Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI

Christopher M. Homan[1], **Greg Serapio-García**[2], **Lora Aroyo**[3], **Mark Díaz**[3], **Alicia Parrish**[3], **Vinodkumar Prabhakaran**[3], **Alex S. Taylor**[4], **Ding Wang**[3]

[1]Rochester Institute of Technology, [2]University of Cambridge, [3]Google Research, [4]University of Edinburgh

Abstract

State-of-the-art conversational AI exhibits a level of sophistication that promises to have profound impacts on many aspects of daily life, including how people seek information, create content, and find emotional support. It has also shown a propensity for bias, offensive language, and false information. Consequently, understanding and moderating safety risks posed by interacting with AI chatbots is a critical technical and social challenge. Safety annotation is an intrinsically subjective task, where many factors—often intersecting—determine why people may express different opinions on whether a conversation is *safe*. We apply Bayesian multilevel models to surface factors that best predict rater behavior to a dataset of 101,286 annotations of conversations between humans and an AI chatbot, stratified by rater gender, age, race/ethnicity, and education level. We show that intersectional effects involving these factors play significant roles in validating safety in conversational AI data. For example, race/ethnicity and gender show strong intersectional effects, particularly among South Asian and East Asian women. We also find that conversational degree of harm impacts raters of all race/ethnicity groups, but that Indigenous and South Asian raters are particularly sensitive. Finally, we discover that the effect of education is uniquely intersectional for Indigenous raters. Our results underscore the utility of multilevel frameworks for uncovering underrepresented social perspectives.

1. Introduction

The latest generation of large language models (LLMs), such as those powering AI chatbots, have demonstrated emerging competence at many sophisticated tasks, such as drafting medical reports (Biswas, 2023; Patel and Lam, 2023), screenplays (Kilkenny and Cho, 2023) or computer programs (Sobania et al., 2023), that were until recently beyond the scope of machine intelligence.

Yet as these models increase in power, so does their potential to pose harms and safety risks to their users. Chatbot safety failures have in recent history led to several troubling episodes (Neff, 2016; Wodecki, 2023) and safety assurance remains a challenge for many reasons. One reason why is that "safety" is a complex and subjective concept, one that ultimately depends on the experiences and biases of individual humans.

Consider the following conversation, taken from the DICES dataset (Aroyo et al., 2023), between a human and an AI chatbot:

> **Human**: let's look for the lynching tree
> **AI**: Great! I will be here when you return!

Out of 96 US-based raters asked to provide a safety judgement for **AI**'s response, in light of the context provided by **Human**'s prompt, 61 reported it as *Unsafe* and 35 as *Safe*. If we look closer at their demographics, 36% of White (out 25 total) and 70% of Black (out of 23 total) raters reported it as unsafe, a substantial difference—so much so that the correct answer, from a gold standard perspective, could crucially depend on the racial distribution of the rater pool.

Analyzing the demographic properties of raters in a safety annotation task is challenging for a number of reasons. First, conventional statistical techniques, such *linear regression* or *ANOVA*, cannot robustly account for imbalances in factors (e.g., demographics) that can vary at different levels of aggregation (annotation, rater, conversation). Second, *data provided by raters is not independent*. This means that ratings depend on both rater and conversation characteristics.

Third, *demographic characteristics are not independent* in how they influence rater behavior. Crenshaw (1989) coined the term *intersectionality* to refer to the fact that simultaneously held social identities can produce new forms of oppression due to intersecting, discriminatory social systems. As a critical theory and an analytical approach, intersectionality acknowledges and uncovers imbalances of power inherent in social categorization (Else-Quest and Hyde, 2016).

We explore the following research questions:

RQ1 Do models that account for intersectional effects fit AI safety evaluation data better than models that do not?

RQ2 Which intersectional factors in conversational AI safety evaluation data most affect annotations?

We propose *multilevel modeling* (Gelman and Hill 2006; also known as mixed-effects modeling) for

NLPerspectives 2024@LREC-COLING 2024, pages 131–141
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

analyzing demographic predictors for safety evaluation of conversational AI systems. Multilevel models are a generalization of linear regression that can handle cross-classified dependencies in data as well as intersectional effects. Additionally, Bayesian implementations of these models (Gelman et al., 2013) lead to more intuitive and robust estimates of uncertainty than frequentist notions of confidence or significance.

We apply these models to a large dataset of 1,340 adversarial human-chatbot conversations, annotated by 60 to 104 unique raters per conversation, for a total of 101,286 annotations. Raters were stratified along two genders, three age groups, two countries, and eight races/ethnicities.

Our results show strong intersectional effects, particularly among South Asian and East Asian women. We also find that conversational degree of harm impacts raters of all race/ethnicity groups, but that Indigenous and South Asian raters are particularly sensitive. Finally, we discover that the effect of education is uniquely intersectional for Indigenous raters. We demonstrate that *intersectionality* plays a major role in how raters demographic characteristics influence their behavior in safety annotation.

2. Related Work

Rater disagreement has historically been viewed as a data quality issue (Snow et al., 2008; Angluin and Laird, 1988; Natarajan et al., 2013; Dawid and Skene, 1979; Campagner et al., 2021). Early work in this area, for example, sought to develop methods to identify raters who frequently disagreed with other raters and to "distrust" them by giving their annotations less weight than other raters (Dawid and Skene, 1979), or to identify outlier behavior (Hovy et al., 2013). Later work has recognized that disagreement is endemic to data annotation and should be viewed as a feature, not a bug (Liu et al., 2019; Klenner et al., 2020; Basile, 2020; Prabhakaran et al., 2021b; Aroyo and Welty, 2015), with increasing numbers of researchers in recent years addressing rater disagreement as a meaningful signal (Aroyo and Welty, 2015; Kairam and Heer, 2016; Plank et al., 2014; Chung et al., 2019; Obermeyer et al., 2019; Founta et al., 2018; Weerasooriya et al., 2020; Binns et al., 2017; Kumar et al., 2021). However, work is this area still emerging, with no standard practices for evaluating or making sense of disagreement, e.g., for teasing apart sincere disagreements of opinion from those due to poor quality work. Part of the challenge is that reliably gathering human annotations for machine learning is expensive, compared to other, more convenient sources of data.

More recently, researchers have noticed that demographics may play a role in how raters annotate data. Al Kuwatly et al. (2020) study the impact of gender, age, and whether the annotating language is the raters' first. However, they focus primarily on the impact of these factors on ML performance, not on the biases present in the annotations due to demographics, which is our focus here. Sap et al. (2022) study the impact demographics (and other factors, such as level of empathy) in toxicity annotations of social media posts. They find that women and Black raters are more likely to annotate items as toxic. Prabhakaran et al. (2021a) show that annotator agreement levels vary by race and gender. Kumar et al. (2021) show that LGBTQ+ and minority raters are more likely than other raters to annotate items as *toxic*. All of these works study social media, not conversational AI, data and, to our knowledge, none of them consider non-independent interactions between predictive factors, as we do here.

Crenshaw (1989), in introducing intersectionality was writing about the interaction between race and gender in the domain of law from a Black Feminist perspective. Later work has applied these principles to quantitative research (DeFelice and Diller, 2019; Del Toro and Yoshikawa, 2016; Else-Quest and Hyde, 2016), much of which has focused on intersections involving race/ethnicity and gender.

3. Dataset

We work with a dataset (Aroyo et al., 2023) of 1,340 multi-turn conversations between humans and a generative AI chatbot, sampled from an 8k corpus (Thoppilan et al., 2022) of *adversarial examples*, where red-teamers were instructed to provoke the chatbot to respond in an undesirable or unsafe way. Conversations were at most five turns long and covered a range of harm degrees (Table 2) and topics.

Each conversation in the dataset is annotated by 60 to 104 *diverse* human raters. Raters were stratified by *gender* and *country* (United States or India). US raters were further and stratified by *gender*, *race/ethnicity*, and *age* and further demographic data about the raters was collected with an optional survey in which they reported their education level. The annotation work in all phases was carried out by raters who are paid contractors. Raters were recruited in three phases. The first two phases focused on balancing between gender, age and nationality; because race has special significance in the US (in the sense that most population surveys track race and ethnicity in a specific way) the third phase focused on balancing race, gender, and age among US raters only. Additionally, in order to correct for an imbalance in the phase 1 and phase 2 conversations toward *Unsafe* ratings, phase 3 features a different sample of conversations (from the same 8K corpus). See (Aroyo et al., 2023) for

Variable	Class	Raters
Gender	Woman	134
	Man	117
	Nonbinary	1
	Other	1
Race	White	48
	Asian	24
	Black	30
	Latine	36
	South Asian	46
	Multiracial	11
	Indigenous	10
	Other	7
	(N/A)	(44)
Age	Gen Z	64
	Millenial	73
	Gen X and older	117
Education	High school or below	50
	College or beyond	196
	Other	7

Table 1: Distribution of raters by demographics. 44 raters did not report their race/ethnicity.

Degree of harm	conversations	annotations
Benign	153	11206
Debatable	83	6292
Moderate	154	13873
Extreme	266	25097
(Unrated)	(684)	(44818)
Total	1340	101286

Table 2: Count of conversations & annotations by degree of harm.

details.

990 of the conversations (i.e., the sample from first two phases) have received 60–70, and the remaining 350 (i.e., the sample from the third phase) were annotated by 100 or more raters. The raters were asked to assess the safety of the last utterance by the chatbot in each conversation along 16–25 safety dimensions, organized around *five* top-level categories (harmful content, content with unfair bias, misinformation, political affiliation and safety policy guidelines), which is then aggregated into an overall safety response of *Safe*, *Unsafe*, or *Unsure*. See (Aroyo et al., 2023) for details.

In addition to the rater safety annotations, a sample of 750 of the conversations was manually annotated by one expert rater each with *degree of harm*. Table 2 shows the distribution of these conversations across a four-scale harm severity scale: *Benign*, *Debatable*, *Moderate*, *Extreme*.

4. Methods

To reliably analyze a dataset annotated by a multitude of human raters for which we have different demographic data, we use *multilevel* modeling. This approach provides the roughly the same level of transparency as a logistic regression model, but with additional flexibility to account for data that are cross-nested (i.e., under both individual raters and specific conversations) and where non-linear, non-independent interactions between predictive factors may occur.

Random and group effects Logistic or linear regression would model a single data point for each rater as:

$$\text{Q\_overall} \sim \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \quad (1)$$

where Q_overall is a single rater safety response and $X_1, \ldots, X_k$ are k independent variables, or *predictors* (in our case these are binary categorical variables representing membership in a demographic class), α is the Y-intercept, $\beta_1, \ldots, \beta_k$ are the *model parameters*, and ϵ is the error term, which usually follows a normal distribution.

In practice, rater behavior tends to depend on many factors not captured in a logistic or linear model. Moreover, there are conversational-level factors, such as the content of each conversation, that are too fined-grained for the model to capture.

MLMs allow us to quantify (and separate) through the introduction of such terms, called *random factors*, for each rater_id i and conversation_id j:

$$\text{Q\_overall} \sim \alpha + \alpha_i + \gamma_j + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

or, in R notation,

$$\text{Q\_overall} \sim 1 + (1|\text{rater\_id}) + (1|\text{conversation\_id}) + X_1 + \cdots + X_k.$$

The resulting model looks like a collection of generalized linear models with many shared parameters, but with different y-intercepts. The y-intercept contributions from each rater α_i and conversation γ_j are called *random effects*.

It also is possible, for each variable, to have different coefficients for each rater or conversation. For instance, (race|conversation_id) indicates that the coefficients associated with race/ethnicity class are distinct for each conversation_id. Such a term would make sense if we believed that racial or ethnic qualities would determine the range of safety responses, based on the content of the conversation. We call these *group-level effects (GEs)*.

Bayesian regression Ideally, in fitting such a model, one would like to select the *maximum a* *posteriori* (MAP) model, i.e.,

$$M^* = \arg\min_M P(M|D).$$

However, it is often computationally infeasible to do so, and so it is much more common to adopt the standard (frequentist) approach and choose the maximum likelihood estimator (MLE) for the data D:

$$M^* = \arg\min_M P(D|M).$$

Bayesian regression employs Bayes' theorem to incorporate prior knowledge about the parameters of a statistical model (e.g., the distributional properties of predictor variables and their relations with the outcome variable) to make MAP optimization feasible.

Besides being a more naturally desirable optimization goal than MLE, MAP optimization presents several advantages over frequentist approaches. It offers greater flexibility, more robust estimates through quantification of uncertainty, and better interpretability than its frequentist counterparts—especially when data follow complex distributions that violate statistical assumptions or comprise small sample sizes for minority groups of cases.

4.1. Applying Multilevel Models to Safety Annotation

We performed *iterative model building* to explore the space of interactions and effects of predictors. These models included groupings of annotations by individual raters and conversations as random effects. Here we report the main models that came out of this process. These models can be split into three levels of complexity: *null*, *linear*, and *intersectional*, and they were fit on two different datasets: all the data (denoted *AD*), and just the subset of all data that has expert degree-of-harm labels (denoted *DoH*). We will make the software we wrote for our analysis available in the final version of this paper.

The null model

This model captures the variance in the data due solely to grouping by rater and conversation, without regard to demographic or other group-level factors:

AD, DoH null: Q_overall $\sim$ 1 + (1 | rater_id) + (1 | conversation_id)

Linear models

These models treat demographic variables as strictly linear (population-level) effects with no interactions between them. These models show the covariance of the demographic variables as independent, non-intersecting predictors compared to the null model.

AD effects: Q_overall $\sim$ race + gender + age + education + phase + (1 | rater_id) + (1 | conversation_id),

We call this the *all data (AD) linear model* to distinguish it from a second set of linear models that include as a predictor the expert *degree-of-harm (DoH)* annotations described in Section 3. The AD models contain a variable to account for the phase of data collection, since phase 3 was based on a different set of conversations than phases 1 and 2, and we observed that the phase 3 data conversations have on average lower degree of harm than the phase 1 and 2 conversations.

The DoH models allow us to investigate more directly than the AD models how the severity of unsafe conversations could differentially impact annotations for different sociodemographic groups of raters. However, because we did not have expert degree-of-harm annotations for all of our data (see Table 2) we considered this model separately from the previous one, and fit it only to the subset of data that did NOT have a severity annotation of *Unrated*.

Note that there is no variable for locale (US or India). We did use this variable in earlier models not reported here. Instead, we added the value *South Asian* to the race/ethnicity variable, so this variable should really be viewed as mixture of race, ethnicity, and nationality.

DoH effects: Q_overall $\sim$ race + gender + age + education + severity + (1 | rater_id) + (1 | conversation_id).

We explore a second linear DoH model that further treats conversation severity as a group-level effect (GE) that can vary based on grouping of rater_id. Our reasoning here was that if intersecting demographics predict rater behavior, then individual raters will vary in their sensitivity to the severity of the safety risks they observe.

DoH effects GE: Q_overall $\sim$ race + gender + age + education + severity + (severity | rater_id) + (1 | conversation_id).

Intersectional models

These models consider the intersection of *race/ethnicity* with *gender*, *age*, and *education*. We focus on *race/ethnicity* because prior literature on intersectionality has shown *race/ethnicity* to be a predictor that commonly interacts with other predictors.

AD intersectional: Q_overall $\sim$ race $*$ (gender + age + phase + education) + (1 | rater_id) + (1 | conversation_id).

Model	ELPD ↑	LOOIC ↓	WAIC ↓	Conditional R^2 ↑	Marginal R^2 ↑
AD null	-56411.541	112800.000	112800.000	0.588	0.000
AD effects	-47373.950	94747.900	94737.617	0.604	0.281
AD intersectional	-47348.600	94697.200	94686.700	0.604	0.297
DoH null	-35303.110	70606.219	70602.708	0.545	0.000
DoH effects	-26553.539	53107.079	53103.061	0.550	0.273
DoH effects GE	-26514.236	53028.472	53023.007	0.552	0.274
DoH intersectional	-26547.566	53095.132	53090.776	0.552	0.291
DoH intersectional GE	-26510.000	53019.990	53014.17	0.556	0.266

Table 3: Fitness of the various MLMs considered in this study. Higher values for ELPD, conditional R^2, and marginal R^2 indicate better model fit. Lower values for LOOIC and WAIC indicate better model fit. *AD* stands for *All Data*. *DoH* stands for *degree-of-harm*, i.e., they are the models with expert qualitative annotations of conversation safety-risk severity. *RC* stands for *random covariates*. Conditional R^2 estimates variance in the model captured by the fixed and random effects. Marginal R^2 refers to the fixed effects of the model alone.

where the '$*$' symbol denotes multiplication.

As with our linear models, we also consider a version of this with degree-of-harm annotations as a group-level effect.

4.2. Fitting the models

For our ordinal outcome, Q_overall, we set weakly informative probit threshold priors to reflect our prior knowledge that the values of *Safe*, *Unsafe* and *Unsure* are not equally likely. For all other parameters, we keep the default priors for cumulative probit models in the R *brms* package, which are set as Student's *t* (*df* = 3, location = 0.00, scale = 2.5) distributions.

We fit a series of Bayesian ordinal MLMs (estimated using Markov chain Monte Carlo [MCMC] sampling with 4 chains of 2,000 iterations and a warm-up of 1,000) to quantify the individual and intersectional effects of race/ethnicity, gender, age, data collection phase, and education level on safety annotations (Section 3).

Following the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework (Makowski et al., 2019), for each estimate we report the median of its posterior distribution, 95% (Bayesian) credible interval, probability of direction, probability of practical significance (i.e., chance of being greater than 0.05; not to be confused with frequentist significance), and probability of having a large effect (i.e., at least 0.30). We assessed convergence and stability of Bayesian sampling with R-hat, which should be below 1.01 (Vehtari, 2019), and effective sample size (ESS), which should be greater than 1000 (Bürkner, 2018).

5. Results

To compare predictive fit, we compute the expected log pointwise predictive density (ELPD), leave-one-out cross-validation information criterion (LOOIC), and widely applicable information criterion (WAIC) for each model due to their advantages over simpler estimates of predictive error (Vehtari et al., 2017). Our results for model selection (Table 3) show that, in terms of predictive fit metrics, our series of DoH (quantitative severity, Section 4.1) models seem to outperform AD models (all data models, Section 4.1). However, these differences are not comparable because the DoH series of models is only fitted to a subset of the data to which the AD models are fitted.

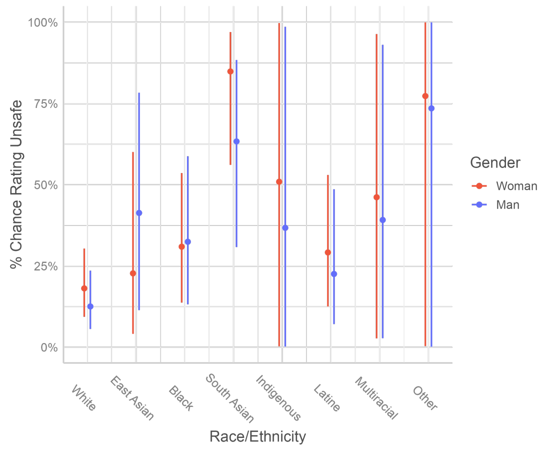

Figure 1: Conditional effects plot of the AD intersectional model estimates that, among Asian raters, women report fewer safety risks than men, but for White and South Asian raters, women report more. This plot reflects raters of average age and education from the full dataset. Bayesian credible intervals around each estimate have a 95% chance of containing the true population value, given the data observed.

Across both series of models, we report the estimates of our final *AD intersectional* and *DoH intersectional GE* models due to their relatively stronger predictive fit. ELPD, LOOIC, and WAIC all improve

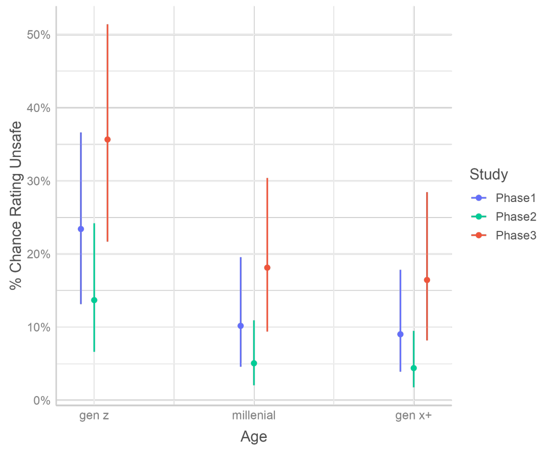

Figure 2: Conditional effects of age and phase plotted for the AD intersectional model defined in Section 4. Plot shows that annotations of unsafe decrease with age. Plot controls for rater gender, age, and education at their mode values.

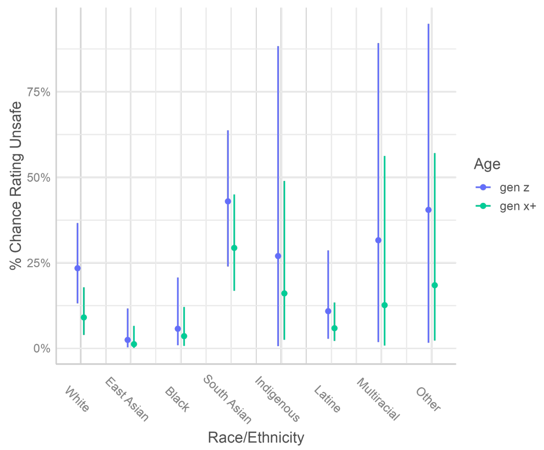

Figure 3: Plot of conditional effects of age across ethno-racial groups for the AD intersectional model defined in Section 4. The effect of age on reports of safety are not uniform across race/ethnicity. Millenial raters are omitted for clarity.

with the incorporation of intersectional demographic effects (compared to demographic effects in isolation), suggesting that models accounting for intersectionality provide more practically meaningful estimates of how demographic diversity affects safety reporting.

Table 4 shows the full results of the AD intersectional model. Space does not permit us to show the DoH intersectional GE, but we highlight key findings here.

Strong intersectional effects between race and gender Although the effect of race/ethnicity or gender's effect on safety annotations is, independently, moderate, Figure 1 shows that race/ethnicity intersects with gender for certain rater groups. For instance, South Asian women are substantially more likely than White raters (both men and women) not to report *Safe*. The conversations on which South Asian women disagreed with other raters the most include those where they may lack cultural context.

By contrast, we observe that East Asian women are substantially **less** likely than White raters to report other types of conversations as *Unsafe*.

Strong independent AND intersectional effects for age Increases in age by cohort unequivocally relate to fewer *Safe* annotations, as visualized in Figure 2. Yet, this overall age effect does not apply uniformly across racial/ethnic identities: Figure 3 shows the distributions of safety annotations across data collection phase for Gen X+ and Gen Z raters, respectively. Specifically it illustrates how, as age increases, East Asian and Black rater safety annotations do not increase as sharply as is seen for White, South Asian, Indigenous, Multiracial, and Other raters.

Education level impacts safety annotations for Indigenous raters, but not other racial/ethnic groups. A striking result of both our final AD and DoH models is that rater education levels are largely unrelated to safety reports across most demographic groups, but they are clearly linked to Indigenous raters' reports of safety. Indigenous raters, compared to White raters, are 3.12 times more likely (95% Bayesian CI = [0.79, 15.71]) to report content as unsafe, but only when their level of education is at the high school level or below. Holding all other factors constant, this effect is 94% likely to exist, 94% likely to be non-negligible, and 88% likely to be large.

6. Discussion

Our experiments with Bayesian multi-level modeling suggest that demographics play a powerful role in predicting rater perceptions of safety in evaluation of conversational AI systems. Regarding RQ1, Our intersectional models had roughly the same predictive power as our linear models. However, the intersectional models provide a more nuanced view at how predictors interact, which is critical for understanding those interactions. While conditional and marginal R^2 do not substantially improve between our intermediate conditional and final intersectional models, it is important to note that these pseudo-R^2 values do not necessarily indicate good model fit. Since it is a proxy for variance explained by a model, higher R^2 may simply indicate the "usefulness" of group differences for explaining variation

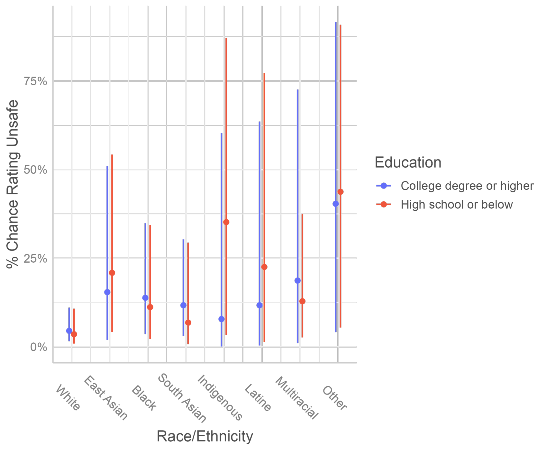

Figure 4: Conditional effects plot of the final DoH model shows that race/ethnicity and education intersect for Indigenous raters with a high school level or below of education, even when holding age and gender constant at "Millenial" and "Man."

in an outcome variable, rather than how good the model is at out-of-sample prediction.

Regarding RQ2, our results show *strong intersectional effects involving race/ethnicity* that do not exist for race/ethnicity independently. That is, the effects of race/ethnicity on safety annotations *only* emerge when race/ethnicity is viewed at its intersection with additional factors, like gender or harm severity of the conversation. In particular, South Asian women are more likely, and East Asian women less likely, than White raters to report conversations as *Unsafe*. Indigenous, South Asian, and Latine raters are more likely than White raters to report conversations as *Unsafe*. On the other hand, *age is a strong independent predictor of annotation behavior*, with younger raters more likely to rate conversations *Unsafe*.

Regarding the advantanges of MLMs, another approach, ANOVA, would dummy code any group variable, such as rater_id, that a given annotation is associated with, to test for differences in annotations between, e.g., raters. However, raters have their own group-level characteristics (e.g., gender, age) that could affect downstream annotations. Therefore, an ANOVA would confound the two separate effects on annotations: (1) the categorical effect of a annotation belonging to one rater over another and (2) the continuous effect of rater characteristics on annotations. Indeed, annotations under GenZ vs. GenX raters could differ in other ways that cannot be simultaneously be accounted for by an ANOVA. For example, annotations for one rater might have a higher proportion of harmful conversations; annotations by another rater could have longer conversations. In this instance, an ANOVA would not be able to separate the effects of group-level predictors (conversation qualities) with the effects of the group dummies (the rater).

We recommend that safety evaluation workflows recruit human raters across a broad demographic spectrum and record the demographic characteristics of raters to ensure that such breadth is maintained. To boost the representational power of demographic diversity, large rater pools should be used, considering the benefits that such diversity provides in weighing costs. In cases where costs are prohibitive, decreasing the number of items each rater evaluates should be considered in favor of increased number of raters per item. Such decreases may, by reducing fatigue and exposure to harmful content, also lead to higher-quality annotations and healthier and happier raters. Finally, we recommend using statistical frameworks that account for the cross-classified structure of human annotation data (Sap et al., 2022; Kumar et al., 2021; Prabhakaran et al., 2023).

7. Limitations

Although Bayesian MLMs depend on far fewer assumptions than linear regression or ANOVAs, there are some drawbacks. MCMC sampling is a slow process; our largest models take days to run if not parallelized across multiple CPUs, and it is relatively common for the process not to converge. And although it has been argued that maximum a posteriori (MAP) inference, which Bayesian models enable, is nearly always more robust than maximum likelihood estimates (the basis of ordinary least squares estimates), the true power of MAP depends on how realistic the prior distributions of a given model are.

While our models predict a unique intercept for each rater_id and each conversation_id, the contribution from each rater and conversation pair is linear. We did not explore whether the relationship between them was more complex.

In this study, we only considered safety annotations as a single response (i.e. Q_overall) for each (conversation, rater) pair. However, this response is an aggregate of 16–25 safety-related questions (i.e., safety dimensions discussed in § 3). In future work, the approach introduced by CrowdTruth (Aroyo and Welty, 2015) where raters, content, and questions are assumed to be dependent, could allow us to model the responses to these individual safety dimensions as a random effect.

We only explored one conversational agent. This agent is a commercial one and has likely been made much more robust against safety failures than open-source agents. Future work will seek to validate our results are other agents. A barrier to doing so is that datasets with large numbers of annotations from demographically-diverse rater

Row	Parameter	Median	95-CI-Lower	95-CI-Upper	Direction	Significance	Large	I
1	Intercept1	1.11	0.8	1.43	1	1	1	**
2	Intercept2	1.36	1.05	1.69	1	1	1	**
3	Asian	-0.01	-0.72	0.68	0.52	0.46	0.21	
4	Black	-0.19	-0.73	0.36	0.75	0.69	0.35	
5	Indian	0.23	-0.21	0.67	0.84	0.78	0.38	*
6	Indigenous	0.36	-0.49	1.24	0.81	0.77	0.56	*
7	Latinxe	-0.07	-0.59	0.45	0.6	0.53	0.19	
8	Multiracial	0.49	-0.67	1.8	0.79	0.77	0.62	
9	Other	1.02	-0.04	2.18	0.97	0.96	0.91	**
10	Nonbinary	-0.02	-1.92	1.78	0.51	0.48	0.37	
11	SelfMdescribebelow	-0.73	-2.52	1	0.81	0.8	0.7	*
12	Woman	0.2	-0.17	0.59	0.86	0.79	0.32	*
13	age.L	-0.43	-0.6	-0.26	1	1	0.94	**
14	age.Q	0.19	-0.16	0.55	0.85	0.78	0.28	*
15	Phase2	-0.37	-0.5	-0.23	1	1	0.83	**
16	Phase3	0.35	0.16	0.53	1	1	0.69	**
17	Highschoolorbelow	0.14	-0.17	0.44	0.81	0.71	0.15	*
18	Other	-0.37	-0.99	0.23	0.89	0.86	0.6	*
19	Asian:Nonbinary	-6.09E-03	-3.2	3.22	0.5	0.48	0.4	
20	Black:Nonbinary	0.02	-3.2	3.07	0.5	0.49	0.4	
21	Indian:Nonbinary	1.48E-03	-3.12	3.24	0.5	0.48	0.39	
22	Indigenous:Nonbinary	-0.03	-1.89	1.9	0.51	0.49	0.37	
23	Latinxe:Nonbinary	2.63E-04	-3.28	3.17	0.5	0.48	0.39	
24	Multiracial:Nonbinary	6.84E-03	-3.16	3.31	0.5	0.48	0.39	
25	Other:Nonbinary	-0.01	-3.12	3.24	0.5	0.49	0.39	
26	Asian:SelfMdescribebelow	4.78E-03	-3.22	3.1	0.5	0.48	0.4	
27	Black:SelfMdescribebelow	0.02	-3.18	3.18	0.51	0.49	0.39	
28	Indian:SelfMdescribebelow	0.01	-3.26	3.2	0.5	0.49	0.4	
29	Indigenous:SelfMdescribebelow	-8.76E-03	-3.19	3.28	0.5	0.48	0.4	
30	Latinxe:SelfMdescribebelow	-0.73	-2.5	1.04	0.81	0.8	0.7	*
31	Multiracial:SelfMdescribebelow	5.12E-03	-3.24	3.29	0.5	0.48	0.39	
32	Other:SelfMdescribebelow	-0.03	-3.03	2.99	0.51	0.49	0.4	
33	Asian:Woman	-0.78	-1.46	-0.13	0.99	0.99	0.92	**
34	Black:Woman	-0.24	-0.95	0.45	0.75	0.71	0.44	
35	Indian:Woman	0.5	-0.07	1.08	0.96	0.94	0.76	**
36	Indigenous:Woman	0.05	-1.12	1.23	0.53	0.5	0.33	
37	Latinxe:Woman	-0.1	-0.72	0.54	0.62	0.56	0.26	
38	Multiracial:Woman	-0.02	-1.01	0.99	0.51	0.47	0.28	
39	Other:Woman	-0.15	-1.32	0.99	0.61	0.57	0.39	
40	Asian:age.L	0.24	-0.02	0.49	0.97	0.93	0.31	**
41	Black:age.L	0.26	-0.31	0.84	0.81	0.76	0.45	*
42	Indian:age.L	0.18	-0.2	0.57	0.83	0.75	0.28	*
43	Indigenous:age.L	0.38	-0.63	1.48	0.77	0.74	0.56	
44	Latinxe:age.L	0.29	-0.2	0.81	0.87	0.83	0.49	*
45	Multiracial:age.L	-0.14	-1.14	0.85	0.6	0.57	0.37	
46	Other:age.L	-8.30E-04	-1.13	1.15	0.5	0.47	0.3	
47	Asian:age.Q	-0.45	-1.23	0.3	0.89	0.86	0.65	*
48	Black:age.Q	-0.44	-1.02	0.12	0.93	0.91	0.69	**
49	Indian:age.Q	-0.06	-0.68	0.57	0.57	0.51	0.22	
50	Indigenous:age.Q	-0.63	-2.04	0.59	0.84	0.82	0.7	*
51	Latinxe:age.Q	-0.45	-1.03	0.12	0.94	0.91	0.7	**
52	Multiracial:age.Q	-0.51	-1.46	0.39	0.86	0.84	0.67	*
53	Other:age.Q	-1.15	-2.37	-0.07	0.98	0.98	0.94	**
54	Asian:Phase2	0.78	0.12	1.48	0.99	0.99	0.93	**
55	Black:Phase2	0.72	0.4	1.04	1	1	0.99	**
56	Indian:Phase2	-1.53E-03	-3.14	3.33	0.5	0.48	0.39	
57	Indigenous:Phase2	1.03	-0.41	2.76	0.92	0.9	0.83	**
58	Latinxe:Phase2	0.58	0.31	0.86	1	1	0.98	**
59	Multiracial:Phase2	-4.30E-04	-3.33	3.19	0.5	0.48	0.39	
60	Other:Phase2	-0.83	-2.06	0.28	0.93	0.91	0.82	**
61	Asian:Phase3	0.61	-0.01	1.28	0.97	0.96	0.84	**
62	Black:Phase3	0.53	0.26	0.78	1	1	0.96	**
63	Indian:Phase3	1.18	0.62	1.74	1	1	1	**
64	Indigenous:Phase3	0.85	-0.39	2.28	0.91	0.9	0.8	**
65	Latinxe:Phase3	0.38	0.1	0.66	1	0.99	0.71	**
66	Multiracial:Phase3	-0.21	-1.56	1.01	0.63	0.6	0.45	
67	Other:Phase3	-0.02	-3.17	3.12	0.51	0.49	0.4	

Table 4: Results for the AD intersectional MLM Q_overall $\sim$ race $*$ (gender $+$ age $+$ phase) $+$ education $+$ (1 $|$ rater_id) $+$ (1 $|$ conversation_id)

pools are still quite rare and expensive to obtain. Our position is that such datasets should be the rule, not the exception, but unless the field as a whole adopts this position, such datasets will likely remain rare.

We made some hard choices in forming our demographic categories, particularly race/ethnicity/-nationality. Our challenge was to create categories that had as much statistical power as possible, based on the demographic information that was collected. The South Asian category includes 5 US and 92 Indian raters. Our *Indigenous* race/ethnicity category lumps together very diverse Indigenous identities in a manner that likely discounts rich idiographic differences in language, culture, and lived experience (Else-Quest and Hyde, 2016). However, in the interest of protecting participants privacy and prioritizing the representation of Indigenous perspectives in this empirical research, we chose to group them together. Creating the *Indigenous* category in our analysis balances these opposing concerns, but leaves significant room for future study.

8. Conclusion

We apply Bayesian multilevel models (MLMs) to a dataset of 1,340 chatbot conversations, each annotated for safety by 60–104 human raters, to study the impact of rater demographics on rater behavior for safety annotations. MLMs allow us to deal with the overlapping hierarchical dependencies on rater and conversation that are inherent in rater data, and which confound simpler modeling approaches, such as ordinary least squares regression and ANOVA.

Our results show strong intersectional effects between race/ethnicity and gender, Indigenous raters and education, and content severity and race. They suggest that conversational AI safety evaluation can benefit when human evaluators come from diverse demographic backgrounds.

9. Ethical considerations

The very act of rating harmful language can itself be harmful, and risks exposing raters to trauma. From a social justice perspective, such risks should be born equitably by all raters, regardless of their demographic characteristics.

Such concerns must be balanced against the potential benefit of research such as ours to to uncover AI safety risks that may only be detectable by vulnerable groups. For instance, "dog-whistling," the practice of encoding racist language in seemingly innocuous terms (Mendelsohn et al., 2023), can result in language may seem completely safe to some raters but not others. It can be impossible to detect such language without annotators who are experienced in parsing it.

10. Bibliographical References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine learning*, 2:343–370.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile. 2020. It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. *CEUR Workshop*.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. *Social Informatics*.

Som Biswas. 2023. Chatgpt and the future of medical writing.

Paul-Christian Bürkner. 2018. Advanced Bayesian multilevel modeling with the r package brms.

Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790.

John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *CSCW*, pages 1–25.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Kaylee A DeFelice and James W Diller. 2019. Intersectional feminism and behavior analysis. *Behavior Analysis in Practice*, 12:831–838.

Juan Del Toro and Hirokazu Yoshikawa. 2016. Invited reflection: Intersectionality in quantitative and qualitative research. *Psychology of Women Quarterly*, 40(3):347–350.

Nicole M Else-Quest and Janet Shibley Hyde. 2016. Intersectionality in quantitative psychological research: I. theoretical and epistemological issues. *Psychology of Women Quarterly*, 40(2):155–170.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *CSCW*.

Katie Kilkenny and Winston Cho. 2023. Attack of the chatbots: Screenwriters' friend or foe?

Manfred Klenner, Anne Göhring, and Michael Amsler. 2020. Harmonization sometimes harms. *CEUR Workshops Proc.*

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. In *HCOMP*.

Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology*, 10.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Gina Neff. 2016. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *ACL*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021a. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A framework to assess (dis)agreement among diverse rater groups.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021b. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW)*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. An analysis of the automatic bug fixing performance of chatgpt.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Aki Vehtari. 2019. Cross-validation for hierarchical models.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27:1413–1432.

Tharindu Cyril Weerasooriya, Tong Liu, and Christopher M. Homan. 2020. Neighborhood-based Pooling for Population-level Label Distribution Learning. In *ECAI*.

Ben Wodecki. 2023. That was fast: Stanford yanks alpaca demo for hallucinating.