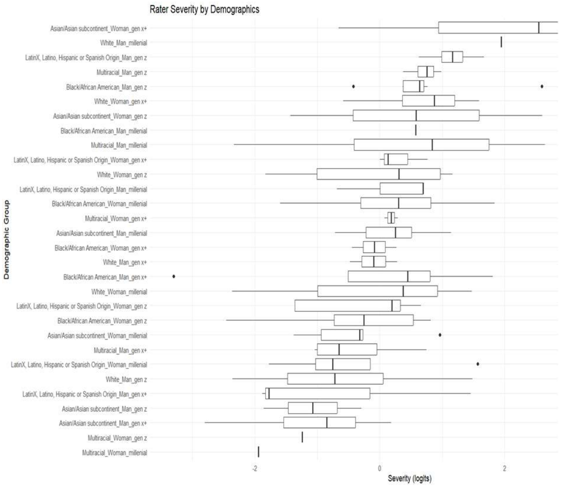

Figure 1: Rater severity by rater demographics

Safety Detection Rate Disparities: Asian raters demonstrated the highest safety detection rates at 39.1%, followed by Other races at 33.9%, and White raters at 28.7%. These differences represent substantial effect sizes, with Asian raters being 36% more likely to identify safety concerns in AI-generated conversations compared to White raters, and 15% more likely than Other race raters.

Figure 2 below shows harm detection rates across various content categories by racial group, revealing substantial variation in detection patterns both within and across demographic groups, with notably higher detection rates for certain harm types like legal issues and violent content.

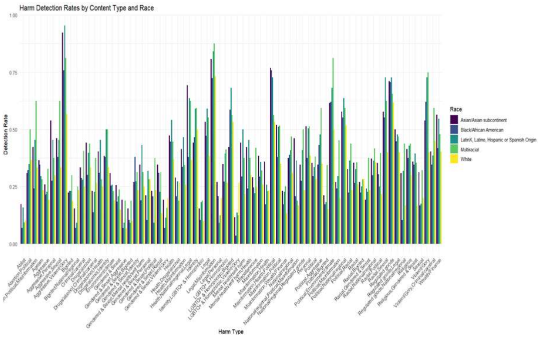

Figure 2: Harm detection rate by content type and race.

Statistical Significance: The racial effects were statistically significant in the expected direction:

- ***Other Race vs Asian***: β = -0.73, SE = 0.38, p = 0.059†
- ***White vs Asian***: β = -1.05, SE = 0.47, p = 0.025*

These findings suggest that raters' race significantly influences the perceived safety of AI-generated conversational content, with important implications for AI safety evaluation team composition.

5.2 Content-Specific Bias

A critical finding is that racial bias in AI safety assessment varies significantly across conversation topics, challenging assumptions of uniform demographic effects across all AI-generated content.

Significant Race × Content Interactions:

- ***Other Race × Miscellaneous content***: β = +0.57, SE = 0.23, p = 0.013* (Non-Asian/White raters more likely to detect safety issues in general AI conversations)
- ***White × Health content***: β = +0.68, SE = 0.39, p = 0.076† (White raters trend toward higher detection in health-related AI conversations)
- ***Other Race × Political content***: β = +0.47, SE = 0.25, p = 0.062† (Non-Asian/White raters are stricter rating political AI content)

Figure 3 below illustrates the significant race × content interactions, where Other Race raters show heightened detection for miscellaneous (β = +0.57) and political content (β = +0.47), while White raters demonstrate increased sensitivity to health-related content (β = +0.68), revealing content-specific deviations from the overall pattern of Asian raters having highest detection rates.

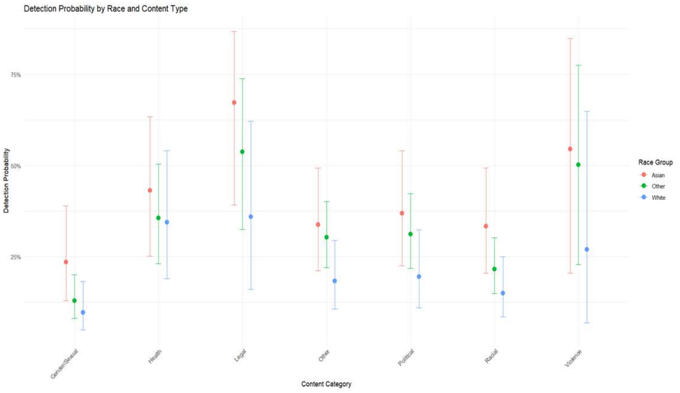

Figure 3: Harm detection probability by race and content type

AI Content Difficulty Hierarchy: Also shown in Figure 3 above, among AI-generated conversations, legal content showed highest baseline safety detection rates (β = +1.91), followed by violence-related (β = +1.52) and health content (β = +0.90), with gender/sexual AI conversations being most difficult to assess for safety violations (baseline category).

These interaction effects suggest that bias in AI safety evaluation is not uniform but depends critically on the topic and content type of AI-generated conversations, requiring content-specific approaches to bias mitigation in AI evaluation workflows.

5.3 Simulation-Based Evidence for Rater Team Configuration and Reliability

Our empirical reliability simulation demonstrates that rater team composition significantly impacts the reliability of AI safety assessments, with clear patterns visible across multiple dimensions.

Team Size Effects for AI Safety Evaluation: The visualization (top left panel) in Figure 4 below reveals a consistent upward trend in reliability as team size increases across all content types, with diminishing returns at larger sizes:

- **Teams of 10**: 70.3% mean reliability (convergence point for all content types)
- **Teams of 9**: 69.5% mean reliability
- **Teams of 8**: 69.0% mean reliability
- **Teams of 6**: 65.8% mean reliability
- **Teams of 3**: 62.2% mean reliability

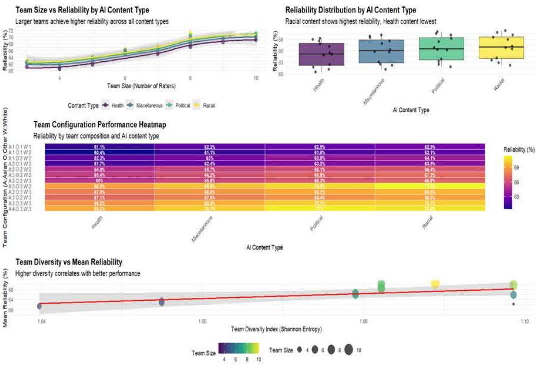

Figure 4: Empirical Reliability Simulation Analysis for Optimal Rater Team Design for AI-Generated Content Moderation.

Optimal Configurations for AI Safety Teams: The heatmap analysis (Figure 4, middle panel) clearly identifies two configurations that achieve the critical ≥70% reliability threshold:

1. **Asian:4 Other:3 White:3** (10 members): 70.3% mean reliability - the top performer across all content types
2. **Asian:3 Other:3 White:2** (8 members): 70.1% mean reliability - demonstrating cost-effective excellence

Content-Specific Reliability Patterns: Box plot analysis (Figure 4, top right panel) reveals systematic differences in evaluation difficulty across AI conversation types:

- **Racial AI content**: 67.0% mean reliability (highest, tightest distribution)
- **Political AI content**: 66.6% mean reliability (moderate variability)
- **Miscellaneous AI content**: 66.0% mean reliability (moderate variability)
- **Health AI content**: 65.2% mean reliability (lowest, highest variability)

The heatmap confirms these patterns, with racial content consistently showing the highest reliability values (yellow/orange cells) across all team configurations, while health content shows the lowest values (purple/blue cells).

Reliability Thresholds for AI Evaluation: Only 17% of tested rater team configurations (2 out of 12) achieved ≥70% reliability, with performance ranging from 60.4% to 71.3%. The top 6 configurations all required balanced demographic representation and achieved 66.4%-70.3% reliability, indicating that 70% represents a practical upper bound for AI safety evaluation.

Diversity-Reliability Relationship: The scatter plot analysis (Figure 4, bottom panel) demonstrates a clear positive relationship between team demographic diversity (Shannon entropy) and mean reliability ($r = 0.579$, $p < 0.01$). Larger, more diverse teams (represented by bigger circles in the upper right) consistently outperform smaller, less diverse configurations, providing quantitative evidence that demographic diversity enhances rather than hinders AI safety evaluation performance.

6 Conclusions

This preliminary study provides the first comprehensive analysis of human rater demographic bias and team reliability in AI-generated conversational content safety assessment using Multi-Facet Rasch Modeling. Our key findings, supported by detailed analysis and visualizations showing clear trends and patterns, demonstrate:

1. **Significant racial disparities** exist in AI safety assessment, with Asian raters 36% more likely to detect safety concerns in AI-generated content than White raters
2. **Content-specific bias patterns** in AI-generated content evaluation require targeted mitigation strategies, with racial content consistently achieving highest reliability and health content presenting greatest challenges
3. **Optimal AI safety rater team composition** involves minimum 8-10 diverse raters to achieve ≥70% reliability, with empirical simulation evidence showing convergence across content types at this threshold
4. **Diversity enhances reliability** in AI safety evaluation, with a strong positive correlation (r = 0.579) between team diversity and performance

Empirical reliability simulation analysis results provide practitioners with actionable guidance for team assembly, clearly demonstrating the reliability benefits of larger, diverse teams and content-specific performance patterns that can inform specialized evaluation strategies.

These findings provide evidence-based guidelines for assembling fair and reliable AI safety evaluation teams. As conversational AI systems scale and become more sophisticated, understanding and optimizing the human evaluation component becomes increasingly critical for maintaining both consistency and equity in AI safety assessment.

Our research also establishes a methodological framework for bias analysis in AI safety evaluation and demonstrates the practical value of psychometric approaches for understanding complex judgment tasks in AI development. Future work should examine intervention strategies for bias reduction and extend this analysis to additional AI systems and conversation domains.

The implications extend beyond academic research to practical AI development: our findings suggest that investing in diverse, appropriately sized evaluation teams is not just an ethical imperative but a technical requirement for reliable AI safety assessment. As the field moves toward more sophisticated conversational AI systems, these insights will become increasingly valuable for ensuring safe and equitable AI deployment.

Acknowledgments

We thank the creators of the Diversity in Conversational AI Evaluation for Safety (DICES) dataset for making this research possible and the anonymous reviewers for their valuable feedback.

References

Aroyo, L., Taylor, A. S., Díaz, M., Homan, C. M., Parrish, A., Serapio-García, G., Prabhakaran, V., & Wang, D. (2023). DICES dataset: Diversity in conversational AI evaluation for safety. *Advances in Neural Information Processing Systems*, *36*, 53330-53342.

Chancellor, S., Blackwell, L., De Choudhury, M., & Davison, L. (2022). Understanding demographic bias in content moderation decisions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1-15.

Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y. L., & Rieser, V. (2022). SafetyKit: First aid for measuring safety in open-domain conversational systems. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 284-299.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Goyal, N., Kivlichan, I. D., Rosen, R., & Vasserman, L. (2022). Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1-28.

Sebok-Syer, S. S., Chahine, S., Watling, C. J., Goldszmidt, M., Cristancho, S., & Lingard, L. (2018). Considering the interdependence of clinical performance: implications for assessment and entrustment. *Medical Education*, *52*(9), 970-980.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., … & Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.