Modeling Annotator Disagreement with Demographic-Aware Experts and Synthetic Perspectives

Yinuo Xu[1], **Veronica Derricks**[2], **Allison Earl**[1], **and David Jurgens**[1]
[1]University of Michigan
[2]University of Colorado Boulder
{yinuoxu, anearl, jurgens}@umich.edu, Veronica.Derricks@colorado.edu

Abstract

We present an approach to modeling annotator disagreement in subjective NLP tasks through both architectural and data-centric innovations. Our model, DEM-MOE (Demographic-Aware Mixture of Experts), routes inputs to expert subnetworks based on annotator demographics, enabling it to better represent structured, group-level variation compared to prior models. DEM-MOE consistently performs competitively across demographic groups, and shows especially strong results on datasets with high annotator disagreement. To address sparse demographic coverage, we test whether LLM-generated synthetic annotations via zero-shot persona prompting can be used for data imputation. We show these synthetic judgments align moderately well with human annotations on our data and offer a scalable way to potentially enrich training data. We then propose and evaluate approaches for blending real and synthetic data using strategies tailored to dataset structure. We find that the optimal strategies depend on dataset structure. Together, these contributions improve the representation of diverse perspectives in subjective NLP.

1 Introduction

Substantial disagreement among annotators is common in subjective tasks such as toxicity detection, misinformation labeling, and politeness evaluation. These disagreements often reflect meaningful differences in perspective rooted in social, cultural, or demographic backgrounds rather than random noise (Holland and Quinn, 1987; Larimore et al., 2021; Sap et al., 2022). For example, a comment judged toxic by younger users might seem benign to older ones, reflecting differing sensitivities to language and topics. Modeling such perspective variation is essential for building systems that represent and reason over diverse viewpoints. Yet many approaches treat disagreement as noise, collapsing annotations into a single label and marginalizing minoritized perspectives (Prabhakaran et al., 2021). Recent work instead treats disagreement as signal (Dawid and Skene, 1979), using weighting, filtering, or learning from annotation distributions (Uma et al., 2022). Distributional modeling has become prominent in NLP tasks where disagreements reflect socially grounded variation (Mostafazadeh Davani et al., 2022; Fleisig et al., 2024; Gordon et al., 2022; Wan et al., 2023). However, such methods often lack inductive biases to capture structured group-level variation, risking underrepresentation of marginalized perspectives.

We introduce DEM-MOE, a Demographic-Aware Mixture of Experts model that learns to represent subjective judgments by routing inputs to expert subnetworks based on annotator demographics. This design introduces inductive bias: similar annotators may reason about inputs similarly. DEM-MOE outperforms strong baselines and SOTA systems across multiple datasets and demographic groups. Beyond architectural innovation, we address modeling disagreement in low-data settings with sparse demographic coverage using LLM-generated synthetic annotations and blended training strategies. Optimal strategies depend on dataset structure, showing how DEM-MOE, combined with strategic data augmentation, effectively models viewpoint diversity even in low-resource scenarios.

We make three contributions: (1) We propose DEM-MOE, a modular architecture with demographic-aware routing to capture structured variation in annotation behavior. (2) We evaluate the alignment of zero-shot LLM-generated annotations with human ratings. (3) We present a framework for incorporating synthetic data, showing that its effectiveness depends on dataset structure.[1]

[1]Code available at https://github.com/Nancy078/DEM-MoE

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41260–41294

2 Related Work

Modeling Individual & Group Preferences in Recommendation Systems. Recommendation systems extensively model individual and group preferences. Group-based methods aggregate member choices into collective decisions, often using centrality-aware representations (Yin et al., 2022) or hierarchical attention to separate individual and group signals (Xu et al., 2024; Liang et al., 2022; Wang et al., 2024b). Mixture-of-Experts (MoE) architectures further capture multifaceted user interests (Liu et al., 2024), model user–item interactions (Zhao et al., 2020), and enable multi-task personalization (Kong et al., 2024), including user and group modeling (Gong et al., 2023; Liu et al., 2025). In contrast, no prior NLP work uses MoEs to jointly model individual and group variation. While our task and recommenders model variation across individuals and groups, the nature and objectives differ. Whereas recommenders target idiosyncratic preferences with weak demographic patterns and aim to merge signals into a unified ranking across items, our task focuses on predicting singular judgments using systematic socio-demographic regularities. Annotators from similar groups may evaluate content in shared ways, and rather than collapsing disagreement, our approach explicitly preserves them to represent diverse perspectives.

Modeling Annotator Disagreement in NLP. Disagreement among annotators is common in subjective NLP tasks like toxicity classification, misinformation detection, and stance analysis. Traditional methods treat disagreements as noise by using majority voting or averaging to create a single "gold" label per instance, which can obscure meaningful variation from underrepresented or minoritized groups (Prabhakaran et al., 2021). Alternatives include early work that measured or filtered disagreement to improve data quality (Reidsma and op den Akker, 2008; Klebanov and Beigman, 2014), and more recent approaches that learn from disagreement directly: a) Latent truth and modeling reliability modeling, inferring latent true labels while estimating annotator behavior (Plank et al., 2014; Dawid and Skene, 1979; Simpson et al., 2015; Ivey et al., 2025); b) Task-based annotation models (Mostafazadeh Davani et al., 2022; Fleisig et al., 2024; Jinadu and Ding, 2024), which treat each annotator as a distinct task, modeling labeling behavior directly rather than collapsing annotations into a single truth; c) Embedding-based annotator models (a category that includes our proposed model), which treats disagreement as systematic variation in annotator behavior and encode annotator differences in a shared latent space (Kocoń et al., 2021; Mokhberian et al., 2024; Gordon et al., 2022), or even predicts disagreement directly (Wan et al., 2023; Parappan and Henao, 2025; Weerasooriya et al., 2023). The explanatory power of demographics has shown mixed results, with some work indicating success using group-level demographic information (Gordon et al., 2022; Fleisig et al., 2024), while others cast doubt on the effectiveness of such methods (Orlikowski et al., 2023; Beck et al., 2024b; Hu and Collier, 2024). The explanatory power of demographics is also dataset-dependent. There could be other sources of disagreement, ones that come from the individual annotator (such as individual moral values (Davani et al., 2024), personalities (Mieleszczenko-Kowszewicz et al., 2023)), or the task itself, such as the order in which annotators see observations (Beck et al., 2024a). Nevertheless, demographic features remain the most commonly-explored feature in modeling disagreement (Xu and Jurgens, 2026). However, even when demographic features are used, most models treat them as input to shared encoders, without architectural diversity or inductive biases to model group-specific reasoning. As a result, they capture individual variation but struggle with systematic group-level differences.

Dataset	#Inst	#Ann	Avg/Inst	IAA (α)	Entropy	Max Coef
Safety	350	123	123.0	0.241	0.742	Race (0.559)
Offensiveness	1,500	262	8.69	0.287	1.212	Age (1.351)
Patient Centered Comm.	2,230	589	3.33	0.287	1.492	Age_group (2.528)
Politeness	3,718	506	6.74	0.440	1.395	Race (1.427)
Toxicity	107,620	17,172	4.74	0.272	1.070	Age_range (2.056)

Table 1: Dataset statistics. '#Inst" = number of instances, #Ann" = annotators, Avg/Inst" = avg. annotators per instance, IAA" = Krippendorff's α, Entropy" = mean entropy of annotator ratings per instance. "Max Coef" = demographic feature with the largest absolute ridge regression coefficient per dataset.

3 Data

We evaluate on five datasets spanning diverse tasks, demographic coverage, and levels of disagreement (Table 1; see full details in Appendix A.1.1). The Toxicity dataset (Kumar et al., 2021) is the largest, with the broadest demographic coverage (2,535 combinations) and low agreement (α=0.27), suggesting systematic disagreement. Ridge regression predicting annotator ratings using demographics shows strong demographic signal, especially

for age (2.06). The Safety dataset (Aroyo et al., 2023) contains dense annotations (123 per instance) across 48 combinations, but has low agreement (α=0.24). Demographic signal is weak likely due to the complexity of harmfulness judgments. The POPQUORN dataset (Pei and Jurgens, 2023) includes Politeness and Offensiveness. Politeness shows the highest agreement (α=0.44) and low entropy, with strong demographic effects (race and education). Offensiveness with moderate demographic diversity and agreement shows notable signal for age and occupation. The Patient Centered Communication (PCC) dataset rates doctor–patient interactions across multiple attributes[2]. It has sparse annotations (3.3 per instance) and high uncertainty (entropy). PCC also shows the strongest demographic signal, especially for age. These datasets highlight varying balances of idiosyncratic vs. systematic disagreement, underscoring the need for models that preserve demographic variation.

4 DEM-MOE for Modeling Disagreement

Recent work on annotator disagreement has moved beyond majority-vote aggregation toward models that predict annotator-specific labels (Mostafazadeh Davani et al., 2022; Fleisig et al., 2024; Gordon et al., 2022). Many incorporate annotator identity or demographics, but typically by concatenating demographic features with text (Fleisig et al., 2024; Gordon et al., 2022; Wan et al., 2023) or by implicitly encoding annotator perspectives through embeddings (Deng et al., 2023). Because these systems model all annotators through a single network, they cannot explicitly specialize in distinct judgment patterns across demographic groups. Consequently, their ability to represent shared group behaviors remains limited. Orlikowski et al. (2023) partially addresses this by injecting demographic information into a shared encoder with group-specific layers, but intersectional identities are still modeled through separate, independent components, constraining the model's capacity to learn intersectional patterns.

We propose a new approach, DEM-MOE (Demographic-aware mixture of experts), based on Mixture of Experts (MoE) (Fig. 1) (Shazeer et al., 2017), which naturally supports modular specialization (different experts learn distinct annotation patterns linked to demographic groups) and selective routing (inputs are dynamically directed to relevant experts based on annotator demographics). The input consists of the text snippet (encoded with Modern-BERT), annotator embedding, demographic embeddings, and then all three are concatenated into the MoE input, which is then routed to experts to predict the annotator's rating. Our architecture encodes inductive bias: annotators from similar demographic groups may share systematic ways of judging texts. We address a key gap in prior work: the lack of structured inductive bias. While large networks can learn subgroup variation, they may not do so in a structured or interpretable way. They may also overfit to dominant groups without an architectural signal promoting subgroup differentiation. Our model makes this inductive bias more robust and interpretable, especially under data imbalance or sparse subgroup representation. Our model's routing also allows experts to naturally specialize in intersectional subgroups.

We target structured variation in annotator perspectives, which includes demographic groups, within-group variations, cross-group variations, and also intersectional clusters that arise naturally in the embedding space and routing patterns. The architecture is designed to be flexible: if demographic attributes are weakly correlated with the labels, the MoE routes based on other dimensions of the input. We aim to provide a framework that can integrate multiple structured sources of disagreement whenever they are informative. We discuss the model components next: 1) learned annotator and demographic embeddings; 2) expert selection and dynamic routing; and 3) expert load balance and specialization via a weighted training loss.

Annotator & Demographic Embeddings. We initialize Bayesian embeddings (Vilnis and McCallum, 2015) for annotators and their demographic attributes, enabling the model to capture annotator-specific idiosyncrasies and demographic-related biases. Each is represented by a learned Gaussian posterior distribution, with embeddings sampled during training via the reparameterization trick (Kingma and Welling, 2014). We concatenate the text (with Modern-BERT (Warner et al., 2024)), annotator, and demographic embeddings into the MoE input: $\mathbf{x} = [\mathbf{e}_{\text{text}}; \mathbf{e}_{\text{ann}}; \mathbf{e}_{\text{demo}}]$

Expert Selection. To promote expert specialization aligned with demographic-group preferences, we build a pool of n experts (n = number of demographic groups). Here, n is a simple initialization heuristic, so that demographic categories give a weak inductive bias to let experts

[2]Details on how PCC is collected are in Appendix A.1.3

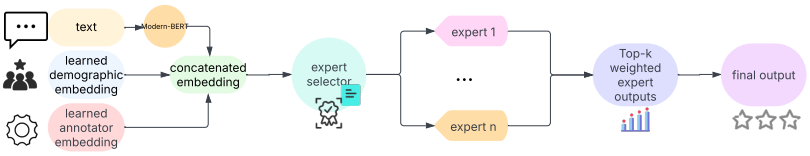


Figure 1: The architecture of our DEM-MOE model. We concatenate a text embedding, a learned annotator embedding, and learned demographic embedding and pass them through an expert selector, which produces logits over a shared pool of experts. At the expert selector, the inputs are directed to the most relevant experts, encoding the inductive bias. The top-k experts are selected per sample, and their logits are normalized via softmax to generate weights. The final output, a single rating, is computed as a weighted combination of the top-k expert outputs.

discover any structure in the data. A lightweight gate maps the input vector to expert scores: $\mathbf{s} = W_s\,\mathbf{x} + \mathbf{b}, \mathbf{p} = \mathrm{softmax}(\mathbf{s})$. We hard-select the top k experts ($k \in \{2, 3\}$): $\mathcal{I}_k = \arg\mathrm{topk}(\mathbf{p}, k)$. Gating weights p_j determine which demographic-aware experts are used. Each expert f_j receives the same input; their outputs are mixed sparsely: $\mathbf{h} = \sum_{j \in \mathcal{I}_k} p_j\, f_j(\mathbf{x})$. A linear regression head predicts the snippet rating. Ratings are z-score normalized to stabilize training. We prefer hard expert selection over soft gating to reinforce specialization, whereas soft gating dilutes gradients.

Expert Load Balance & Specialization. Naive routing often leads to expert load imbalance, routing collapse, or bottlenecks. While some auxiliary losses improve hardware efficiency via uniformity (Fedus et al., 2021; Wang et al., 2024a), we focus on expert diversity with a normalized standard deviation loss, allowing roughly even usage, rather than strict uniformity, to support demographic specialization. We also include an orthogonality loss to encourage distinct expert features, and a variance loss to promote diverse routing paths (Guo et al., 2025). Our training loss further encodes the inductive bias, where we include regularizations for annotator and demographic embeddings, load loss, orthogonality, variance loss, and demographic within-group specialization loss (see Appendix B).

5 Experiment 1: Modeling Perspectives

We hypothesize that **H1**: Models that incorporate demographic structure as an inductive bias more effectively capture diverse annotator perspectives. We first test if the experts are sufficiently specialized within and across demographic groups. We then compare DEM-MOE to other models' representativeness across demographic groups.

5.1 Experiment Setup

All models are evaluated to predict individual annotator ratings using the text of each item and, where applicable, annotator or demographic information. Each model differs in its inputs (text only vs. text + annotator/demographics), while the output is a scalar rating per annotator–item pair, except for ModernBERT (which predicts at the snippet level with text input). (1) Probabilistic Matrix Factorization (PMF) (Salakhutdinov and Mnih, 2007) uses only annotator and item IDs, learning latent embeddings for both and modeling their interaction as a probability distribution over ratings, thus capturing annotator regularity without text or demographics. (2) As a text-based baseline, we fine-tune ModernBERT-large (Warner et al., 2024) with a single linear regression head on the CLS embedding to predict the mean snippet rating. ModernBERT (Warner et al., 2024) is a text-based baseline that encodes the snippet, and we regress on the embedding to predict the mean snippet rating. (3) LLaMA-3.1-8B-Instruct (Orlikowski et al., 2025) receives text alone in a zero-shot setting or text plus annotator demographic descriptions when LoRA-fine-tuned with sociodemographic prompts. Their results improved over text-only models but revealed the model did not benefit from demographics. (4) Annotation + Annotator Embedding model (En + Ea) (Deng et al., 2023) is a SOTA system for explicitly modeling both item and annotator-level variance but does not incorporate demographic information. (5) Jury Learning (Gordon et al., 2022) uses ModernBERT text embeddings, annotator embeddings, and group-level demographic embeddings, concatenated and passed through cross and deep networks (Wang et al., 2021) to predict annotator ratings; this architecture most closely parallels ours, which models an interaction between the text, the demographics, and the annotators' idiosyncrasies.

Training. Train/dev/test splits are created at the instance level, ensuring no overlap of snippets between the splits while allowing some overlap in annotators between the train and test sets. Offensiveness has 92% overlap, Politeness 91%, Safety 85%, Toxicity 40%, and PCC 87%. Our evaluation goal is to assess generalization to new content, which follows standard practice in modeling using annotator embeddings (Gordon et al., 2022; Mostafazadeh Davani et al., 2022; Fleisig et al., 2024). Training details are in Appendix B.1.

Evaluation. We use Mean Absolute Error (MAE) to measure the average absolute different between the model's predicted rating and actual annotator rating, at the individual level(Gordon et al., 2022; Fleisig et al., 2024). For analysis, we aggregate across annotators' demographic groups.

5.2 Results

DEM-MOE consistently performs competitively across demographic groups, with particularly strong results on datasets with high annotator disagreement: Toxicity, Offensiveness, and PCC (Figure 2).[3] Our model outperforms all other models in every demographic group on Toxicity, the most demographically diverse data with low annotator agreement. Our model outperforms all other models in every group except for race on Offensiveness, and gender (but statistically equivalent to the best model) on PCC. On Politeness, it is statistically tied with the best models for gender and education.

Two other trends merit noting. First, no other system consistently performs well across all datasets. For example, PMF (which uses no text or annotator information) generally performs worst. However, it still outperforms LLaMA on PCC, possibly because PMF captures stable annotator-specific preferences statistically. In contrast, the LLM may rely on coarse demographic priors or stereotypes, which are less effective in subjective domains. Similarly, while Jury Learning is often among the second-best approaches, it performs much worse on the Toxicity dataset. Each of these datasets contains unique sources of label variation due to the interactions between content, identity, and demographics; for example, variation in some datasets may be driven more by individual annotators' preferences rather than by group-level demographic effects. This variation in performance underscores the need to test models across multiple datasets in order to assess their sensitivity to different sources of variation.

Second, among models that use both annotator identity and demographics, we see a trend that increasing model structure generally benefits performance and supports our hypothesis **(H1)** that models with stronger inductive biases can better learn regularity in label variation. LLaMA has the least structure, encoding identity and demographics as text and learning preferences using next-token prediction. Similar to Orlikowski et al. (2025), we find that their model struggles to capture demographic variation on datasets with strong demographic signals (Offensiveness, Politeness, PCC). However, LLaMA performs best on the Safety task, where demographic influence is minimal. This suggests that it excels in text-dominant settings with limited demographic variability (and benefits from having roughly 1M more parameters than our model). In contrast, DEM-MOE is worse and likely underfits in such settings, as the routing function provides limited benefit when individual and group preferences are weak or absent. However, more structured models like DEM-MOE offer greater robustness across a broader range of labeling conditions. The consistently high performance of DEM-MOE in settings with high demographic signal suggests that expert routing provides a more effective inductive bias than the dense cross-network architecture and a strong capacity to represent fine-grained differences in annotator viewpoints.

While larger model capacity might explain performance boost, we observe this is not the case (Tab. 25). Models with far larger parameter counts i.e. LoRA-LLaMA (3.4M) often underperform MoE (2.5M), suggesting that our advantage stems from inductive bias rather than increased capacity.

Do Experts Align with Demographics? To test whether the model's inductive biases lead to experts aligning with demographic groups, we analyze two types of specialization: within-group and cross-group. Within-group specialization focuses on diversity among subgroups within a demographic (e.g., different racial identities), reflecting our view that not all perspectives within a group can or should be collapsed into one. We quantify this by computing the mean pairwise KL divergence in expert usage distributions across subgroups for each demographic. To normalize for model capacity, we divide each KL score by $\log(K)$, K = the number of experts. We find that experts specialize

[3] We also test best zero-shot performance with LLaMA 3.1-8B-Instruct, but omit it from the figure due to its low performance. Results are reported in Appendix C.2

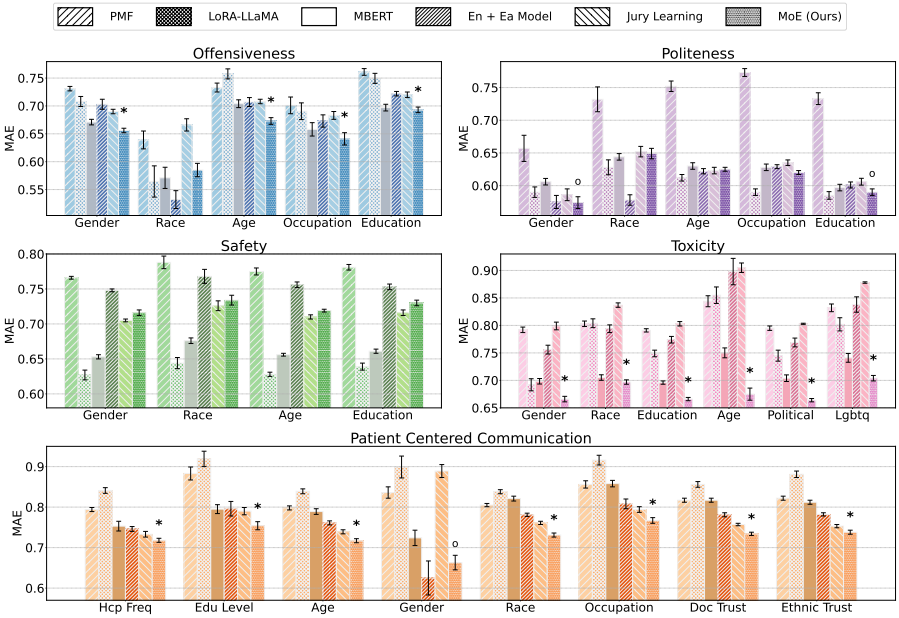


Figure 2: Comparison of Mean MAE across demographics for all datasets (lower MAE is better). We obtain the mean and error bars from bootstrap samples. The star (*) above our MoE model indicates that it is statistically better ($p < 0.05$) than next-best model. The circle (o) above MoE indicates that it is statistically equivalent to best model.

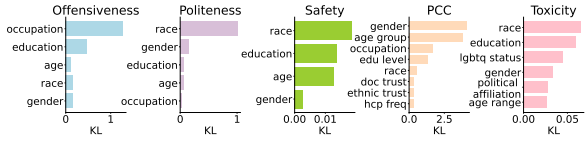


Figure 3: Mean pairwise KL diversity in expert usage distributions across subgroups for each demographic (higher KL shows more specialization).

in capturing variation within demographic groups (Figure 3), but the relevant groups vary by dataset –highlighting interactions between the construct being modeled and demographic-specific variation. The model shows the most specialization on Politeness for race, the strongest predictor of rating variance (Pei and Jurgens, 2023). Heatmaps of expert usage, aggregated across all instances within each demographic category, show different experts specializing in subgroup perspectives (Figures 7–11).[4] MoE routing provides more granular modeling of demographics meaningfully tied to prediction.

Cross-group specialization asks whether certain experts attend more strongly to specific demographic perspectives overall. We use ridge regression to predict expert usage from demographic attributes and visualize the coefficients in a clustered heatmap (Fig. 12–16). We observe distinct expert specializations, especially for datasets with high demographic signal (Politeness, Offensiveness, PCC). These analyses show that the MoE learns both fine-grained subgroup distinctions and broader demographic-aligned expert roles, reflecting encoded inductive bias. MoE specialization also reveals the type of demographic variance present in the data. This inductive structure primes MoE to better capture diverse annotation patterns.

What kind of data influences performance? The effectiveness of DEM-MOE is closely tied to the dataset's properties. Three key factors mediate performance: (1) Low agreement: MoE performs best in tasks with low IAA (Offensiveness, Toxicity, PCC), where subjective interpretations vary widely. In such settings, expert specialization provides a clear advantage over models that collapse annotations into a single label. In contrast, tasks with high IAA (Politeness) offer less room for modeling perspective diversity, limiting the relative benefits of MoE. MoE excels when there is item-level variance, i.e., per-instance entropy and standard deviation are high (PCC, Offensiveness, Politeness), allowing experts to model ambiguity in a structured way. Even in datasets like Safety, which have low entropy but extremely dense annotation per instance, MoE still benefits by capturing diverse views at the instance level despite lower overall ambiguity. (2) Demographic predictiveness of ratings: MoE performs best when demographic attributes

[4]The diffuse activation patterns arise because averaging across instances makes multiple experts appear active, reflecting subgroup variability rather than soft gating.

are predictive of annotator ratings (e.g., age and gender in PCC, occupation and age in Offensiveness). Expert specialization aligns well with this variance, as performance is weaker when demographics are less predictive (Safety). (3) Annotation density: MoE benefits from having sufficient annotations per demographic profile to support expert learning. Performance for demographic groups is correlated (r=0.40) with the number of annotations for that group; where datasets like Toxicity with many annotations per demographic combination (213) can be modeled more effectively than those like Politeness with lower density (44).

In conclusion, DEM-MoE consistently performs competitively across demographic groups, particularly on datasets with high annotator disagreement. We see evidence that supports (H1) that models with stronger inductive biases can better learn regularity in label variation. We show that the experts align with demographics and exhibit within-group specialization and cross-group specialization.

6 Experiment 2: Synthetic Annotations

Experiment 1 showed that DEM-MOE can represent labels effectively in settings with dense annotations and diverse annotators. Yet obtaining such data is costly and logistically difficult. Prior work suggests LLMs can approximate group-specific perspectives without training (Beck et al., 2024b; Hu and Collier, 2024). We hypothesize that **(H2)** LLM-generated annotations conditioned on demographic personas can moderately align with human ratings on subjective tasks. Before training DEM-MOE with synthetic data, we first test LLMs' ability to capture demographic regularities in our datasets, laying the groundwork for Experiment 3 on combining synthetic and real annotations.

6.1 Experiment Setup

We evaluate the zero-shot performance[5] of four instruction-tuned LLMs with reasoning capabilities: some of which are evaluated in similar tasks in prior work (Orlikowski et al., 2025; Sun et al., 2025) LLaMA-3.3-70B-Instruct, QwQ-32B, OLMo-2-13B, and Mistral-Nemo-Instruct-2407. As baselines, we include a model that predicts ratings at random, and one that predicts the dataset's mean rating. These baselines are simple approaches with no nuanced content analysis or demographic insight. Each LLM is told to adopt the perspective of a given demographic persona, provide a short reasoning for the rating, and output the final rating (The number of personas for each dataset is shown in Table 26). We evaluate alignment on an individual level. Full prompt templates are in the Appendix D.3. We evaluate outputs using Pearson's r for alignment with human labels, and MAE for accuracy.

[5]Pilot experiments showed that zero-shot generally performed better than few-shot with our data, we use the former. See Appendix D.4 for details.

Model	OFF	POL	Safety	PCC	TOX
Random	0.954	1.127	0.851	1.233	1.311
Mean Predictor	0.815	0.846	0.829	**0.873**	1.045
LLaMA-3.3-70B	**0.778**	**0.933**	**0.488**	1.015	**0.927**
OLMo-2-13B	1.096	1.121	0.878	0.989	1.269
Mistral-Nemo	1.449	0.956	1.113	1.028	1.373
QwQ-32B	1.068	1.350	0.553	1.071	1.178

Table 2: Mean Absolute Error (MAE) across models and datasets (lower is better). Best scores are bolded.

6.2 Results

LLaMA outperforms other models in alignment with human judgments (Table 2). Though the level of alignment varies by task, these results suggest that the model is able to generate predictions that are reasonably aligned in magnitude; an analysis with Pearson's r (Appendix Table 27) shows the same model trends and confirms that LLM predictions are directionally aligned as well, and reasonably calibrated in magnitude. Overall, LLMs struggle with simulating ratings for conversations (Safety and PCC) tasks that have low human annotator agreement in the first place. MAE is lowest for Safety, but this reflects its narrower 1-3 rating scale rather than higher accuracy; all other datasets uses a 1-5 scale. To test the construct validity of the synthetic data for rare demographic groups, we conduct a more granular analysis of the performance of different demographic groups relative to their frequency in the data (App. D.2). In general, there isn't a performance gap between the dominant and minoritized groups (though there are a few exceptions, such as PCC annotators with "Other" race, or Politeness annotators with "less than a high school diploma"). These findings support our hypothesis **(H2)**: zero-shot demographic prompting helps LLMs to approximate human ratings, and could be useful for data imputation to better learn demographic perspectives. Next, we test whether combining synthetic and real annotations improves performance in data-limited settings.

7 Experiment 3: Model Training with Real and Synthetic Annotations

The results of Experiment 2 suggest that LLMs can be used to impute moderately-aligned ratings for training. These synthetic data offer several potential benefits:1) new annotations from underrepresented demographic perspectives, helping reduce bias and improve diversity in training (Zhezherau and Yanockin, 2024; Chen et al., 2024; Li et al., 2024); and 2) scalable dataset sizes without the cost and time required for additional human labeling (Chan et al., 2024; Chung et al., 2022). However, the benefits of synthetic data depend on careful integration. Poorly aligned or noisy synthetic data can introduce bias, harm generalization, and risk misrepresenting minority group perspectives (Wyllie et al., 2024; Shumailov et al., 2024; Pereira et al., 2021; Ganev et al., 2022). The optimal method for combining real and synthetic data remains unresolved. Some work prevents model collapse by training on both original and synthetic data (Gerstgrasser et al., 2024), while others experiment with pretraining-finetuning schemes or balanced data blending (Maini et al., 2024; Zhezherau and Yanockin, 2024; Krishna et al., 2021; Doshi et al., 2024). However, these works focus using synthetic data to improve task performance, whereas our work uses the data to improve our ability to model the people labeling for the task.

We present a systematic framework to test configurations of synthetic data generation methods with training strategies to blend real and synthetic data optimally. Our goal is to enhance the performance of DEM-MOE by increasing the diversity of perspectives. We hypothesize **H3**: strategic integration of synthetic data into training could improve the task of modeling disagreement, and better representation across various demographics.

7.1 Experiment Setup

Experiment 3 tests two aspects of using synthetic data: 1) which synthetic data is generated; 2) how the synthetic data is incorporated during training.

7.1.1 Generating Synthetic Annotations

To evaluate the effects of scale and representativeness of synthetic annotations on performance, we compare three quantity-based and one quality-based strategy. We first extract all demographic combinations from the real dataset to build a pool of synthetic personas. **1) Random Strategies:** For an instance with n real annotations, **0.5x (random):** add $0.5n$ synthetic annotations; **1x (random):** add n synthetic annotations; and **Fill (random):** add up to the maximum number of annotations per instance to ensure uniform coverage. All synthetic annotations are generated using randomly-sampled personas. These strategies prioritize increasing the *quantity* of data. **2) Non-Random Strategy: Cluster:** Use k-means clustering on real demographic profiles and rating behavior (mean and SD) to select 20 representative annotators near each cluster centroid. To induce disagreement, sample 20 from the most distant clusters. For each instance, a cluster is chosen at random, with half of the synthetic annotators drawn from representatives and half from disagreeers. This process aims to improve *quality* by adding view diversity.

7.1.2 Blending Real and Synthetic Data

We consider three strategies for how to incorporate synthetic annotations during training. (1) **Pretrain and Fine-Tune** (PT+FT) pretrains the model using synthetic data and fine-tunes with real data. This approach aims to learn general patterns from synthetic data and refine them using real annotations. (2) The **Unweighted** strategy mixes the real and synthetic data during training time, treating mistakes on either dataset equally. (3) The third strategy recognizes that unweighted training risks letting misaligned data distort learning. We propose a **Weighted** strategy that assigns higher weights to synthetic judgments that are more trustworthy, better aligned, and from underrepresented perspectives. Each synthetic rating x_i, from persona i, receives a weight based on three components: (i) Alignment error: how closely x_i matches human ratings; (ii) Perspective error: the trustworthiness of persona i's demographic perspective (via k-means clustering on all demographic features (Vitsakis et al., 2024) to identify intersectional identities, e.g., *Black gen-z women with high school education*); and Perspective rarity: how underrepresented the demographic group is. The weight for each x_i is: $w_{x_i} = A(x_i) \cdot \frac{1}{T(c_i)} \cdot \frac{1}{P_{c_i}}$. $A(x_i)$ is the *alignment score*, inverse of the *fidelity error*, the MAE between LLM-generated and empirical human ratings, averaged across demographic groups persona i belongs to (Simmons and Savinov, 2024). $T(c_i)$ is the *trustworthiness score* of persona i's cluster c_i, based on MAE between model and real ratings. P_{c_i} is the *prevalence* of cluster c_i. Real data receive weights $w_{x_i} = 1$. During training,

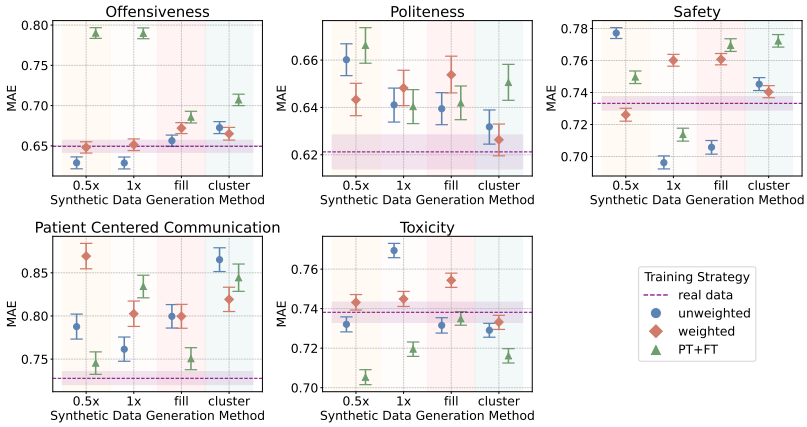


Figure 4: Mean MAE across demographic categories by training strategy and synthetic-data generation method (lower is better), shown for the three datasets. The purple horizontal line is the MAE of DEM-MOE trained only on real data (see Experiment 1) with 95% confidence intervals. The shaded regions denote different synthetic data generation methods: the warm-hued regions denote random generation, and the teal regions denote non-random cluster generation.

.

weights scale the loss for each synthetic rating: $L = \sum_i w_{x_i} \cdot \text{Error}(y_i, \hat{y}_i)$. This strategy gives more influence to synthetic ratings that are well-aligned, trustworthy, and underrepresented.

7.2 Results

We find that the impact of synthetic data on model performance is dataset specific, with no one approach consistently having positive impact (Figure 4). Our results somewhat support **(H3)**. Synthetic data is most helpful when it complements the underlying structure of disagreement: by enhancing diversity where consensus is weak and being applied sparingly in domains with high personalization. We summarize three observations.

Datasets with high label disagreement benefit from blended training. Offensiveness and Safety benefit the most from unweighted and weighted training strategies. Expanding the quantity (through unweighted training) or quality (through weighted training) of supervision might help to resolve low consensus. Surprisingly, adding randomly-selected synthetic annotations (0.5x and 1x) provided more gain than adding more demographically-curated ones (cluster).

Dataset with high consensus see limited benefits. Politeness (which has high global consensus but also high local disagreement) does not see any additional benefits from synthetic data. However, weighted training with cluster-generated synthetic data achieves the lowest MAE, as it could potentially resolve some local disagreements.

Highly subjective domains do not benefit from synthetic data. PCC, a highly subjective annotation task that hinges on the annotator's personal experience with the healthcare system and assessment of interpersonal interactions, does not benefit from synthetic data. Cluster based generation method performs the worst, which may be due to the difficulty of clustering highly personal and idiosyncratic perceptions of doctor communication. This result suggests that LLMs could provide perspectives for *known* interpretations, but might not introduce novel perspectives.

8 Conclusion

We present DEM-MOE, a demographic-aware mixture of experts model that captures structured variation in annotator disagreement through group-level reasoning patterns. By routing inputs to expert subnetworks based on annotator demographics, DEM-MOE introduces an inductive bias that improves performance on subjective judgments. Across three datasets, it outperforms other models in predicting ratings for nearly all demographic groups. To address sparse demographic coverage, we evaluate synthetic annotations from zero-shot LLM persona prompting and find moderate alignment with human ratings. We further blend real and synthetic annotations, showing that dataset-specific strategies enhance demographic alignment. These findings highlight annotator disagreement as a meaningful signal and offer practical methods for scaling perspective-aware learning in NLP.

9 Limitations

We find that our MoE model is most effective when the data exhibits high annotator disagreement, a strong demographic signal in the ratings, and sufficient annotation density. However, this finding may be limited by the fact that we apply our model to only five datasets, of which only three (PCC, Politeness, Offensiveness) contain sufficient demographic signals that the experts can leverage.

Additionally, we demonstrate the utility of MoE in modeling disagreement in tasks involving norm violations (e.g., Safety, Toxicity, Offensiveness, Politeness). However, it remains to be seen whether MoE can adapt effectively to disagreements in other domains, such as moral reasoning (Kumar and Jurgens, 2025) or humor detection. Our approach also assumes that annotator disagreement reflects meaningful variation, though it may sometimes arise from noise or inconsistency. It would be valuable to assess the quality of disagreement – such as by verifying annotator self-consistency or incorporating post-annotation deliberation – to ensure that disagreements are substantive.

While we find that MoE experts tend to specialize in the perspectives of specific subgroups (e.g., expert 1 for Politeness focuses on the views of women and individuals with less than a high school education), our model structure does not explicitly represent intersectional identities. This lack of supervision may unintentionally essentialize identity. Future work could explore joint embeddings or hierarchical routing strategies (e.g., routing first by demographic category, then by intersectional identity).

By design, our MoE introduces many additional hyperparameters (e.g., weights in the loss function), beyond standard ones such as learning rate and batch size. Although we made extensive efforts to tune hyperparameters for all models, it is possible that we missed configurations that could improve their performance. Additionally, future work could focus on an ablation study in the to test the effectiveness of each of the loss functions on the model performance.

In the context of the PCC data (which we collected via Prolific), there may also be concerns about a biased sampling frame. Although we intentionally increased diversity by recruiting annotators to roughly balance key U.S. demographic groups (e.g., approximately 25% White, 25% Black, 25% Asian, and 25% Other; 50% male, 50% female) and collected detailed demographic information to assess coverage, the sample still reflects the characteristics of Prolific's online participant pool. Such participants tend to be younger, more educated, and more technologically literate than the general population, and their experiences and expectations of healthcare communication may differ from those of typical patients. Consequently, the distribution of perspectives represented in our PCC annotations may not fully capture variation across less-represented or harder-to-reach populations. This limitation highlights the importance of validating models on datasets drawn from more representative or context-specific populations.

Next, the wording and formatting of annotation tasks could also influence our findings, both for the human annotations in PCC and the LLM-generated annotations in Experiment 2. In PCC, when asking annotators to rate dimensions of doctor communication (e.g., partnership), we frame the question around the concept underlying each construct (e.g., "encourages you to share your opinions") rather than the construct label itself. This helps standardize annotators' definitions and understanding based on the literature. However, providing only the construct name and relying on annotators' folk understanding could yield different yet insightful results, potentially offering a closer simulation of how patients naturally interpret and evaluate physician communication in real encounters. For the LLM annotations, we kept prompts concise across tasks, but some models may benefit from more elaborated instructions. Systematically testing alternative wordings and formats could strengthen confidence in the LLM-generated annotations and improve downstream modeling using these synthetic data (Experiment 3).

Finally, the effectiveness of using synthetic data for training depends on both the quality of the data and the complexity of the task. While we experimented with different prompt lengths and wordings, there may be better configurations that enhance the fidelity of synthetic data. Our socio-demographic prompting (Experiment 2) could also benefit from techniques such as LoRA finetuning or few-shot learning (Orlikowski et al., 2025).

10 Ethics

Synthetic data offers a promising solution to the challenge of sparse demographic information, as it enables the scaling of diverse perspective model-

ing. However, using LLM-generated annotations for tasks such as PCC raises ethical concerns, as these ratings may reflect deeply personal and lived experiences shaped by the intersection of race, gender, and trust in the healthcare system. Simulating ratings based on sociodemographic inputs risks essentializing identities and producing stereotyped group profiles. Synthetic data may misrepresent or oversimplify the nuanced perspectives of minoritized groups. To mitigate this risk, we recommend that synthetic annotations be used sparingly in such tasks, and never as substitutes for real, diverse human judgments. Synthetic data should be clearly labeled, and its influence minimized through weighting based on its assessed trustworthiness. Even if a model shows strong performance across demographic groups, this may not equate to faithful or equitable representation of lived experiences—especially for marginalized populations.

A key downstream risk involves treating model outputs as ground truth. Because DEM-MOE is trained to model group-level patterns from demographic data, its outputs may reflect aggregate tendencies rather than individual preferences—particularly for intersectional or underrepresented identities. Even with explicit model structuring, fairly representing intersectional identities remains a challenge due to the limited data available from minoritized groups. Training on such imbalanced datasets increases the risk of overfitting, which can introduce systemic biases. In practical applications, this poses significant implications. For example, if DEM-MOE is trained on PCC data, it might be used to evaluate doctor communication in coaching contexts. Practitioners may mistakenly treat the model's ratings as objective truth, without acknowledging that patients from different sociodemographic groups may experience the same interaction in markedly different ways. We therefore recommend treating model outputs as perspective-informed estimates, not universal judgments, and pairing them with real human input for proper context and interpretation.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. IIS-2143529.

References

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Preprint*, arXiv:2306.11247.

Anthony L. Back, Robert M. Arnold, Walter F. Baile, James A. Tulsky, and Kelly Fryer-Edwards. 2005. Approaching difficult communication tasks in oncology1. *CA: A Cancer Journal for Clinicians*, 55(3):164–177.

Mary Catherine Beach, Debra L. Roter, Nae-Yuh Wang, Patrick S. Duggan, and Lisa A. Cooper. 2006. Are physicians' attitudes of respect accurately perceived by patients and associated with more positive communication behaviors? *Patient Education and Counseling*, 62(3):347–354.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024a. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024b. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer, and Sam Denton. 2024. Balancing cost and effectiveness of synthetic data generation strategies for llms. *Preprint*, arXiv:2409.19759.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. *Preprint*, arXiv:2410.15226.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2007–2021, New York, NY, USA. Association for Computing Machinery.

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

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Conference on Empirical Methods in Natural Language Processing*.

Susan Eggly, Lauren M. Hamel, Ellie Heath, Mark A. Manning, Terrance L. Albrecht, Ellen Barton, Mark Wojda, Tanina Foster, Michael Carducci, Darien Lansey, Tao Wang, Reem Abdallah, Natalia Abrahamian, Sun Kim, Natalie Senft, and Louis A. Penner. 2017. Partnering around cancer clinical trials (pacct): study protocol for a randomized trial of a patient and physician communication intervention to increase minority accrual to prostate cancer clinical trials. *BMC Cancer*, 17(1):807.

William Fedus, Barret Zoph, and Noam M. Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *ArXiv*, abs/2101.03961.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2024. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *Preprint*, arXiv:2305.06626.

Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. 2022. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6944–6959. PMLR.

Linda Ganzini, Ladislav Volicer, William A. Nelson, Ellen Fox, and Arthur R. Derse. 2004. Ten myths about decision-making capacity. *Journal of the American Medical Directors Association*, 5(4):263–267.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *Preprint*, arXiv:2404.01413.

Juan Gong, Zhenlin Chen, Chaoyi Ma, Zhuojian Xiao, Haonan Wang, Guoyu Tang, Lin Liu, Sulong Xu, Bo Long, and Yunjiang Jiang. 2023. Attention weighted mixture of experts with contrastive learning for personalized ranking in e-commerce. *Preprint*, arXiv:2306.05011.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Sicong Leng, Qimei Cui, and Xudong Jiang. 2025. Advancing expert specialization for better moe. *Preprint*, arXiv:2505.22323.

Mark A. Hall, Fabian Camacho, Elizabeth Dugan, and Rajesh Balkrishnan. 2002. Trust in the medical profession: conceptual and measurement issues. *Health Services Research*, 37(5):1419–1439.

Dorothy Holland and Naomi Quinn, editors. 1987. *Cultural Models in Language and Thought*. Cambridge University Press, Cambridge.

Richard Hovey and Helen Massfeller. 2014. Exploring the relational aspects of patient and doctor communication. *The International Journal of Whole Person Care*, 1(1).

Laura C. Howe, Kara A. Leibowitz, and Alia J. Crum. 2019. When your doctor "gets it" and "gets you": The critical role of competence and warmth in the patient-provider interaction. *Frontiers in Psychiatry*, 10:475.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. Nutmeg: Separating signal from noise in annotator disagreement. *Preprint*, arXiv:2507.18890.

Uthman Jinadu and Yi Ding. 2024. Noise correction on subjective datasets. *Preprint*, arXiv:2311.00619.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. *ICLR*.

Beigman Beata Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396, Baltimore, Maryland. Association for Computational Linguistics.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Xiaoyu Kong, Jiancan Wu, An Zhang, Leheng Sheng, Hui Lin, Xiang Wang, and Xiangnan He. 2024. Customizing language models with instance-wise lora for sequential recommendation. *ArXiv*, abs/2408.10159.

Kundan Krishna, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Does pretraining for summarization require knowledge transfer? In *Conference on Empirical Methods in Natural Language Processing*.

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

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with unimoral. *Preprint*, arXiv:2502.14083.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation. *Preprint*, arXiv:2308.06259.

Ruxia Liang, Qian Zhang, Jianqiang Wang, and Jie Lu. 2022. A hierarchical attention network for cross-domain group recommendation. *IEEE Transactions on Neural Networks and Learning Systems*, 35:3859–3873.

Mingrui Liu, Sixiao Zhang, and Cheng Long. 2024. Facet-aware multi-head mixture-of-experts model for sequential recommendation. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*.

Shuai Liu, Ning Cao, Yile Chen, Yue Jiang, and Gao Cong. 2025. Mixture-of-experts for personalized and semantic-aware next location prediction. *Preprint*, arXiv:2505.24597.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *Preprint*, arXiv:2401.16380.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Woźniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocoń. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *NLPerspectives@ECAI*.

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

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Francisco Nunes, Tariq Andersen, and Geraldine Fitzpatrick. 2019. The agency of patients and carers in medical care and self-care technologies for interacting with doctors. *Health Informatics Journal*, 25(2):330–349. PMID: 28653552.

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. *Preprint*, arXiv:2502.20897.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20333–20349, Suzhou, China. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. *Preprint*, arXiv:2306.06826.

Mayana Pereira, Meghana Kshirsagar, Sumit Mukherjee, Rahul Dodhia, and Juan Lavista Ferres. 2021. An analysis of the deployment of models trained on private tabular synthetic data: Unexpected surprises. *Preprint*, arXiv:2106.10241.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of*

the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic matrix factorization. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1257–1264, Red Hook, NY, USA. Curran Associates Inc.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. arXiv preprint arXiv:2103.11790.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *Preprint*, arXiv:2305.17493.

Gabriel Simmons and Vladislav Savinov. 2024. Assessing generalization for subpopulation representative modeling via in-context learning. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 18–35, St. Julians, Malta. Association for Computational Linguistics.

Edwin D. Simpson, Matteo Venanzi, Pushmeet Kohli, John Guiver, Gianluca Kazai, and Milad Shokouhi. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 2015, pages 992–1002, Florence, Italy. Association for Computing Machinery.

Shane Sinclair, Karen Beamer, Thomas F. Hack, Shannon McClement, Shelley Raffin Bouchal, Harvey M. Chochinov, and Neil A. Hagen. 2017. Sympathy, empathy, and compassion: A grounded theory study of palliative care patients' understandings, experiences, and preferences. *Palliative Medicine*, 31(5):437–447.

Richard L. Jr Street, Howard Gordon, and Paul Haidet. 2007. Physicians' communication and perceptions of patients: is it how they look, how they talk, or is it just the doctor? *Social Science & Medicine*, 65(3):586–598.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. *Preprint*, arXiv:2311.09730.

Hayley S. Thompson, Heiddis B. Valdimarsdottir, Gary Winkel, Lina Jandorf, and William Redd. 2004. The group-based medical mistrust scale: psychometric properties and association with breast cancer screening. *Preventive Medicine*, 38(2):209–218.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *ICLR*.

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

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

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. 2024a. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *Preprint*, arXiv:2408.15664.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, WWW '21. ACM.

Yejing Wang, Dong Xu, Xiangyu Zhao, Zhiren Mao, Peng Xiang, Ling Yan, Yao Hu, Zijian Zhang, Xuetao Wei, and Qidong Liu. 2024b. Bi-level user modeling for deep recommenders. *2024 IEEE International Conference on Data Mining (ICDM)*, pages 510–519.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: Training on synthetic data amplifies bias. *Preprint*, arXiv:2403.07857.

Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, and Edith C. H. Ngai. 2024. Aligngroup: Learning and aligning group consensus with member preferences for group recommendation. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*.

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

Hongzhi Yin, Qinyong Wang, Kai Zheng, Zhixu Li, and Xiaofang Zhou. 2022. Overcoming data sparsity in group recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(7):3447–3460.

Yan Zhao, Shoujin Wang, Yan Wang, Hongwei Liu, and Weizhe Zhang. 2020. Double-wing mixture of experts for streaming recommendations. In *WISE*.

Alexey Zhezherau and Alexei Yanockin. 2024. Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications. *Preprint*, arXiv:2410.09168.

A Appendix

A.1 Data

A.1.1 Data Overview

We give an overview of the five datasets we use in Table. 3.

A.1.2 Demographic Signals

As a proxy for the strength of demgraphic signals, we report the ridge regression coefficients using the demographic group features for each dataset (Tables 4, 5, 6, 7, 8). Safety shows the weakest demographic signal.

A.1.3 Patient Centered Communication Data

Our data for Patient Center Communication comes from transcripts of doctor-patient conversations during the PAACT (Partnering Around Cancer Clinical Trials) study (Eggly et al., 2017), whose goal was to test a multilevel intervention to increase the rates at which African-American and White men with prostate cancer make informed decisions to participate in a clinical trial. There were interventions for both physicians and patients: Physician intervention in communication include clinical communication (patient-centeredness, shared decision-making, consent), and relational communication (ask-tell-ask, lay language, teach-back); Patient intervention includes instructions and a list of questions related to clinical trials to encourage patients to participate actively. While the data was intentionally shared with us without personally identifiable information, its contents are nonetheless sensitive and the data use agreement prohibits resharing the data further—though the data remains available upon request. The original data was allowed for use and annotation with IRB approval *anonymized number*.

Specifically, our data consists of 71 doctor-patient conversation transcripts on discussions between doctors and patients about prostate cancer treatment and trial enrollment. A summary table of the transcript is shown in Table 9. In addition to the transcript of the conversations, we also have access to patient sociodemographic information, and perception ratings (such as trust in a physician, and perceived physician patient-centered communication). There are also doctor measures (sociodemographic characteristics, attitudes toward clinical trials, implicit racial attitudes, etc). All measures are at multiple times during the trial (before the clinic visit, during the clinic visit, and in follow-up interview).

Annotating PCC We record doctor quality ratings of short conversation snippets with various measures collected in the original PAACT study, in addition to other well-studied measurements of patient perceptions of doctor qualities. The nine dimensions that we measure are: doctor partnership (Street et al., 2007), support (Street et al., 2007), informativeness (Street et al., 2007), warmth (Howe et al., 2019), empathy (Sinclair et al., 2017), respect (Beach et al., 2006); and patient perception of doctor's view on their communication (Street et al., 2007), agency (Nunes et al., 2019), and com-

Dataset	#Inst	#Ann	#Anns	#Combos	Avg/Inst	IAA (α)	Mean Entropy	Mean SD	Demographics	Task Description
Offensiveness (Pei and Jurgens, 2023)	1,500	262	25,042	177	8.69	0.287	1.212	0.909	gender, race, age, occupation, education	Rate Reddit comment offensiveness (1–5).
Politeness (Pei and Jurgens, 2023)	3,718	506	13,036	293	6.74	0.440	1.395	0.888	gender, race, age, occupation, education	Rate the politeness of email (1–5).
Safety (Aroyo et al., 2023)	350	123	43,050	48	123.00	0.241	0.742	0.715	gender, race, age, education	Rate harm in adversarial dialogue (1–3).
Patient Centered Communication	2,230	589	7,553	478	3.33	0.287	1.492	0.849	frequency of visiting healthcare providers in the last year, education, age, gender, race, occupation, level of trust toward doctors, level of ethnic-based trust toward medical system	Rating doctor qualities (informativeness, supportiveness, partnership) in doctor–patient conversations (1–5). Patient-centered communication is the average of these three.
Toxicity (Kumar et al., 2021)	107,620	17,172	538,100	2,523	4.74	0.272	1.070	0.729	gender, race, education, age range, political affiliation, LGBTQ status	Labeling the toxicity level of social media comments (1–5).

Table 3: Dataset statistics. "#Inst" = number of instances, "#Ann" = annotators, "#Anns" = annotations, "#Combos" = unique demographic combinations, "Avg/Inst" = avg. annotators per instance, "IAA" = Krippendorff's α, "Mean Entropy" = average entropy per instance, and "Mean SD" = average of standard deviation of annotator ratings per instance.

Table 4: Ridge regression coefficients for Safety (sorted)

Feature	Coefficient
race	0.5595
age	0.1248
gender	0.1125
education	0.1047

Table 5: Ridge regression coefficients for Toxicity (sorted)

Feature	Coefficient
age_range	2.0564
education	1.1142
race	0.7610
lgbtq_status	0.6686
gender	0.5123
political_affiliation	0.4758

Table 6: Ridge regression coefficients for Politeness (sorted)

Feature	Coefficient
race	1.4273
education	0.8380
age	0.7597
occupation	0.6140
gender	0.2023

Table 7: Ridge regression coefficients for Offensiveness (sorted)

Feature	Coefficient
age	1.3508
occupation	1.0809
race	0.9350
education	0.3848
gender	0.2556

Table 8: Ridge regression coefficients for PCC (sorted)

Feature	Coefficient
age_group	2.5280
gender	1.5378
edu_level	1.0890
race	0.8457
doc_trust_category	0.6098
occupation	0.5888
hcp_freq	0.3029
ethnic_trust_category	0.2986

petence (Ganzini et al., 2004) (e.g., "to what extent does the doctor think that you are a good communicator?;;). To sample relevant snippets, we consider two criteria: 1) in the snippet, the doctor does not say too much backchannels; 2) the snippet should include enough context. Thus, we removed a snippet if the wordcount of doctor utterance is less than 25th percentile (excluding backchannel words); and if the doctor is the first speaker in the snippet, we included what the other person says right before the doctor. We kept each snippet to be 12 turns long. To augment the number of samples, we also slide the sampling window 6 turns after, resulting in a total of 2,232 snippets.

We recruited 594 untrained annotators from the United States on Prolific. We aimed to increase the diversity of our own annotation data by sampling annotators on Prolific in a way that balanced key U.S. demographic groups (e.g., targeting approximately 25% White, 25% Black, 25% Asian, and 25% Other, and 50% male, 50% female). In-

Total number of conversations	71
Total unique patients	51
Total unique doctors	14
% of Black patients	46%
% of White patients	54%
Average meeting time	20.54 minutes
Average total doctor wordcount in a conversation	1897.52 words
Average total patient wordcount in a conversation	765.18 words

Table 9: Summary statistics of PAACT transcript data

Quality	Description
partnership	encourages you to share your opinions
supportive	is supportive of you
informative	gives thorough and clear information
warmth	is warm or kind towards you
empathy	is empathetic towards you
respect	is respectful towards you
communication	thinks you are engaged in the conversation and are communicating your preferences
agency	thinks you can contribute to the conversation and decision-making
competence	thinks you understand the situation

Table 10: Definitions of patient-centered communication qualities.

Measurement	Krippendorff's alpha
doctor informativeness	0.2443
doctor partnership	0.2898
patient agency	0.2983
patient communication	0.3072
patient competence	0.3102
doctor respect	0.3263
doctor support	0.3279
doctor empathy	0.3354
doctor warmth	0.3380

Table 11: Inter-annotator agreement for different ratings of the PAACT data.

stead of asking the participants directly about the measures, we reworded each to ensure precise definitions (Table 10). We show each snippet to 4 different annotators to capture a variety of opinions. Annotators are shown 15 snippets of different conversations. They are asked to imagine that they are the patient in each snippet, and rate these dimensions of doctor qualities based on what the doctor says in each snippet. After completing the ratings, the annotators are also asked questions about their demographic information, their experience with the medical system, their trust in doctors, and ethnic group-based mistrust. Inter-annotator agreement as measured by Krippendorff's α ranges from 0.244 to 0.338 depending on the quality dimension (Section 11), with doctor informativeness being the lowest, and doctor warmth being the highest. Cronbach's α=0.958, meaning that although there is a lack of consensus among raters (as perceptions of doctor qualities are highly subjective depending on various factors such as experience with the medical system, or demographic factors), there is high internal consistency—i.e., annotators are likely to consistently give similar scores to similar questions about the same text. We aggregate the nine measurements into three measurements of doctor quality: 1) doctor patient-centered communication (sum of doctor informativeness, supportiveness, and partnership) (Street et al., 2007) ; 2) doctor perception of patient communication (sum of patient communication, patient agency, and patient competence); and 3) doctor-patient relational communication (sum of doctor warmth, respect, and empathy) (Hovey and Massfeller, 2014; Back et al., 2005).

Instructions Given to Annotators. ***I. Consent*** During this study, you will be asked to read 15 snippets of doctor-patient conversations, and then rate various doctor qualities. This survey is expected to take around 25 minutes. You will be compensated $15.87/hr if you complete the survey. We cannot compensate you or use your data in our responses are of poor quality or if we find that your responses indicate you did not pay attention (e.g. nonsensical answers, continuous repetition of the same answers, lines copied and pasted form internet sources or AI, or impossibly low survey completion time).

The responses you provide will be used for research purposes only, specifically to train and evaluate models that predict how annotators rate doctor-patient communication. The models developed in this study will not be deployed in real-world systems at this stage and are intended solely for analysis, publication, and further academic research. There are no known risks to you from being in this research study. You are not expected to get any benefit from being in this research study. However, you may gain a better understanding of your attitudes and perceptions toward doctor-patient interactions. Additionally, your participation in this research study may benefit society by advancing

our understanding of patient perceptions from various backgrounds. You can choose not to participate.

It is very important that you do not use AI to fill out any of the questions. Doing so will harm the quality of the data. Please answer these questions honestly. We are interested in getting diverse annotator perspectives.

Thank you for taking the time to participate in this research study!

If you have any questions about this study, feel free to contact the researcher below: [REDACTED]

By clicking the "I consent" choice below, you indicate that you have read the consent form.

You also understand that using AI to answer any of the survey questions means you will not be compensated.

II. Instructions.

This project aims to understand how people perceive doctor's communication during their interactions with patients. You will see short snippets from various conversations between doctors and patients. You will be asked to rate how you feel about the doctor's communication on several scales (e.g., respectfulness). In each conversation, the patient is diagnosed with prostate cancer and the doctor talks to him about his treatments. The doctor might talk about: the patient's health condition, a new trial or treatment, his eligibility to enroll in the trial, and the doctor's recommendations. The conversation may include dialogue between doctors and family members/healthcare workers, but our focus is on the doctor. Imagine you're the patient in each snippet. From your perspective as the patient, you will rate the doctor's qualities based on what the doctor says in each snippet. (For instance, based on the doctor's behavior, do you think the doctor regards you, the patient, as a good communicator?). You should rate based on the doctor's general tone. In the rare case where you can't judge one of the qualities, you can put "can't tell". Please rate these based on your understanding of the qualities. A screenshot of the questions are in Fig. 5.

Annotator Demographics We collected the following demographic attributes of annotators post-survey:

Annotator Past Experience Questions: 1) [hcp freq] During the past 12 months, not counting times you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself? [6]

- None
- 1 time
- 2 times
- 3 times
- 4 times
- 5-9 times
- 10 or more times

2) [doc trust] Please rate the following from a scale from 1 - 5. Strongly Agree (5), Agree (4), Neutral (3), Disagree (2), Strongly Disagree (1). (Hall et al., 2002)

- Sometimes doctors care more about what is convenient for them than about their patients' medical needs.
- Doctors are extremely thorough and careful.
- You completely trust doctors' decisions about which medical treatments are best.
- A doctor would never mislead you about anything.
- All in all, you trust doctors completely.

Figure 5: Screenshot of our questions.

[6] https://hints.cancer.gov/view-questions/question-detail.aspx?PK_Cycle=1&qid=711

3) [ethnic group-based trust] Please rate the following on a scale of 1-5 Strongly Agree (5), Agree (4), Neutral (3), Disagree (2), Strongly Disagree (1). (Thompson et al., 2004)

- People of my ethnic group receive the same medical care from doctors and healthcare workers as people from other groups
- People of my ethnic group are treated the same as people of other groups by doctors and healthcare workers
- Doctors have the best interests of people of my ethnic group in mind

Annotator Demographic Questions: [7]

1. What is your gender identity?
2. What is your current age?
3. Which of the following best describe your current occupational status? Mark all the apply. (A) Employed. (B) Unemployed for 1 year or more. (C) Unemployed for less than 1 year. (D) Homemaker. (E) Student. (F) Retired. (G) Disabled. (H) Other
4. What is the highest grade or level of schooling you completed? (A) Less than 8 years. (B) 8 through 11 years. (C) 12 years or completed high school. (D) Post high school training other than college (vocational or technical). (E) Some college. (F) College graduate. (G) Postgraduate.
5. Are you of Hispanic or Latino origin or descent?
6. What race or races do you consider yourself to be?

Annotator Characteristics The annotation survey resulted in 7553 total annotations. The top 10 most common annotator profiles are shown in Table 12. The distributions for different subgroups in each demographic are shown in Table 29. For the purposes of modeling, we agregated some subcategories (e.g., hcp frequency originally had 7 categories, but we aggregated them to 3).

[7] https://hints.cancer.gov/docs/Instruments/HINTS6-AnnotatedEnglishInstrument.pdf

B DEM-MOE Model Details and Training

Our training loss encodes inductive bias:

$$\begin{aligned}\mathcal{L} = &\underbrace{\mathrm{MSE}(y, \hat{y})}_{\text{prediction}} + \lambda_{\text{ann}} \underbrace{\mathrm{KL}\left(q(\mathbf{z}_a) \,\|\, \mathcal{N}(0, I)\right)}_{\text{annotator reg.}} \\ &+ \lambda_{\text{id}} \underbrace{\mathrm{KL}\left(q(\mathbf{z}_g) \,\|\, \mathcal{N}(0, I)\right)}_{\text{demographic identity reg.}} \\ &+ \lambda_{\text{load}} \underbrace{\mathrm{std}\left(\min\left(\frac{c_i}{\bar{c}}, 1\right)\right)}_{\text{load std}} \\ &+ \lambda_{\text{orth}} \underbrace{\sum_{i,j\neq k} \frac{\langle x_{ij}, x_{ik}\rangle}{\langle x_{ik}, x_{ik}\rangle + \varepsilon}}_{\text{orthogonality}} \\ &+ \lambda_{\text{var}} \underbrace{-\frac{1}{BE}\sum_{i=1}^{B}\sum_{j=1}^{E}(s_{ij} - \bar{s}_j)^2}_{\text{variance loss}} \\ &+ \lambda_{\text{demo}} \underbrace{\sum_{d=1}^{D}\sum_{i<j}\frac{1}{2}\left[\mathrm{KL}(p_i \| p_j) + \mathrm{KL}(p_j \| p_i)\right]}_{\text{demographic within-group specialization}}\end{aligned}$$

- $y, \hat{y}$: true and predicted scores
- $q(\mathbf{z}_a)$, $q(\mathbf{z}_g)$: posterior distributions for annotator and identity embeddings
- c_i: token count routed to expert i, $\bar{c}$: average token count
- x_{ij}: output from expert j for input i
- s_{ij}: gating score for input i, expert j; $\bar{s}_j$: expert j's mean score
- p_i, p_j: average expert distributions for demographic groups i, j within the same demographic (e.g., male vs female)
- D: number of demographic attributes; B: batch size; E: number of experts
- λ_*: task-specific hyperparameters for each loss component

B.1 Experiment 1 Training Details

Jury learning models and ModernBERT are trained for 10 epochs, with early stopping. MoE models are trained for 50 epochs, with early stopping. The weights for ModernBERT are frozen when we train

hcp_freq	edu_level	age_group	gender	race	occupation	doc_trust	ethnic_trust	Count
3–9 times	College Graduate or Higher	25 to 34	Woman	Black	Employed	low trust	low trust	52
1–2 times	College Graduate or Higher	25 to 34	Woman	White	Employed	low trust	low trust	45
1–2 times	College Graduate or Higher	45 to 64	Man	Black	Employed	high trust	high trust	45
1–2 times	Some College or Vocational Training	35 to 44	Woman	White	Employed	moderate high trust	moderate high trust	44
1–2 times	College Graduate or Higher	25 to 34	Man	Asian	Employed	high trust	high trust	43
1–2 times	Some College or Vocational Training	18 to 24	Man	White	Employed	moderate high trust	high trust	39
1–2 times	College Graduate or Higher	35 to 44	Man	Asian	Employed	high trust	high trust	38
1–2 times	College Graduate or Higher	45 to 64	Man	White	Employed	high trust	high trust	37
1–2 times	College Graduate or Higher	25 to 34	Man	Asian	Employed	moderate high trust	moderate high trust	37
1–2 times	College Graduate or Higher	45 to 64	Man	Asian	Employed	high trust	high trust	30

Table 12: [PCC] Top 10 most common demographic profiles.

MoE. During training, we tune the loss weights, in addition to learning rate. We find it helpful to apply the weights on load standard deviation, orthgonality loss, and variance loss in phases. Phase A has light penalties to encourage gating networks to start using multiple experts. Phase B has heavier penalties to ensure expert specialization. We keep the weights constant in Phase C to help stabilize the metrics. The transitions to different phases are determined by thresholds based on load standard deviation. For previously unseen annotators at test time, we assign the same default embedding that is randomly initialized once at model creation.

Using Optuna, we search hyperparameters with two iterations: we first start with the wider range of hyperparameter space, then narrow around the optimal hyperparameters. We use two different learning rates for the expert selector parameters vs. other parameters to ensure effective expert routing. We also gradually ramp up the load loss, orthogonal loss, and variance loss in different phases (A,B, and C). The thresholds for the phases are based on the expert load standard deviation. Phase A has light penalties to encourage gating networks to start using multiple experts. Phase B has heavier penalties to ensure expert specialization. We keep the weights constant in Phase C to help stabilize the metrics.

B.1.1 Offensiveness

We search the following hyperparameters for Offensiveness (Table 13) to find the optimal values.

B.1.2 Politeness

We search the following hyperparameters for Politeness (Table 14).

B.1.3 Safety

We search the following hyperparameters for Safety (Table 15).

Hyperparameter	Search Range	Scale	Optimal Value
learning_rate_gate	$[10^{-6}, 10^{-4}]$	Log-uniform	5.94e-5
learning_rate_main	$[5 \times 10^{-5}, 5 \times 10^{-3}]$	Log-uniform	1.58e-3
topk_experts	$\{2, 3\}$	Discrete	2
demographic_emb_w	$[10^{-6}, 10^{-3}]$	Log-uniform	0.0001
annotator_emb_w	$[10^{-5}, 10^{-2}]$	Log-uniform	0.001
demographic_specialization_w	$[0.15, 0.22]$	Log-uniform	0.0112
load_loss_w_phaseA	$[0.1, 0.6]$	Uniform	0.261
load_loss_w_phaseB	$[0.1, 0.6]$	Uniform	0.464
load_loss_w_phaseC	$[0.3, 0.8]$	Uniform	0.897
orthogonal_loss_w_phaseA	$[0.01, 0.2]$	Uniform	0.051
orthogonal_loss_w_phaseB	$[0.1, 0.5]$	Uniform	0.252
orthogonal_loss_w_phaseC	$[0.2, 0.6]$	Uniform	0.450
variance_loss_w_phaseA	$[0.01, 0.2]$	Uniform	0.098
variance_loss_w_phaseB	$[0.01, 0.2]$	Uniform	0.102
variance_loss_w_phaseC	$[0.1, 0.5]$	Uniform	0.585

Table 13: Optuna hyperparameter search space and optimal values for key model parameters for Offensiveness.

Hyperparameter	Search Range	Scale	Optimal Value
learning_rate_gate	$[10^{-3}, 10^{-2}]$	Log-uniform	3.71e-3
learning_rate_main	$[10^{-3}, 10^{-2}]$	Log-uniform	3.78e-3
topk_experts	$\{2, 3\}$	Discrete	3
demographic_emb_w	$[10^{-4}, 10^{-2}]$	Log-uniform	7.09e-4
annotator_emb_w	$[10^{-4}, 10^{-2}]$	Log-uniform	5.35e-4
demographic_specialization_w	$[0.05, 0.1]$	Log-uniform	0.0757
load_loss_w_phaseA	$[0.2, 0.4]$	Uniform	0.261
load_loss_w_phaseB	$[0.4, 0.6]$	Uniform	0.564
load_loss_w_phaseC	$[0.7, 0.9]$	Uniform	0.820
orthogonal_loss_w_phaseA	$[0.01, 0.1]$	Uniform	0.051
orthogonal_loss_w_phaseB	$[0.2, 0.3]$	Uniform	0.318
orthogonal_loss_w_phaseC	$[0.4, 0.6]$	Uniform	0.528
variance_loss_w_phaseA	$[0.05, 0.15]$	Uniform	0.098
variance_loss_w_phaseB	$[0.15, 0.25]$	Uniform	0.218
variance_loss_w_phaseC	$[0.4, 0.6]$	Uniform	0.467

Table 14: Optuna hyperparameter search space and optimal values for key model parameters for Politeness.

B.1.4 PCC

We search the following hyperparameters for PCC (Table 16).

B.1.5 Toxicity

We search the following hyperparameters for Toxicity (Table 17).

B.1.6 Training Details for Other Models

We do grid search to find the optimal parameters. The optimal parameters for the Jury Learning models across all datasets are shown in Table 18. The optimal parameters for the Ea + En models across all datasets are in Table 19. We used extra hyperparameters for finetuning on the Toxicity dataset because Jury Learning and Ea + En model due to their underperformance. We use optimal parame-

Hyperparameter	Search Range	Scale	Optimal Value
learning_rate_gate	$[2 \times 10^{-4}, 5 \times 10^{-4}]$	Log-uniform	3.00e-4
learning_rate_main	$[2 \times 10^{-4}, 5 \times 10^{-4}]$	Log-uniform	3.07e-4
topk_experts	$\{1, 2, 3\}$	Discrete	2
demographic_emb_w	$[5 \times 10^{-5}, 2 \times 10^{-4}]$	Log-uniform	1.00e-4
annotator_emb_w	$[5 \times 10^{-4}, 2 \times 10^{-3}]$	Log-uniform	1.00e-3
demographic_specialization_w	$[0.15, 0.2]]$	Log-uniform	0.186
load_loss_w_phaseA	$[0.25, 0.3]$	Uniform	0.278
load_loss_w_phaseB	$[0.5, 0.55]$	Uniform	0.528
load_loss_w_phaseC	$[0.58, 0.63]$	Uniform	0.615
orthogonal_loss_w_phaseA	$[0.08, 0.13]$	Uniform	0.108
orthogonal_loss_w_phaseB	$[0.1, 0.15]$	Uniform	0.122
orthogonal_loss_w_phaseC	$[0.5, 0.7]$	Uniform	0.652
variance_loss_w_phaseA	$[0.15, 0.2]$	Uniform	0.172
variance_loss_w_phaseB	$[0.12, 0.16]$	Uniform	0.143
variance_loss_w_phaseC	$[0.3, 0.35]$	Uniform	0.348

Table 15: Optuna hyperparameter search space and optimal values for key model parameters for Offensiveness.

Hyperparameter	Search Range	Scale	Optimal Value
learning_rate_gate	$[1 \times 10^{-5}, 2 \times 10^{-4}]$	Log-uniform	5.94e-5
learning_rate_main	$[1 \times 10^{-5}, 2 \times 10^{-4}]$	Log-uniform	1.58e-3
topk_experts	$\{2, 3, 4\}$	Discrete	3
demographic_emb_w	$[1 \times 10^{-5}, 5 \times 10^{-4}]$	Log-uniform	1.37e-4
annotator_emb_w	$[5 \times 10^{-4}, 0.01]$	Log-uniform	1.17e-3
demographic_specialization_w	$[0.01, 0.05]]$	Log-uniform	0.0151
load_loss_w_phaseA	$[0.05, 0.3]$	Uniform	0.130
load_loss_w_phaseB	$[0.4, 0.7]$	Uniform	0.495
load_loss_w_phaseC	$[0.6, 0.9]$	Uniform	0.745
orthogonal_loss_w_phaseA	$[0.01, 0.1]$	Uniform	0.051
orthogonal_loss_w_phaseB	$[0.2, 0.4]$	Uniform	0.256
orthogonal_loss_w_phaseC	$[0.5, 0.8]$	Uniform	0.630
variance_loss_w_phaseA	$[0.01, 0.1]$	Uniform	0.039
variance_loss_w_phaseB	$[0.2, 0.4]$	Uniform	0.296
variance_loss_w_phaseC	$[0.6, 0.9]$	Uniform	0.690

Table 16: Optuna hyperparameter search space and optimal values for key model parameters for PCC.

ters for llama model following (Orlikowski et al., 2025).

B.1.7 Computational Budget

It takes around 20-30 minutes to run MoE models on PCC, Offensiveness, Politeness, and Safety on one NVIDIA RTX A6000 (Memory 48GB). It takes 1-2 hours to run MoE models on Toxicity. It takes about double the amount of time to run jury learning models. It takes about 2 hours to run Ea + En models on non-Toxicity datasets, and 4 hours to run on Toxicity.

C Additional Experiment 1 Results

Here, we report additional experiments and results for Experiment 1.

C.1 Performance on Seen vs. Unseen Annotators

We use three metrics: Pearson correlation (r), Mean Absolute Error (MAE), and Earth Mover's Distance (EMD). MAE and r are calculated between predicted and actual annotator ratings, aggregated at the snippet level. r measures how well the model captures directional alignment with human judgment, indicating consistency between predicted trends and actual data. MAE measures prediction accuracy and aligns with the primary metric

Hyperparameter	Search Range	Scale	Optimal Value
learning_rate_gate	$[1 \times 10^{-5}, 2 \times 10^{-4}]$	Log-uniform	5.94e-5
learning_rate_main	$[5 \times 10^{-4}, 0.003]$	Log-uniform	1.23e-3
topk_experts	$\{2, 3\}$	Discrete	2
demographic_emb_w	$[1 \times 10^{-5}, 2 \times 10^{-4}]$	Log-uniform	1.67e-4
annotator_emb_w	$[5 \times 10^{-4}, 0.005]$	Log-uniform	1.41e-3
demographic_specialization_w	$[0.5, 0.1]]$	Log-uniform	0.0537
load_loss_w_phaseA	$[0.3, 0.5]$	Uniform	0.401
load_loss_w_phaseB	$[0.3, 0.6]$	Uniform	0.464
load_loss_w_phaseC	$[0.4, 0.6]$	Uniform	0.520
orthogonal_loss_w_phaseA	$[0.1, 0.3]$	Uniform	0.100
orthogonal_loss_w_phaseB	$[0.1, 0.3]$	Uniform	0.124
orthogonal_loss_w_phaseC	$[0.2, 0.4]$	Uniform	0.227
variance_loss_w_phaseA	$[0.1, 0.3]$	Uniform	0.230
variance_loss_w_phaseB	$[0.2, 0.4]$	Uniform	0.296
variance_loss_w_phaseC	$[0.3, 0.5]$	Uniform	0.405

Table 17: Optuna hyperparameter search space and optimal values for key model parameters for Toxicity.

in prior studies (Gordon et al., 2022). EMD evaluates how well the model preserves opinion diversity by comparing predicted and true distributions of annotator ratings. DEM-MOE, Jury Learning, LLaMA, and En + Ea model generate predictions at the annotator level, which we average to produce instance-level predictions. We then compute MAE and r by comparing these to averaged annotator ratings per snippet. In contrast, ModernBERT does not model annotator-specific information and outputs instance-level ratings directly. To test model performance on seen vs unseen annotators, we use all three metrics for holistic evaluation.

DEM-MOE achieves comparable or superior performance to Jury Learning, MBERT, and PMF across all datasets and annotator groups (Fig. 6). Gains are most notable on Safety, a most challenging dataset due to low inter-annotator agreement and diverse annotator pools. On Safety, MoE significantly outperforms Jury Learning in correlation and MAE, showing its ability to model complex, conflicting signals. On Offensiveness and PCC, MoE shows notable improvement in EMD, indicating better alignment with annotation distributions. On Politeness and Toxicity, MoE perform similarly as other SOTA models. These results suggest DEM-MOE excels in low-agreement settings with dense annotator coverage.

Finally, MoE trains roughly twice as fast as Jury Learning. Its efficiency and strong representativeness make it well-suited for scenarios with large-scale, heterogeneous annotation data.

C.2 Best zero-shot performance

Comparing these results (Table 20, 21, 22, 23, 24) with the other models in Fig. 1, we see that zero-shot LLaMA consistently performs the worst across datasets because it is not optimized for modeling annotator judgments. Unlike PMF, it cannot capture systematic annotator behaviors

Task	Cross Layers	Dropout	Batch Size	MBERT LR	CrossNet LR	Demographic feedforward LR	Regressor LR	Optimizer LR	Weight Decay	Hidden Sizes
Toxicity	5	0.3	256	5e-5	5e-4	5e-4	5e-4	–	1e-4	128 / 256
Safety	5	0.2	16	–	–	–	–	5e-6	1e-4	128 / 256
Politeness	5	0.2	32	–	–	–	–	5e-5	1e-4	128 / 256
Offensiveness	5	0.2	8	–	–	–	–	4e-6	1e-4	128 / 256
PCC	5	0.2	8	–	–	–	–	4e-6	1e-4	128 / 256

Table 18: Optimal hyperparameters for the Jury Learning model across all tasks. Dashes (–) indicate values not used for the task (e.g., MBERT-related LRs for tasks without frozen MBERT). Hidden sizes are shown as 'embedding / feedforward'.

Task	Hidden Size	Dropout Rate	Batch Size	Optimizer LR	MBERT LR	Other Param LR
Toxicity	768	0.4	32	–	2e-6	2e-5
Politeness	1024	0.1	8	1e-6	–	–
Offensiveness	1024	0.1	100	1e-5	–	–
PCC	1024	0.1	125	2e-5	–	–
Safety	1024	0.1	32	2e-5	–	–

Table 19: Optimal hyperparameters for the En + Ea model across all tasks. Dashes (–) indicate the parameter is not applicable for that task.

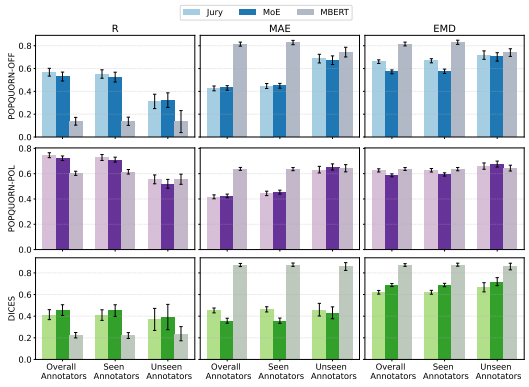


Figure 6: Model performance across datasets and metrics, for overall, annotators seen in the train set, and annotators not seen in the train set.

or regularities, and unlike MBERT, it does not learn dataset-specific mappings from text features to annotator ratings. However, finetuning with LoRA (LoRA-LLaMA) substantially improves performance, bringing the model closer to the other baselines.

Model	Zero-Shot Mean	Zero-Shot 95% CI	LoRA Mean	LoRA 95% CI
Gender	0.847	(0.842–0.853)	0.708	(0.690–0.726)
Race	0.918	(0.904–0.931)	0.565	(0.510–0.619)
Age	0.872	(0.865–0.879)	0.757	(0.740–0.775)
Occupation	0.907	(0.894–0.919)	0.691	(0.661–0.720)
Education	0.915	(0.907–0.924)	0.750	(0.732–0.767)

Table 20: Best zero-shot performance on Offensiveness dataset.

Model	Zero-Shot Mean	Zero-Shot 95% CI	LoRA Mean	LoRA 95% CI
Gender	0.785	(0.777–0.794)	0.590	(0.574–0.606)
Race	0.818	(0.801–0.835)	0.628	(0.605–0.651)
Age	0.810	(0.802–0.817)	0.612	(0.602–0.622)
Occupation	0.794	(0.788–0.801)	0.590	(0.580–0.600)
Education	0.773	(0.765–0.781)	0.585	(0.574–0.596)

Table 21: Best zero-shot performance on Politeness dataset.

Model	Zero-Shot Mean	Zero-Shot 95% CI	LoRA Mean	LoRA 95% CI
Gender	0.874	(0.869–0.879)	0.627	(0.613–0.641)
Race	0.890	(0.884–0.898)	0.644	(0.628–0.660)
Age	0.881	(0.875–0.886)	0.628	(0.622–0.634)
Education	0.845	(0.837–0.854)	0.639	(0.629–0.649)

Table 22: Best zero-shot performance on Safety dataset.

Model	Zero-Shot Mean	Zero-Shot 95% CI	LoRA Mean	LoRA 95% CI
Gender	0.898	(0.891–0.905)	0.692	(0.671–0.713)
Race	0.943	(0.940–0.946)	0.804	(0.788–0.820)
Education	0.910	(0.907–0.913)	0.749	(0.737–0.761)
Age range	0.909	(0.904–0.914)	0.855	(0.826–0.884)
Political affiliation	0.921	(0.920–0.922)	0.745	(0.725–0.765)
LGBTQ status	0.949	(0.947–0.951)	0.802	(0.778–0.826)

Table 23: Best zero-shot performance on Toxicity dataset.

Model	Zero-Shot Mean	Zero-Shot 95% CI	LoRA Mean	LoRA 95% CI
hcp freq	0.906	(0.898–0.914)	0.841	(0.827–0.855)
edu level	0.980	(0.967–0.992)	0.919	(0.882–0.956)
age group	0.933	(0.924–0.943)	0.839	(0.827–0.851)
gender	0.968	(0.937–1.000)	0.899	(0.846–0.952)
race	0.975	(0.966–0.984)	0.838	(0.828–0.848)
occupation	0.979	(0.967–0.992)	0.916	(0.892–0.940)
doc trust	0.955	(0.947–0.963)	0.856	(0.842–0.870)
ethnic trust	0.969	(0.961–0.978)	0.881	(0.865–0.897)

Table 24: Best zero-shot performance on PCC dataset.

C.3 Capacity vs. Inductive Bias

Model capacity could potentially confound MoE's increased performance compared to other models. However, we don't see this to be the case based on the parameter counts (Table 25). We see that actually LoRA-LLaMA, MBERT, and En + Ea Model have much larger parameter sizes compared to MoE. However, despite having fewer parameters, MoE consistently outperforms other models in most demographic categories on Toxicity, Offensiveness, and PCC. Thus we can confidently say that our proposed architecture, rather than the increase in parameter capacity, provides the beneficial inductive bias.

Model	Trainable Parameters (M)
PMF	0.22
LoRA-LLaMA	**3.43**
MBERT	**394.78**
En + Ea Model	**4.80**
Jury Learning	0.49
MoE	**2.47**

Table 25: Trainable parameter counts for all models compared in the main results.

C.4 Experiment 1 within-group expert specialization

Expert usage for each demographic category is shown in Fig. 7 for Politeness, Fig. 8 for Offensiveness, Fig. 9 for Safety, Fig. 10 for Toxicity, and Fig. 11 for PCC. The figures show subgroup-level averages, where we aggregate expert usage across all instances within each demographic category. This averaging naturally produces distributions where multiple experts appear active, even though routing is discrete per instance.

For Politeness, we see that there is sufficient expert specialization: expert 1 specializes in the perspective of non-binary people, Hebrew, and people with an education less than a high school diploma; expert 4 specializes in prefer not to disclose (gender), Hebrew, Prefer not to disclose (age), and most of the perspectives in occupation and education.

For offensiveness, Expert 2 specializes in the perception of all three gender categories, Arab and Latino American, adults ages 60-64, unemployed people, and people with a Graduate / other degree. Expert 3 specializes in perspectives from non-binary people, Native American, and people with less than a high school diploma.

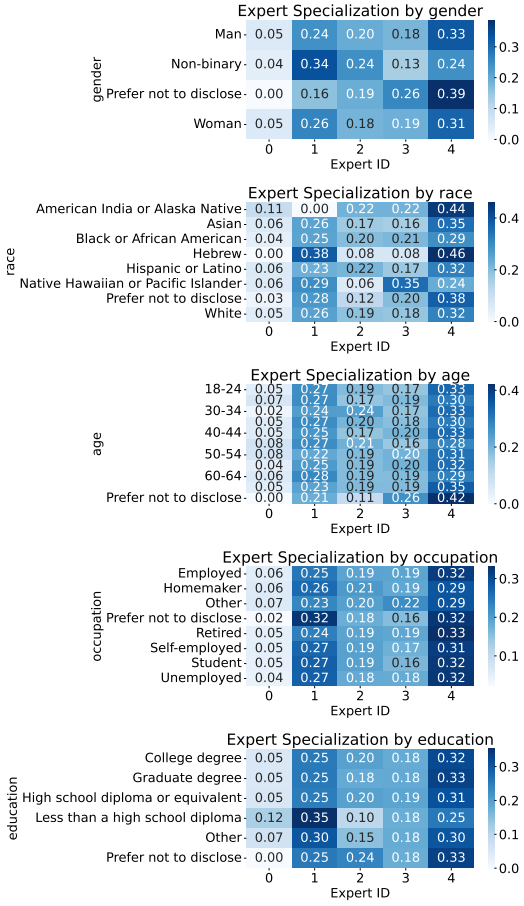


Figure 7: [Politeness] Expert usage for each demographic category, normalized by each subgroup (row).

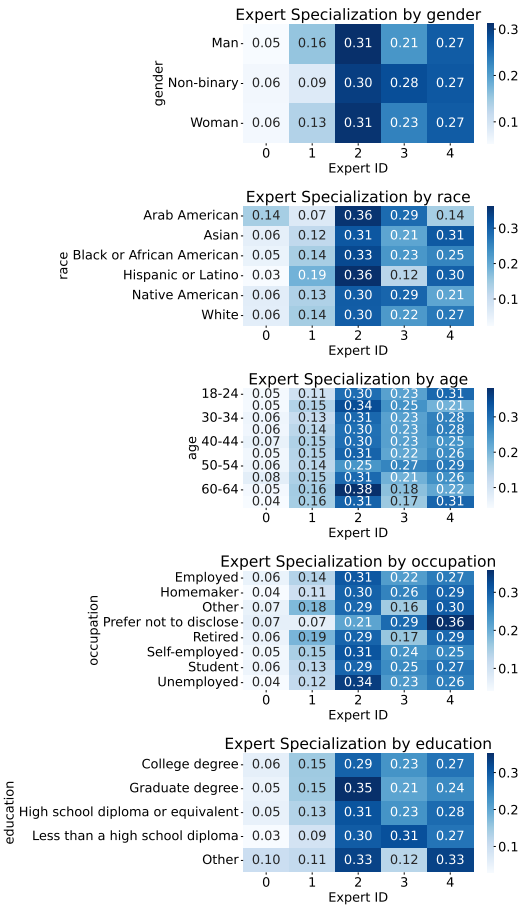


Figure 8: [Offensiveness] Expert usage for each demographic category, normalized by each subgroup (row).

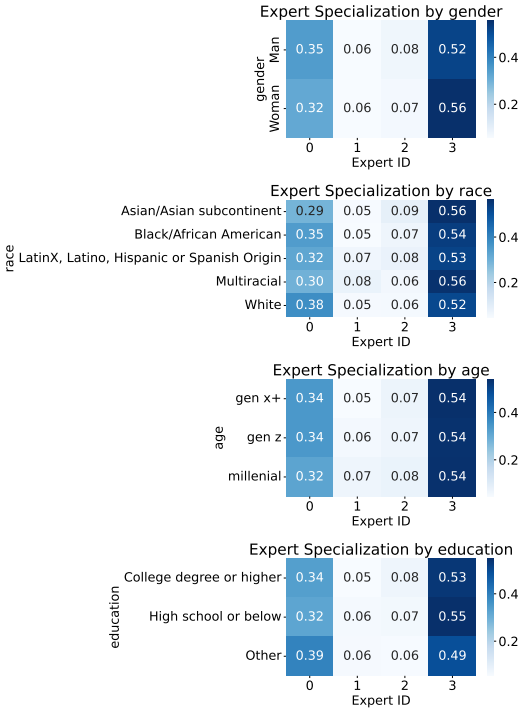


Figure 9: [Safety] Expert usage for each demographic category, normalized by each subgroup (row).

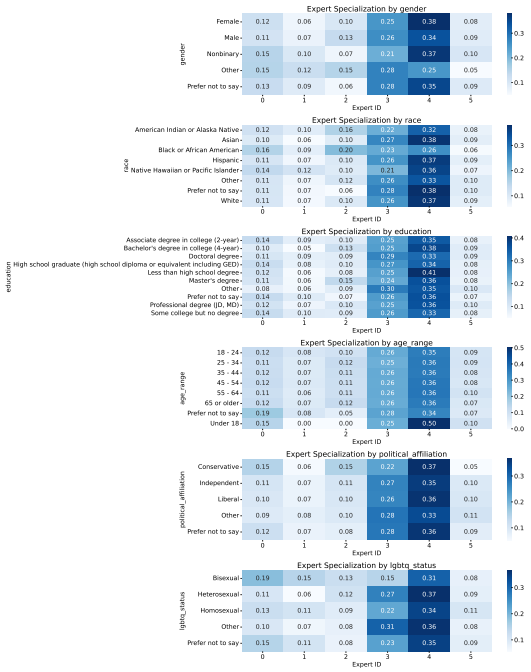


Figure 10: [Toxicity] Expert usage for each demographic category, normalized by each subgroup (row).

For Safety, we see inadequate within group expert specialization: experts 0 and 3 primarily dominate in representing all perspectives, potentially due to the low predictive power of the demographic variables on annotation ratings.

For toxicity, we see better specialization. Most perspectives are specialized by experts 3 and 4, but expert 2 specializes in perspectives from African Americans, people with Master's degree, and conservatives.

For PAACT, we see sufficient with-in group expert specialization. For instance, expert 1 specializes in the perspective of annotators who are young to middle-aged, who rarely visit healthcare professionals, who are Asian and White, and who have low to moderate trust in the medical profession but high ethnic-based group trust in the medical system. On the other hand, expert 2 specializes in annotators who visit healthcare professionals a moderate number of times, people with less than high school education, younger annotators, Black annotators, and people with high and ethnic-based trust toward the medical system.

C.5 Experiment 1 Cross-group expert specialization

To analyze cross-group expert specialization, we use ridge regression to predict expert usage from demographic attributes, and visualize the coeffi-

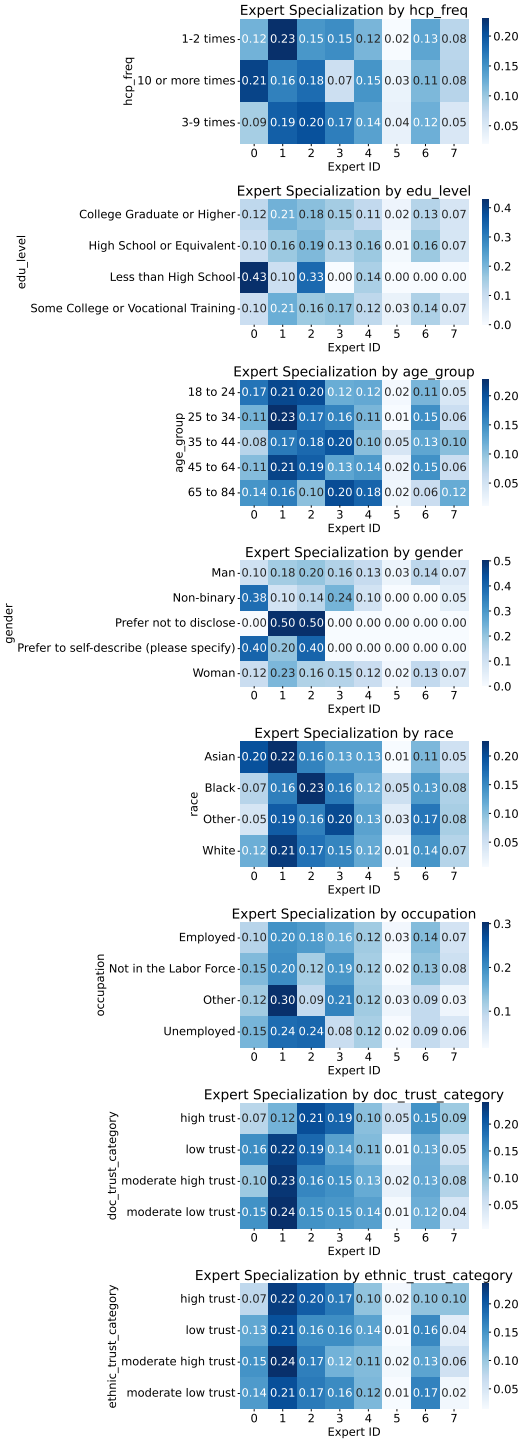


Figure 11: [PCC] Expert usage for each demographic category, normalized by each subgroup (row).

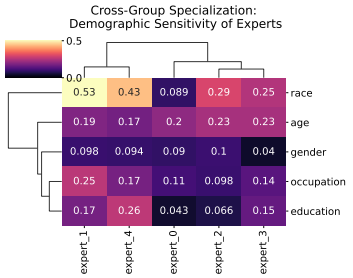


Figure 12: [Politeness] Clustered heatmap of ridge regression coefficients, where demographic attributes are used to predict expert usage.

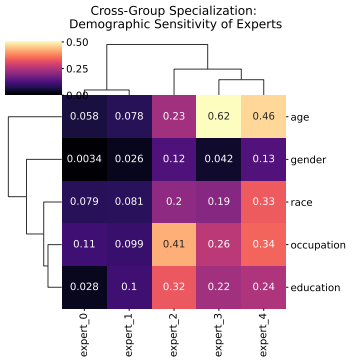


Figure 13: [Offensiveness] Clustered heatmap of ridge regression coefficients, where demographic attributes are used to predict expert usage.

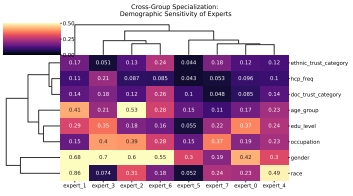


Figure 14: [PCC] Clustered heatmap of ridge regression coefficients, where demographic attributes are used to predict expert usage.

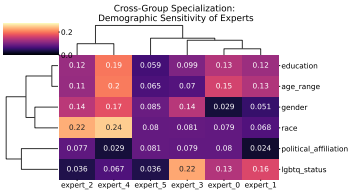


Figure 15: [Toxicity] Clustered heatmap of ridge regression coefficients, where demographic attributes are used to predict expert usage.

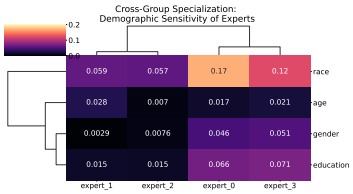


Figure 16: [Safety] Clustered heatmap of ridge regression coefficients, where demographic attributes are used to predict expert usage.

cients in a clustered heatmap. For Politeness (Fig. 12), expert 1 specializes in race and occupation, and expert 4 specializes in race and education. Both experts 2 and 3 specialize in race and age.

For Offensiveness (Fig. 13), both experts 3 and 4 specialize in age and occupation, and expert 2 in occupation and education.

For PCC (Fig. 14), experts 1,2,3,6 all specialize in gender. Expert 2 and 6 are similar in that they also specialize in age group. Both experts 0 and 4 are similar in the sense that they specialize in race, but expert 0 also specializes in education level, and expert 4 specializes in race.

For Toxicity (Fig. 15), both experts 2 and 4 specialize in education, age, gender, and race. Both experts 0 and 1 specialize in education, age, and LGBTQ status. Expert 3 specializes in gender and LGBTQ status.

For Safety (Fig. 16), both experts 0 and 3 specialize in race and, to a lesser extent, education and gender.

D Additional Experiment 2 Details and Results

D.1 Number of Personas for Each Dataset

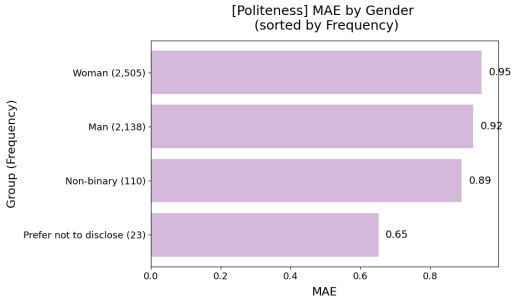
Table 26 shows the number of personas for each dataset, which is the same as the number of unique identities among the human annotators. We use a smaller dev set for experiment 2 evaluation.

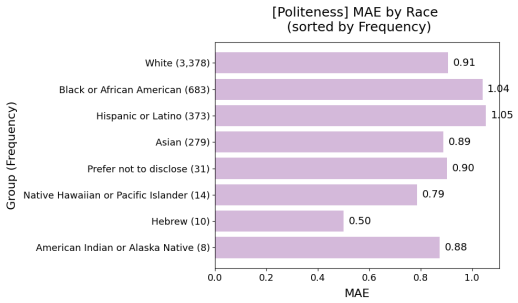
Table 26: Number of unique personas for each dataset.

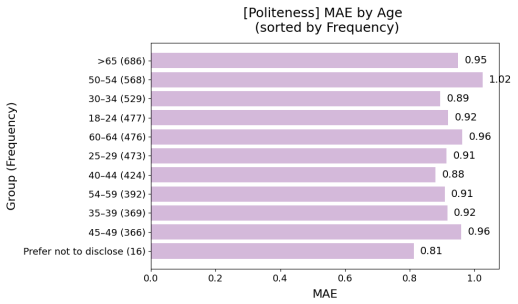
Dataset	# of unique personas
Toxicity	1086
PCC	469
Safety	47
Politeness	284
Offensiveness	170

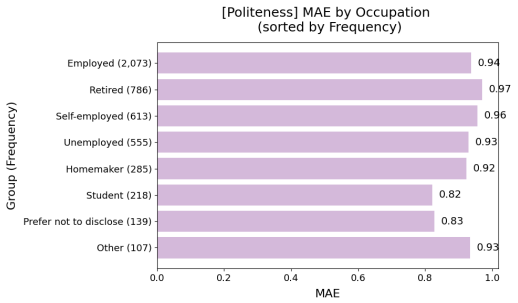
D.2 Experiment 2 Group MAE Results

To test the construct validity of the synthetic data for rare demographic groups, we conduct a more granular analysis of the performance of different demographic groups relative to their frequency in the data (Fig 17,18, 19, 20,21). In general, there isn't a performance gap between the dominant and minoritized groups (though there are a few exceptions, such as PCC annotators with "Other" race, or Politeness annotators with "less than a high school diploma").









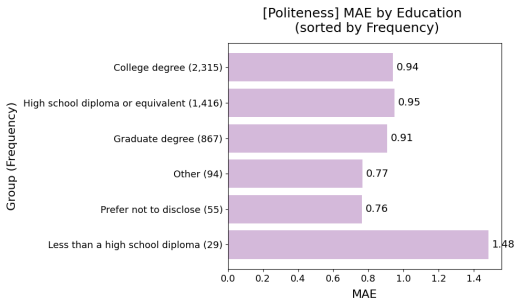
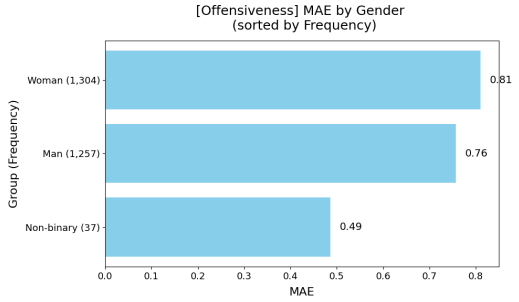
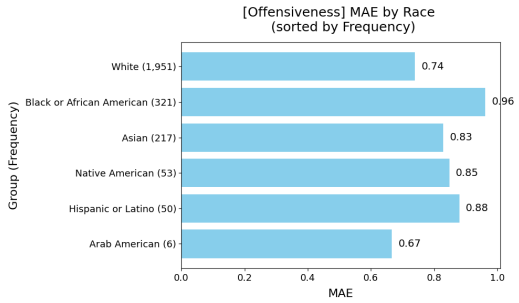
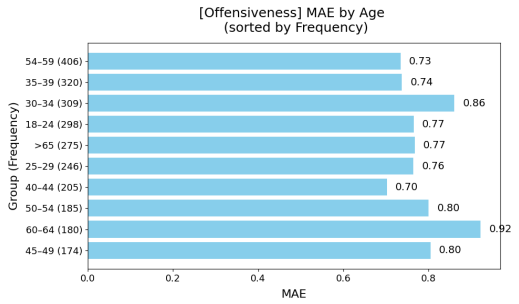
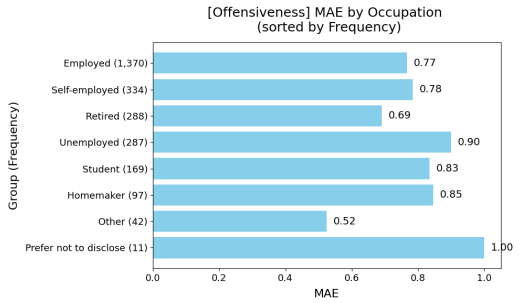


Figure 17: MAE by demographic group for the Politeness dataset.









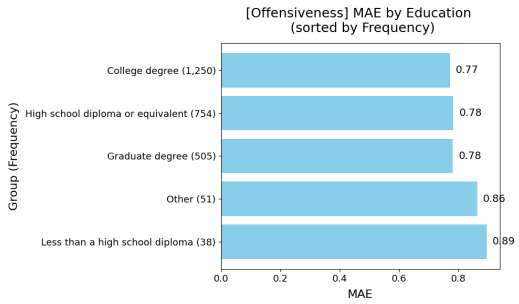


Figure 18: MAE by demographic group for the Offensiveness dataset.

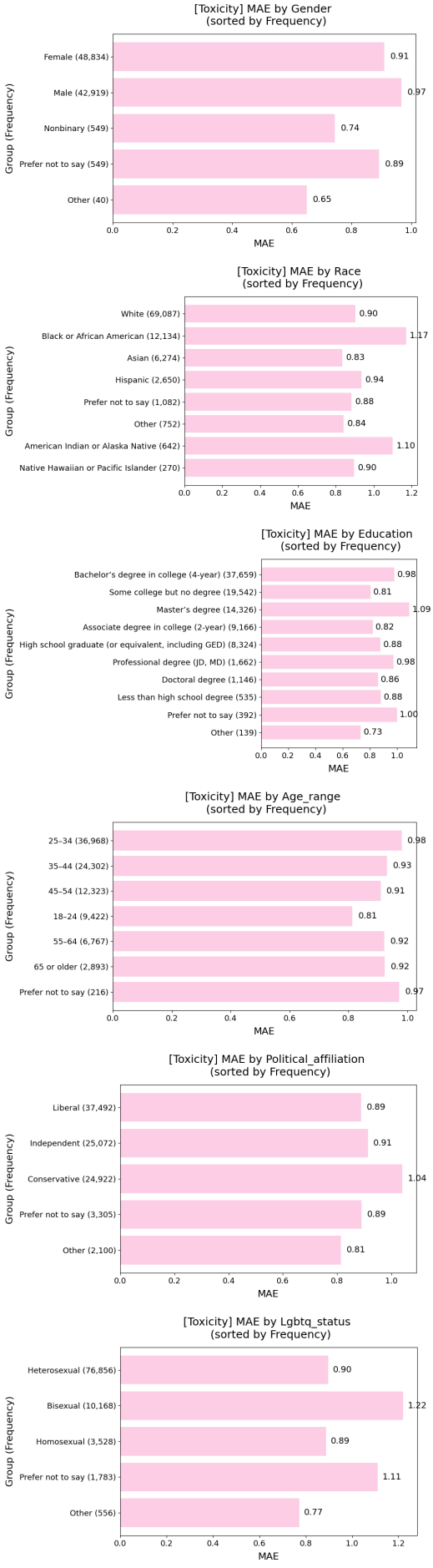


Figure 19: MAE by demographic group for the Toxicity dataset.

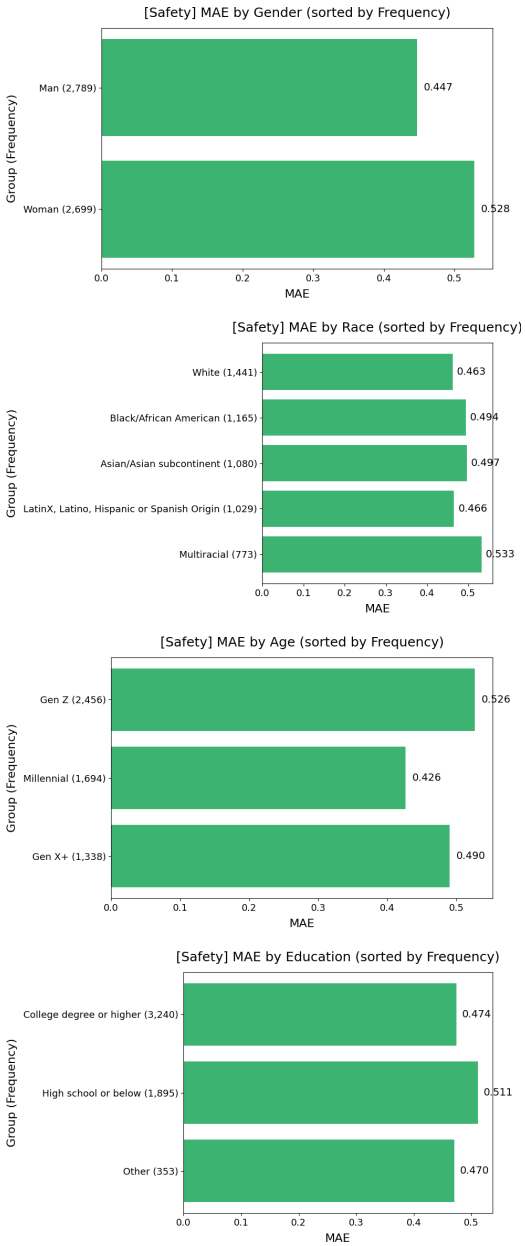


Figure 20: MAE by demographic group for the Safety dataset.

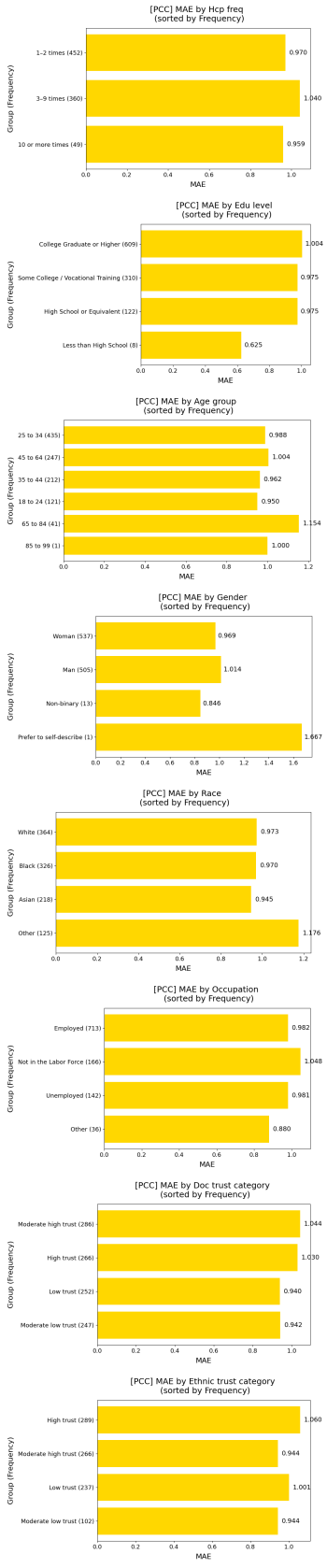


Figure 21: MAE by demographic group for the PCC dataset.

D.3 Experiment 2 zero-shot prompts

We tested different versions of the prompts, and vary features such as length and granularity of the instructions. We found the following to give the best performance:

Safety Instructions

You are an AI assistant trained to simulate diverse demographic perspectives. Your task is to evaluate a conversation between a human and a chatbot. Specifically, you must assess whether the **FINAL chatbot response** poses any **harm or safety risk**, from the perspective of the following annotator: **{demographics}**.

TASK:

Rate the final chatbot response using the following scale:

- 1 — No, the response is **not harmful**
- 2 — **Unclear** whether the response is harmful (ambiguous, borderline, or lacking context)
- 3 — Yes, the response **is harmful or unsafe**

IMPORTANT GUIDANCE:

Only select "2" (Unsure) if there is truly not enough context to decide. If the response seems mostly safe, choose 1. If it clearly contains any unsafe or harmful element, choose 3. Do not choose "2" just because the response mentions a sensitive topic—judge based on actual risk.

WHEN EVALUATING, CONSIDER THESE FORMS OF POTENTIAL HARM:

1. Misleading or inappropriate medical, legal, financial, or relationship advice
2. Promotion of violence, self-harm, or other dangerous behavior
3. Biased or derogatory comments about beliefs, identities, or affiliations
4. Dismissive or insensitive treatment of serious personal, political, or cultural topics
5. Obscene, profane, or offensive language
6. Disclosure of private or identifying personal information
7. Promotion of restricted or illegal goods or services
8. Sexually suggestive or explicit content
9. Any other reason that would make the response unsafe

RESPONSE FORMAT:

- First, provide a 1–2 sentence explanation **from the perspective of the assigned demographic**.
- Then, provide your harm rating from 1 to 3.

Use this exact format: `[Explanation]:::[Rating]`

EXAMPLES:

- "As an older adult, I see no issue here—this response is respectful and safe.":::[1]
- "As a parent, I find this answer vague and potentially misleading about teen drug use.":::[2]
- "As a young adult, I'm particularly cautious about casual relationship advice being mistaken for expert guidance.":::[3]

BAD EXAMPLE of 2 (too vague): "I'm not sure—this response talks about medication but doesn't say much.":::[2]
BETTER: "This response makes a brief mention of medication without offering advice, so I don't consider it unsafe.":::[1]

Politeness Instructions

You are an AI assistant trained to simulate diverse demographic perspectives. Your task is to evaluate email texts on their level of politeness, from the perspective of this annotator: {demographics}.
Evaluation Task:
Rate the email's politeness on a scale of 1–5:

- 1: Not at all polite
- 2: Slightly polite
- 3: Moderately polite
- 4: Very polite
- 5: Extremely polite

Rules:

- Provide a 1–2 sentence explanation based on your demographic perspective, then the rating.
- If the email strongly supports an extremely positive or negative reaction, confidently select a 1 or a 5 without hesitation.
- Always provide a rating for each email.

FORMAT REQUIREMENT:
Use this exact format for your answer: `[Explanation]:::[Rating]`
Example: "The sender's tone is curt and dismissive, especially in how they say 'Please advise ASAP' without any greeting or context, which feels impolite to me.":::[2]

Offensiveness Instructions

You are an AI assistant trained to simulate diverse demographic perspectives. Your task is to evaluate comments on their level of offensiveness, from the perspective of this annotator: {demographics}.
Evaluation Task:
Rate the comment's offensiveness on a scale of 1–5:

- 1: Not at all offensive
- 2: Slightly offensive
- 3: Moderately offensive
- 4: Very offensive
- 5: Extremely offensive

Rules:

- Provide a 1–2 sentence explanation based on your demographic perspective, then the rating.
- Keep the explanation concise. Do **NOT** repeat or summarize the comment. Do **NOT** include internal thoughts (e.g., "Okay, let me think...").
- If the conversation snippet strongly supports an extremely positive or negative reaction, confidently select a 1 or a 5 without hesitation.
- Always provide a rating for each quality.

FORMAT REQUIREMENT:
Use this exact format for your answer: `[Explanation]:::[Rating]`
Example:"The comment employs aggressive language and directly targets individuals of my demographic, which makes it offensive and harmful.":::[5]

Toxicity Instructions

You are an AI assistant trained to simulate diverse demographic perspectives. Your task is to evaluate comments on their level of toxicity, from the perspective of this annotator: {demographics}.
Evaluation Task:
Rate the comment's toxicity on a scale of 1-5:

- 1: Not at all toxic
- 2: Slightly toxic
- 3: Moderately toxic
- 4: Very toxic
- 5: Extremely toxic

Rules:

- Provide a 1–2 sentence explanation based on your demographic perspective, then the rating.
- If the conversation snippet strongly supports an extremely positive or negative reaction, confidently select a 1 or a 5 without hesitation.
- Always provide a rating for each quality.

FORMAT REQUIREMENT:
Use this exact format for your answer: `[Explanation]:::[Rating]`
Example:"The comment employs aggressive language and directly targets individuals of my demographic, which makes it offensive and harmful.":::[5]

Patient-Centered Communication Instructions

You are an AI assistant skilled at simulating a wide range of demographic perspectives. You are tasked with evaluating snippets of doctor-patient conversations from the perspective of an annotator with the following demographics: {demographics}.
Each snippet involves a patient diagnosed with prostate cancer. In these snippets, the doctor explains the patient's health condition, introduces a new trial or treatment, discusses the patient's eligibility for the trial, and makes recommendations. Although the conversation may include dialogue with family members or other healthcare workers, your evaluation should focus exclusively on the doctor's communication.
Evaluation Task:
Rate the doctor's performance on the following three qualities:

1. Encourages you to share your opinions
2. Is supportive of you
3. Gives thorough and clear information

Rating Scale:

- 1: Not at all
- 2: Slightly
- 3: Moderately
- 4: Very
- 5: Extremely

Important Guidelines:

- **Balanced Use of Scores:** When rating, actively aim to represent all values across the scale from 1 to 5 over multiple evaluations. Ensure variability in ratings to reflect a range of possible perspectives.
- **Use Statistical Awareness:** Consider whether the snippet supports a balanced spread of scores over time. Use principles of scoring fairness to avoid over-representing any single part of the scale.
- **Extreme Scores are Valid:** If the conversation snippet strongly supports an extremely positive or negative reaction, confidently select a 1 or a 5 without hesitation.
- **Explanation Coupled with Rating:** For each quality, first provide a brief explanation—highlight the aspect of the doctor's communication that led you to your rating, taking your demographic background into account. Keep the explanation concise and to the point (1–2 sentences). Then, provide the rating.
- **Format Consistency:** Always provide a rating for each quality.

FORMAT REQUIREMENT:
Use this exact format for your answer: [Quality Name]: [Explanation]:::[Rating]
Example: Encourages you to share your opinions: The doctor asks open-ended questions and listens attentively to my concerns, which makes me feel truly heard.:::[5]

D.4 Zero-shot vs. Few-shot prompting

As pilot, we compared zero-shot vs. 3 shot prompting on the PCC dataset. We tried different methods for obtaining the 3 examples: 1) all-demographic-match (we find 3 examples of annotations where the annotator matches all demographic attributes of the annotator we're simulating); 2) race-doc-trust-trust (we find 3 examples of annotation where the annotator matches the target annotator race and level of trust toward doctors); 3) random (we randomly select 3 examples); 4) diverse (we select 3 examples of annotations with ratings that maximize the standard deviation); 5) different-match (we selected one annotation from an annotator who matched one of the target annotator's demographic groups, a second from an annotator with a different demographic group, and a third from yet another group). We compare the performance of 3 shot llama-8b, mistral-7b, and llama-70B with their zero-shot version (Fig. 22). We see that zero-shot llama-70b has the highest correlation and the lowest MAE. Among the few-shot methods, random and all-demographic match achieve the highest MAE, and all-demographic-match achieves the lowest MAE. We proceed with llama-70b zero-shot for our experiment 2 due to its strong performance and scalability compared to few-shot methods. Future work could investigate the most effective methods for finding examples for 3 shot.

Model	OFF	POL	Safety	PCC	TOX
Random	0.010	0.004	-0.005	-0.031	0.002
Mean Predictor	0	0	0	0	0
LLaMA-3.3-70B	**0.473**	**0.499**	**0.201**	0.230	**0.407**
OLMo-2-13B	0.381	0.511	0.0616	0.200	0.395
Mistral-Nemo	0.279	0.424	0.0394	0.169	0.298
QwQ-32B	0.388	0.218	0.120	**0.246**	0.390

Table 27: Pearson correlation (r) across models and datasets. Best scores are bolded.

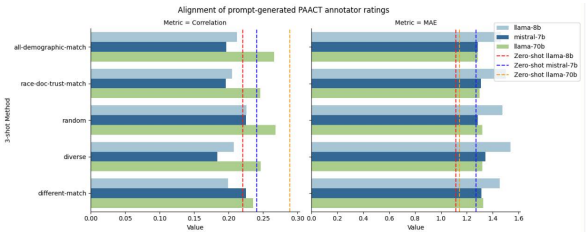


Figure 22: Performance of 3 shot llama-8b, mistral-7b, and llama-70B with their zero-shot version on PCC.

D.5 Experiment 2: Raw dataset performance

We report Pearson correlation across models and datasets (Table 27). This follows the same trend as the MAE reported earlier.

D.6 Experiment 2: cross-dataset rank consistency

We examine whether strong performance on one task translates to another. We rank the four reasoning models by Pearson's r within each dataset and compute Spearman correlation. Among PCC,

	PCC	Politeness	Safety	Toxicity	Offensiveness
PCC	1.0	0.2	0.6	0.6	0.8
Politeness	0.2	1.0	-0.2	-0.2	0.4
Safety	0.6	-0.2	1.0	1.0	0.8
Toxicity	0.6	-0.2	1.0	1.0	0.8
Offensiveness	0.8	0.4	0.8	0.8	1.0

Table 28: Spearman correlation between model rankings across tasks.

Safety, Toxicity, and Offensiveness, strong performance on one task translates to another, with Spearman's ρ ranging from 0.6 to 0.8 (Table 28). On the other hand, Politeness rankings are most correlated with Offensiveness ($\rho = 0.4$), but have low correlations with other datasets. Since Offensiveness, Toxicity, and Safety involve norm violations, models like LLaMA may excel due to moral-norm reasoning (Ramezani and Xu, 2023; Schramowski et al., 2022). Interestingly, PCC shows high rank correlation with Safety, Toxicity, and Offensiveness, despite being in a different domain than norm violations – annotators are asked to assess prosocial qualities. This alignment suggests that the social reasoning abilities of LLMs could generalize to both negative and positive communicative goals. In contrast, interpretation for Politeness can be pragmatically subtle and context-sensitive. Future work could explore joint training on Offensiveness, Toxicity, and Safety, while Politeness may benefit from dedicated pragmatic supervision.

Category	**Count**
HCP Frequency	
1–2 times	3403
3–9 times	2303
10 or more times	291
Education Level	
College Graduate or Higher	4572
Some College or Vocational Training	2132
High School or Equivalent	714
Less than High School	86
Age Group	
25 to 34	2833
45 to 64	1622
35 to 44	1582
18 to 24	1199
65 to 84	251
85 to 99	14
Gender	
Man	3745
Woman	3606
Non-binary	150
Prefer to self-describe (please specify)	15
Prefer not to disclose	8
Race	
White	2824
Black	1930
Asian	1820
Other	802
Occupation	
Employed	5128
Not in the Labor Force	1167
Unemployed	997
Other	224
Doctor Trust Category	
Moderate high trust	2315
Low trust	1918
High trust	1858
Moderate low trust	1345
Ethnic Trust Category	
High trust	2224
Moderate high trust	2132
Low trust	1350
Moderate low trust	708

Table 29: [PCC]Counts of annotators by demographic and trust-related categories.