THE ANATOMY OF SOCIO-DEMOGRAPHICS IN PLURALISTIC OFFENSIVENESS MODELING: WHEN AND WHERE IT HELPS, AND THE ROLE OF REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Subjective judgments such as offensiveness often vary substantially across individuals, motivating pluralistic modeling approaches that incorporate annotators' socio-demographic information to capture this variation. However, prior findings on the benefits of socio-demographic information for pluralistic modeling are mixed, calling for a systematic analysis of both its effectiveness and the mechanisms through which it improves prediction. In this work, we study **when** and **where** socio-demographics benefit pluralistic offensiveness modeling, as well as **the role of explicit reasoning** in leveraging this information. Across three multi-annotator datasets with prompted and trained LLMs, we first show that demographic gains are small and inconsistent for individual predictions but become clearer at the group and population levels, because aggregation attenuates the idiosyncratic variation that masks small but systematic demographic signals. Second, using a reasoning and prediction format that explicitly decomposes individual judgment into an overall baseline and a demographic-specific deviation, we show that the gains arise from both components: demographic information improves baseline estimation and enables meaningful demographic-specific shifts, and interventions show that better component estimates improve final predictions. Finally, comparing direct demographic conditioning with explicit reasoning about how the text interacts with annotator demographics, we find that reasoning yields more accurate baseline and shift estimates as well as final predictions, whereas conditioning alone rarely produces informative shifts. These results suggest that the mixed evidence in prior work may partly stem from individual-level evaluation, and that the value of socio-demographic information depends on both how it is evaluated and how models are guided to use it.

1 INTRODUCTION

Subjective judgments such as sentiment, bias, and offensiveness often vary substantially across annotators. Rather than treating such disagreement as noise, pluralistic modeling aims to preserve and model the diversity of human judgments (Sorensen et al., 2024; Mostafazadeh Davani et al., 2022; Fleisig et al., 2023). A common approach is to incorporate annotators' socio-demographic information, motivated by the observation that social and cultural backgrounds can systematically shape how the same text is perceived. However, prior studies provide mixed evidence on its benefits for pluralistic modeling: demographic information improves prediction in some settings or along particular demographic axes, while providing limited benefits or even degrading performance in others (Mokhberian et al., 2024; Zhang & Coltekin, 2026; Orlikowski et al., 2023; 2025; Cai & Zafarani, 2026). These conflicting findings call for a systematic analysis of when socio-demographic information is beneficial, where its benefits arise, and how modeling choices affect its use.[1]

We first ask **when** socio-demographic information is beneficial across three levels of evaluation: individual annotators, socio-demographic groups, and the overall population. Existing work has evaluated demographic-aware modeling at these levels, but largely treats them as separate targets rather than examining how the benefits change across them (Fleisig et al., 2023; Parappan & Henao,

[1] A more detailed discussion of related work is provided in Appendix A.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025; Xu et al., 2026). This gap can obscure small but systematic demographic signals that are difficult to detect at the individual level yet become apparent after group or population aggregation. We find that demographic gains are indeed limited at the individual level and become more pronounced at the group and population levels. A simple statistical analysis explains why: aggregation attenuates residual individual variation, which demographics cannot explain, while preserving systematic demographic effects.

We next ask **where** in the prediction process these benefits arise. An individual judgment can be viewed as an overall baseline plus a demographic-specific deviation, but existing approaches leave this decomposition implicit and expose only the final prediction (Xu et al., 2026; Zhang & Coltekin, 2026). It is therefore unclear whether socio-demographics help by improving the baseline, the demographic-specific deviation, or both. We instead use a reasoning and prediction format that exposes both estimates: as illustrated in Figure 1, the model first estimates how a broad population would perceive the text (`crowd`) and then whether the target annotator would deviate from this baseline (`shift`) before making its final prediction. This decomposition makes each component inspectable: in the example, we can verify that the model correctly estimates a non-offensive `crowd` baseline and a positive `shift` for the Japanese annotator, reflecting the historical sensitivity of the topic, and thereby see why it makes the correct final offensive prediction despite the non-offensive majority. We find that socio-demographic information improves both components, and that interventions on them change final predictions in the expected direction.

Finally, we study **the role of reasoning** in leveraging demographic information. Throughout our analyses, we compare direct demographic conditioning with a variant that explicitly reasons about how the text interacts with the annotator's demographics. We find that explicit reasoning yields more accurate baseline and shift estimates than conditioning alone, which in turn improve aggregated predictions. Reasoning thus does not merely improve the final prediction; it strengthens how demographic information is used within each component.

In summary, our contributions are: (1) a multi-level evaluation showing that the value of socio-demographic information is underestimated by individual-level metrics, together with a statistical account of why; (2) a component-level analysis, enabled by explicitly decomposing predictions into a baseline and a demographic-specific shift, showing through evaluation and interventions that demographic gains arise from both improved baseline estimation and meaningful deviation estimation; and (3) a controlled comparison of demographic conditioning and demographic reasoning across three datasets and both prompted and trained LLMs, showing that explicit reasoning is key to effectively using socio-demographic information, yielding more accurate component estimates and improved predictions.

2 Incorporating Socio-Demographics in Offensiveness Modeling

We describe the task, the decomposition underlying our analyses, the model variants, and the experimental setup shared by Sections 3 and 4.

Problem Formulation Consider a dataset $\mathcal{D} = \{(x_i, a_i^j, y_i^j)\}$, where $x_i \in \mathcal{X}$ is the text of item i, $a_i^j \in \mathcal{A}_i \subseteq \mathcal{A}$ is the j-th annotator assigned to item i, and $y_i^j \in \{0, 1\}$ indicates whether annotator j labels x_i as offensive. Here, $\mathcal{A}$ denotes the annotator space of the dataset, and $\mathcal{A}_i$ the set of annotators of item i. Each annotator a_i^j is associated with a multi-attribute socio-demographic profile $\mathbf{z}_i^j$. We use $g \in \mathcal{G}$ to denote a demographic group, i.e., a particular value along a demographic axis.

An individual judgment can be decomposed into three components: an item-level baseline capturing how a broad pool of annotators would perceive the text, a demographic-specific deviation capturing how the annotator's demographic profile systematically shifts this perception for the item, and a zero-mean residual capturing idiosyncratic differences not explained by the profile:

$$\underbrace{y_i^j}_{\text{individual}} = \underbrace{c^\star(x_i)}_{\text{baseline}} + \underbrace{\delta^\star(x_i, \mathbf{z}_i^j)}_{\text{demographic-specific deviation}} + \underbrace{\varepsilon_i^j}_{\text{zero-mean residual}}, \tag{1}$$

Existing approaches model y_i^j directly, leaving the decomposition implicit. If a model instead outputs estimates of the baseline and the deviation, we can evaluate not only its final predictions $\hat{y}_i^j$ against y_i^j, but also the effect of demographic information on each component separately.

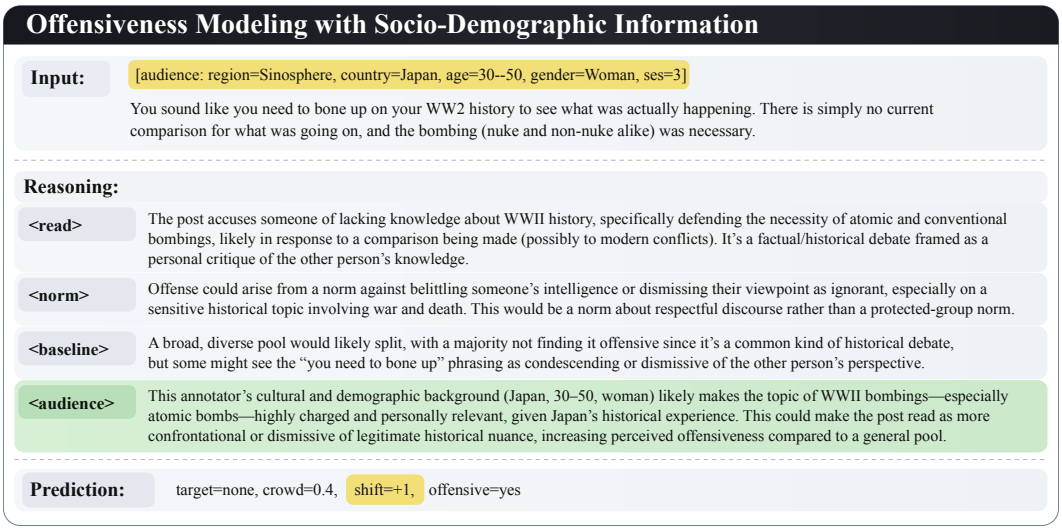

Figure 1: Example from D3CODE using the trained Qwen-4B model. Variants differ in whether they incorporate the annotator's socio-demographic profile and the demographic-specific deviation, and socio-demographics-aware reasoning. Here, the model correctly estimates a non-offensive baseline and, given the historical sensitivity of the topic, a positive shift for the Japanese annotator, leading to the correct offensive individual prediction despite the non-offensive majority judgment.

Variants of Reasoning and Prediction Format We therefore prompt or train LLMs to follow the reasoning and prediction format in Figure 1, which emits the baseline, the demographic-specific shift, and the final individual prediction as separate outputs. We consider three variants that differ in whether and how they use annotators' socio-demographic information. The basic variant, **NoDemo**, is annotator-blind and receives only the post. It reasons about the content through `read`, which summarizes the post, `norm`, which identifies potentially violated social norms, and `baseline`, which estimates how a broad population would perceive it. It then predicts the relevant `target`, the baseline (`crowd`$\in [0, 1]$), and the binary individual judgment (`offensive`$\in \{$`yes`, `no`$\}$). **DemoCond** (demographic conditioning) additionally receives the annotator's socio-demographic profile and predicts a demographic-specific `shift`$\in \{-1, 0, +1\}$ relative to the baseline, but does not explicitly reason about the effect of demographics. **DemoReason** (demographic reasoning) further adds an `audience` reasoning block that explicitly reasons about how the target demographic features may shift the baseline before producing the `shift` and `offensive` prediction.[2]

Prompting and Training We evaluate both prompted and trained LLMs. In the prompting setting, LLMs receive task instructions together with the reasoning and prediction format of each variant. For training, we adopt a standard two-stage post-training pipeline (Fan et al., 2025; Chen et al., 2026). Claude Sonnet 5 first generates cold-start reasoning trajectories for each variant using its corresponding inputs, the gold individual `offensive` label, and the empirical item-level mean of individual annotations as the `crowd` target. To avoid leaking demographic information into the reasoning, DemoCond reuses the reasoning blocks generated for NoDemo (without access to demographics), paired with the prediction line of the corresponding DemoReason trajectory. We then perform SFT on the resulting trajectories, followed by GRPO with rewards based on the correctness of the final individual predictions.

Datasets and Models We conduct experiments on three multi-annotator corpora: DICES (Aroyo et al., 2023), D3CODE (Davani et al., 2024), and POPQUORN (Pei & Jurgens, 2023). Each dataset provides multiple annotations per item (81.0/32.6/8.7 on average), enabling empirical estimates of overall baselines and proxy targets for demographic-specific deviations. Table 5 in Appendix B summarizes the dataset statistics and socio-demographic coverage. We train four open-source Llama-3.x and Qwen-3.5 models ranging from 3B to 9B parameters, and prompt nine models: open-source Llama and Qwen models ranging from 3B to 122B, and two commercial models, GPT-4o-mini

[2]We encode only the direction of the demographic effect in `shift`, as magnitude estimation is unstable and prone to collapse. Zero-shot and free-form CoT baselines are reported in Appendix C.

and Haiku-4.5. Model configurations, training and prompting details, and additional baselines are provided in Appendix B.

3 When Do Socio-Demographics Help? The Role of Aggregation

We first ask whether the benefits of incorporating annotator socio-demographics differ across levels of aggregation: individual, group, and population.

Evaluation Setup We evaluate model predictions at the individual, group, and population levels. At the individual level, each prediction is compared directly with its corresponding annotation. At the group level, model predictions and annotations are aggregated within each item–group pair to obtain model-estimated and empirical judgments for a specific socio-demographic group. At the population level, predictions are aggregated across all annotators for each item to obtain model-estimated and empirical population judgments.[3]

We then evaluate these predictions using both discrete and continuous metrics. For discrete evaluation, the individual predictions are evaluated directly, while group- and population-level estimates are binarized. We report F1 score (**F1**) and balanced accuracy (**Acc**). For continuous evaluation at the group and population levels, we directly compare the aggregated model estimates with their empirical targets using mean squared error (**MSE**) and Spearman correlation (ρ).

Table 1: Individual-, group- and population-level results for prompted and trained models. Scores are averaged across all dataset–model pairs. * / ** indicate significant improvement over NoDemo / over both NoDemo and the other demographic-aware variant (paired t-test, $p < 0.05$). **The benefits of incorporating socio-demographics become more apparent at higher levels of aggregation.**

Setting	Variant	Individual		Group				Population			
		F1 ↑	Acc ↑	F1↑	Acc↑	MSE↓	ρ ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑
Prompted	NoDemo	0.410	0.587	0.382	0.630	0.263	0.269	0.384	0.673	0.191	0.434
	DemoCond	0.414	**0.589**	0.390	0.636	**0.230***	0.295*	0.389	0.680*	0.132*	0.526*
	DemoReason	**0.426****	0.588	**0.394***	**0.639**	0.235*	**0.301***	**0.394***	**0.690****	**0.128***	**0.543****
Trained	NoDemo	**0.357**	**0.583**	0.375	**0.631**	0.198	0.291	0.412	0.684	0.120	0.473
	DemoCond	0.355	0.580	**0.375**	0.629	**0.179***	**0.303***	0.411	0.673	0.095*	0.512*
	DemoReason	0.351	0.565	0.375	0.628	0.181*	0.298*	**0.421***	**0.686**	**0.075****	**0.519***

3.1 Socio-Demographic Gains Become More Visible with Aggregation

Table 1 compares annotator-blind and demographic-aware variants and shows that the benefits of demographic modeling become increasingly visible at higher levels of aggregation.

At the individual level, demographic information provides little consistent advantage. DemoCond and DemoReason achieve balanced accuracy comparable to NoDemo for prompted models, while both underperform NoDemo in F1 and Acc after training. At the group level, the gains of demographic modeling become clearer. On discrete metrics, demographic-aware variants begin to outperform NoDemo for prompted models and perform comparably to NoDemo after training, while significantly improving continuous metrics in both settings.

The advantage is strongest at the population level. For prompted models, the F1 gains of DemoCond/DemoReason persist, while their Acc gains increase from 0.006/0.009 at the group level to 0.007/0.017, with DemoReason significantly outperforming NoDemo on both metrics. For trained models, DemoReason also surpasses NoDemo in both F1 and Acc. Continuous metrics show an even clearer aggregation effect. For prompted models, the MSE gains of DemoCond/DemoReason at the population level increase to approximately 1.8×/2.3× their group-level gains, while the corresponding ρ gains increase to 3.5×/3.4×. Similarly, for trained models, the corresponding MSE gains at the population level increase by approximately 1.3×/2.6×, while the ρ gains increase to 3.3×/6.6× their group-level improvements.

[3]Note that the model-emitted `crowd` baseline and aggregated individual predictions both estimate broad population perception but are computed differently. We refer to the former as baseline or `crowd` baseline, and to the latter as population or population-level judgments. Their modeling performance is compared in Table 2.

Taken together, socio-demographic information captures useful systematic variation whose contribution is difficult to observe at the individual level but becomes increasingly visible as aggregation attenuates residual variation. This pattern is particularly clear for the continuous metrics, which avoid information loss from binarization. Similar patterns are observed across the other baselines and when transferring to unseen datasets. Complete per-model results are reported in Appendix C, with cross-dataset and held-out demographic-group generalization results in Appendix D.

Compared with simple demographic conditioning, explicit demographic reasoning generally yields larger gains, particularly at the group and population levels. Accordingly, DemoReason typically outperforms DemoCond at these levels, suggesting that explicit reasoning helps LLMs better exploit the systematic information conveyed by socio-demographic profiles.

3.2 Aggregation Reveals Demographic Signal Obscured by Individual Noise

Why are demographic gains hard to observe individually but clear after aggregation? Recall Eq. 1: $y_i^j = c^\star(x_i) + \delta^\star(x_i, \mathbf{z}_i^j) + \varepsilon_i^j$, with $\mathbb{E}[\varepsilon_i^j \mid x_i, \mathbf{z}_i^j] = 0$. Let $p_i^j := c^\star(x_i) + \delta^\star(x_i, \mathbf{z}_i^j) = \mathbb{E}[y_i^j \mid x_i, \mathbf{z}_i^j]$ denote the systematic component, such that $y_i^j = p_i^j + \varepsilon_i^j$. Socio-demographic information may improve estimation of p_i^j, but cannot explain ε_i^j. Assume that the residuals are conditionally independent with a common variance σ^2 and independent of the model predictions. When estimating the mean judgment $\bar{y}_i$ of n annotators with the mean prediction $\bar{\hat{y}}_i$, the expected squared loss decomposes into a reducible model error and an irreducible residual variance term:

$$\mathbb{E}[(\bar{\hat{y}}_i - \bar{y}_i)^2] = \underbrace{\mathbb{E}[(\bar{\hat{y}}_i - \bar{p}_i)^2]}_{\text{reducible}} + \underbrace{\mathrm{Var}(\bar{\varepsilon}_i)}_{\text{irreducible}} = \underbrace{\mathbb{E}[(\bar{\hat{y}}_i - \bar{p}_i)^2]}_{\text{reducible}} + \underbrace{\sigma^2/n}_{\text{irreducible}} . \tag{2}$$

In subjective tasks, the residual captures idiosyncratic preferences and other unsystematic variation, which can constitute a substantial portion of prediction error. Thus, at the individual level ($n = 1$), residual variation can mask improvements in the systematic component. From individual to group and population levels, the residual variance decreases at rate $1/n$, while systematic differences in expected judgments are preserved. Aggregation therefore does not make socio-demographic information more informative; it makes its contribution easier to observe.

Empirical Evidence Two empirical observations support this explanation. First, a bias-corrected decomposition of within-item label variation shows that no single demographic axis explains more than 6.2% of the variation across datasets (D3CODE: $\leq$ 6.2%; DICES: $\leq$ 1.7%; POPQUORN: $\leq$ 2.6%), indicating that demographics capture only a small systematic component of individual variation (Hu & Collier, 2024; Sun et al., 2025). Second, inter-annotator agreement differs little between annotator pairs from the same versus different demographic groups: the largest gap on the test split is only 3.1 percentage points, compared with overall agreement rates of 63–71%. Demographic effects are thus systematic but small relative to residual individual variation.

Takeaway: Socio-demographic information improves pluralistic offensiveness modeling, but its benefits are obscured by residual variation at the individual level and emerge clearly at the group and population levels. Explicit reasoning yields larger gains than simple conditioning.

4 Where Do Socio-Demographics Help in the Prediction Process?

Leveraging the decomposition of individual predictions into a baseline and a demographic-specific shift, we next examine whether socio-demographics improve baseline estimation (Sections 4.1–4.2), demographic-specific shift estimation (Section 4.3), or both, and whether these improvements carry over to final predictions (Section 4.4). We provide a qualitative comparison of the variants in terms of both prediction components and final predictions in Appendix E.

4.1 Incorporating Socio-Demographics Improves Baseline Estimation

Somewhat surprisingly, incorporating socio-demographics leads to more accurate estimates of the baseline, a quantity that should not depend on the target annotator. We evaluate the model-emitted `crowd` against an empirical baseline $c(x_i)$, computed by averaging the individual annotations for

item i. Since these annotations come from demographically diverse annotator pools, The empirical baseline $c(x_i)$ should reliably approximate the underlying $c^\star(x_i)$. Because demographic-aware variants may emit different `crowd` values for different annotators of the same item, we evaluate them at two levels: directly at the annotation level and, after averaging across annotators, at the item level. We report mean squared error (**MSE**) and mean absolute error (**MAE**) for point-wise quality, together with Spearman correlation (ρ) and Wasserstein-1 distance (**W1**) for recovering the relative ordering and overall distribution of empirical baselines, respectively.

Table 2: Annotation- and item-level baseline estimation results for prompted and trained models. Scores are averaged across all dataset–model pairs. * / ** indicate significant improvement over NoDemo / over both NoDemo and the other demographic-aware variant (paired t-test, $p < 0.05$). *Avg. Individual Pred.* averages individual predictions across annotators for each item and reports the best score across variant–setting pairs for each metric. **Incorporating socio-demographic information, particularly through explicit reasoning, consistently improves baseline estimation.**

Setting	Variant	Baseline Estimation (Ann.)				Baseline Estimation (Item)			
		MSE ↓	MAE ↓	ρ ↑	W1 ↓	MSE ↓	MAE ↓	ρ ↑	W1 ↓
Prompted	NoDemo	0.067	0.195	0.524	0.114	0.065	0.194	0.542	0.115
	DemoCond	0.061	0.186	0.529	0.102*	0.051*	0.172*	**0.603***	0.101*
	DemoReason	**0.059**	**0.183***	**0.538**	**0.097***	**0.050***	**0.169***	0.601*	**0.097**
Trained	NoDemo	0.050	0.168	0.505	0.098	0.045	0.160	0.566	0.094
	DemoCond	0.047	0.164	**0.510**	0.090	0.038*	0.148*	**0.604***	0.086
	DemoReason	**0.047**	**0.163**	0.502	**0.083****	**0.037***	**0.146***	0.599*	**0.080***
Avg. Individual Pred.		0.136	0.299	0.417	0.242	0.075	0.210	0.546	0.139

Table 2 shows that incorporating annotator demographics improves `crowd` baseline estimation, with the clearest gains at the item level. In the prompted setting, both DemoCond and DemoReason improve all annotation- and item-level metrics over NoDemo, with almost all item-level gains reaching significance; the trained setting shows largely the same pattern. The emitted `crowd` estimates also substantially outperform *Avg. Individual Pred.*, despite having no access to the empirical target distribution, suggesting an advantage of explicitly reasoning of the baseline and modeling the baseline directly as a continuous output rather than recovering it by averaging binary predictions.

Between the two demographic-aware variants, DemoReason matches or improves on DemoCond in MSE, MAE, and W1 at both levels and in both settings. Its advantage is most evident at the annotation level: DemoReason significantly improves prompted MAE over NoDemo where DemoCond does not, and significantly outperforms both NoDemo and DemoCond on trained W1. Although DemoReason's improvement in prompted item-level W1 does not reach significance while DemoCond's does, its average W1 remains better. Explicit demographic reasoning thus strengthens baseline estimation beyond simple conditioning. Per-model results are provided in Appendix F.

4.2 Why Does Socio-Demographic Information Improve Baseline Estimation?

Since both DemoCond and DemoReason improve baseline estimation even for prompted LLMs, these gains cannot be explained solely by exposure to empirical baseline distributions during training. Instead, they suggest that socio-demographic information changes how models reason about and estimate the baseline. To understand this effect, and the additional role of explicit reasoning, we analyze model behavior and identify two contributing mechanisms.

Incorporating Socio-Demographics Captures Heterogeneity and Reduces Over-Consensus. Conditioning on a specific demographic profile leads to less extreme `crowd` estimates, reducing the model's tendency to assume excessive agreement across annotators. NoDemo, DemoCond, and DemoReason predict an item-level `crowd` below 0.1 for 37.1%, 26.9%, and 25.6% of items, respectively, compared with only 16.8% in the empirical data. DemoReason reduces this over-consensus most, suggesting that explicitly reasoning about text–annotator interactions makes the model more sensitive to heterogeneous perceptions of the same content. Consistent with this explanation, Table 3 shows that demographic-aware modeling yields substantially larger improvements

on high-disagreement items, suggesting that the baseline gains arise from better modeling annotator heterogeneity. DemoReason further strengthens these gains over DemoCond.

Table 3: Gain over NoDemo in item-level baseline estimation, stratified by disagreement (low/high disagreement: empirical baseline outside/inside $[0.4, 0.6]$, 83.1%/16.9% of items). Positive values indicate improvement over NoDemo, with larger values indicating greater gains. **Demographic-aware variants show larger gains on high-disagreement items.**

Regime	Variant	Low disagreement				High disagreement			
		MSE	MAE	ρ	W1	MSE	MAE	ρ	W1
Prompted	DemoCond	+0.013	+0.019	**+0.075**	+0.013	+0.020	+0.041	**+0.022**	+0.042
	DemoReason	**+0.014**	**+0.019**	+0.069	**+0.014**	**+0.026**	**+0.052**	+0.017	**+0.052**
Trained	DemoCond	+0.003	+0.006	+0.034	+0.009	+0.012	+0.025	+0.024	**+0.024**
	DemoReason	**+0.005**	**+0.008**	**+0.048**	**+0.013**	**+0.014**	**+0.031**	**+0.091**	+0.023

Reasoning Makes Baseline Estimates More Stable Across Annotators. A second mechanism helps explain why DemoReason outperforms DemoCond on annotation-level baseline metrics, particularly for prompted LLMs. By reasoning about demographic-specific perceptions in a separate block, DemoReason better separates the annotator-specific component from the `crowd` baseline, which should be invariant to the target annotator. It therefore produces more consistent `crowd` estimates across annotators of the same item, as reflected by the lower within-item standard deviation in Figure 2. The greater stability allows item-level improvements to carry over to the annotation level. This advantage is not observed for trained models, whose within-item `crowd` variation is already lower than that of their prompted counterparts.

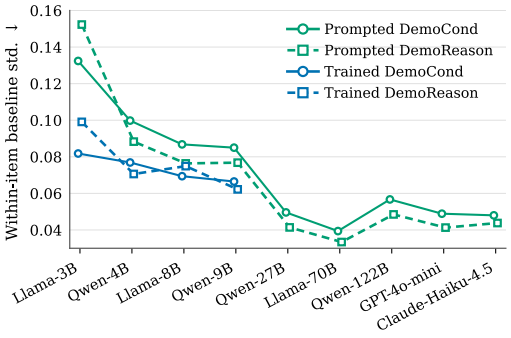

Figure 2: Within-item standard deviation of baseline estimates for DemoCond and DemoReason across trained and prompted LLMs. **DemoReason shows lower within-item variation than DemoCond for prompted models.**

Takeaway: Incorporating annotator socio-demographics improves baseline estimation by making models more sensitive to heterogeneous perceptions, while explicit reasoning further improves estimation by making the inferred baseline more stable across annotators.

4.3 LLMs Capture Meaningful Demographic-Specific Deviations

We next ask whether the model-emitted `shift` captures systematic demographic-specific deviations from the baseline. Since neither the true deviation $\delta^\star(x_i, \mathbf{z}_i^j)$ nor its empirical counterpart $\delta(x_i, \mathbf{z}_i^j)$ is directly observable, we evaluate `shift` against three increasingly fine-grained proxy targets: global group tendencies, individual–majority deviations, and item–group interactions.

Reasoning-Enhanced Shift Predictions Recover Global Group Tendencies We first test whether `shift` predictions recover dataset-level group tendencies. For each annotation of item i by annotator j, we define a noisy individual-level shift proxy as $y_i^j - c(x_i)$, i.e., the deviation of the individual annotation from the empirical baseline. Averaging these shifts over annotations from demographic group g yields its empirical global tendency Δ_g, indicating whether members of g tend to judge their assigned items as more or less offensive than the corresponding baselines. We analogously obtain $\hat{\Delta}_g$ by averaging the model-emitted `shift` for DemoCond and DemoReason; for NoDemo, which does not explicitly predict a `shift`, we instead use $\hat{y}_i^j - \hat{c}(x_i)$. We then compare $\hat{\Delta}_g$ with Δ_g using Pearson correlation (r), Spearman correlation (ρ), and balanced sign agreement (**Sign**) across groups within each demographic axis.

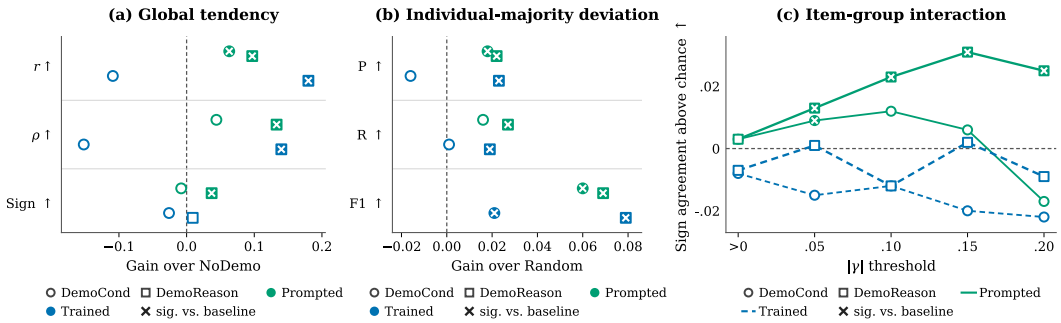

Figure 3: Model-emitted `shift` results for (a) global group tendency, shown as gain over NoDemo; (b) individual–majority deviations, shown as gain over random; and (c) item–group interactions, shown as sign agreement above chance across $|\gamma|$ thresholds. Scores are averaged over dataset–model pairs; significance is assessed with paired t-tests ($p < 0.05$). **DemoReason consistently captures meaningful demographic effects across all three analyses.**

Figure 3(a) shows that DemoReason recovers dataset-level group tendencies: it outperforms both NoDemo and DemoCond across all metrics and settings, with significant gains in all cases except trained sign agreement. In contrast, DemoCond yields only limited and inconsistent gains for prompted models and even underperforms NoDemo after training. One possible explanation is that DemoCond predicts `shift`=0 for the vast majority of instances, suggesting that demographic conditioning alone often fails to elicit demographic-specific deviations. These results suggest that LLMs can leverage socio-demographic information to recover systematic group tendencies, but that doing so relies on explicit reasoning rather than conditioning alone.

Reasoning-Enhanced Shifts Capture Individual–Majority Deviations Beyond Chance. We next examine `shift` at a finer granularity using a three-way proxy target based on individual–majority prediction deviation to approximate the underlying demographic-specific deviation. The proxy is 0 when the individual annotation agrees with the empirical majority, $+1$ when the majority is non-offensive but the individual annotation is offensive, and -1 in the opposite direction.[4] Performance above a random baseline would indicate that `shift` captures item-specific demographic signal informative of individual–majority deviations. Figure 3(b) reports macro-averaged precision, recall, and F1 over the three proxy classes.

In line with the global tendency results, DemoReason consistently outperforms the random baseline, whereas DemoCond shows no consistent improvement, particularly after training. The gains are nevertheless modest, consistent with Section 3.2, where socio-demographic attributes explain only a small fraction of individual-level variation.

Reasoning Identifies Item–Group Interactions Beyond Global Group Tendency An observed demographic-specific deviation may reflect both a group's global tendency and an item-specific *interaction* with the text. We therefore ask whether `shift` captures these finer-grained interactions beyond global tendencies. For example, although annotators from Japan are globally less likely to rate texts as offensive, DemoReason predicts a positive shift for the example in Figure 1, capturing how the specific content interacts with the annotator's demographics. Since such interactions are sparse and lack explicit ground-truth labels, we focus on item–group pairs whose demographic effects deviate substantially from the corresponding global tendency.

To quantify such interactions, for each item i and group g with at least $m = 5$ annotations, we first define the observed deviation as $\delta_{i,g} = c_g(x_i) - c(x_i)$, where $c_g(x_i)$ and $c(x_i)$ are the group-specific and overall empirical baseline, respectively. We then subtract the group's global tendency Δ_g to obtain the item-specific interaction $\gamma_{i,g} = \delta_{i,g} - \Delta_g$. Analogously, from the model-emitted `shift`, we compute $\hat{\gamma}_{i,g} = \bar{\texttt{shift}}_{i,g} - \hat{\Delta}_g$, where $\bar{\texttt{shift}}_{i,g}$ is the mean predicted shift for the item–group pair. We evaluate balanced sign agreement between $\gamma_{i,g}$ and $\hat{\gamma}_{i,g}$ at increasing $|\gamma|$ thresholds, progressively focusing on pairs with stronger item–group interactions.

[4] This is a noisy proxy for the demographic-specific deviation $\delta^\star(x_i, \mathbf{z}_i^j)$, as it inherits the individual residual ε_i^j and captures only effects large enough to change the final discrete prediction.

Figure 3(c) shows that prompted DemoReason is the only variant to perform significantly above chance on item–group interactions. Its performance further improves as the $|\gamma|$ threshold increases, consistent with stronger interactions being less affected by annotation noise. However, its improvement over chance remains below 0.04, indicating that the captured interaction signal is modest. These results suggest that, with explicit reasoning, prompted LLMs can identify fine-grained interactions between text content and annotator socio-demographics.

Takeaway: With explicit reasoning about socio-demographic information, LLMs capture meaningful demographic-specific deviations that do not consistently emerge under simple conditioning.

4.4 Better Baseline and Deviation Estimates Improve Final Predictions

Finally, we examine whether better baseline and deviation estimates translate into better final predictions, using observational and interventional evidence on POPQUORN.

Observational Evidence We first provide observational evidence by correlating prediction correctness with the quality of `crowd` and `shift` estimates. We measure `crowd` quality as the negative squared error from the empirical baseline, and `shift` correctness against the three-way individual–majority deviation proxy. Across trained and prompted DemoCond and DemoReason models, individual prediction correctness is positively correlated with both `crowd` quality ($r = 0.24$–0.27) and `shift` correctness ($r = 0.40$–0.57).

Interventional Evidence We intervene on the `crowd` and `shift` estimates of a trained and a prompted LLM (Llama3.1-8B and Llama3.3-70B) and measure the resulting changes in final predictions. An **oracle** intervention replaces `crowd` with the empirical baseline and `shift` with the individual–majority deviation proxy, whereas an **adversarial** intervention replaces both with estimates from a randomly sampled item. As shown in Table 4, oracle interventions mostly improve prediction performance, while adversarial interventions generally degrade it. Component-wise interventions (Appendix G) show a clearer effect for `crowd`, whereas the original `shift` predictions do not consistently outperform adversarial shifts, consistent with the modest shift effects in Section 4.3. Together, these results indicate that more accurate baseline and shift estimates are associated with better predictions, and that the quality of these components directly influences final predictions.

Table 4: Performance under joint interventions on `crowd` and `shift` estimates. Each cell reports *Original / Oracle / Adversarial* performance. Bold marks cells following the expected ordering Oracle $>$ Original $>$ Adversarial. **Better baseline and shift estimates are not only associated with, but also lead to, improved final predictions.**

Model	Variant	Individual	Group		Population	
		F1 ↑	F1 ↑	MSE ↓	F1 ↑	MSE ↓
Llama-8B (Trained)	DemoCond	**0.453/0.540/0.364**	**0.442/0.589/0.424**	0.190/0.105/0.170	**0.549/0.667/0.444**	**0.094/0.084/0.106**
	DemoReason	**0.474/0.600/0.470**	**0.450/0.522/0.439**	**0.188/0.145/0.196**	0.526/0.600/0.541	0.059/0.065/0.068
Llama-70B (Prompted)	DemoCond	**0.438/0.554/0.312**	**0.425/0.544/0.257**	**0.178/0.140/0.200**	**0.435/0.652/0.387**	0.095/0.068/0.077
	DemoReason	**0.458/0.557/0.401**	**0.443/0.550/0.379**	**0.180/0.166/0.201**	**0.500/0.630/0.465**	**0.084/0.068/0.085**

5 Conclusion

We analyzed when and where socio-demographic information benefits pluralistic offensiveness modeling, together with the role of explicit reasoning. The value of socio-demographics lies more in capturing systematic group- and population-level patterns than in recovering individual judgments, whose idiosyncratic variation can obscure these gains until it is reduced through aggregation. These benefits arise from both improved baseline estimation and meaningful, though modest, demographic-specific deviation estimates, while explicit reasoning about socio-demographic information strengthens both components beyond direct conditioning and leads to better predictions. Limited individual-level gains should thus not be interpreted as evidence that socio-demographic information is uninformative. We encourage future work to evaluate socio-demographic modeling across aggregation levels and, where possible, across distinct prediction components. [5]

[5] Limitations of our work are discussed in Appendix H.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development, related work discovery, and the editing and polishing of the manuscript. We did not use generative AI tools to formulate the research ideas or interpret the experimental results; these were developed and analyzed by the authors. All AI-assisted code and text were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the details necessary to reproduce our experiments in the main paper, Appendix, and supplementary materials, including model configurations, prompting and training procedures, evaluation protocols, and implementation details. The complete codebase will be publicly released upon acceptance of the paper.

ETHICS STATEMENT

Using socio-demographic information in predictive models warrants careful consideration. Demographic attributes may introduce broad group-level associations or stereotypes that do not always reflect individual annotators accurately. Such information should therefore be used cautiously and should not be interpreted as deterministically implying an individual's beliefs or judgments. Our explicit reasoning format makes the use of demographic information more transparent by exposing the model's baseline and demographic-specific reasoning, which can facilitate inspection and auditing. However, transparency does not guarantee that the reasoning is unbiased or appropriate, and the generated rationales themselves may reproduce stereotypes. We therefore recommend careful evaluation and human inspection when demographic-aware models are deployed or used to support real-world decisions.

REFERENCES

Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53330–53342. Curran Associates, Inc., 2023. doi: 10.52202/075280-2321. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a74b697bce4cac6c91896372abaa8863-Paper-Datasets_and_Benchmarks.pdf.

Weibin Cai and Reza Zafarani. When does demographic information help? data and modeling regimes for perspective-aware hate speech detection. *arXiv preprint arXiv:2605.27313*, 2026.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru WANG, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. Rm-r1: Reward modeling as reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88313–88342, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8e3b8de251afd887fb4589c1e3a3c793-Paper-Conference.pdf.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18511–18526, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1029. URL https://aclanthology.org/2024.emnlp-main.1029/.

Zhiting Fan, Ruizhe Chen, and Zuozhu Liu. BiasGuard: A reasoning-enhanced bias detection tool for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics:*

ACL 2025, pp. 9753–9764, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.506. URL https://aclanthology.org/2025.findings-acl.506/.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.415. URL https://aclanthology.org/2023.emnlp-main.415/.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3502004. URL https://doi.org/10.1145/3491102.3502004.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7337–7349, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.407. URL https://aclanthology.org/2024.naacl-long.407/.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL https://aclanthology.org/2022.tacl-1.6/.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1017–1029, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.88. URL https://aclanthology.org/2023.acl-short.88/.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2092–2111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.104. URL https://aclanthology.org/2025.acl-long.104/.

Mohammed Fayiz Parappan and Ricardo Henao. Learning subjective label distributions via sociocultural descriptors. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and

Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20322–20338, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1026. URL https://aclanthology.org/2025.emnlp-main.1026/.

Jiaxin Pei and David Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In Jakob Prange and Annemarie Friedrich (eds.), *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pp. 252–265, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.law-1.25. URL https://aclanthology.org/2023.law-1.25/.

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. Position: A roadmap to pluralistic alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2047–2095, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.106. URL https://aclanthology.org/2025.emnlp-main.106/.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 845–854, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.71. URL https://aclanthology.org/2025.naacl-short.71/.

Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41260–41294, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1914. URL https://aclanthology.org/2026.acl-long.1914/.

Leixin Zhang and Cagri Coltekin. Modeling human perspectives with socio-demographic representations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40781–40794, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2026. URL https://aclanthology.org/2026.findings-acl.2026/.

A Related Work

Socio-Demographics in Pluralistic Modeling Prior work has explored incorporating annotators' socio-demographic information into pluralistic modeling at multiple granularities, with mixed results. Some approaches report gains for certain demographic axes or experimental settings (Gordon et al., 2022; Fleisig et al., 2023; Xu et al., 2026; Zhang & Coltekin, 2026), whereas others find limited benefits or even negative effects (Orlikowski et al., 2023; 2025; Sorensen et al., 2025). These findings are consistent with our observation that demographic information does not reliably improve modeling at lower levels of aggregation. Cai & Zafarani (2026) further investigates the

Table 5: Dataset statistics and socio-demographic coverage.

Dataset	Test Annots. / Items	# Annots. / Item	Axes (# groups)
D3CODE	2,507 / 77	32.6	region (8), country (21), age (3), gender (3), SES (10)
DICES	10,849 / 134	81.0	gender (4), race (6), age (3), education (3)
POPQUORN	1,308 / 150	8.7	gender (3), race (6), age (10), education (5)

data and training conditions under which demographic modeling benefits individual-level prediction. However, existing work largely treats individual, group, and population modeling as separate evaluation targets, rather than systematically characterizing when the benefits of demographic information become more visible across levels of aggregation. Moreover, prior work typically evaluates only the resulting predictions, leaving unclear where the benefits of socio-demographic information arise within the prediction process: whether they come from better estimation of a shared overall baseline, demographic-specific shifts, or both. Consequently, it provides limited insight into how demographic information contributes to improved pluralistic modeling.

Reasoning in LLMs LLMs with reasoning capabilities have achieved substantial gains in tasks such as mathematics and coding (Jaech et al., 2024; Guo et al., 2025). More recently, reasoning has also been introduced into subjective tasks, including reward modeling and bias detection (Chen et al., 2026; Fan et al., 2025). These approaches typically first distill reasoning traces from stronger reasoning models to establish the desired format, and then further optimize the model with RL objectives such as DPO or GRPO. However, existing work generally evaluates reasoning as a complete modeling paradigm, without isolating whether reasoning about a particular source of information changes how that information is used by the model. Our work instead examines how explicit reasoning about socio-demographics affects distinct components of pluralistic modeling, including baseline estimation, demographic-specific shift estimation, and the final prediction.

B Details on Experimental Setup

Datasets The dataset statistics and socio-demographic coverage of the three multi-annotator datasets are summarized in Table 5. We report only test-set statistics, since the effective amount of training data varies with the number of SFT and GRPO optimization steps.

Models We use both open-source and commercial models for training and prompting. For training, we consider Llama3.2-3B, Llama3.1-8B, Qwen3.5-4B, and Qwen3.5-9B. For prompting, we additionally evaluate Qwen3.5-27B, Llama3.3-70B, Qwen3.5-122B-A10B, GPT-4o-mini, and Haiku-4.5. We use the instruction-tuned variants of all Llama models. For Qwen models, thinking mode is disabled during training. In the prompting setting, we likewise disable thinking mode for all models except Qwen3.5-122B-A10B, as enabling it substantially reduces output parseability for the smaller models.

DemoReason Prompt We present the complete DemoReason prompt below. DemoCond and NoDemo are constructed by removing specific components from this full prompt, as described afterward.

System prompt.

```
You are a careful fairness and toxicity judge. Given a social media post preceded by an
[audience: ...] annotator profile, produce structured reasoning in tagged blocks ---
<read>, <norm>, <baseline>, <audience> --- and end with a single line in exactly this
format:
=== Verdict: target=<group|none>, crowd=<0.0-1.0>, shift=<-1|0|+1>, offensive=<yes|no>
shift encodes DIRECTION ONLY: +1 if this annotator's group finds the post MORE offensive
than the broad pool, -1 if LESS, 0 if no meaningful difference.
```

User turn.

```
You are annotating social media posts for offensiveness on behalf of a specific
annotator. Posts are rated by many annotators, who do not always agree. Decide how this
post would be rated by the annotator described below.
POST: {text}
ANNOTATOR: [audience: {axis_1}={val_1}, {axis_2}={val_2}, ...]
<read> what is said, literally and by implication, and about whom
<norm> what norm would make this offensive, if any, and whose norm
<baseline> how a broad, diverse pool of annotators would split on this, and why
<audience> whether this annotator's group tends to read THIS post differently from the
broad pool, and why. shift encodes DIRECTION ONLY and must be exactly one of three
values: +1 if the group would find it MORE offensive than the broad pool, -1 if LESS, 0
if no meaningful difference. It is a direction, not an intensity: never write +2, -2, or
any other number, no matter how large the difference feels. 0 is the correct, expected
answer whenever the group reads the post about the same as the pool.
=== Verdict: target=<jewish|christian|muslim|LGB|transgender|none>, crowd=<0.0-1.0>,
shift=<-1|0|+1>, offensive=<yes|no>
Rules:
- Reason as genuine analysis.
- Never infer an opinion from an identity ("as a woman, she would...").
Reason about norms, context, and what registers as ordinary speech.
- Report crowd as the fraction of a broad, diverse pool that would find the post
offensive, rounded to one decimal place: one of 0.0, 0.1, 0.2, ..., 1.0.
- In the verdict line write bare values, never the angle brackets from the specification
above.
- Use 1--3 sentences per block and use the tags exactly.
- The verdict line must be the final line of the response.
```

DemoCond Prompt DemoCond removes the explicit <audience> reasoning block from DemoReason while retaining the annotator socio-demographic profile and the shift prediction. Accordingly, the required reasoning blocks are <read>, <norm>, and <baseline>; all other task instructions, demographic inputs, output fields, and rules remain unchanged.

NoDemo Prompt NoDemo further removes all annotator-specific information. The ANNOTATOR profile is omitted, the task is framed without reference to a specific annotator, and both the <audience> block and shift field are removed. The remaining reasoning blocks are <read>, <norm>, and <baseline>, followed by target, crowd, and offensive.

Training Configuration We use identical training hyperparameters across all datasets and variants. For SFT, we apply QLoRA with 4-bit quantization (LoRA rank 16, $\alpha = 32$), a learning rate of 2×10^{-4}, 3 epochs, an effective batch size of 32, and a maximum sequence length of 1024. GRPO is initialized from the corresponding SFT checkpoint and trained for 150 steps with a learning rate of 1×10^{-6}, batch size 8 with gradient accumulation 16, 8 generations per prompt with temperature=1.0, $\beta = 0.001$, and a maximum completion length of 768. Importantly, GRPO uses only final-label correctness as the reward, based on the emitted offensive prediction, without any explicit reward on crowd or shift. At inference time, we use greedy decoding.

Baselines We compare against several baselines to contextualize our results. For prompted LLMs, we consider zero-shot and free-form CoT baselines, both with and without user profiles in the input. All prompted baselines are required to output a crowd field as their estimate of the baseline. For training, we compare BERT models that are annotator-blind with majority-vote or individual labels (Mostafazadeh Davani et al., 2022), and annotator-aware with textual demographic descriptions (Fleisig et al., 2023) or learned demographic embeddings (Gordon et al., 2022). We also include DEM-MoE (Xu et al., 2026), a more recent method that uses external LLM judgments during training. The predicted positive-class probability is used as the baseline estimate.

C Full Performance Results

Tables 6 and 7 report per-model performance at the individual, group, and population levels. The baseline models exhibit the same overall pattern: the benefits of incorporating socio-demographics are limited at the individual level but become increasingly apparent with aggregation.

Table 6: Performance for all *prompted* models and baselines across different levels. Metric scores are averaged over three datasets, and *Overall* averages across backbones. * denotes significant improvement over NoDemo; ** denotes significant improvement over both NoDemo and the other variant across model–dataset pairs (paired t-test, $p < 0.05$).

Model	Variant	Individual		Group				Population			
		F1 ↑	Acc ↑	F1↑	Acc↑	MSE↓	ρ ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑
Llama-3B	NoDemo	**0.455**	0.583	0.376	0.623	0.377	0.247	0.324	0.645	0.302	0.401
	DemoCond	0.453	**0.584**	**0.380**	**0.624**	0.320	**0.271**	0.334	0.639	0.194	**0.533**
	DemoReason	0.425	0.569	0.360	0.608	**0.280**	0.227	**0.337**	**0.663**	**0.134**	0.498
Qwen-4B	NoDemo	**0.426**	**0.593**	**0.391**	**0.638**	0.260	**0.282**	**0.384**	**0.682**	0.189	0.443
	DemoCond	0.364	0.575	0.362	0.618	0.211	0.270	0.360	0.664	0.115	0.494
	DemoReason	0.346	0.567	0.342	0.609	**0.200**	0.262	0.335	0.648	**0.095**	**0.542**
Llama-8B	NoDemo	0.406	**0.570**	0.352	0.605	0.271	0.210	**0.377**	0.659	0.197	0.356
	DemoCond	0.397	0.561	0.348	0.602	**0.221**	0.223	0.361	0.659	**0.089**	0.490
	DemoReason	**0.474**	0.553	**0.371**	**0.609**	0.342	**0.277**	0.343	**0.665**	0.202	**0.564**
Qwen-9B	NoDemo	0.358	0.572	0.361	0.618	0.229	0.229	0.356	0.651	0.162	0.381
	DemoCond	0.436	0.602	**0.420**	**0.663**	**0.227**	0.314	**0.424**	0.714	0.133	0.548
	DemoReason	**0.456**	**0.610**	0.418	0.662	0.232	**0.326**	0.414	**0.723**	**0.128**	**0.568**
Qwen-27B	NoDemo	0.441	0.609	0.422	0.660	0.228	0.317	0.428	0.718	0.158	0.507
	DemoCond	0.449	0.611	0.437	0.668	**0.219**	0.338	0.453	0.730	0.132	0.553
	DemoReason	**0.457**	**0.614**	**0.442**	**0.678**	0.222	**0.347**	**0.454**	**0.734**	**0.130**	**0.576**
Llama-70B	NoDemo	0.411	0.596	0.412	0.655	0.219	0.296	0.427	0.701	0.150	0.471
	DemoCond	0.417	0.599	0.420	0.657	0.201	**0.331**	0.413	0.690	0.127	0.526
	DemoReason	**0.420**	**0.600**	**0.421**	**0.660**	**0.199**	0.330	**0.432**	**0.704**	**0.118**	**0.533**
Qwen-122B	NoDemo	0.358	0.579	0.389	0.637	0.195	0.291	0.420	0.683	0.134	0.433
	DemoCond	0.377	0.582	0.386	0.628	0.191	0.312	0.392	0.659	**0.098**	0.510
	DemoReason	**0.400**	**0.589**	**0.407**	**0.646**	**0.189**	**0.325**	**0.438**	**0.686**	0.100	**0.518**
GPT-4o-mini	NoDemo	**0.476**	0.601	0.372	0.612	0.375	0.281	0.329	0.647	0.291	0.495
	DemoCond	0.466	**0.607**	0.390	0.638	0.285	**0.312**	0.383	0.700	0.181	**0.583**
	DemoReason	0.460	0.600	**0.394**	**0.640**	**0.257**	0.298	**0.405**	**0.716**	**0.135**	0.542
Haiku-4.5	NoDemo	0.356	0.579	0.366	0.621	0.209	0.269	**0.415**	0.670	0.141	0.421
	DemoCond	0.368	0.579	0.368	0.625	0.200	0.282	0.383	0.664	0.116	0.500
	DemoReason	**0.398**	**0.588**	**0.390**	**0.637**	**0.198**	**0.317**	0.388	**0.672**	**0.111**	**0.544**
Overall	NoDemo	0.410	0.587	0.382	0.630	0.263	0.269	0.384	0.673	0.191	0.434
	DemoCond	0.414	**0.589**	0.390	0.636	**0.230***	0.295*	0.389	0.680*	0.132*	0.526*
	DemoReason	**0.426****	0.588	**0.394***	**0.639**	0.235*	**0.301***	**0.394***	**0.690****	**0.128***	**0.543****
Baselines											
GPT-4o-mini											
Zero-shot		**0.450**	0.586	**0.411**	**0.650**	0.265	0.296	**0.389**	**0.690**	0.188	0.485
Zero-shot + Demo.		0.436	**0.598**	0.405	0.644	0.211	**0.333**	0.349	0.643	0.108	**0.586**
CoT		0.422	0.568	0.345	0.590	0.312	0.202	0.335	0.656	0.233	0.327
CoT + Demo.		0.432	0.577	0.401	0.642	**0.209**	0.324	0.369	0.654	**0.083**	0.573
Haiku-4.5											
Zero-shot		0.300	0.563	0.345	0.615	0.189	0.260	0.343	0.619	0.130	0.372
Zero-shot + Demo.		0.242	0.546	0.296	0.593	**0.175**	0.261	0.373	0.629	0.095	0.423
CoT		**0.395**	**0.582**	**0.379**	**0.627**	0.232	0.257	**0.382**	**0.661**	0.164	0.408
CoT + Demo.		0.378	0.577	0.364	0.621	0.185	**0.305**	0.366	0.640	**0.079**	**0.533**

D GENERALIZATION RESULTS

We report generalization results of trained models for both cross-dataset transfer and transfer to held-out demographic groups.

Cross-dataset transfer Cross-dataset transfer results of trained LLMs in Table 8 show a pattern broadly consistent with the in-domain results, with overall performance remaining comparable despite evaluation on unseen datasets. NoDemo is still strongest at the individual level, whereas the benefits of socio-demographic modeling become clearer with aggregation. DemoCond and especially DemoReason improve group- and population-level performance, with DemoReason performing best on most population metrics and the continuous group-level metrics. This suggests that the systematic signal captured by socio-demographic modeling generalizes across datasets rather than being specific to the training distribution.

Table 7: Performance for all *trained* models and baselines across different levels. Metric scores are averaged over three datasets, and *Overall* averages across backbones. * denotes significant improvement over NoDemo; ** denotes significant improvement over both NoDemo and the other variant across model–dataset pairs (paired t-test, $p < 0.05$).

Model	Variant	Individual		Group				Population			
		F1 ↑	Acc ↑	F1↑	Acc↑	MSE↓	ρ ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑
Llama-3B	NoDemo	**0.342**	**0.576**	0.339	0.610	0.204	0.261	0.340	0.634	0.124	0.424
	DemoCond	0.308	0.562	0.312	0.592	0.188	0.239	0.310	0.600	0.097	0.439
	DemoReason	0.334	0.570	**0.356**	**0.611**	**0.188**	**0.282**	**0.378**	**0.657**	**0.073**	**0.497**
Qwen-4B	NoDemo	0.381	0.587	0.404	**0.645**	0.197	0.317	0.467	**0.720**	0.117	0.512
	DemoCond	0.372	**0.589**	0.402	0.642	**0.171**	**0.331**	0.444	0.686	0.094	**0.540**
	DemoReason	**0.389**	0.586	**0.405**	0.645	0.185	0.314	**0.526**	**0.761**	**0.079**	0.537
Llama-8B	NoDemo	0.403	**0.599**	0.395	0.648	0.215	0.304	0.411	0.712	0.137	0.495
	DemoCond	**0.408**	0.597	**0.411**	**0.656**	0.185	**0.331**	**0.439**	**0.726**	0.098	**0.567**
	DemoReason	0.365	0.534	0.384	0.635	**0.176**	0.316	0.366	0.649	**0.069**	0.519
Qwen-9B	NoDemo	0.303	0.570	0.360	0.621	0.175	0.282	0.430	0.668	0.102	0.461
	DemoCond	**0.333**	**0.574**	**0.374**	**0.626**	**0.170**	**0.309**	**0.451**	**0.680**	0.090	0.502
	DemoReason	0.315	0.569	0.353	0.621	0.176	0.279	0.416	0.679	**0.080**	**0.523**
Overall	NoDemo	**0.357**	**0.583**	0.375	**0.631**	0.198	0.291	0.412	0.684	0.120	0.473
	DemoCond	0.355	0.580	**0.375**	0.629	**0.179***	**0.303***	0.411	0.673	0.095*	0.512*
	DemoReason	0.351	0.565	0.375	0.628	0.181*	0.298*	**0.421***	**0.686**	**0.075****	**0.519***
Baselines											
BERT											
w/ Majority Vote		0.145	0.532	0.170	0.553	0.179	0.215	0.235	0.578	0.122	—
w/ Indiv. Label		0.192	0.543	0.249	0.574	0.180	0.225	0.353	0.620	0.120	0.279
w/ Demo. Description		0.312	**0.573**	0.277	**0.583**	**0.163**	**0.310**	**0.390**	**0.636**	**0.088**	**0.482**
w/ Demo. Embedding		0.197	0.544	0.241	0.572	0.178	0.227	0.370	0.631	0.111	0.357
DEM-MoE		**0.391**	0.505	**0.285**	0.507	0.371	0.017	0.214	0.488	0.232	0.046

Table 8: Cross-dataset transfer performance of trained models. For each source dataset and model, performance is averaged over the two target datasets as transfer targets; each cell then reports the mean across all source-dataset–model pairs.

Variant	Individual		Group				Population			
	F1 ↑	Acc ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑
NoDemo	**0.362**	**0.572**	**0.363**	0.602	0.222	0.247	0.391	0.642	0.150	0.387
DemoCond	0.340	0.566	0.354	**0.604**	0.199	0.259	0.381	0.642	0.108	0.466
DemoReason	0.345	0.561	0.361	0.600	**0.198**	**0.261**	**0.394**	**0.659**	**0.088**	**0.484**

Held-out demographic-group transfer To evaluate generalization to unseen demographic groups, we train a Llama-8B DemoReason variant on D3CODE while holding out one group from each demographic axis (region=Oceania, country=Brazil, age=50+, gender=Other, ses=8). We then evaluate the model separately on demographic groups seen during training and the held-out groups. Table 9 shows that DemoReason performs comparably across seen and unseen groups overall. On individual- and group-level discrete metrics, it even performs better on the held-out groups, whereas performance is lower for unseen groups at the population level. The continuous metrics reveal a clearer generalization gap, with DemoReason achieving better calibration on seen than unseen groups. Overall, these results suggest that DemoReason can generalize to unseen demographic groups.

E Qualitative Comparison

We provide a qualitative comparison of NoDemo, DemoCond, and DemoReason in estimating the `crowd` baseline, `shift` deviation, and final judgment. We select a representative example from D3CODE using outputs from the trained Qwen-9B variants. The empirical crowd baseline is 0.4306, while the individual judgment is positive (offensive). Accordingly, the individual–majority deviation

Table 9: Unseen demographic-group transfer results of Llama-8B DemoReason on D3CODE. We report both ID results, where demographic groups are observed during training, and OOD results, where the evaluated groups are held out and unseen during training.

Variant	Split	Individual		Group				Population			
		F1 ↑	Acc ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑	F1 ↑	Acc ↑	MSE ↓	ρ ↑
DemoReason	ID	0.349	0.546	0.327	0.570	**0.255**	**0.179**	**0.408**	**0.632**	**0.070**	**0.482**
	OOD	**0.352**	**0.556**	**0.342**	**0.572**	0.310	0.157	0.397	0.628	0.122	0.368

proxy is positive, since the individual judgment differs from the binarized majority judgment implied by the empirical baseline.

Figures 4, 5 and 6 illustrate how the three variants differ in estimating the `crowd` baseline and annotator-specific `shift`. Without access to the annotator's socio-demographic profile, NoDemo substantially underestimates the crowd baseline (0.1 vs. the empirical 0.4306) and predicts the judgment as non-offensive. Conditioning on socio-demographics, DemoCond recognizes greater disagreement in how the post may be perceived and produces a more accurate baseline estimate of 0.3. However, without explicitly reasoning about how the specific annotator may deviate from the crowd, it predicts `shift=0` and still yields an incorrect non-offensive prediction. DemoReason further improves the baseline estimate to 0.4 and, through explicit socio-demographic reasoning, identifies that a Canadian annotator may be more sensitive to a post targeting Canadians. It consequently predicts `shift=+1`, correctly capturing the individual–majority deviation and yielding the correct offensive judgment. This example illustrates the mechanisms identified in Section 4: incorporating socio-demographics improves baseline estimation, with explicit reasoning yielding a more accurate estimate than demographic conditioning alone. Explicit reasoning also captures a meaningful demographic-specific deviation, whereas simple conditioning fails to recover this shift. Together, the improved estimates of both components lead to a more accurate final prediction. However, the reasoning also reveals a potential risk: DemoReason may rely on stereotypical assumptions about demographic groups, such as suggesting that women are more sensitive to certain forms of offensiveness.

NoDemo Example

Input: You're telling me Canadians are bipolar! They need their Lithium dosage quickly.

Reasoning:

<read> The post makes a joke that Canadians are "bipolar," implying they have mood swings, and then puns on the medical term "lithium" (used to treat bipolar disorder) as if it were a literal dosage they need. It's a stereotype-based joke about a nationality, not about any protected identity group.

<norm> This could be seen as offensive under a norm against national or ethnic stereotyping, where mocking a group's perceived traits (here, moodiness) is treated as demeaning. However, this is a lighthearted, pun-based joke rather than a hostile or dehumanizing statement.

<baseline> Most annotators would see this as a harmless, punny joke about a national stereotype, similar to common "Canadians are nice/polite" or "Canadians are obsessed with hockey" jokes. A small minority might flag it as reinforcing a negative stereotype about mental health or nationality, but the majority would find it benign humor.

Prediction: target=none, crowd=0.1, offensive=no

Figure 4: NoDemo example from a trained Qwen-9B model on D3CODE. Without access to the annotator's socio-demographic profile, NoDemo substantially underestimates the crowd baseline (0.1 vs. 0.4306) and incorrectly predicts the judgment as non-offensive.

F Full Baseline Estimation Results

Tables 10 and 11 report per-model annotation- and item-level baseline estimation results. Incorporating socio-demographics generally improves baseline estimation in the baseline models as well.

DemoCond Example

Input: [audience: region=North America, country=Canada, age=18-30, gender=Woman, ses=6]

You're telling me Canadians are bipolar! They need their Lithium dosage quickly.

Reasoning:

<read> The post makes a joke about Canadians being "bipolar," implying they have mood swings, and adds a pun about lithium (a medication for bipolar disorder). It's a stereotype-based joke targeting a nationality, not a protected identity group.

<norm> Offensiveness here would depend on whether mocking a nationality's perceived traits counts as a norm violation; some might see it as a harmless joke, while others could view it as reinforcing a stereotype about Canadians.

<baseline> A diverse pool would likely split, with many seeing it as a common, lighthearted joke about national stereotypes, while a minority would flag it as mildly offensive due to the bipolar/lithium pun and generalization.

Prediction: target=none, crowd=0.3, shift=0, offensive=no

Figure 5: DemoCond example from a trained Qwen-9B model on D3CODE. Compared with NoDemo, DemoCond recognizes greater disagreement in how annotators may perceive the post and therefore provides a more accurate crowd baseline estimate (0.3 vs. 0.4306). However, without explicitly reasoning about the perception of the specific annotator, DemoCond predicts `shift=0`, fails to capture the individual–majority deviation, and still makes an incorrect final prediction.

DemoReason Example

Input: [audience: region=North America, country=Canada, age=18-30, gender=Woman, ses=6]

You're telling me Canadians are bipolar! They need their Lithium dosage quickly.

Reasoning:

<read> The post makes a joke comparing Canadians to people with bipolar disorder, implying they are emotionally unstable or "crazy," and references lithium (a mood stabilizer) as a punchline. It targets Canadians as a national group by using a mental health condition as a stereotype/joke.

<norm> This would be considered offensive under norms that treat mental health conditions as serious medical issues, not punchlines, and that discourage using a nationality as a vehicle for mocking someone's perceived temperament or stability. The norm here is broadly shared among those sensitive to mental health stigma and national stereotyping.

<baseline> A diverse pool would split, with many finding it a harmless, common nationality joke (similar to "Canadians are polite" or "Canadians are obsessed with hockey"), while others would flag it as trivializing mental illness or reinforcing a negative stereotype about a national group. This explains a moderate but not overwhelming offense rate.

<audience> Canadian annotators, especially younger women, may be more attuned to national stereotypes and mental health sensitivities, and could perceive this as a more pointed mockery of their country's identity rather than a generic joke. This could push their reaction slightly more toward offense than the broad pool.

Prediction: target=canadians, crowd=0.4, shift=+1, offensive=yes

Figure 6: DemoReason example from a trained Qwen-9B model on D3CODE. Compared with DemoCond, DemoReason provides a more accurate `crowd` estimate (0.4 vs. 0.4306). With explicit socio-demographic reasoning, DemoReason identifies that a Canadian annotator may be more sensitive to a post targeting Canadians and therefore predicts `shift=+1`, correctly capturing the individual–majority deviation and yielding the correct offensive final prediction.

G Baseline- and Shift-only Intervention Tests

Beyond the joint intervention on `crowd` and `shift`, Tables 12 and 13 report separate interventions on the two components. The baseline-only intervention more consistently follows the expected Oracle $>$ Original $>$ Adversarial ordering, providing stronger evidence that baseline quality directly affects final predictions. In contrast, this ordering is less consistent for shift-only interventions, where a common failure mode is that adversarial shifts outperform the original ones. This is consistent with our earlier finding that model-emitted `shift` captures only modest effect sizes.

Table 10: Annotation- and item-level baseline estimation results for all *prompted* models and baselines. Metric scores are averaged over three datasets, and *Overall* averages across backbones. * denotes significant improvement over NoDemo; ** denotes significant improvement over both NoDemo and the other variant across model–dataset pairs (paired t-test, $p < 0.05$).

Model	Variant	Baseline Estimation (Ann.)				Baseline Estimation (Item)			
		MSE ↓	MAE ↓	ρ ↑	W1 ↓	MSE ↓	MAE ↓	ρ ↑	W1 ↓
Llama-3B	NoDemo	**0.060**	**0.186**	**0.320**	0.098	0.059	0.185	0.333	**0.097**
	DemoCond	0.060	0.193	0.286	0.098	**0.041**	**0.161**	**0.501**	0.112
	DemoReason	0.068	0.204	0.236	**0.078**	0.044	0.168	0.394	0.098
Qwen-4B	NoDemo	0.093	0.238	**0.555**	0.155	0.089	0.234	0.565	0.154
	DemoCond	0.083	0.223	0.516	0.129	0.066	0.198	0.588	0.113
	DemoReason	**0.076**	**0.210**	0.552	**0.117**	**0.061**	**0.187**	**0.619**	**0.103**
Llama-8B	NoDemo	0.049	0.167	0.468	0.077	0.047	0.164	0.501	0.079
	DemoCond	**0.048**	**0.167**	0.516	**0.074**	**0.038**	**0.152**	0.617	**0.076**
	DemoReason	0.050	0.175	**0.565**	0.092	0.044	0.167	**0.620**	0.095
Qwen-9B	NoDemo	0.110	0.266	0.442	0.190	0.111	0.269	0.457	0.195
	DemoCond	0.091	0.236	0.509	0.154	0.078	0.217	0.597	0.143
	DemoReason	**0.082**	**0.219**	**0.533**	**0.134**	**0.067**	**0.197**	**0.614**	**0.117**
Qwen-27B	NoDemo	0.081	0.220	**0.602**	0.145	0.078	0.217	0.622	0.144
	DemoCond	**0.081**	**0.217**	0.597	**0.142**	**0.073**	**0.208**	0.641	0.137
	DemoReason	0.082	0.220	**0.602**	0.142	0.074	0.210	**0.652**	**0.137**
Llama-70B	NoDemo	0.045	0.157	**0.601**	**0.059**	0.042	0.154	0.613	**0.063**
	DemoCond	0.041	0.153	**0.601**	0.072	0.037	0.147	**0.634**	0.074
	DemoReason	**0.039**	**0.149**	0.584	0.069	**0.035**	**0.143**	0.625	0.068
Qwen-122B	NoDemo	0.064	0.187	0.541	0.124	0.061	0.184	0.570	0.122
	DemoCond	0.056	0.175	0.583	0.103	0.050	0.165	0.623	0.102
	DemoReason	**0.052**	**0.167**	**0.588**	**0.087**	**0.046**	**0.159**	**0.626**	**0.087**
GPT-4o-mini	NoDemo	0.051	0.179	**0.594**	0.109	0.050	0.179	0.621	0.112
	DemoCond	0.044	0.157	0.580	**0.083**	0.038	0.151	0.627	**0.085**
	DemoReason	**0.039**	**0.151**	0.592	0.085	**0.035**	**0.146**	**0.641**	0.090
Haiku-4.5	NoDemo	0.047	0.156	0.589	0.069	0.046	0.158	0.594	0.073
	DemoCond	0.044	0.153	0.576	**0.066**	0.040	**0.146**	0.602	**0.067**
	DemoReason	**0.043**	**0.153**	**0.590**	0.075	**0.039**	0.148	**0.614**	0.079
Overall	NoDemo	0.067	0.195	0.524	0.114	0.065	0.194	0.542	0.115
	DemoCond	0.061	0.186	0.529	0.102*	0.051*	0.172*	**0.603***	0.101*
	DemoReason	**0.059**	**0.183***	**0.538**	**0.097***	**0.050***	**0.169***	0.601*	**0.097**
Baselines									
GPT-4o-mini									
Zero-shot		0.067	0.197	**0.449**	0.105	0.069	0.203	0.470	0.113
Zero-shot + Demo.		**0.063**	**0.192**	0.428	**0.095**	0.047	0.167	0.575	**0.083**
CoT		0.084	0.221	0.402	0.114	0.083	0.220	0.402	0.116
CoT + Demo.		0.090	0.225	0.371	0.112	**0.040**	**0.163**	**0.587**	0.106
Haiku-4.5									
Zero-shot		0.054	0.171	0.531	0.137	0.052	0.169	0.545	0.132
Zero-shot + Demo.		**0.051**	**0.168**	0.545	0.117	0.045	0.157	0.600	0.095
CoT		0.054	0.175	0.509	**0.094**	0.053	0.173	0.509	0.094
CoT + Demo.		0.053	0.176	**0.546**	0.100	**0.040**	**0.152**	**0.616**	**0.071**

H LIMITATIONS

First, some of our analyses, particularly those involving demographic-specific `shift`, rely on empirical or proxy targets because the underlying quantities are not directly observable. These proxies are inevitably affected by annotation noise and may therefore underestimate or distort the true signal. Moreover, the empirical baseline reflects the annotator population available in each dataset and may not perfectly represent a broader population. To mitigate this issue, we evaluate the relevant effects at multiple levels of aggregation and focus on datasets or annotation pools with sufficiently

Table 11: Annotation- and item-level baseline estimation results for all *trained* models and baselines. Metric scores are averaged over three datasets, and *Overall* averages across backbones. * denotes significant improvement over NoDemo; ** denotes significant improvement over both NoDemo and the other variant across model–dataset pairs (paired t-test, $p < 0.05$).

Model	Variant	Baseline Estimation (Ann.)				Baseline Estimation (Item)			
		MSE ↓	MAE ↓	ρ ↑	W1 ↓	MSE ↓	MAE ↓	ρ ↑	W1 ↓
Llama-3B	NoDemo	0.063	0.198	0.424	0.135	0.058	0.189	0.485	0.133
	DemoCond	**0.054**	0.184	**0.452**	0.116	0.044	0.171	0.545	0.115
	DemoReason	0.054	**0.180**	0.440	**0.089**	**0.041**	**0.158**	**0.565**	**0.093**
Qwen-4B	NoDemo	**0.039**	**0.144**	**0.594**	**0.055**	0.034	0.137	**0.653**	**0.046**
	DemoCond	0.041	0.151	0.559	0.056	**0.033**	**0.135**	0.640	0.051
	DemoReason	0.042	0.153	0.559	0.066	0.033	0.138	0.643	0.057
Llama-8B	NoDemo	0.044	0.157	**0.515**	0.072	0.039	0.148	0.586	0.072
	DemoCond	0.044	0.155	0.508	0.080	0.034	0.138	**0.601**	0.074
	DemoReason	**0.040**	**0.150**	0.506	**0.067**	**0.032**	**0.133**	0.575	**0.071**
Qwen-9B	NoDemo	0.054	0.175	0.488	0.131	0.049	0.167	0.541	0.124
	DemoCond	**0.049**	**0.164**	**0.522**	**0.107**	**0.041**	**0.150**	**0.628**	0.102
	DemoReason	0.051	0.170	0.501	0.109	0.042	0.153	0.614	**0.099**
Overall	NoDemo	0.050	0.168	0.505	0.098	0.045	0.160	0.566	0.094
	DemoCond	0.047	0.164	**0.510**	0.090	0.038*	0.148*	**0.604***	0.086
	DemoReason	**0.047**	**0.163**	0.502	**0.083****	**0.037***	**0.146***	0.599*	**0.080***
Baselines									
BERT									
w/ Majority Vote		0.059	0.189	0.522	0.146	0.059	0.187	0.541	0.142
w/ Indiv. Label		0.030	**0.123**	**0.632**	0.045	0.029	**0.122**	0.647	**0.045**
w/ Demo. Description		0.040	0.147	0.529	**0.042**	0.030	0.123	**0.650**	0.048
w/ Demo. Embedding		**0.029**	0.124	0.625	0.043	**0.028**	0.122	**0.650**	0.046
DEM-MoE		0.091	0.266	0.017	0.260	0.090	0.265	0.027	0.262

Table 12: Performance under baseline-only interventions on baseline and shift estimates for Llama-8B (trained) and Llama-70B (prompted). Each cell reports *Original / Oracle / Adversarial* performance. Bold marks cells following the expected ordering Oracle > Original > Adversarial.

Model	Variant	Individual	Group		Population	
		F1 ↑	F1 ↑	MSE ↓	F1 ↑	MSE ↓
Llama-8B (Trained)	DemoCond	0.453/0.441/0.337	**0.442/0.477/0.410**	0.190/0.152/0.168	**0.549/0.650/0.457**	**0.094/0.081/0.096**
	DemoReason	**0.474/0.478/0.443**	**0.450/0.453/0.421**	**0.188/0.185/0.197**	0.526/0.595/0.556	**0.059/0.059/0.063**
Llama-70B (Prompted)	DemoCond	**0.438/0.510/0.335**	**0.425/0.511/0.270**	**0.178/0.164/0.195**	**0.435/0.667/0.353**	0.095/0.075/0.078
	DemoReason	**0.458/0.549/0.432**	**0.443/0.534/0.413**	0.180/0.186/0.196	**0.500/0.632/0.468**	0.084/0.077/0.073

large numbers of annotations. Nevertheless, some findings should still be interpreted with these measurement and sampling limitations in mind.

Second, our findings are specific to the prompting templates, reasoning format, and training pipeline considered in this work. Although we observe consistent patterns across multiple datasets and models, and report performance of several baseline models in Appendices C and F, it remains unclear whether they generalize to other ways of incorporating socio-demographic information, other subjective tasks beyond offensiveness, or different annotator populations. Furthermore, although our format explicitly elicits `crowd` baseline and `shift` estimates, LLMs are not constrained to derive their final predictions from these components. Our intervention results establish their functional relevance to prediction, but do not demonstrate that the model internally follows the assumed decomposition.

Finally, the available socio-demographic attributes provide only a coarse representation of annotator identity and the factors underlying subjective judgments. Our analyses primarily consider groups defined along individual demographic axes, while intersectional effects are difficult to estimate reliably because of data sparsity. Socio-demographics also explain only a limited portion of individual

Table 13: Performance under shift-only interventions on baseline and shift estimates for Llama-8B (trained) and Llama-70B (prompted). Each cell reports *Original / Oracle / Adversarial* performance. Bold marks cells following the expected ordering Oracle $>$ Original $>$ Adversarial.

Model	Variant	Individual	Group		Population	
		F1 ↑	F1 ↑	MSE ↓	F1 ↑	MSE ↓
Llama-8B (Trained)	DemoCond	0.453/0.412/0.333	**0.442/0.483/0.398**	0.190/0.131/0.165	0.549/0.378/0.432	0.094/0.110/0.108
	DemoReason	0.474/0.582/0.492	0.450/0.527/0.450	**0.188/0.149/0.200**	0.526/0.500/0.600	0.059/0.075/0.071
Llama-70B (Prompted)	DemoCond	0.438/0.538/0.499	0.425/0.495/0.463	**0.178/0.167/0.183**	0.435/0.480/0.519	0.095/0.095/0.088
	DemoReason	0.458/0.525/0.500	0.443/0.495/0.482	**0.180/0.175/0.188**	0.500/0.528/0.538	**0.084/0.086/0.084**

variation, leaving many potentially relevant individual, contextual, and experiential factors unobserved.