Multi-Perspective LLM Annotations for Valid Analyses in Subjective Tasks

Navya Mehrotra[⋆] Johns Hopkins University nmehrot2@jhu.edu
Adam Visokay[⋆] University of California, Berkeley avisokay@berkeley.edu
Kristina Gligorić Johns Hopkins University gligoric@jhu.edu

Abstract

Large language models are increasingly used to annotate texts, but their outputs reflect some human perspectives better than others. Existing methods for correcting LLM annotation error assume a single ground truth. However, this assumption fails in subjective tasks where disagreement across demographic groups is meaningful. Here we introduce PERSPECTIVE-DRIVEN INFERENCE, a method that treats the distribution of annotations across groups as the quantity of interest, and estimates it using a small human annotation budget. We contribute an adaptive sampling strategy that concentrates human annotation effort on groups where LLM proxies are least accurate. We evaluate on politeness and offensiveness rating tasks, showing targeted improvements for harder-to-model demographic groups relative to uniform sampling baselines, while maintaining coverage.

1 Introduction

LLMs are promising tools for reducing annotation burden in computational social science (CSS) (Ziems et al., 2024; Manning et al., 2024). Yet, LLM annotations can be wrong (Dillion et al., 2023), encode particular positionalities (Santy et al., 2023), and reproduce stereotypes (Shrawgi et al., 2024).

Methods like Prediction-Powered Inference (Angelopoulos et al., 2023a) and Design-Based Supervised Learning (Egami et al., 2023) offer a repair (Hullman et al., 2026), addressing these limitations by combining potentially inaccurate LLM proxies with a small set of ground truth human labels, yielding valid downstream inference even when the LLM proxy is imperfect.

However, these repair methods assume a ground truth exists. In many CSS settings, ground truth does not exist. Reasonable people disagree about whether language is toxic, empathetic, or polite, and that disagreement is not an error to be corrected, but *a signal to be preserved*. Collapsing annotator disagreement into a majority label discards the very variation that matters for understanding how people use and perceive language.

A common method to address this is persona prompting (Truong et al., 2025): conditioning LLMs on demographic attributes to simulate the perspectives of specific groups. However, persona prompting does not always improve performance of large language models (Zheng et al., 2024; Sun et al., 2025): it can exaggerate group differences (Cheng et al., 2023), misrepresent minority viewpoints (Wang et al., 2025), and in many settings performs no better than a generic prompt (Zheng et al., 2024; Sun et al., 2025; Xu and Jurgens, 2026).

Here, we introduce PERSPECTIVE-DRIVEN INFERENCE, a framework for valid corpus inference in subjective annotation tasks where no single ground truth exists (illustrated in Figure 1). Our goal is not to train LLMs to make more accurate predictions, but to leverage their annotations to conduct robust downstream statistical inference that faithfully retains annotator disagreement.

Our contributions are:

1. **Framework.** We formalize multi-perspective corpus inference as estimation of a vector estimand (one entry per demographic group), rather than a single scalar, and show how standard PPI-based correction extends to this setting.

2. **Sampling method.** We develop a sampling strategy that concentrates the human annotation budget on demographic groups for which LLM annotations are predicted to be less reliable, with the goal of improving inference for underrepresented or poorly modeled perspectives.

[⋆]Equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 537–560
November 6-10, 2026 ©2026 Association for Computational Linguistics

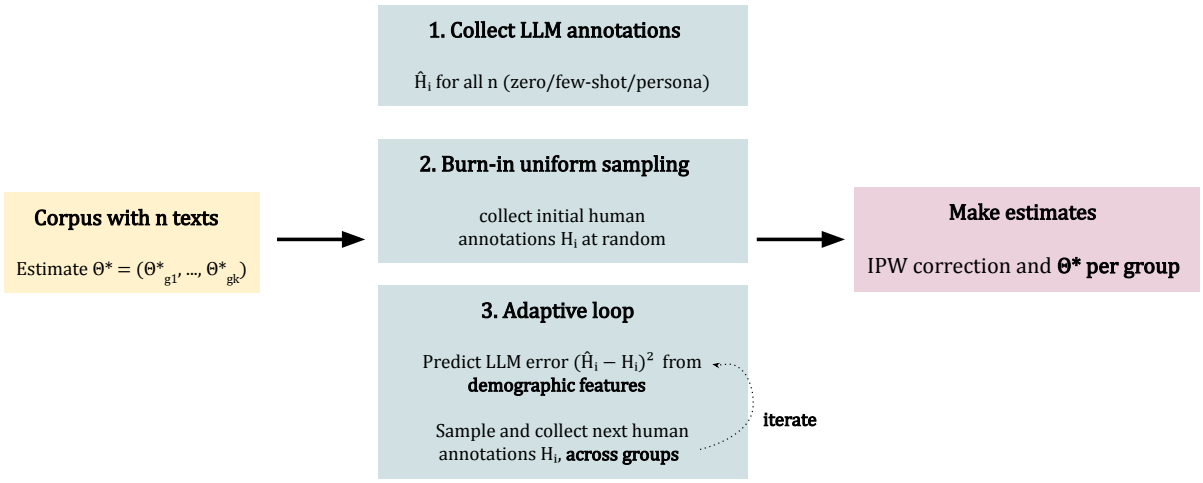


Figure 1: **Overview of PERSPECTIVE-DRIVEN INFERENCE.** Starting from a corpus of n texts, we collect LLM annotations, initialize human annotation via uniform sampling, and then enter an adaptive loop that predicts LLM error from demographic features, sampling human annotations across groups. The core idea is to prioritize human annotation for texts where LLMs poorly represent the target demographic group. Once the annotation budget is exhausted, we apply inverse probability weighting (IPW) to produce group-level estimates θ^*.

3. **Evaluation.** We evaluate PERSPECTIVE-DRIVEN INFERENCE on three datasets: politeness ratings, offensiveness ratings, and synthetic binary annotations, finding improved accuracy for targeted demographic groups, while maintaining coverage.

We release code and data at `https://github.com/JHU-CSS/PDI` to support tasks where retaining differences in annotator perspectives is desired.

2 Background

Annotators' background A growing body of NLP literature views annotation disagreement not as noise, but as a valuable signal (Uma et al., 2021; Hovy, 2015). Prior work has shown considerable heterogeneity in annotator preferences on subjective tasks, ranging from toxicity (Sap et al., 2022) and sexism detection (Tahaei and Bergler, 2024), to theory of mind modeling (Fan et al., 2025) and counterspeech generation (Mun et al., 2024), with demographic axes like gender (Biester et al., 2022; Chen et al., 2025a) and race (Lee et al., 2024) shifting annotations systematically (Al Kuwatly et al., 2020). Modeling this disagreement explicitly, rather than collapsing it into majority vote, has become a concern in NLP (Fleisig et al., 2023; Xu and Jurgens, 2026; Xu et al., 2025). Recent efforts push toward characterizing how demographic factors shape human preferences in LLM outputs (Movva et al., 2025; Chen et al., 2025b), and using that understanding to build systems that represent multiple perspectives (Sorensen et al., 2024; Feng et al., 2024; Röttger et al., 2022). What is different about our work is that we ask: how can we efficiently leverage limited budgets of human input, given that models do not capture different perspectives equally well (Sen et al., 2025; Santurkar et al., 2023)? In doing so, our goal is not to use LLMs to make better predictions, but to use LLM annotations as inputs to downstream analyses, while preserving diverse perspectives on subjective tasks.

Valid inference with proxy annotations Corpus-level inference from text has roots in literary analysis (Yavuz, 2019) and causal inference in NLP (Feder et al., 2022). Early work formalized the task of inferring document class prevalence under uncertainty (Keith and O'Connor, 2018). A separate thread examined whether annotations themselves are valid (Valentino et al., 2021; Ruder et al., 2025), and how annotation error and genuine label variation can be disentangled (Weber-Genzel et al., 2024; Baledent et al., 2022). More recent work in CSS turns to downstream inference under annotation imperfection, using LLM outputs as surrogates for human labels and developing statistically rigorous methods for correction (Wang et al., 2020; Egami et al., 2023; Angelopoulos et al., 2023b). We contribute to this line of work by complicating the assumption that a ground truth exists at all: in subjective tasks, disagreement is constitutive, and we show how these repair methods apply in settings where no single correct label can be defined, and

the target estimand is a vector, rather than a scalar.

Adaptive sampling Active learning methods reduce annotation cost by selecting the most informative instances for labeling (Baldridge and Osborne, 2004; Arora et al., 2009), with established tools to support these workflows (Lin et al., 2019; Tsvigun et al., 2022; Nghiem and Ananiadou, 2018). Work in the multi-annotator setting specifically targets instances where annotators are likely to disagree (Baumler et al., 2023), and recent methods allocate annotation effort between humans and LLMs based on uncertainty (Li et al., 2023; Rouzegar and Makrehchi, 2024; Zhang et al., 2023). Most related to our approach is sampling based on LLM annotation confidence (Gligoric et al., 2025), and stratified PPI (Fisch et al., 2024), which improves inference efficiency by partitioning samples into strata. In contrast, our stratification is along *perspectives* rather than strata with data subsets: we sample human annotators to cover the distribution of viewpoints, not to reduce label uncertainty around a single ground truth.

3 Methods

3.1 Problem setup

We have a corpus of n independent and identically distributed (i.i.d.) text instances $T_1, \ldots, T_n$. We consider a *perspective-aware* setup: each instance is annotated by multiple annotators, each coming from a particular perspective. Specifically, each observation is a (text, annotator) pair (T_i, d_i), where $d_i \in \mathcal{D}$ encodes the demographic profile of the annotator (e.g., gender, race, age, education level).

The human annotation H_i therefore captures the judgment of an annotator with demographics d_i on text T_i (i.e., index i runs over (text, annotator) pairs). Note that we focus on demographic features as a proof of concept. In principle, d_i features could include any axis along which human perspectives vary, informed by theoretical background knowledge, or exploratory analyses.

We partition annotators into K demographic groups $\mathcal{G} = \{g_1, \ldots, g_K\}$. We wish to estimate a *vector* of group-specific quantities

$$\theta^* = (\theta^*_{g_1}, \ldots, \theta^*_{g_K}),$$

where each $\theta^*_{g_k} = \mathbb{E}[H_i \mid d_i \in g_k]$ is the expected annotation value within demographic group g_k, i.e., the group mean. Note that, beyond the group means, the framework applies for standard statistics such as group medians or regression coefficients (Angelopoulos et al., 2023b).

For instance, $\theta^*_{g_k}$ can be the fraction of texts rated as polite by annotators in group g_k, or their mean perceived offensiveness rating. Estimating θ^* separately per group reveals how a quantity of interest differs across demographic perspectives.

In addition to human annotations, we have access to LLM annotations $\hat{H}_i$ for all n instances. These serve as a cheap but potentially inaccurate proxy for human judgments. We define the *LLM error* for instance i as $(\hat{H}_i - H_i)^2$.

Our goal is to estimate θ^* with a valid confidence interval at a pre-specified level $(1-\alpha)$, using only $n_{\text{human}} \ll n$ human annotations in total. Two challenges make this goal difficult: (i) rare demographic groups are likely to receive too few annotations under uniform sampling, yielding imprecise group-level estimates; and (ii) LLM annotations may be systematically biased for certain demographic perspectives.

3.2 Perspective-driven inference

To overcome these challenges, we combine LLM annotations with *adaptive* sampling of human annotations, to produce an unbiased estimate $\hat{\theta}_{\text{PDI}}$ of each $\theta^*_{g_k}$, together with a valid confidence interval.

The adaptive design concentrates the human annotation budget on demographic groups for which LLM annotations are *least reliable*, improving precision for poorly-modeled groups.

Adaptive sampling rule We collect a human annotation H_i for observation (T_i, d_i) with probability π_i. Let $\xi_i = \mathbf{1}\{H_i \text{ collected}\}$ denote the corresponding indicator. Motivated by active inference (Zrnic and Candès, 2024; Gligoric et al., 2025), we set

$$\pi_i \propto \widehat{\texttt{err}}_i(d_i),$$

normalized so that $\sum_{i=1}^{n} \pi_i = n_{\text{human}}$. The function $\widehat{\texttt{err}}_i$ is a black-box predictor of the LLM squared error trained on annotator demographics. Crucially, H_i is unknown before annotation, so π_i is approximated by learning a predictor of the LLM error from annotator demographic features.

As annotations are collected, we use $\{(d_j, (\hat{H}_j - H_j)^2)\}_{j<i,\, \xi_j=1}$ as feature–label pairs to fit $\widehat{\texttt{err}}_i$ via a black-box regressor (gradient-boosted regression tree, XGBoost). That is, we train a model to predict the squared LLM error from the annotator's demographic profile, and then allocate human annotations to groups with high predicted error.

To bootstrap the sampling rule, we first collect n_{burnin} annotations uniformly at random (referred to as the *burn-in* phase), which are part of the total annotation budget. We then proceed in B batches. Within each batch b, we normalize the predicted errors into a discrete probability vector, and smooth it with the uniform distribution over the batch to prevent zero-probability allocations. We refine $\widehat{\texttt{err}}_i$ after each batch using all annotations collected so far.

Perspective-driven estimator After collecting human annotations according to π_i, we estimate each group mean $\theta^*_{g_k}$ via an inverse-probability-weighted (IPW) rectified estimator (Appendix, Sec. B). We form a valid $(1-\alpha)$-level confidence interval for each $\theta^*_{g_k}$ via bootstrap (Zrnic and Candès, 2024).

3.3 Baselines

LLM only (zero-shot) This baseline prompts the LLM with the task description and text T_i directly, with no examples, and treats the resulting annotation as ground truth. No human annotation budget is required, but the estimate may be biased if the LLM misrepresents certain demographic perspectives, or is inaccurate, depending on which demographics LLMs default to during annotation (Schäfer et al., 2025).

LLM only (few-shot) This baseline augments the zero-shot prompt with a small number of randomly chosen labeled examples. The few-shot examples may improve accuracy relative to zero-shot prompting, but demographic bias can persist.

LLM only (persona prompt) This baseline instructs the LLM to adopt a specific demographic persona before annotating. Persona prompting is designed to steer the LLM toward a particular demographic perspective, but fidelity to that perspective is not guaranteed.

PPI (uniform sampling) This baseline uses the same rectified estimator as PDI but collects human annotations *uniformly at random* rather than adaptively, i.e., $\pi_i = n_{\text{human}}/n$ for all i. It is a special case of $\hat{\theta}_{\text{PDI}}$ with uniform π_i and optimally chosen λ. Since uniform sampling ignores group membership, minority demographic groups may receive few annotations. This baseline corresponds to the estimator of Angelopoulos et al. (2023b).

For both PPI and PDI, for each model, we use the LLM variant (zero-shot, few-shot, or persona) that achieves the highest coverage for the LLM-only estimator, selected separately for each model. We note that we are aiming to achieve targeted sub-improvements, rather than perform universally superior compared to PPI.

3.4 Evaluation

Datasets We evaluate on both real-world and synthetic data. For real-world data, we use two annotation tasks from POPQUORN (Pei and Jurgens, 2023), a large-scale dataset that collects demographically diverse human annotations via Prolific. POPQUORN contains 45,000 annotations from 1,484 annotators across multiple tasks. The texts come from three sources: the Enron email corpus (emails), Reddit (social media comments), and SQuAD (Wikipedia-based reading comprehension questions posed by crowdworkers). We use the politeness rating and offensiveness rating subsets, taking 1,000 unique annotations from each task. For additional robustness experiments, we construct a synthetic dataset of $N = 10,000$ items drawn from two demographic groups with selected label rates and group-specific LLM error rates (see Appendix, Section H for details).

Annotation tasks In the politeness task, annotators rate how polite a statement would be in an email from a colleague on a Likert scale from 1.0 (not polite at all) to 5.0 (very polite). In the offensiveness task, annotators rate how offensive a statement would be as a Reddit comment on the same scale. Synthetic annotations are binary.

Inference tasks Our goal is to estimate the vector $\theta^* = (\theta^*_{g_1}, \ldots, \theta^*_{g_K})$, where each $\theta^*_{g_k}$ is the mean annotation rating within demographic group g_k. For the politeness task, $\theta^*_{g_k}$ is the mean perceived politeness rating among annotators in group g_k; for the offensiveness task, it is the mean perceived offensiveness rating. We stratify separately by gender, age, and education, yielding K groups per demographic axis, used as features in the black box regressor.

Descriptive statistics Annotation ratings vary significantly across demographic groups in both real-world datasets, motivating the multi-perspective estimation. We verify this using chi-squared tests on ratings. In the offensiveness dataset, we find significant variation by gender ($\chi^2(4) = 13.305$, $p = 0.0099$), with men rating statements as more offensive on average than

women (1.77 vs. 1.58), as well as by age ($\chi^2(28) = 90.261$, $p < 0.001$) and education ($\chi^2(16) = 49.837$, $p < 0.001$). The politeness dataset shows the same pattern such that ratings differ significantly by gender, age, and education. Since the differences between age groups are most pronounced in both politeness and offensiveness ratings (Figure 2), we focus on age strata in the evaluation. For illustration, Tables 7 and 8 list example texts, along with annotations. For example, "Women's studies is dumb though, stop defending your degree?" was generally seen as more offensive by younger annotators: those aged 18–24 and 25–29 rated it 5, while most annotators aged 50 and above rated it 1, with one 50–54-year-old exception who rated it 5. A similar pattern appears for "Still not as scary as autism.", which received two ratings of 5 from 18–24-year-olds, whereas most annotators aged 45 and above rated it 1, although one 54–59-year-old rated it 5.

Models We evaluate across eight LLMs spanning a range of model families and scales: GPT-5.2, GPT-OSS-120B, Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, Gemini 3.1 Pro Preview, Llama 3 8B Instruct, and Mistral Large 3. Each model is prompted to produce offensiveness and politeness ratings on the same statements presented to human annotators, using identical task descriptions and rating scales. Full prompt texts and prompting parameters such as temperature are listed in Appendix, Section A.

Evaluation metrics We evaluate methods in terms of *coverage* and *mean absolute error*.

Coverage. Coverage is the empirical rate at which the produced confidence intervals contain the true $\theta^*_{g_k}$. Since $\theta^*_{g_k}$ is unknown, we approximate it by the group mean computed on the full dataset. For PDI and PPI, coverage is provably valid by construction for large n (Angelopoulos et al., 2023b); we report it as a sanity check and to evaluate the LLM-only baselines, which have no coverage guarantees.

Delta. For each estimator, we measure how far its point estimate deviates from the human-rated group mean $\theta^*_{g_k}$, approximated by the full-dataset group mean. Specifically, we report the average absolute error in percentage points, averaged across demographic groups. Lower Δ indicates that the method's estimates are closer to ground truth on average. This metric captures estimation accuracy directly in the units of the quantity of interest (percentage points). Coverage asks whether the interval is in the right place, while delta asks whether the estimate itself is. Metrics are reported across 20 trials.

4 Results

4.1 Politeness

In the main text we focus on GPT-5.2 (results with the other models are reported in the Appendix, Table 5). Across prompting variants, the few-shot configuration yields the smallest mean prediction error on the politeness task. The subpopulation with the largest error under the best-performing variant is **Age 50+**, with a few-shot error of 0.426.

Coverage First, we find that coverage consistently falls below 90% for several LLM-only methods on older age groups (Table 1). Zero-shot GPT-5.2 misses the target for Age 35–49 (coverage = 0.800). For Age 50+, both the few-shot variant (coverage = 0.750) and the persona variant (coverage = 0.850) fall short. In contrast, PDI maintains coverage at or above 90% across all three age groups, reaching 0.950 for Age 50+.

Delta Averaged across age groups, the LLM-only methods are comparable to PPI and non-rectified baselines: zero-shot and few-shot GPT-5.2 produce average deltas of 7.48 and 7.52 percentage points, respectively, versus 6.35 for PPI and 7.10 for PDI. Again, the critical divergence appears for **Age 50+**: The few-shot variant is inaccurate for this group (coverage = 0.750), producing the worst delta at 16.31%. In contrast, PDI achieves the lowest delta for Age 50+ at 11.23%, outperforming PPI (13.63%) and all three LLM-only variants.

Upsampling lift The lower delta for Age 50+ under PDI reflects error-driven upsampling. Because the error regressor accurately identifies this group as high-error, PDI directs a disproportionate share of the annotation budget there (Table 2). Averaged across 20 trials, PPI allocates 9.2 annotations to Age 50+ under uniform sampling; PDI allocates 12.2, a 33% increase. This confirms that the predicted error signal is useful: the regressor correctly identifies where annotation effort matters most, and the resulting upsampling translates into improved coverage and accuracy for the most challenging subpopulation.

Crucially, this reallocation does not compromise overall validity—averaged across all three age groups, PDI maintains 0.933 coverage.

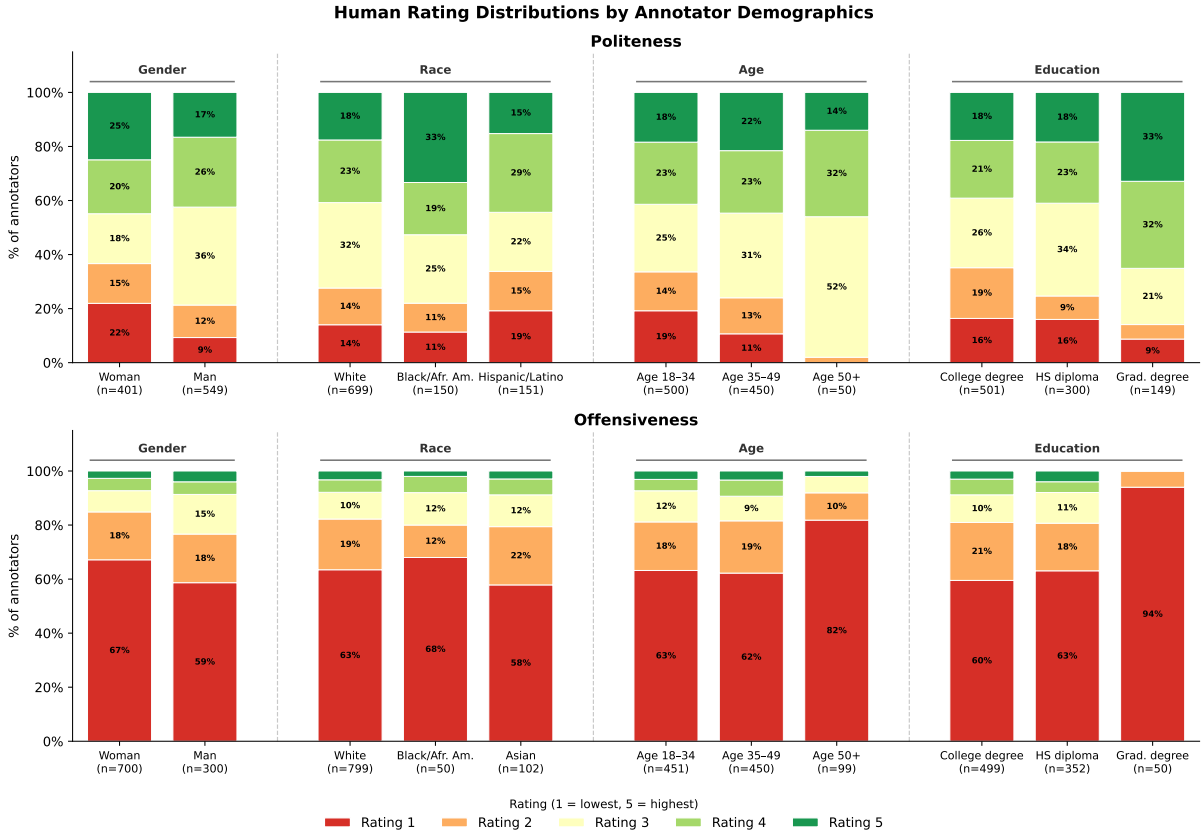


Figure 2: **Annotation distributions vary across demographic groups.** Human ratings for politeness (top) and offensiveness (bottom) broken down by annotator demographics. Variation across groups motivates estimating a vector of group-specific means rather than a single aggregate.

4.2 Offensiveness

For the offensiveness task, the persona variant results in the smallest mean prediction error for GPT-5.2. As with politeness, age is the demographic axis with the greatest variation in performance, and Age 50+ again emerges as the hardest subgroup (Table 3).

Coverage The coverage for offensiveness is substantially worse than for politeness. All three LLM-only methods fail: averaged across age groups, zero-shot GPT-5.2 achieves only 11.7% coverage, the persona variant 13.3%, and few-shot variant only 53.3% (substantially below the 90% coverage across every age group). For age 35–49, the persona variant achieves 0% coverage. In contrast, both PPI and PDI maintain 95.0% average coverage across all three age groups, including Age 50+.

Delta The LLM-only delta values reflect the same breakdown: zero-shot and persona produce average deltas above 19 percentage points, and even the best LLM-only variant (few-shot, 14.21%) falls far short of PPI (3.86%) and PDI (6.04%). For Age 50+, all three LLM-only variants exceed 24% delta, with persona reaching 28.40%. PDI achieves the lowest delta for Age 50+ at 5.24%, outperforming PPI (5.64%), while also maintaining coverage.

Upsampling lift As with politeness, the PDI performance for Age 50+ is supported by error-driven upsampling (Table 4). PDI allocates on average 25.8 annotations to Age 50+ across 20 trials, compared to 21.6 under uniform PPI sampling (a 19% increase). The upsampling lift surpasses PPI while preserving validity: averaged across all three age groups, PDI maintains coverage.

4.3 Synthetic data

To understand the impact of the human annotation budget, size of the burn-in, group skew, and LLM error, we analyze PDI on synthetic data, under controlled conditions. The simulated data consist of $N = 10{,}000$ items drawn from two demographic groups: a large group (Group 1, $N_1 = 9{,}000$, 90%) and a small group (Group 2, $N_2 = 1{,}000$, 10%). Each item carries a binary human label $Y_i \in \{0, 1\}$ and a binary LLM prediction $\hat{Y}_i$. LLM predic-

Method	Avg (Age)		Age 18–34		Age 35–49		Age 50+	
	Cov.	Δ	Cov.	Δ	Cov.	Δ	Cov.	Δ
GPT-5.2 (zero shot)	88.3	7.48	95.0	4.10	80.0	6.88	90.0	11.46
GPT-5.2 (few shot)	91.7	7.52	100.0	2.45	100.0	3.81	75.0	16.31
GPT-5.2 (persona)	93.3	7.02	100.0	3.58	95.0	4.60	85.0	12.87
PPI	95.0	6.35	100.0	2.93	100.0	2.49	85.0	13.63
PDI	93.3	7.10	90.0	5.51	95.0	4.58	95.0	11.23

Table 1: **Politeness.** Coverage (Cov.) and delta (Δ) by age group on the politeness task (GPT-5.2). Values below 90% coverage are highlighted.

Group	PDI samples	Uniform samples
Gender		
Woman	78.7	85.1
Man	121.3	114.9
Race		
White	135.8	140.2
Black/Afr. Am.	34.3	28.8
Hispanic/Latino	29.9	31.1
Age		
Age 18–34	105.0	106.8
Age 35–49	82.9	84.0
Age 50+	12.2	9.2
Education		
College degree	83.8	94.8
HS diploma	73.2	63.6
Grad. degree	32.8	31.9

Table 2: **Politeness.** Average number of human annotations allocated per group by PDI and PPI, averaged across 20 trials (total budget 200 per trial). PDI allocates 33% more annotations to Age 50+ than PPI (12.2 vs. 9.2), reflecting error-driven upsampling toward the hardest subgroup.

tions are generated by flipping the true label independently with group-specific error rates (see Appendix, Sec. H for details). Overall, we find that adaptive sampling is advantageous over uniform sampling when the budget of human annotations is at least 20%, as sufficient labels are needed to learn the sampling rule (Figure 3), while the choice of precise burn-in size is not consequential, as long as it is under 50% of the complete budget (Figure 4). As expected, the stronger the disparity in LLM performance between groups and the more severe the group skew, the more PERSPECTIVE-DRIVEN INFERENCE deviates from uniform sampling (Figures 5 and 6). These differences are particularly pronounced for group sizes under 10%, and for error differences above 20%, where upsampling is particularly valuable, when compared to uniform sampling.

4.4 Limitations of persona prompting

Lastly, overall across the experiments, we find that persona prompting is not a reliable method. Across models, it is the highest-variance prompting strategy and frequently the worst (Appendix, Sec. F). Persona prompting sometimes matches few-shot performance, but it can also shift predictions systematically in the wrong direction, suggesting the effect depends heavily on model-specific patterns, rather than on the demographic content of the persona itself. Persona prompting is the most common method for capturing demographic perspectives with LLMs, yet it fails precisely for the groups it is meant to represent, and in several cases leaves coverage at zero. The results confirm that conditioning an LLM on a demographic label is not a substitute for human annotation from that group, and underscore why a correction framework that treats LLM outputs as potentially inadequate proxies (rather than steerable perspectives) can be useful.

5 Discussion

Overall, our results show that correcting LLM annotation error can be important when the target estimand is a distribution over perspectives rather than a single ground truth. LLM-only methods fail for specific groups that are poorly modeled by persona prompting (such as Age 50+). PDI addresses this by directing human annotation effort where it matters most. The gains can be substantial: for offensiveness, PDI reduces delta for Age 50+ to 5.24%, while LLM-only variants exceed 24%. Coverage holds across all groups. Practitioners should expect the largest gains over uniform sampling when LLM performance disparities between groups are large, and groups are skewed.

The evaluation therefore demonstrates that PDI can be viewed as a targeted subgroup-improvement method (rather than a universally superior alternative to uniform PPI). On aggregate metrics, PDI does not consistently outperform uniform PPI. But

Method	Avg (Age)		Age 18–34		Age 35–49		Age 50+	
	Cov.	Δ	Cov.	Δ	Cov.	Δ	Cov.	Δ
GPT-5.2 (zero shot)	11.7	20.11	15.0	14.05	5.0	18.38	15.0	27.90
GPT-5.2 (few shot)	53.3	14.21	65.0	8.01	60.0	10.06	35.0	24.54
GPT-5.2 (persona)	13.3	19.68	25.0	13.12	0.0	17.50	15.0	28.40
PPI	95.0	3.86	100.0	2.38	95.0	3.57	90.0	5.64
PDI	95.0	6.04	95.0	7.31	95.0	5.57	95.0	5.24

Table 3: **Offensiveness.** Coverage (Cov.) and delta (Δ) by age group on the offensiveness task (GPT-5.2). Values below the 90% coverage are highlighted.

Group	PDI samples	PPI samples
Gender		
Woman	152.3	143.7
Man	47.6	56.4
Race		
White	164.7	162.9
Black/Afr. Am.	6.8	9.0
Asian	17.4	20.2
Age		
Age 18–34	85.1	87.0
Age 35–49	89.2	91.4
Age 50+	25.8	21.6
Education		
College degree	102.4	103.2
HS diploma	67.5	68.5
Grad. degree	12.2	11.4

Table 4: Average number of human annotations allocated per group by PDI and PPI, averaged across 20 trials (total budget 200 per trial, offensiveness task). PDI allocates 19% more annotations to Age 50+ than PPI (25.8 vs. 21.6), reflecting error-driven upsampling toward the hardest subgroup.

PDI is designed for settings where the main concern is not average performance across all groups, but better inference for those groups that are hardest for LLM proxies to capture. Our findings support this interpretation, with PDI showing its clearest benefits on harder-to-model groups, such as Age 50+, while maintaining coverage.

The framework may generalize beyond the tasks studied here. More broadly, corpus inference problems where annotator identity shapes labels are promising candidates for this approach. The demographic axes we use are a proof of concept. In practice, the stratification could follow any theoretically motivated partition, including ideological orientation, professional background, or lived experience (Orlikowski et al., 2025). The key requirement is that the error predictor receives informative features. Instead, where demographic proxies are weak, the adaptive gains will be smaller, and the method degrades to uniform PPI.

The problem we study also appears in three high-stakes settings that have received attention in NLP. In RLHF, preference labels reflect the perspectives of whoever provided feedback, yet models are trained as if those preferences were universal (Movva et al., 2025; Chen et al., 2025b). In LLM-as-judge evaluation, a single model scores outputs according to implicit standards that may not generalize across demographic groups (Dong et al., 2024). In LLM simulations, the task is often to predict distributions of opinions as self-reported in survey answers (Krsteski et al., 2026; Suh et al., 2025), and LLM predictions do not represent each respondent equally well. All three settings involve collapsing perspective-dependent judgments into a scalar, discarding variation that is not noise. Extending our framework to these settings is a natural direction, and one with implications for how models are trained or evaluated.

Lastly, several limitations remain. Our evaluation uses two tasks, and performance on tasks with subtler demographic structure is a promising avenue for future work. The burn-in phase requires uniform samples before adaptation begins, which may be inefficient when the human budget is very small. The error predictor is a gradient-boosted tree trained on demographic features alone; richer features, including linguistic properties of the text, could improve predictions and allocate the budget more efficiently.

6 Conclusion

LLMs are widely used as annotation proxies, but their outputs do not represent all demographic perspectives equally. We introduced a framework that treats the distribution of group-specific annotations as the estimand of interest, and allocates a limited human annotation budget adaptively to where LLM error is highest. Across two subjective rating tasks, PDI maintains valid coverage and reduces estimation error for the hardest subpopulations. The

framework is task-agnostic and can extend to any setting where annotator identity shapes labels, including RLHF, LLM-as-judge evaluation, and survey simulation. Correcting for LLM bias at the aggregate level is not enough; valid inference requires knowing *whose* perspective the model fails to capture, and directing human effort there.

Limitations

First, we stratify by gender, age, and education independently, but these axes correlate in practice. An annotator who is older and identifies as a particular gender occupies an intersection that our per-axis treatment does not capture. Modeling intersectional subgroups directly would require larger annotation budgets and raises challenges for the error predictor when intersectional cells are sparse. We leave joint stratification to future work. Because our predictor targets overall squared error, it does not disentangle reducible error (driven by underrepresentation) from irreducible noise (driven by inherent, intra-group heterogeneity). Consequently, while the adaptive rule correctly allocates budget to minority groups to improve precision, it will also oversample adequately represented but highly heterogeneous groups. While this irreducible noise could theoretically be mitigated by further partitioning the data into more homogeneous subgroups, doing so requires robust domain knowledge and experimentation to guarantee that the newly defined strata capture meaningful perspectives.

Second, our evaluation focuses on zero-shot, few-shot, and persona-prompted LLMs without fine-tuning. PDI is agnostic to how the proxy is constructed, and applies equally to fine-tuned models. A fine-tuned proxy that better tracks demographic variation would reduce the correction needed and likely improve efficiency further; our results represent a conservative setting where the proxy is cheap and off-the-shelf.

Third, we evaluate on two subjective rating tasks from a single dataset. Demographic axes are used as a proof of concept: in principle, any partition along which human perspectives vary systematically (like ideological orientation, professional background, lived experience) can serve as the stratification. Broader evaluation across tasks, domains, and non-demographic perspective axes remains an important direction.

Lastly, the adaptive component is deliberately simple: the error predictor relies only on demographic features, rather than richer textual signals that might improve targeting accuracy. In principle, this likely leaves some performance on the table. The current design is a proof of concept: it shows that even a minimal adaptive strategy can deliver meaningful subgroup gains.

Ethical considerations

We caution against using demographic persona prompting as a substitute for actual human input. Conditioning LLMs on demographic attributes can misrepresent and exaggerate group identities (Wang et al., 2025), and our results show that persona prompting often fails precisely for the groups it is meant to represent. Our framework instead treats LLM outputs as potentially inadequate proxies and uses human annotations to correct for that inadequacy.

Our approach estimates group-level means, not individual-level behavior. A finding that Age 50+ annotators rate statements differently on average should not be interpreted as a claim about how any particular older person reasons or perceives language. We aim to capture diversity in aggregate judgments, and caution against misreading our framework as a tool for stereotyping or as evidence of fixed group-level dispositions.

The demographic categories we use (gender, age, education) are coarse proxies for perspective, not deterministic accounts of it. These categories are themselves socially constructed and unevenly operationalized across datasets. Results should be interpreted in light of how demographic attributes were collected in POPQUORN, and may not transfer to populations where those categories have different meaning. The dataset we leverage is publicly available and does not contain personally identifying information, nor does it uniquely identify individual people.

Our framework requires collecting human annotations from demographically identified participants. This raises concerns about participant privacy, informed consent, and the potential for re-identification from demographic metadata. Any deployment should ensure that annotators are fairly compensated, particularly those from underrepresented groups who may be asked to contribute disproportionately under adaptive sampling.

Finally, adaptive upsampling concentrates annotation effort on groups where LLMs perform worst. In practice, these are often already marginalized

groups. While this is a feature of our design, it also means those annotators bear more of the annotation labor. This tradeoff should be made explicit to participants, and compensation should reflect it.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

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

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023a. Prediction-powered inference. *Science*, 382(6671):669–674.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. 2023b. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

Shilpa Arora, Eric Nyberg, and Carolyn P. Rosé. 2009. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 18–26, Boulder, Colorado. Association for Computational Linguistics.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, agreement, consensuality and annotated data quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2940–2948.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 10–19.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and 1 others. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021.

Hongyu Chen, Neele Falk, Michael Roth, and Agnieszka Falenska. 2025a. "feels feminine to me": Understanding perceived gendered style through human annotations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31447–31468.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025b. Steer-bench: A benchmark for evaluating the steerability of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36:68589–68601.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. 2025. SoMiToM: Evaluating multi-perspective theory of mind in embodied social interactions. *arXiv preprint arXiv:2506.23046*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, and 1 others. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLMm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Adam Fisch, Joshua Maynez, R Hofer, Bhuwan Dhingra, Amir Globerson, and William W Cohen. 2024. Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37:111489–111514.

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.

Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, Albuquerque, New Mexico. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long papers)*, pages 752–762.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2026. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! 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 4238–4249.

Katherine Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585.

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. 2026. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10887–10906.

Michelle S. Lam, Janice Teoh, James Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Cinoo Lee, Kristina Gligorić, Pratyusha Ria Kalluri, Maggie Harrington, Esin Durmus, Kiara L Sanchez, Nay San, Danny Tse, Xuan Zhao, MarYam G Hamedani, and 1 others. 2024. People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences*, 121(38):e2322764121.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505.

Bill Yuchen Lin, Dong-Ho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. 2019. AlpacaTag: An active learning-based crowd annotation framework for sequence tagging. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 58–63, Florence, Italy. Association for Computational Linguistics.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. 2025. What's in my human feedback? learning interpretable descriptions of preference data. *arXiv preprint arXiv:2510.26202*.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and AI needs in the fight against online hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Minh-Quoc Nghiem and Sophia Ananiadou. 2018. APLenty: annotation tool for creating high-quality datasets using active and proactive learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 108–113, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 84–94.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 175–190.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through LLM-driven active learning and human annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 98–111, St. Julians, Malta. Association for Computational Linguistics.

Deniss Ruder, Andero Uusberg, and Kairit Sirts. 2025. Assessing the reliability and validity of GPT-4 in annotating emotion appraisal ratings. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 1–11.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. 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.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wuehrl, Sabine Weber, and Roman Klinger. 2025. Which demographics do LLMs default to during annotation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17331–17348, Vienna, Austria. Association for Computational Linguistics.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. 2025. Missing the margins: A systematic literature review on the demographic representativeness of LLMs. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24263–24289.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170.

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.

Narjes Tahaei and Sabine Bergler. 2024. Analysis of annotator demographics in sexism detection. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 376–383, Bangkok, Thailand. Association for Computational Linguistics.

Kimberly Truong, Riccardo Fogliato, Hoda Heidari, and Steven Wu. 2025. Persona-augmented benchmarking: Evaluating LLMs across diverse writing styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Olga Shahmatova, Dmitry V. Dylov, Igor Galitskiy, and Artem Shelmanov. 2022. ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Marco Valentino, Ian Pratt-Hartmann, and André Freitas. 2021. Do natural language explanations represent valid logical arguments? verifying entailment in explainable nli gold standards. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 76–86.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. 2020. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. 2025. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. *arXiv preprint arXiv:2508.02853*.

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in NLP. *Preprint*, arXiv:2601.09065.

Mehmet Can Yavuz. 2019. Analyses of literary texts by using statistical inference methods. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 410–416, Bari, Italy. CEUR Workshop Proceedings.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2023. Data-efficient active learning for structured prediction with partial annotation and self-training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12991–13008, Singapore. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Tijana Zrnic and Emmanuel J Candès. 2024. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62993–63010.

A Prompting details

Table 6 lists the complete prompt texts used for annotation. Models were accessed using OpenRouter, using default parameters such as temperature and max token outputs.

B PDI Estimator

For each text i in group g_k, we define labeled and unlabeled IPW weights

$$w_i^{\text{lab}} = \frac{\xi_i}{\pi_i}, \qquad w_i^{\text{unlab}} = \frac{1-\xi_i}{1-\pi_i},$$

where $\xi_i = \mathbf{1}\{H_i \text{ collected}\}$ and π_i is the annotation probability.

For any quantity V, let $\bar{V}_{g_k}^{(w)} = \frac{\sum_{i\in g_k} w_i V_i}{\sum_{i\in g_k} w_i}$ denote the IPW-weighted mean over the group. We estimate $\theta^*_{g_k}$ using PPI++ (Angelopoulos et al., 2023b) applied within each group, with IPW-weighted means replacing the standard unweighted means to account for selective annotation.

We draw bootstrap resamples and compute $\hat{\theta}_{\text{PDI},g_k}$ on each resample. The confidence interval is

$$C_{1-\alpha,g_k} = \left[\hat{\theta}^{(\alpha/2)},\ \hat{\theta}^{(1-\alpha/2)}\right],$$

where $\hat{\theta}^{(p)}$ is the p-th quantile of the bootstrap distribution of $\hat{\theta}_{\text{PDI},g_k}$. The validity of the interval holds under adaptive (non-i.i.d.) sampling because of the IPW correction.

C LLM-Only coverage is consistently low

No model–prompting combination achieves the target coverage on either task. The best observed coverage across all 48 combinations is 0.73 (claude-haiku-4.5 few-shot on politeness; gemini-3.1-pro and gpt-oss-120b few-shot on offensiveness). The median coverage is 0.45 for politeness and 0.09 for offensiveness. The root cause is prediction bias rather than sampling variance: with all n predictions available, confidence intervals are narrow, but they are centered on the wrong value. This motivates methods such as PPI++ and PDI that rectify LLM estimates with human labels.

D Offensiveness is harder than politeness

Overall, across models and prompting variants, coverage and Δ are worse on offensiveness than on politeness. Several models that perform reasonably on politeness (e.g., llama-3-8b, mistral-large-2512) exhibit near-zero coverage and $\Delta > 30\%$

on offensiveness. Two factors contribute: A lower base rate (32% vs. 44%) makes errors costlier on a percentage-point scale. The concept of "offensive" is more subjective and socially situated than "polite," so models trained on different corpora disagree more with the human ground truth.

E Few-shot prompting is consistently best

Few-shot prompting is the best or tied-best variant for 7 of 8 models on politeness and for 7 of 8 models on offensiveness (Table 5). The gains can be large: claude-haiku-4.5 improves from $\Delta = 15.5\%$ (zero-shot) to 6.0% (few-shot) on politeness, and from 28.9% to 5.3% on offensiveness.

F Persona prompting often fails

Persona prompting sometimes matches few-shot performance but can also hurt it, particularly on offensiveness: claude-haiku-4.5: 0% coverage, $\Delta = 48.4\%$ (persona) vs. 54.5% coverage, 5.3% (few-shot); llama-3-8b: $\Delta = 64.3\%$ (persona) vs. 21.2% (few-shot); mistral-large-2512: $\Delta = 68.7\%$ (persona) vs. 34.1% (few-shot). The persona instruction appears to shift predictions systematically in the wrong direction for these models, suggesting that persona-based framing is sensitive to how a model was trained.

G Models

Politeness Under the best prompting variant, models cluster into two tiers. *Strong performers* ($\Delta < 7\%$): gpt-oss-120b (5.3%), mistral-large-2512 (5.7%), gemini-3.1-pro (5.7%), claude-haiku-4.5 (6.0%), gpt-5.2 (6.4%). *Weaker performers* ($\Delta \geq 7\%$): llama-3-8b (9.5%), claude-opus-4.6 (8.2%), claude-sonnet-4.6 (6.6% with persona but 10.5% with few-shot).

Offensiveness *Strong performers* (few-shot): gemini-3.1-pro ($\Delta = 5.0\%$), claude-haiku-4.5 (5.3%), claude-sonnet-4.6 (6.0%), gpt-oss-120b (6.2%). *Weakest*: mistral-large-2512 (0% coverage even with few-shot, $\Delta = 34.1\%$) and llama-3-8b (0% coverage, $\Delta = 21.2\%$). These open-weight models appear poorly calibrated for detecting offensiveness.

H Robustness tests

Figures 3 to 6 show robustness tests. Across four controlled experiments on synthetic data, PDI consistently outperforms PPI for the minority group while maintaining valid coverage for both methods. The precision advantage is largest when the human annotation budget is small, when the LLM makes more errors on the minority group, and when the minority group is small. These are the conditions where uniform sampling fails. Burn-in size has a modest effect on PDI's precision, with diminishing returns beyond 100 samples. On the majority group, PPI retains a slight edge throughout, since the LLM is already accurate there and uniform sampling suffices. PDI's adaptive allocation is most valuable when disparities in LLM performance across groups are large and groups are skewed.

I Disagreement analysis

To identify the sources of annotation disagreement, we first applied the Fightin' words method (Monroe et al., 2008) to uncover the words that are associated with the most disagreement between two groups of annotators. (≥ 50 vs. other strata). We then applied LLooM (Lam et al., 2024) to tokens with the highest human disagreement. Rather than analyzing individual words in isolation, this step helps us understand high-level themes that explain the disagreements.
We used this method to:

1. Identify themes that annotators aged ≥ 50 years rated differently in terms of politeness than other strata, and

2. Identify themes that GPT-5.2 was differing in annotations from human annotators.

The resulting themes are summarized in Tables 9 and 10. In particular, annotators in the older stratum (>= 50 years) perceive differently texts that embed requests for information (e.g., "if you have any questions, please feel free to call me"), hopeful sentiment ("look forward to seeing you tomorrow"), or polite closings.

J Additional datasets

To test whether our findings extend to other datasets with similar structure, we ran a supplementary analysis on three additional perspectivist datasets that release annotator-level labels together with annotator metadata:

- **MultiPICo** (Casola et al., 2024): irony detection for post–reply pairs from Reddit and Twitter. We use the English subset. Labels are

binary (ironic vs. not ironic). The annotator attributes are gender, age, ethnicity, and nationality (UK, US, Ireland, Australia, India).

- **Conversational Sarcasm Corpus (CSC)** (Jang and Frassinelli, 2024): the sarcasm of a response given a described situation, rated by external evaluators on a 1–6 scale. We binarize ratings at ≥ 4 (sarcastic). The evaluator attributes are gender, age, and ethnicity.

- **HS-Brexit** (Akhtar et al., 2021): hate speech detection in tweets about Brexit, one of the LeWiDi datasets used by Rizzi et al. (2024). Labels are binary. Each tweet is labeled by six annotators: a *target group* of three Muslim immigrants in the UK and a *control group* of three other annotators.

From each dataset we sample 1,000 (text, annotator) annotations. This matches the size of the POPQUORN subsets in the main experiments. Age is bucketed into 18–34, 35–49, and 50+, as in the main text. Annotators with missing gender or age, or with non-binary or undisclosed gender, are excluded, as in the main experiments.

We annotate every sampled instance with a single LLM, `gpt-oss-120b`, which we accessed through OpenRouter with default sampling parameters. We use the three prompting variants of Table 6, adapted to each task: zero-shot, few-shot, and persona. The few-shot variant uses 10 examples drawn at random from texts outside the sample. Persona prompts describe the annotator's demographic profile. For HS-Brexit, they describe the annotator as either a Muslim immigrant living in the UK or a UK resident who is not a Muslim immigrant.

J.1 Exploratory analysis: does LLM performance vary across groups?

PERSPECTIVE-DRIVEN INFERENCE can only improve over uniform sampling if the LLM's error depends on the annotator's group, because this dependence is what the error predictor learns and leverages. Before running PDI, we therefore test on each dataset whether LLM error varies across demographic groups. For each demographic dimension and prompting variant, we compute the per-group error rate: the fraction of annotations on which the binarized LLM label disagrees with the individual human label. We then test whether error is independent of group membership with a χ^2 test. For reference, we also test whether the human labels themselves vary across groups. Groups with fewer than 30 annotations in the sample are excluded.

Table 11 shows the results. Across all three datasets, the LLM's error rate is relatively flat across groups:

- **MultiPICo:** per-group error rates lie between 0.25 and 0.39 for every dimension and variant.
- **CSC:** they lie between 0.23 and 0.39.
- **HS-Brexit:** they lie between 0.15 and 0.23.

None of the 21 tests on LLM error remains significant after Bonferroni correction ($\alpha = 0.05/21 \approx 0.002$). On MultiPICo and HS-Brexit, LLM error does not differ across groups. There, PDI's error predictor can only fit noise, so adaptive sampling reduces to uniform sampling (PPI) in expectation, with extra variance from the importance weights. We therefore do not evaluate PDI on these two datasets.

We evaluate PDI on CSC only, since it has the largest error gap across groups: under few-shot prompting, which is the variant PPI and PDI use on CSC, error is 0.38 for Black evaluators and 0.23 to 0.30 for the other ethnicity groups (Table 12). Of the three datasets, CSC is therefore the most relevant test case for error-driven upsampling.

J.2 Setup on CSC

We follow the protocol of main evaluation exactly:

- total human annotation budget $n_{\text{human}} = 200$ out of $n = 999$ annotations (one instance whose zero-shot label could not be parsed is dropped);
- a uniform burn-in of $n_{\text{burnin}} = 100$, followed by $B = 3$ adaptive batches;
- smoothing with the uniform distribution at weight $\tau = 0.1$;
- an XGBoost error predictor trained on one-hot encoded evaluator gender, age, and ethnicity;
- 95% confidence intervals, with results averaged over 20 trials.

The target $\theta^*_{g_k}$ is the share of annotations in group g_k that rate the response as sarcastic (≥ 4). PPI and PDI use the prompting variant with the

highest overall accuracy. This is few-shot (accuracy 0.708, vs. 0.705 for zero-shot and 0.683 for persona).

J.3 Results

Table 12 reports coverage and Δ for the ethnicity dimension, where group differences in LLM error are largest. Table 13 reports how many human annotations PDI and uniform PPI allocate to each group.

Coverage. As on POPQUORN, the LLM-only baselines frequently miss the target. Averaged across ethnicity groups, coverage is 76.3% for zero-shot, 41.3% for few-shot, and 66.3% for persona prompting. Few-shot, the most accurate variant overall, covers the White group in only 5% of trials. PDI maintains coverage of at least 95% for every ethnicity group (97.5% on average). PPI's coverage falls to 85% for Asian evaluators and 75% for evaluators of mixed ethnicity.

Delta. For Black evaluators, the group with the highest LLM error, PDI attains the lowest Δ of all methods (8.21, vs. 8.67 for PPI and at least 8.64 for the LLM-only baselines). Averaged across ethnicity groups, PPI has a slightly lower Δ than PDI (8.59 vs. 9.51). This mirrors the main results: PDI's benefit is targeted at the hardest group rather than uniform across groups.

Upsampling lift. The error predictor successfully identifies Black evaluators as the high-error group. PDI allocates them 31.7 human annotations on average, compared with 22.9 under uniform PPI, a 38% increase (Table 13). For the other groups, PDI's allocation stays within -13% and $+8\%$ of uniform sampling, consistent with their near-identical error rates. Along the gender and age dimensions, where error rates are nearly flat (0.29–0.30 and 0.24–0.31), PDI and PPI both maintain coverage. PDI's coverage ranges from 85% (Age 50+) to 100%, and PPI has a lower Δ in every group, as expected when there is no error difference to exploit.

J.4 Takeaway

On CSC, PDI behaves as it does on the POPQUORN tasks. The LLM-only baselines, including persona prompting, often fail to cover the group-level targets. PDI upsamples the group where the LLM is least accurate (Black evaluators, +38% annotations) and achieves the lowest Δ and full coverage for that group. Averaged over groups, uniform PPI remains competitive or better on Δ. The additional datasets also make the method's precondition concrete. PDI's gains depend on LLM error varying across annotator groups. On MultiPICo and HS-Brexit it does not, and there we expect PDI to offer no advantage over uniform PPI. Checking whether LLM error varies across groups, which requires only the burn-in sample, is therefore a useful first step before choosing between PDI and uniform sampling.

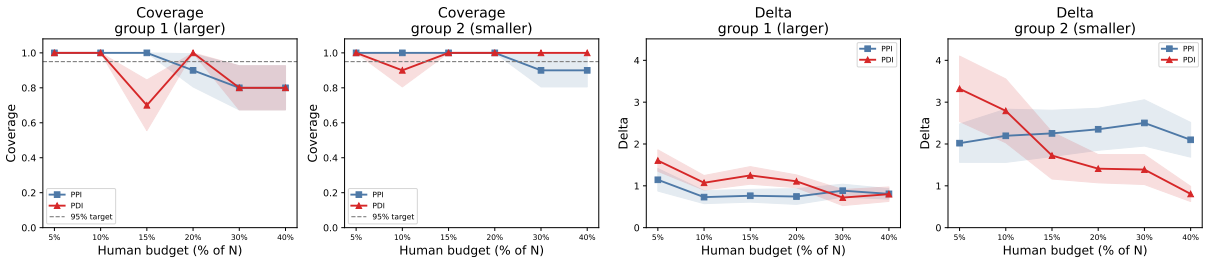


Figure 3: **Human annotation budget.** Coverage and delta of PPI and PDI as a function of human annotation budget (as a percentage of total sample size N), shown separately for group 1 (larger) and group 2 (smaller). Shaded regions indicate 95% confidence bands across simulation runs.

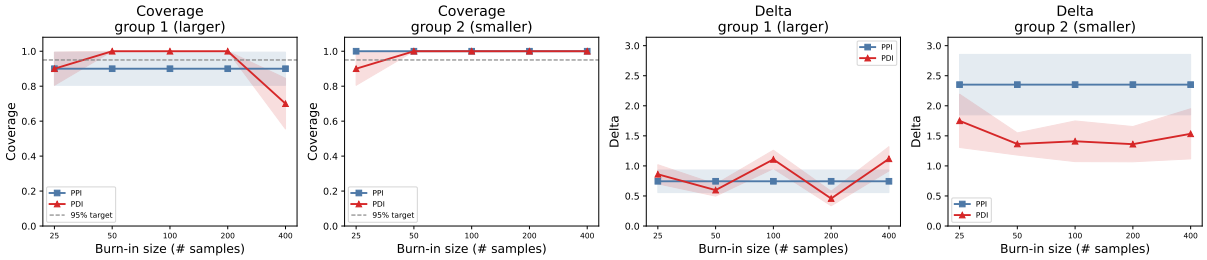


Figure 4: **Burn-in size.** Coverage and delta of PPI and PDI as a function of burn-in size, shown separately for group 1 (larger) and group 2 (smaller). Shaded regions indicate 95% confidence bands across simulation runs.

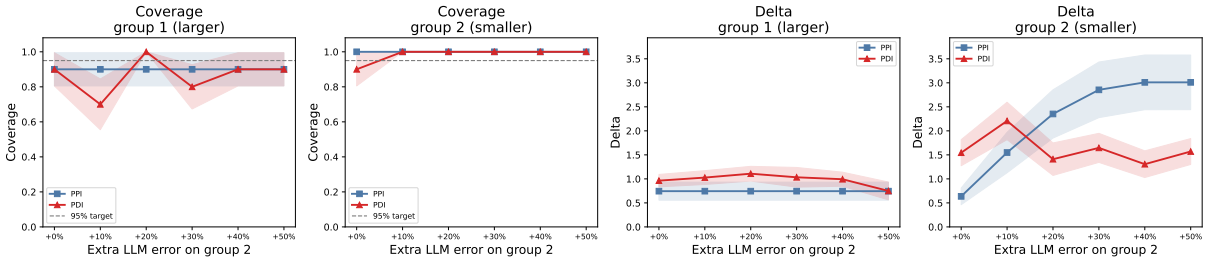


Figure 5: **LLM extra error on smaller group.** Coverage and delta of PPI and PDI as a function of LLM error discrepancy, shown separately for group 1 (larger) and group 2 (smaller). Shaded regions indicate 95% confidence bands across simulation runs.

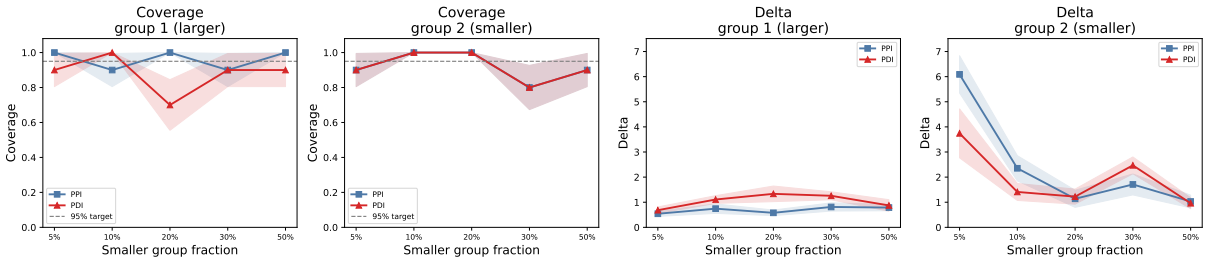


Figure 6: **Size of smaller group**. Coverage and delta of PPI and PDI as a function of the group skew, shown separately for group 1 (larger) and group 2 (smaller). Shaded regions indicate 95% confidence bands across simulation runs.

Model	Prompting	Politeness		Offensiveness	
		Cov.	Δ	Cov.	Δ
gpt-5.2	ZERO-SHOT	0.36	6.5	0.09	18.7
	FEW-SHOT	0.64	6.6	0.09	13.5
	PERSONA	0.55	6.4	0.09	18.4
claude-sonnet-4.6	ZERO-SHOT	0.09	12.0	0.00	37.7
	FEW-SHOT	0.27	10.5	0.64	6.0
	PERSONA	0.64	6.6	0.00	41.4
claude-opus-4.6	ZERO-SHOT	0.45	8.2	0.45	8.0
	FEW-SHOT	0.45	9.1	0.27	13.1
	PERSONA	0.45	8.3	0.45	8.2
gemini-3.1-pro	ZERO-SHOT	0.55	5.7	0.27	9.4
	FEW-SHOT	0.64	7.7	0.73	5.0
	PERSONA	0.55	6.3	0.18	14.3
claude-haiku-4.5	ZERO-SHOT	0.00	15.5	0.00	28.9
	FEW-SHOT	0.73	6.0	0.55	5.3
	PERSONA	0.09	14.5	0.00	48.4
llama-3-8b	ZERO-SHOT	0.00	11.2	0.00	45.9
	FEW-SHOT	0.27	9.5	0.09	21.2
	PERSONA	0.00	13.2	0.00	64.3
mistral-large-2512	ZERO-SHOT	0.55	5.7	0.00	67.3
	FEW-SHOT	0.64	6.4	0.00	34.1
	PERSONA	0.45	7.4	0.00	68.7
gpt-oss-120b	ZERO-SHOT	0.55	6.6	0.18	15.1
	FEW-SHOT	0.55	6.3	0.73	6.2
	PERSONA	0.55	5.3	0.27	13.1

Table 5: LLM-only evaluation across 8 models and 3 prompting strategies on politeness and offensiveness annotation tasks. Cov. = coverage (higher is better). Δ = mean absolute error in percentage points (lower is better), averaged over all non-empty demographic subgroups.

Prompt Type	Prompt Text
Politeness Zero Shot	Consider you read this email from a colleague. How polite is it? Answer using a 1–5 Likert scale where 1 means 'not polite at all' and 5 means 'very polite'. Return ONLY one integer.
Offensiveness Zero Shot	Consider you read this comment on Reddit. How offensive is it? Answer using a 1–5 Likert scale where 1 means 'not offensive at all' and 5 means 'very offensive'. Return ONLY one integer.
Politeness Few Shot	Here are example politeness ratings (1=very impolite, 5=very polite): {Followed by Text: and Rating: for 10 examples. Rest of the prompt is the same as zero shot.}
Offensiveness Few Shot	Here are example offensiveness ratings (1=not offensive, 5=extremely offensive): {Followed by Text: and Rating: for 10 examples. Rest of the prompt is the same as zero shot.}
Politeness Persona Prompt	You are {persona}. Answer all questions from the perspective of {persona}. Be consistent with this perspective, but still follow the instructions. {followed by same prompt as Politeness Zero Shot}. persona = a gender aged age with education
Offensiveness Persona Prompt	You are {persona}. Answer all questions from the perspective of {persona}. Be consistent with this perspective, but still follow the instructions. {followed by same prompt as Offensiveness Zero Shot}.

Table 6: Prompts used for LLM annotation.

Table 7: Example annotations illustrating inter-annotator disagreement on offensiveness. Each row corresponds to one annotator. A ≥50 annotator rating of 5.0 was required for selection. **Bold** rating = annotator aged ≥50 who rated 5.0 (selection criterion). Warning: content in this table may be upsetting or offensive.

Offensiveness (1 = not offensive, 5 = very offensive)		
Text	**Age**	**Rating**
"Women's studies is dumb though, stop defending your degree?"	18–24	5
	25–29	5
	30–34	3
	40–44	5
	45–49	1
	50–54	1
	50–54	**5**
	54–59	1
	mean = 3.25, std = 1.98	
"Still not as scary as autism."	18–24	5
	18–24	5
	35–39	3
	35–39	2
	40–44	1
	45–49	1
	54–59	**5**
	54–59	1
	>65	3
	>65	1
	mean = 2.70, std = 1.77	
"Let's ban baths and haircuts for non-vaxxed people so everyone else knows to stay the hell away!"	18–24	5
	25–29	2
	30–34	1
	40–44	1
	45–49	4
	54–59	1
	60–64	**5**
	>65	1
	>65	5
	mean = 2.78, std = 1.92	

Table 8: Example annotations illustrating inter-annotator disagreement on politeness. Each row corresponds to one annotator. A $\geq$50 annotator rating of 5.0 was required for selection. **Bold** rating = annotator aged $\geq$50 who rated 5.0 (selection criterion). Warning: content in this table may be upsetting or offensive.

Politeness (1 = very impolite, 5 = very polite)		
Text	**Age**	**Rating**
"No, tell her that it is off of Buffalo Speedway."	18-24	1
	25-29	3
	30-34	1
	30-34	5
	50-54	4
	60-64	**5**
	>65	1
	mean = 2.86, std = 1.86	
"I would appreciate it if you didn't blame me for spills in your cube."	18-24	1
	18-24	1
	18-24	1
	30-34	4
	54-59	3
	60–64	**5**
	mean = 2.50, std = 1.76	
"You can find my draft of the filing below if you are not already familiar with it."	18-24	5
	25-29	5
	30-34	2
	45-49	4
	54-59	1
	60-64	**5**
	>65	2
	mean = 3.43, std = 1.72	

Concept	Prevalence	Match Count	Meaning	Example
Request for Information	0.26	64	Does the text include a request for information or clarification?	“I HAVE NOT RECEIVED YOUR DATA FORMS… IF YOU HAVE ANY QUESTIONS, PLEASE FEEL FREE TO CALL ME. Thank you”
Hopeful Sentiment	0.18	45	Does the text express hope or a positive outlook?	“Look forward to seeing you tomorrow.”
Polite Closing	0.09	23	Does the text include a polite closing or sign-off, such as ’best regards’?	“I will leave the legal review up to your team. Please let me know if you have any questions about these CPs. Regards”
Attachment Mention	0.07	17	Does the text mention an attachment or include the word ’attached’?	“I am send you the following attached documents per request.”
Expression of Gratitude	0.04	9	Does the text express gratitude or thanks?	“I would appreciate if it you didn’t blame me for the spills in your cube”

Table 9: Themes with age-dependent differences ($\geq$ 50 vs. other strata). Warning: content in this table may be upsetting or offensive.

Concept	Prevalence	Match Count	Meaning	Example
Request for Action	0.28	8	Does the text contain a request for action or follow-up, indicated by words like 'please', 'let', or 'thanks'?	"Please find attached the Summary and Hot List as of 11/16. Please contact me if you have any questions/comments. Thanks"
Emotional Sharing	0.21	6	Does the text involve sharing emotions or personal connections?	"Are you in love with him?"
Evaluation and Quality	0.14	4	Does the text involve evaluating the quality or value of something?	"Let me tell you who won't win - RU. They are arguably the worst division 1 football team I've ever seen. I saw them for the 1st time this weekend on TV. They are horrible. Hope everything's going well. Life at Enron sucks. I'm sure you've been following the situation. I'm sure there'll be a #1 bestseller written about this circus. Talk to you soon. KR"
Motivation and Support	0.10	3	Does the text discuss the importance of encouragement or support?	"Congrats on keeping your turf. I hope this question doesn't offend you, but would you trade it for an $80 ENE price? Thanks."
Travel Plans	0.07	2	Does the text mention any travel plans or locations?	"Hey give me a call! You should come to NYC this weekend to party with your friends!"

Table 10: Themes with Model-Human Disagreement (GPT-5.2 vs human annotators). Warning: content in this table may be upsetting or offensive.

			Human label		LLM error (zero shot)		LLM error (few shot)		LLM error (persona)	
Dataset	Dimension	K	Range	p	Range	p	Range	p	Range	p
MultiPICo	Gender	2	0.30–0.34	0.272	0.29–0.32	0.299	0.30–0.33	0.391	0.29–0.32	0.318
	Age	3	0.29–0.36	0.249	0.30–0.35	0.398	0.28–0.39	0.044	0.27–0.35	0.180
	Ethnicity	4	0.25–0.39	0.135	0.26–0.35	0.546	0.31–0.36	0.752	0.25–0.36	0.223
	Nationality	5	0.28–0.37	0.406	0.27–0.36	0.352	0.28–0.36	0.517	0.25–0.38	0.031
CSC	Gender	2	0.31–0.35	0.140	0.28–0.31	0.323	0.29–0.30	0.839	0.29–0.35	0.083
	Age	3	0.26–0.36	0.048	0.27–0.31	0.363	0.24–0.31	0.229	0.30–0.33	0.689
	Ethnicity	4	0.28–0.41	0.231	0.29–0.37	0.477	0.23–0.38	0.161	0.27–0.39	0.457
HS-Brexit	Annotator group	2	0.05–0.21	$<$0.001	0.17–0.21	0.115	0.15–0.17	0.614	0.19–0.23	0.138

Table 11: **Exploratory analysis on additional datasets.** For each demographic dimension (K groups with ≥ 30 annotations in the 1,000-annotation sample), we show the range of per-group rates across groups and the p-value of a χ^2 test of independence between group and the quantity. *Human label* is the per-group positive rate. *LLM error* is the per-group rate at which the binarized `gpt-oss-120b` label disagrees with the human label. No test of LLM error survives Bonferroni correction across the 21 tests.

	Avg (Ethnicity)		White		Black		Asian		Mixed	
Method	Cov.	Δ	Cov.	Δ	Cov.	Δ	Cov.	Δ	Cov.	Δ
Group size n / LLM error (few shot)	999 / 0.29		746 / 0.28		117 / 0.38		70 / 0.30		43 / 0.23	
gpt-oss-120b (zero shot)	76.3	11.63	35.0	8.86	95.0	8.64	80.0	14.73	95.0	14.28
gpt-oss-120b (few shot)	41.3	17.16	5.0	12.82	35.0	20.22	45.0	19.90	80.0	15.69
gpt-oss-120b (persona)	66.3	13.43	20.0	11.52	90.0	8.84	75.0	13.90	80.0	19.46
PPI	86.3	8.59	90.0	2.62	95.0	8.67	85.0	8.82	75.0	14.25
PDI	97.5	9.51	100.0	3.14	100.0	8.21	95.0	12.44	95.0	14.26

Table 12: **CSC (sarcasm).** Coverage (Cov., %) and delta (Δ, percentage points) by evaluator ethnicity (`gpt-oss-120b`, $n_{\text{human}} = 200$, 20 trials). Values below 90% coverage are highlighted. Bold indicates the best performance for Black evaluators, the group with the highest LLM error. PPI and PDI use the few-shot variant. The *Avg* group-size cell shows the total number of annotations and the overall few-shot error rate.

Dimension	Group	LLM err.	PDI	PPI
Gender	Woman	0.30	110.8	105.2
	Man	0.29	89.2	94.8
Age	Age 18–34	0.31	95.6	96.8
	Age 35–49	0.30	75.2	69.8
	Age 50+	0.24	29.2	33.5
Ethnicity	White	0.28	141.5	149.9
	Black	0.38	31.7	22.9
	Asian	0.30	13.0	13.9
	Mixed	0.23	8.3	9.1

Table 13: **CSC.** Average number of human annotations allocated per group by PDI and PPI, averaged across 20 trials (total budget 200 per trial). *LLM err.* is the per-group error rate of the few-shot variant. PDI allocates 38% more annotations to Black evaluators than PPI (31.7 vs. 22.9), the group with the highest LLM error.