# Aligning NLP Models with Target Population Perspectives using PAIR: Population-Aligned Instance Replication

**Stephanie Eckman**[*,♣] **Bolei Ma**[*,♡] **Christoph Kern**[♡] **Rob Chew**[♠]
**Barbara Plank**[♡] **Frauke Kreuter**[♣,♡]

[♣]University of Maryland, College Park
[♡]LMU Munich & Munich Center for Machine Learning
[♠]RTI International

[*]Equal contributions.

{steph, fkreuter}@umd.edu, {bolei.ma, christoph.kern, b.plank}@lmu.de, rchew@rti.org

## Abstract

Models trained on crowdsourced annotations may not reflect population views, if those who work as annotators do not represent the broader population. In this paper, we propose **PAIR**: **P**opulation-**A**ligned **I**nstance **R**eplication, a post-processing method that adjusts training data to better reflect target population characteristics without collecting additional annotations. Using simulation studies on offensive language and hate speech detection with varying annotator compositions, we show that non-representative pools degrade model calibration while leaving accuracy largely unchanged. PAIR corrects these calibration problems by replicating annotations from underrepresented annotator groups to match population proportions. We conclude with recommendations for improving the representativity of training data and model performance.[1]

## 1 Introduction and Inspiration

When a hate speech detection model flags harmless expressions as toxic, or a content moderation system fails to identify genuinely harmful content, the root cause often lies not in the model architecture, but in who annotated the training data. While Natural Language Processing (NLP) models aim to serve broad populations, the human judgments used to train these systems often come from crowdworkers and convenience samples. And the demographics, cultural contexts, and worldviews of these annotators often differ from those of the communities the models ultimately impact (Sorensen et al., 2024; Fleisig et al., 2024). These non-representative annotator pools can have real consequences, because annotator characteristics like age, education level, and cultural background impact how content is annotated (Sap et al., 2022; Fleisig et al., 2023; Kirk et al., 2024). The influence of annotator characteristics underscores that language understanding is not a single objective truth but a constellation of equally valid interpretations anchored in different lived experiences. When this perspectivist interpretation is ignored, models trained on non-representative data can perpetuate the biases and blind spots of their limited training data (Berinsky et al., 2012; Hebert-Johnson et al., 2018; Mehrabi et al., 2021; Rolf et al., 2021; Hüllermeier and Waegeman, 2021; Ouyang et al., 2022; Favier et al., 2023; Smart et al., 2024).

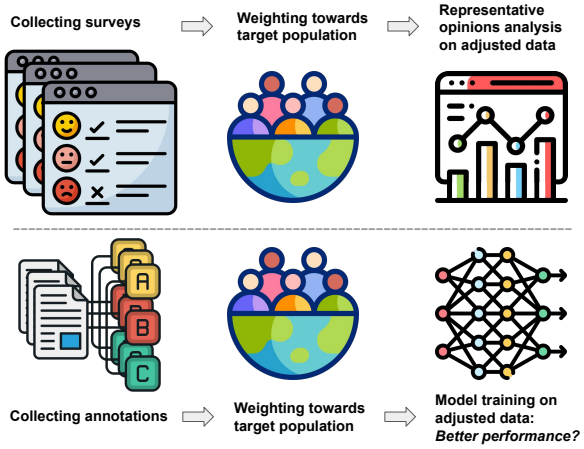


Figure 1: **Top**: Adjusting survey data to match population produces high quality results.
**Bottom**: Can a similar adjustment in data annotations also improve model performance?

Fortunately, survey researchers have developed statistical techniques to produce population-level estimates from non-representative samples (Bethlehem et al., 2011). The top panel of Figure 1 shows a simple survey workflow: collecting survey data, creating statistical weights to match the sample to the population, and estimating population parameters. We adapt this approach to the machine learning context, enabling models to better align

[1]The code for experiments is available at https://github.com/soda-lmu/PAIR.

with target populations even when trained on non-representative annotator pools (bottom panel).

Our Population-Aligned Instance Replication (PAIR) method post-processes training data to better reflect target populations without collecting additional annotations. We test the approach with a simulation study (Burton et al., 2006; Valliant, 2019; Morris et al., 2019) and answer two questions:

- **RQ1:** How do non-representative annotator pools impact model calibration and accuracy?
- **RQ2:** Can our proposed weighting method (PAIR) mitigate these annotator pool effects?

Our results demonstrate that models trained on non-representative annotator pools perform worse than those trained on representative pools. However, simple adjustment methods can improve performance without collecting additional data. These findings suggest that insights from survey methodology can help artificial intelligence (AI) systems better represent the populations they serve.

## 2 Related Work

Several strands of related work inform our approach to identifying and mitigating bias due to the use of non-representative annotators:

**Annotator Impact on Data and Models.** Annotator characteristics and attitudes significantly influence annotation quality, particularly for subjective tasks like toxicity detection (Giorgi et al., 2025; Prabhakaran et al., 2021; Fleisig et al., 2023; Sap et al., 2022). For example, annotators' political views and racial attitudes affect their toxicity judgments (Sap et al., 2022). Models trained on non-representative annotator pools inherit these biases and generalize poorly (Berinsky et al., 2012; Mehrabi et al., 2021; Rolf et al., 2021; Ouyang et al., 2022; Favier et al., 2023; Smart et al., 2024; Mokhberian et al., 2024).

**Annotator Demographics.** Several researchers advocate collecting annotator demographics to assess representation and identify biases (Bender and Friedman, 2018; Prabhakaran et al., 2021; Plank, 2022; Wan et al., 2023; Santy et al., 2023; Pei and Jurgens, 2023).[2] However, collecting and releasing these data can raise privacy concerns (Fleisig et al., 2023). Recent works have also used demographics to prompt the large language models (Argyle et al., 2023), and some find that these are less effective in subjective contexts (Sun et al., 2025; Orlikowski et al., 2025).

**Debiasing & Data Augmentation Methods.** Prior work has proposed various approaches to reduce bias in training data features and annotations. Most similar to our work is the resampling and reweighting approaches of Calders et al. (2009) and Kamiran and Calders (2012), imputation (Lowman-stone et al., 2023), and the oversampling of minority class cases of Ling and Li (1998). PAIR adapts these methods to balance *annotator* characteristics rather than class labels or sensitive observation-level features. PAIR retains the simplicity and interpretability of earlier resampling methods while extending them to a "Learning with Disagreement" (Uma et al., 2021; Leonardelli et al., 2023) setting with multiple annotations per observation, by replicating annotations from underrepresented annotator groups.

## 3 PAIR Algorithm: Adjustment via Pseudo-Population

To adjust an annotator pool to better reflect a target population, we propose the PAIR algorithm, which constructs a pseudo-population through post-stratification, weight normalization, and deterministic replication. This adjustment strategy is inspired by established methods in survey sampling (Quatember, 2015).

Post-stratification aligns a sample more closely with population-level distributions (Bethlehem et al., 2011; Valliant et al., 2013). Annotators are grouped into strata based on demographic or behavioral characteristics. For each unit $i$ in stratum $s$, a post-stratification weight is computed as:

$$w_{s,i} = \frac{P_s}{S_s} \tag{1}$$

where $P_s$ and $S_s$ denote the share of the population in stratum $s$ and the share of the sample (or annotator pool), respectively. The $P$ values come from official statistics or surveys. The $S$ values likely come from the annotators themselves and researchers may have to collect them. This technique can accommodate multiple stratification variables; it is only limited by the availability of population or reference data and data about the annotators.

[2] In our context, these characteristics are used only to analyze bias. Because they are not available for unannotated text, they are not features that the model can use.

These weights have only relative meaning and are invariant to multiplication by a constant ($K$):

$$w_i^{\text{normalized}} = w_i^{\text{initial}} \times K \tag{2}$$

Normalization useful if research teams have a target number of annotations per observation in mind, for either computational or design reasons, or if some weights given by Eq. 1 are very small and round to one.

To generate a pseudo-population, we apply deterministic replication: each unit is replicated $n_i$ times where

$$n_i = \text{round}(w_i^{\text{normalized}}) - 1 \tag{3}$$

ensuring integer replication counts. This approach produces a dataset that reflects population proportions while maintaining interpretability and reproducibility.

While we focus in this initial study on deterministic replication, alternative implementations are possible, including resampling-based replication or direct incorporation of weights into model training.

## 4 Annotation Simulation and Model Training

To address our research questions, we conduct a simulation study on offensive language and hate speech detection. We imagine a population made up of equal shares of two types of people: those more likely to perceive offensive language and hate speech and those less likely. We create three datasets of simulated annotations which differ in the mix of the annotator types. We then create a fourth dataset, using the PAIR algorithm, to fix the imbalance in the annotators. We fine-tune RoBERTa models on the four datasets and evaluate the effect of annotator composition on model performance (RQ1) and the ability of the PAIR algorithm to improve performance (RQ2).

### 4.1 Simulating Annotations

We use our previously collected dataset on tweet annotation sensitivity (Kern et al., 2023)[3], which is a dataset of 3,000 English-language tweets, each with 15 annotations of both offensive language (OL: yes/no) and hate speech (HS: yes/no). We chose this dataset because the high number of annotations of each tweet gives us a diverse set of labels to work with. We select the 2,500 training subset of the dataset. We randomly select (without replacement) 12 annotations (of both OL and HS) of each tweet in the original dataset.[4] Let $p_{i,OL}$ be the proportion of the 12 annotators who annotated tweet $i$ as OL and $p_{i,HS}$ defined similarly. Figure 2 shows the distribution of these proportions across the 3,000 tweets. The HS annotations are clustered near 0, whereas the OL annotations are more spread out between 0 and 1.

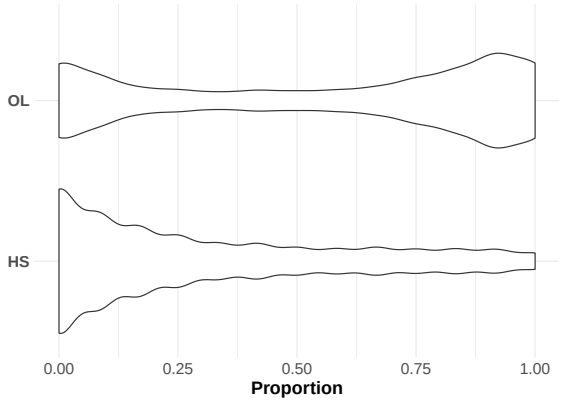


Figure 2: Distribution of $p_{i,OL}$ and $p_{i,HS}$ in original data

The population contains two types of people (50% each). **Type A** people are *less likely* to say a tweet contains OL. **Type B** people are *more likely*:

$$p_{i,OL}^A = \max(p_{i,OL} - \beta, 0) \tag{4}$$

$$p_{i,OL}^B = \min(p_{i,OL} + \beta, 1) \tag{5}$$

Here $\beta$ captures the magnitude of the bias. We vary $\beta$ from $[0.05, 0.3]$ by $0.05$, corresponding to an increase or decrease in the probability to judge a tweet as OL by five to 30 percentage points. This range is large on the probability scale and covers most reasonable situations. With these six values of $\beta$, we create six vectors of probabilities $(p_{i,OL}^A, p_{i,OL}^B)$ for each tweet.

We then create four datasets, each with 3,000 tweets (Table 1), for each value of $\beta$. The **Representative** Dataset contains OL annotations from six A annotators (drawn from Bernoulli$(p_{i,OL}^A)$) and six B annotators (drawn from Bernoulli$(p_{i,OL}^B)$). The proportion of A and B annotators in this dataset matches the simulated population we created.

[3] https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2

[4] As shown in Table 1, we can more carefully control the construction of our datasets when the number of annotations per tweet is even.

| Dataset | Annotations per tweet | A annotations | B annotations |
|---|---|---|---|
| Representative | 12 | 6 | 6 |
| Non-representative 1 | 9 | 6 | 3 |
| Non-representative 2 | 12 | 9 | 3 |
| Adjusted | 12 | 6 | 3 + 3* |

* 3 B annotations replicated

Table 1: Four training datasets for each bias value ($\beta$)

We next create two unbalanced datasets. **Non-representative 1** randomly deletes three B annotations for each tweet from the Representative Dataset. **Non-representative 2** adds three additional A annotations, drawn from $p^A_{i,OL}$, to the Non-representative 1 dataset. The Non-representative 2 Dataset is more unbalanced than Non-representative 1, but contains the same number of annotations as the Representative dataset.

### 4.2 Applying PAIR Algorithm

Finally, we use the PAIR algorithm to create the **Adjusted** Dataset. Starting with the Non-representative 1 Dataset, we calculate the share of the annotator pool that is in the A and B strata: $S_A = \frac{2}{3}, S_B = \frac{1}{3}$. The population proportions, by construction, are $P_A = 0.5, P_B = 0.5$. Applying (1), we get $w_{A,i} = 0.75, w_{B,i} = 1.5$. We multiply these weights by $K = \frac{4}{3}$ to get $w_{A,i} = 1, w_{B,i} = 2$. These weights give us $n_{A,i} = 0, n_{B,i} = 1$, which leads us to replicate all B annotations in the Non-representative 1 Dataset (see Table 1).

The HS probabilities for the A and B annotators are defined in the same way: $p^A_{i,HS} = \max(p_{i,HS} - \beta, 0)$, $p^B_{i,HS} = \min(p_{i,HS} + \beta, 1)$. We also construct the four datasets (Representative, Non-representative 1, Non-representative 2, Adjusted) in the same way we did in the OL case.

Figures 3 and 4 show the percentage of instances annotated OL and HS in the four datasets for each value of $\beta$. In both, the percentage of OL/HS annotations in the Adjusted dataset is similar to that in the Representative dataset for all values of $\beta$. The percentage in the two unbalanced datasets is lower, because those datasets overrepresent the A annotators, who are less likely to annotate OL/HS.

HS is rare in our dataset (16.7% of instances were annotated as HS), and our simulation strategy overrepresents A annotators in the two Non-representative datasets, who are less likely to perceive HS (Table 1). For these reasons, as $\beta$ increases, more $p^A_{i,HS}$ are 0 while the $p^B_{i,HS}$ probabilities increase. This issue leads the proportion of HS annotations in the Representative and Adjusted datasets to increase with $\beta$ in the HS dataset, which have more B annotations than the unadjusted datasets (Figure 4).

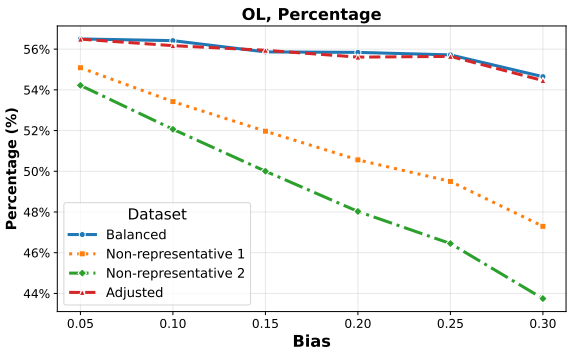


Figure 3: Percentage of instances annotated as OL, by dataset and bias ($\beta$)

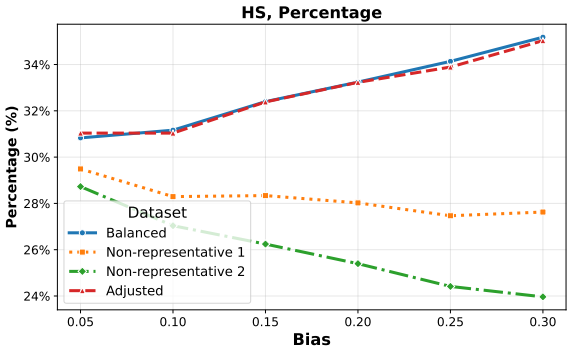


Figure 4: Percentage of instances annotated as HS, by dataset and bias ($\beta$)

### 4.3 Model Training and Evaluation

**Training and Test Setup.** We train models on each of the eight datasets: four for OL, four for HS. We divide each dataset, at the tweet level, into training (2,000 tweets), development (500), and test (500) sets. Each tweet appears 12 times in the Representative, Non-representative 2, and Adjusted datasets and nine times in the Non-representative 1 set.

**Model Selection and Training.** We used RoBERTa base (Liu et al., 2019) as our text classifier, training for five epochs on each dataset, with development set optimization. To ensure reliable results, we trained five versions with different random seeds and averaged their performance.

Our implementation of RoBERTa models was based on the libraries `pytorch` (Paszke et al., 2019) and `transformers` (Wolf et al., 2020). During training, we used the same hyperparameter settings (Table 2) for the five training conditions to keep these variables consistent for comparison purposes. We trained each model variation with five random seeds $\{10, 42, 512, 1010, 3344\}$ and took the average across the models. All experiments were conducted on an NVIDIA® A100 80 GB RAM GPU.

| Hyperparameter | Value |
|---|---|
| encoder | `roberta-base` |
| epochs_trained | 5 |
| learning_rate | $3e^{-5}$ |
| batch_size | 32 |
| warmup_steps | 500 |
| optimizer | `AdamW` |
| max_length | 128 |

Table 2: Hyperparameter settings of RoBERTa models

**Performance Metrics.** We evaluate models using **calibration and accuracy metrics** on the test set. While accuracy metrics directly measure classification performance, calibration metrics provide crucial insights into model reliability by assessing probability estimate quality – particularly important for high-stakes applications requiring trustworthy confidence measures.

For **calibration**, we report Mean Calibration Bias (MCB, Equation 6), which measures the difference between the model's average predicted probability and the average annotation frequency in the population. Specifically, we compare the mean predicted probability of offensive language across all tweets ($\text{preds}_{i,OL}$) with the mean proportion of annotators labeling tweets as offensive ($p_{i,OL}$).

MCB captures systematic over- or underestimation of class probabilities at the corpus level. Unlike instance-level calibration metrics, MCB does not penalize errors that cancel out across samples, but instead reflects global shifts in predicted probability mass relative to human annotation distributions.

$$\text{MCB}_{OL} = \left| \frac{1}{n} \sum_{i=1}^{n} \text{preds}_{i,OL} - \frac{1}{n} \sum_{i=1}^{n} p_{i,OL} \right| \quad (6)$$

For **accuracy**, we report the F1 score.

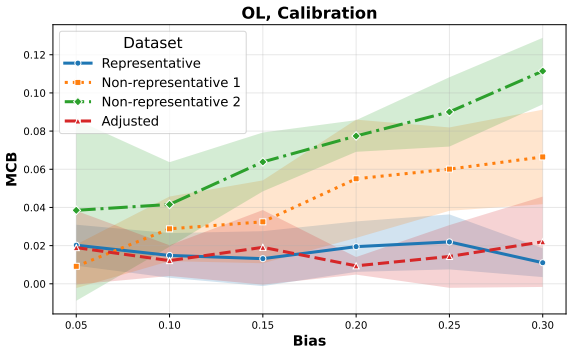


Figure 5: MCB scores for OL Models, by dataset and bias ($\beta$)

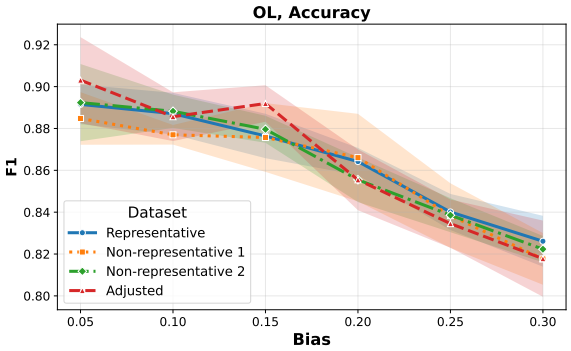


Figure 6: F1 scores for OL Models, by dataset and bias ($\beta$)

## 5 Results

We show results separately for OL and HS models.

### 5.1 OL Models

**Calibration.** Figure 5 compares the MCB in the test set for models trained on the four datasets. The dark lines show average MCB across the five training runs and the shading shows the standard deviation. The MCB for the models trained on the Adjusted dataset closely tracks that for the Representative dataset and does not increase with $\beta$. MCB for the models trained on the two unbalanced datasets is greater and grows with $\beta$. These results demonstrate the effectiveness of our adjustment method. Replicating the annotations from the underrepresented annotator type to match population proportions improves model calibration.

**Accuracy.** Figure 6 compares the models' F1 scores. In contrast to Figure 5, we do not see strong differences between the models trained on the different datasets. For all datasets, model performance declines with $\beta$: as the amount of bias in the annotations increases, the models are less able to predict

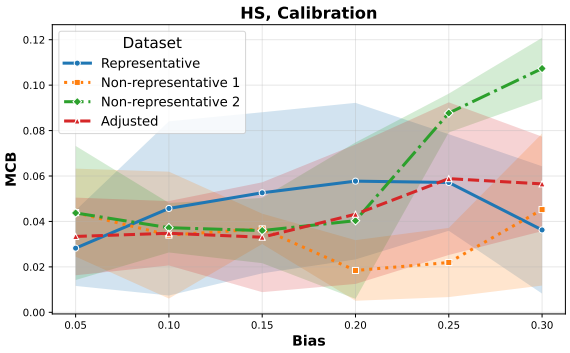


Figure 7: MCB scores for HS Models, by dataset and bias ($\beta$)

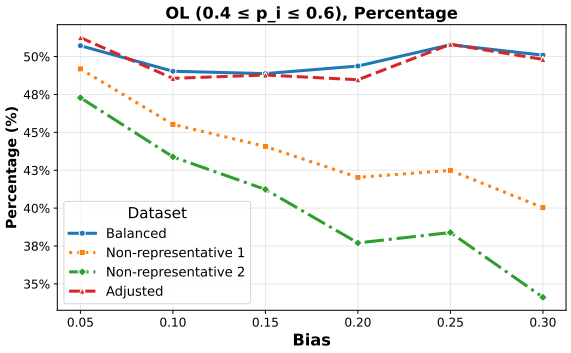


Figure 9: Percentage of OL instances on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$) by dataset and bias ($\beta$)

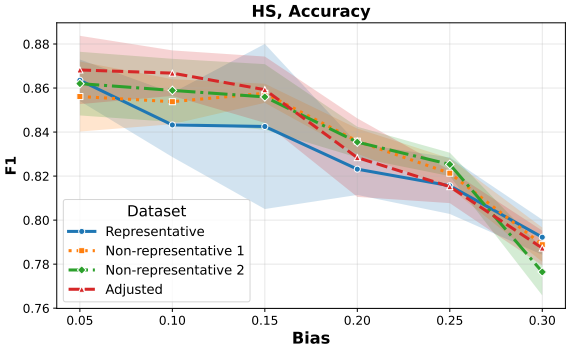


Figure 8: F1 scores for HS Models, by dataset and bias ($\beta$)

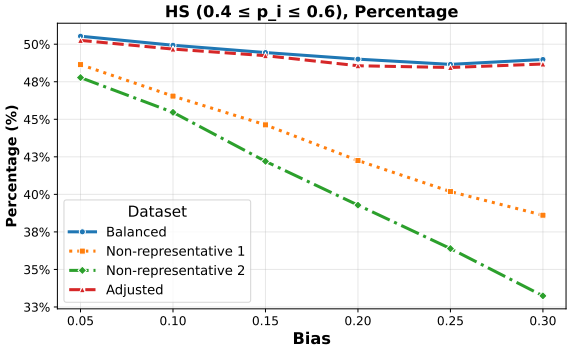


Figure 10: Percentage of HS instances on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$) by dataset and bias ($\beta$)

the binary OL label.

Because the F1 metric focuses on binary predictions, it is less sensitive to training biases than calibration metrics like MCB, which more explicitly capture biases through prediction scores. In decision-making, miscalibrated predictions can have harmful consequences when, for example, hateful content remains undetected (Van Calster et al., 2019). These findings suggest that calibration metrics provide a clearer view of the impact of annotators on models: binary classification metrics can obscure such effects.

## 5.2 HS Models

Figure 7 contains the MCB results and Figure 8 the F1 score results for the HS models trained on each dataset. Though the adjusted model roughly tracks the representative models for MCB, there is instability in the results. All models show lower average MCB values than the representative model across a wide range of the bias offset (0.10 - 0.20). The PAIR approach does not improve calibration or accuracy: the adjusted model performs similarly to the Non-representative models. This effect is likely due to the combination of label rarity and our simulation design. With few positive annotations to begin with, the impact of the $\beta$ parameter and the overrepresentation of the A annotators may be overwhelmed by the baseline scarcity of hate speech annotations. Calibration metrics can be less reliable with rare classes (Zhong et al., 2021).

## 5.3 Sensitivity Analysis: Difficult Tweets

Our simulations assumed that all annotator type impacts all tweets the same way (Eq. 5), which is an oversimplification. More likely, **annotator characteristics have more impact for ambiguous tweets**. For example, prior research in the psychology literature on judgment under uncertainty suggests that people draw more heavily on personal heuristics when interpreting unclear or underspecified information (Tversky and Kahneman, 1974). For this reason, we repeat model training and recompute metrics for those tweets where $0.4 \leq p_i \leq 0.6$. Subsetting the tweets in this way also eliminates the floor and ceiling effects in Eq. 5. The filtered datasets contain 267 (OL) and 360 (HS) tweets. The proportions of OL and HS annotations are sta-

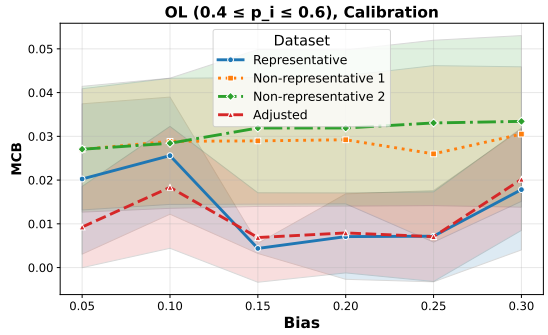


Figure 11: MCB scores for OL Models, on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$), by dataset and bias ($\beta$)

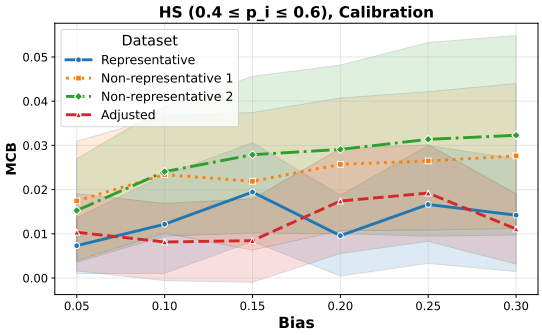


Figure 13: MCB scores for HS Models, on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$), by dataset and bias ($\beta$)

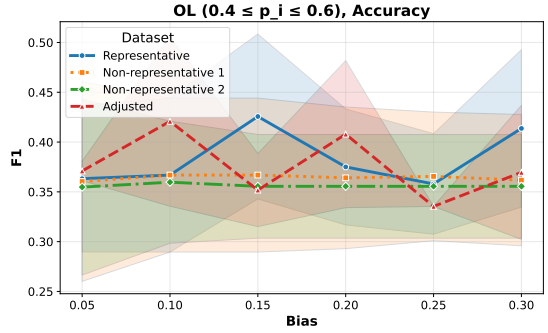


Figure 12: F1 scores for OL Models, on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$), by dataset and bias ($\beta$)

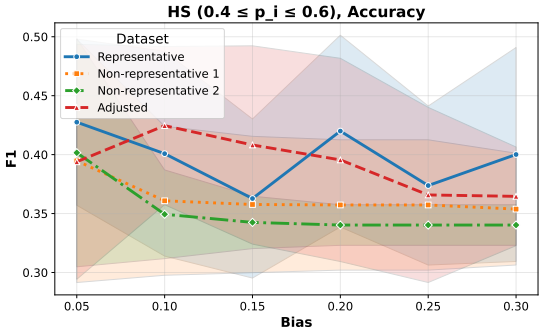


Figure 14: F1 scores for HS Models, on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$), by dataset and bias ($\beta$)

ble for the Representative and Adjusted sets and decrease for the Non-representative sets as we increase the bias offset (Figures 9, 10). This mimics the trend for OL in the full dataset (Figure 3).

Figures 11, 12, 13, and 14 show results for two metrics (MCB, F1) for filtered OL and HS annotations. In Figure 11, the Representative and Adjusted models have similar MCB and are lower than the Non-representative models. The F1 scores do not show differences between the models. These results are similar to those on the full set of tweets (Figures 5 and 6). In the two HS figures (13, 14), we see signs that the Representative and Adjusted models perform similarly, and better than the two Non-representative models, on both metrics. These results are more promising than those on the full set of tweets (Figures 7 and 8) and support our hypothesis that the rarity of HS annotations contributed to the lack of positive results for the PAIR approach in §5.2. The PAIR algorithm works well with difficult tweets, which is where it is likely most needed.

## 6 Discussion & Recommendations

Our experimental results show the OL prediction models perform less well when trained on data from non-representative annotator pools (RQ1), and simple statistical adjustments can improve model calibration without collecting additional annotations or involving additional annotators (RQ2). While PAIR's impact was harder to assess for the rare HS class, PAIR did improve calibration of both the OL and HS models when trained on difficult tweets. These findings establish a promising bridge between survey statistics and machine learning – offering a practical approach to make AI systems more representative of and responsive to the populations they serve, particularly for tasks involving subjective human judgments.

We recommend the following four steps to reduce bias due to non-representative annotator pools:

1) **Use social science research** to identify the annotator characteristics that influence the propensity to engage in annotation and the annotations provided (Eckman et al., 2024).
2) **Collect these characteristics** from annotators and gather corresponding population-level data from national censuses or high-quality surveys.[5]

[5]Collection and release of annotator characteristics or weights derived from them may raise confidentiality concerns.

3) **Calculate weights** that match the annotators to the population on those characteristics (Bethlehem et al., 2011; Valliant et al., 2013).
4) **Use these weights in model training**. Our simple replication approach showed promise, and future work should test more sophisticated weighting approaches.

## 7 Limitations

We outline key areas where future work could broaden the applicability and robustness of our findings.

**Stylized Biases and Simulated Data.** Our simulation makes strong assumptions about annotator behavior: there are only two types of annotators, and, within each type, annotators behave similarly. Real-world annotator biases may be more nuanced or context-dependent. The simulated annotators might not be representative of a stable opinion group (Mokhberian et al., 2024; Vitsakis et al., 2024). Future work could incorporate more realistic biases and refine the proposed simulations and statistical techniques.

**Sampling Variability.** We have created only one version of the four datasets for each annotation type and value of $\beta$, each of which contains random draws from the Bernoulli distribution. A more traditional statistical approach would create multiple versions of the datasets and train models on each one, to average over the sampling variability. Though limited by computational constraints in this work, future work could take on a more expansive simulation.

**Need for Population Benchmarks and Annotator Characteristics.** PAIR requires high quality benchmark information about the relevant population. These benchmarks might come from national statistical offices or national surveys. Annotators must provide accurate data on the same characteristics available in the benchmark data. Unfortunately, annotators sometimes do not provide accurate information (Chandler and Paolacci, 2017; Huang et al., 2023). In addition, theory demonstrates that bias will be reduced only when the characteristics used in weighting correlate with the annotations (Eckman et al., 2024). In our simulation, differences in annotations were driven solely by group membership ($A$, $B$). In the real world, it is challenging to know what characteristics impact annotation behavior for a given task and to find good benchmarks for those characteristics.

**Generalization Beyond Task Types.** The study focuses only on binary classification tasks. Many real-world annotation tasks involve multiple classes or labels, which may show different bias patterns. Additional research is needed to extend these methods to more complex classification scenarios.

**Evaluation Metrics.** While we measured calibration and accuracy, we did not examine other important metrics such as fairness across subgroups or robustness to adversarial examples. Future work on training data adjustment should assess a broader range of performance measures.

## 8 Ethical Considerations

In this simulation study, we experiment on a publicly available dataset collected in our previous study (Kern et al., 2023), which contains offensive and hateful tweets. We do not support the views expressed in these tweets. The simulation study itself does not collect any new data or raise any ethical considerations.


## Acknowledgments

We acknowledge use of the Claude model to edit the text of the paper and to assist in coding. We thank the members of SODA Lab and MaiNLP labs from LMU Munich, and the members of the Social Data Science Group from University of Mannheim for their constructive feedback. This research is partially supported by RTI International, MCML, and BERD@NFDI. BP is supported by ERC Consolidator Grant DIALECT (101043235).

We thank Narjes Tahaei for identifying an error in the metric used in an earlier version of this paper. Specifically, the calibration metric reported in the paper was not calculated as described. We have kept the calculation and updated the formula to reflect the calculation performed. We have retitled the metric to Mean Calibration Bias (MCB) to reflect the calculation used. The earlier version gave the name and formula for Absolute Calibration Bias (ACB) but in fact reported MCB. In this revised version, we have corrected the metric name and formula in Section 4.3. The results in the main body of the paper are unchanged. For transparency, we give results for the correct calculation of ACB in Appendix A of the current version.


The survey literature offers advice for sharing sensitive data (see Karr, 2016, for a review). Collecting annotator characteristics may also require involvement of Institutional Review Boards or other participant protection organizations (Kaushik et al., 2024).

## References


Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):339–355.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368.

Jelke Bethlehem, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Wiley.

Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

Jesse J. Chandler and Gabriele Paolacci. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5):500–508.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from survey methodology can improve training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12268–12283. PMLR.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):653–670.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Alan F. Karr. 2016. Data sharing and access. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):113–132.

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. 2024. Resolving the human-subjects status of ml's crowdworkers. *Commun. ACM*, 67(5):52–59.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Charles X. Ling and Chenghui Li. 1998. Data mining for direct marketing: problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 73–79. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. 2023. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. In *NLPerspectives@ECAI*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–36.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Tim P. Morris, Ian R. White, and Michael J. Crowther. 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jiaxin Pei and David Jurgens. 2023. 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)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andreas Quatember. 2015. *Pseudo-Populations: A Basic Concept in Statistical Surveys*. Springer.

Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9040–9051. PMLR.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Andrew Smart, Ding Wang, Ellis Monk, Mark Díaz, Atoosa Kasirzadeh, Erin Van Liemt, and Sonja Schmer-Galunder. 2024. Discipline and label: A weird genealogy and social theory of data annotation. *Preprint*, arXiv:2402.06811.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Richard Valliant. 2019. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Richard Valliant, Jill A Dever, and Frauke Kreuter. 2013. *Practical tools for designing and weighting survey samples*, volume 1. Springer.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra Macaskill, David J. McLernon, Karel G. M. Moons, Ewout W. Steyerberg, Ben Van Calster, Maarten van Smeden, and Andrew J. Vickers. 2019. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving calibration for long-tailed recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16484–16493.

# A Additional Metric: ACB

We introduce an additional calibration metric Absolute Calibration Bias (ACB, Equation 7) to measure the instance-level $L_1$ distance between predicted probabilities and true label frequencies. Unlike MCB, which measures global shifts, ACB computes the mean absolute error between the predicted probability $preds_{i,OL}$ and the observed annotation proportion $p_{i,OL}$ for each tweet $i$:

$$\text{ACB}_{OL} = \frac{1}{n}\sum_{i=1}^{n}\left|\text{preds}_{i,OL} - p_{i,OL}\right| \qquad (7)$$

As shown in Figure 15, the ACB results exhibit a counter-intuitive trend: the error increases as the bias $\beta$ increases. This pattern closely mirrors the F1 score trends rather than the MCB calibration trends. This occurs because higher bias pushes the simulated label proportions toward the boundaries (0 and 1). As the model predicts these polarized values more confidently, the absolute distance mechanically shrinks, even if the model is poorly calibrated in a statistical sense.

This finding highlights a critical limitation of instance-level calibration metrics like ACB in biased settings: they may conflate predictive confidence with calibration quality. Consequently, we argue that while ACB provides a granular view of model-annotator agreement, it must be interpreted alongside MCB to distinguish between global calibration shifts and local prediction errors. Future work should include a more diverse set of measures to ensure robustness across different data distributions.

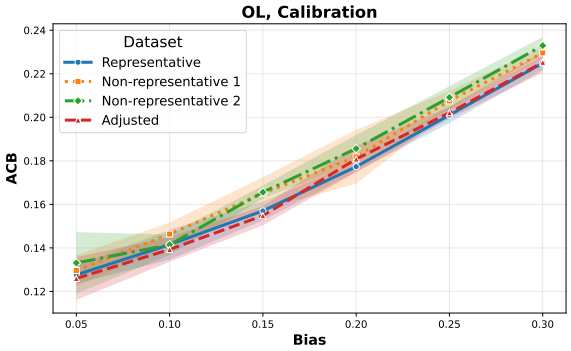


Figure 15: ACB scores for OL Models, by dataset and bias ($\beta$)