Demographic Prompting at Scale: When More Attributes Hurt LLM–Human Agreement

Mahammed Kamruzzaman, Shrabon Kumar Das, Gene Louis Kim
Bellini College of AI, Cybersecurity and Computing
University of South Florida
{kamruzzaman1, das157, genekim}@usf.edu

Abstract

We investigate how annotator demographic attributes, supplied as prompt cues, shape the alignment between large language model (LLM) predictions and human annotations across five tasks. Using five open-source LLMs, we systematically vary the number and composition of demographic components in the prompt, spanning every combination from single-attribute through full-attribute configurations. Our experiments reveal three principal findings. First, demographic prompting is highly model- and task-dependent. The baseline is best in 12 of 25 model–task pairs. Among the remaining 13, 11 peak with 1–3 attributes, while two require higher-order combinations. The full-attribute prompt is never the best configuration. Second, the overall magnitude of demographic influence on human annotations does not predict which attributes improve LLM alignment; instead, both the learnability and the directional coherence of each attribute's annotation signal need to be considered jointly. Third, neuron probing reveals that specialized activation correlates with alignment gains only under coherent annotation signals, and that activation volume alone does not imply steerability. Together, these results demonstrate that demographic prompting is not a monolithic intervention: its utility is highly context-dependent, shaped by attribute signal quality, task characteristics, and model architecture.[1]

1 Introduction

LLMs are increasingly used as substitutes for, or complements to, human annotators on subjective NLP tasks such as toxicity detection, sentiment analysis, and offensiveness rating (Gilardi et al., 2023; Törnberg, 2023). Because these tasks reflect annotator subjectivity, a natural question arises: how well can an LLM actually model a given demographic perspective when prompted to do so? A growing body of work shows that LLM outputs are not demographically neutral, predictions tend to align more closely with certain demographic groups in the absence of demographic cues, and explicitly incorporating such cues into the prompt can shift model behaviour in ways that are neither uniform nor always beneficial (Beck et al., 2024; Sun et al., 2025; Alipour et al., 2025; Kamruzzaman et al., 2025; Schäfer et al., 2025).

Despite this progress, the existing literature leaves several important questions unresolved. Most prior studies examine only one or two specific demographic attributes at a time, or compare a no-demographic baseline against a single all-attributes prompt, without exploring the combinatorial space in between (Beck et al., 2024; Sun et al., 2025; Alipour et al., 2025). None systematically chart the incremental trajectory from single-attribute through multi-attribute to full-attribute prompting, nor investigate the signal to noise relationship along this trajectory. To address this, we pose our first research question: **RQ1:** *To what extent do individual versus combined annotator demographic attributes shape LLM–human alignment, which demographic features (or combinations) most significantly affect alignment, and is there a threshold beyond which additional demographic information no longer benefits alignment?*

While several studies report that demographic prompting sometimes helps and sometimes hurts alignment (Gupta et al., 2023; Kamruzzaman et al., 2024), the field lacks a principled account of *why*: what properties of a demographic attribute, at the dataset level, predict whether prompting with it will improve a given model's agreement with human annotators? An attribute may strongly predict variation in human labels yet carry internally opposed subgroup signals that no single persona prompt can resolve. Understanding this requires moving beyond aggregate importance measures to characterise the structural quality of each attribute's an-

[1]Our code is available https://github.com/ShrabonDas/demographic-prompting-at-scale.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 1373–1399
October 24-29, 2026 ©2026 Association for Computational Linguistics

notation signal. We therefore ask: **RQ2:** *To what extent do dataset-level demographic signals predict when demographic prompting improves LLM–human alignment, and do LLM outputs reflect the learnability and directional coherence of those signals?*

Finally, the internal mechanisms by which LLMs process demographic cues remain poorly understood. Neuron-level interpretability methods have been applied to multilingual and cultural knowledge (Ying et al., 2025; Zhao et al., 2024), but no prior work has connected neuron activation patterns to demographic alignment in subjective language understanding annotation tasks. If a demographic cue activates specialized neurons within the model, does that internal engagement translate to better alignment with the prompted group's annotation norms, or can high activation volume coexist with poor steerability? This motivates our third research question: **RQ3:** *To what extent do specialized neuron activations explain the variations in LLM–human alignment across demographic configurations and tasks?*

To answer these questions, we conduct experiments across five subjective language understanding task datasets, each annotated with demographic metadata, on five open-source LLMs. Overall, our work makes the following contributions:

1. **Full combinatorial analysis of demographic prompting.** To the best of our knowledge, we are the first to systematically enumerate every combination of demographic attributes, from single-component through full-attribute configurations, across five tasks and five models. This reveals a strongly model- and task-dependent pattern. The no-demographic baseline is best in 12 of the 25 model–task pairs. Among the 13 pairs where at least one demographic configuration numerically exceeds baseline, 11 achieve their optimum with 1–3 attributes, while two require higher-order combinations. Across all 25 pairs, the full-attribute configuration is never the best-performing configuration.
2. **A three-level framework linking dataset-side signal to LLM alignment.** We show that the magnitude of demographic influence on human annotations is uninformative for predicting alignment, while the learnability of word-demographic interaction patterns is a better but still insufficient predictor. The missing piece is directional coherence: attributes whose subgroups agree on which words signal the task label enable alignment gains, while attributes with opposed subgroup signals invert the relationship between learnability and alignment. This three-level characterisation (magnitude, learnability, coherence) provides a principled diagnostic to decide which attributes to include in demographic prompts.
3. **First application of neuron probing to demographic alignment.** By adapting the specialised-neuron methodology of Ying et al. (2025) to the demographic prompting setting, we show that neuron activation proportion correlates with alignment only when the underlying annotation signal is directionally coherent. We also document the DeepSeek high-volume paradox, where the model activating the most specialised neurons is the least demographically steerable, demonstrating that activation quantity does not dictate alignment utility.

2 Related Work

Sociodemographic prompting. The idea of conditioning LLM predictions on annotator demographics has been explored in several recent studies, though each addresses only a subset of the design space we consider. Beck et al. (2024) examine five attributes across seven datasets and find that single-attribute prompts alter predictions while the all-attributes prompt causes the most label changes, but no intermediate combinations are tested. Sun et al. (2025) find that models align most with White annotators on politeness and offensiveness, and that demographic cues rarely improve alignment. Similarly, Schäfer et al. (2025) report that default model behaviour already leans toward White and younger viewpoints. Alipour et al. (2025) show that confounders such as document difficulty and annotator sensitivity explain more variance than demographics alone, using logistic regression to model alignment as a function of both which is methodologically closest to our RQ2, though we focus on the structural properties of the demographic signal itself. Personas can also trigger biases, mirroring gender-emotion stereotypes (Plaza-del Arco et al., 2024) and producing regional disparities in emotion attribution (Kamruzzaman et al., 2025).

Dataset annotator demographic analysis. A complementary line of research examines how annotator demographics shape the annotation data itself, independent of LLMs. Hovy et al. (2013) present methods to model annotator reliability,

while Davani et al. (2022) argue that aggregation can suppress minority viewpoints. Orlikowski et al. (2023) demonstrate that individual annotator differences often exceed demographic-group differences, cautioning against ecological fallacy in demographic alignment claims. Our SHAP and LinearSVC analyses (Section 4.2) operate at the dataset level, which are connected to LLM behavior through rank correlation and Fisher-averaged coherence, bridging the gap between annotation analysis and prompting outcomes.

Mechanistic interpretability. Our neuron probing approach builds on work linking neuron activations to knowledge use in LLMs. Prior studies identify language-selective neurons that improve multilingual accuracy (Tang et al., 2024; Zhao et al., 2024) and show that activations encode recoverable parametric knowledge (Hong et al., 2025). Most closely aligned to our paper, Ying et al. (2025) show that activating more specialized neurons corresponds to better cultural knowledge use, while Cao et al. (2025) argue that activation quantity does not necessarily reflect knowledge quality. Internal activations have also been used to steer model behavior (Rimsky et al., 2024) and mitigate bias through neuron editing (Liu et al., 2024; Yu and Ananiadou, 2025). Unlike these approaches, we use neuron activations purely as a diagnostic signal to explain when demographic prompting succeeds or fails.

3 Experimental Setup

3.1 Datasets

We use five datasets where annotator demographic information is recorded alongside their task-specific annotations.

1. **Toxicity Detection:** For toxicity detection, we use the *Diverse Perspectives (DP)* dataset (Kumar et al., 2021). We use eight annotator demographic attributes for this dataset: *age_range, education, gender, is_parent, lgbtq_status, political_affiliation, race, and religion_importance.*
2. **Sentiment Analysis:** Díaz et al.'s (2018) sentiment dataset studies age-related bias in sentiment analysis. The nine demographic attributes in this dataset are: *age, education, employment_status, gender, hispanic_latino, income, living_situation, political_id, and race.*
3. **Politeness:** Pei and Jurgens's (2023) politeness dataset is a subset of the POPQUORN dataset (Pei and Jurgens, 2023), with five annotator demographic attributes: *age, education, gender, occupation, and race.*
4. **Offensiveness:** We use the offensiveness subset of the POPQUORN dataset. This dataset has the same five demographic annotator attributes as the politeness task.
5. **Emotion Attribution:** We use the International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Wallbott, 1994), which has seven demographic attributes: *gender, religion, father's_occupation, mother's_occupation, field_of_study, country, and age.*

All of these demographic attributes are originally included in their respective datasets; we did not create or modify them. See Appendix A for dataset details.

3.2 LLMs

We use five open-source LLMs: LLama-3.2-3B, Mistral-7B, Gemma3-12B, Qwen2.5-7B, and DeepSeek-R1-7B. For model details, see Appendix B.

3.3 Prompts

We prompt the LLMs both with demographic attributes and without (baseline). Prompts follow a common structure, with task-specific variation reflecting the nature of the task and available annotation labels: *"Given the following text: {text}, how would a person of {demographic_attributes} rate this {task}?"*. {demographic_attributes} can be instantiated with any combination of demographic attributes available in the dataset. The {task} component is adjusted depending on the specific task being evaluated. All task-specific prompts are provided in Appendix C. To ensure the robustness of our findings, experiments include two additional paraphrases of the prompting template and report the average results across all prompting templates.

3.4 Model-human Alignment Metrics

We quantify model–human alignment[2] with quadratic-weighted *Cohen's* κ (Cohen, 1968), a chance-corrected agreement coefficient for ordered

[2]Ongoing alignment research debates whether models should represent a broad "average user" or multiple demographic groups (Schwerzmann and Campolo, 2025). We **do not** investigate which alignment scale is normatively preferable; rather, we ask: when demographic attributes are provided in the prompt, to what extent do LLM outputs actually reflect the associated annotation patterns?

rating scales. The quadratic penalty assigns partial credit to near misses and treats distant disagreements as total mismatches, capturing the ordinal structure of our five-point toxicity, sentiment, politeness, and offensiveness labels while allowing direct comparison with human–human reliability standards (Landis and Koch, 1977). Cohen's κ ranges from -1 to $+1$, where $+1$ indicates perfect agreement, 0 agreement no better than chance, and -1 systematic disagreement. We report 95% bootstrap confidence intervals and additionally compute *macro* and *micro* accuracy. ***For the Emotion dataset only, we omit weighted κ as the labels in this dataset have no natural ordinality.*** We use weighted κ as the primary metric for discussing the results, as it largely correlates with macro accuracy.

3.5 Dataset-Side Demographic Signal Analysis

To quantify the demographic signal present in each dataset, we employ **three** complementary analyses that characterize each attribute's signal at increasing levels of structural detail. **First**, to measure the ***magnitude*** of each attribute's influence on human label variation, we train a Logistic Regression classifier on one-hot-encoded annotator demographic attributes and apply SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017), aggregating absolute SHAP values of binary indicator features back to their parent demographic groups (e.g., combining `gender_Female` and `gender_Male` into a single `gender` importance score). This analysis is conducted entirely on the annotation data; no LLM outputs are involved. Full details are in Appendix E. **Second**, to assess the ***learnability*** of each attribute's lexical signal, we train a LinearSVC (LSVC) with explicit *word×demographic interaction* features, computed as the element-wise product:

$$X_{\text{inter},j} = X_{\text{words}} \odot d_j \quad (1)$$

where X_{words} is the TF-IDF matrix of text features and d_j is the demographic indicator vector for category j. This yields a separate learned weight for every (word, demographic category) pair, so the resulting LSVC κ reflects how much discriminative lexical signal each attribute contributes beyond text alone. **Third**, to characterize the ***directional coherence*** of each attribute's signal, we extract the top-200 most influential words per attribute from the fitted LSVC interaction weights and compute Spearman rank correlations between the weight vectors of every pair of demographic categories (i.e., subgroups) within each attribute (e.g., Female vs. Male within `gender`, or Liberal vs. Conservative within `political_affiliation`), summarized via the Fisher z-transformation into a single Fisher-averaged $\bar{\rho}$ per attribute. A positive $\bar{\rho}$ means subgroups weight the same words similarly (a coherent signal a prompted LLM could in principle exploit) while a negative $\bar{\rho}$ means subgroups diverge, creating structurally opposed signal that no single persona prompt can simultaneously satisfy. To connect these dataset-side quantities to LLM alignment, we compute Spearman rank correlations between each quantity's per-attribute ranking and the corresponding single-component LLM alignment ranking. Full construction details for the interaction features and coherence analysis are in Appendix F.

3.6 Neuron Probing

To understand how demographic prompting affects internal model behavior, we probe neuron activations following the interpretability framework of Ying et al. (2025). Their work traces neuron activations across *languages* to explain cultural-linguistic synergy in multilingual LLMs. We adapt this to compare across *prompting conditions* instead: for the same text in the same language, we compare which neurons fire when the model receives a demographic persona versus no persona (baseline). Following Geva et al. (2021), we define the i-th neuron at layer l as the i-th element of the post-activation vector $\sigma(\mathbf{W}^l_{\text{up}} \cdot \mathbf{h}^l) \in \mathbb{R}^{d_m}$, where σ is the MLP activation function (SiLU, GELU, etc.). We register forward hooks on the activation function of every MLP block to record these values. We only consider activations at *response token positions* (i.e., from where the model's generated answer begins to where it ends) since our interest is in how demographic context changes generation behavior, not prompt encoding.

Key Neuron selection. For each sample, we select the top-k neurons per layer by activation magnitude across all response positions, giving us a Key Neuron set N_{sample}:

$$N_{\text{sample}} = \left\{(i,l) \mid v^{r_t}_{(i,l)} \geq V^{\text{top-}k}_l,\ r_t \in R\right\} \quad (2)$$

where $V^{\text{top-}k}_l$ is the k-th largest activation at layer l, $R = \{r_1, \ldots, r_n\}$ are the response tokens, and $v^{r_t}_{(i,l)}$ is the activation of neuron i at layer l for

token r_t. Following the ablation study (in Appendix H.1), we evaluated threshold values of $k \in \{5, 10, 15, 20, 30, 40, 50, 100\}$ and found that selecting the top-10 neurons (k = 10) most effectively captures the model's knowledge for our given task and thus set k = 10 for our main results.

Specialized neuron set. Each base text appears under both a baseline (no-demographic) prompt and one or more identity-conditioned prompts. For a matched pair sharing the same base text, the *specialized neuron set* is $S = N_{\text{demo}} \setminus N_{\text{baseline}}$, i.e., the neurons active under the demographic prompt but not under the baseline. The *specialization proportion* is:

$$p = \frac{|N_{\text{demo}} \setminus N_{\text{baseline}}|}{|N_{\text{demo}}|} \quad (3)$$

A high p means the demographic prompt engages a substantially different set of neurons. We compute p at the individual pair level and then average across all matched pairs within each (task, model, attribute) combination.

4 Results and Discussion

4.1 Combinatorial Effects of Demographic Attributes on LLM Alignment (RQ1)

Across the 25 model–task pairs, demographic prompting does not exhibit a universal low-order optimum. The no-demographic baseline is the best-performing configuration in 12 pairs. Among the remaining 13 pairs, where at least one demographic prompt numerically exceeds the baseline, the best configuration uses 1–3 attributes in 11 cases, while DeepSeek on Sentiment and Llama on Emotion peak with higher-order combinations. ***Across all 25 model–task pairs, the full-attribute configuration is never the best-performing configuration.*** Thus, the results indicate a tendency toward compact demographic prompts when prompting is beneficial, rather than a universal 1–3 attribute threshold. Figures 1–5 trace the best achievable Cohen's κ (or accuracy for Emotion) with the number of demographic components in the prompt; full per-model breakdowns are in Tables 12 to 16 (Appendix D).

Toxicity. Toxicity is the task on which demographic prompting is most broadly beneficial. As Figure 1 shows, four of the five models peak above their no-demographic baseline with a compact set of attributes: Gemma with `religion_important` alone (κ = 0.271); Llama with `lgbtq_status + race` (κ = 0.299); Qwen with `age_range + lgbtq_status + religion_important` (κ = 0.294); and Mistral with `lgbtq_status + political_affiliation` (κ = 0.323, a gain of nearly five points over its baseline of 0.276). DeepSeek is the sole exception: its baseline represents its highest performance (see Table 12). Beyond the optimal point, extending the prompt consistently hurts: the 8-component *"all together"* prompt is never the best configuration and, for most models, represents a statistically significant degradation (see Table 12).

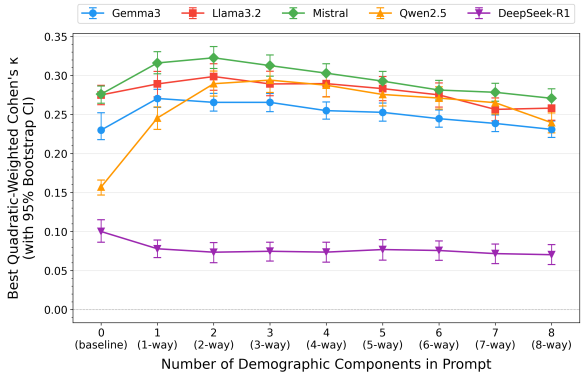

Figure 1: Best quadratic-weighted Cohen's κ on the **Toxicity task** by the number of demographic prompt components across five LLMs.

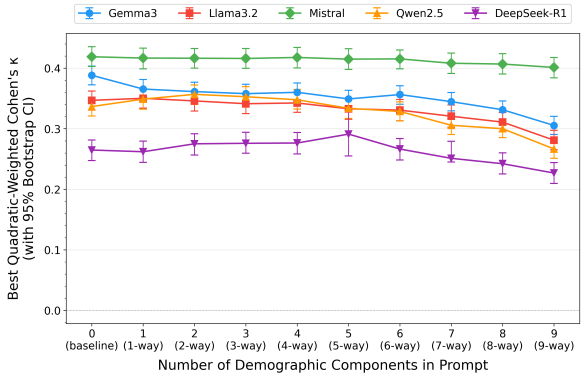

Figure 2: Best quadratic-weighted Cohen's κ on the **Sentiment task** by the number of demographic prompt components across five LLMs.

Sentiment Analysis. Sentiment presents a more mixed picture (Figure 2), with demographic prompting offering only modest and model-specific benefits. The three models that do gain are Llama with `gender` alone, Qwen with `education + hispanic_latino`, and DeepSeek with a complex 5-way combination. For Gemma and Mistral, no demographic prompt surpasses their respective baselines. The curves in Figure 2 are largely flat

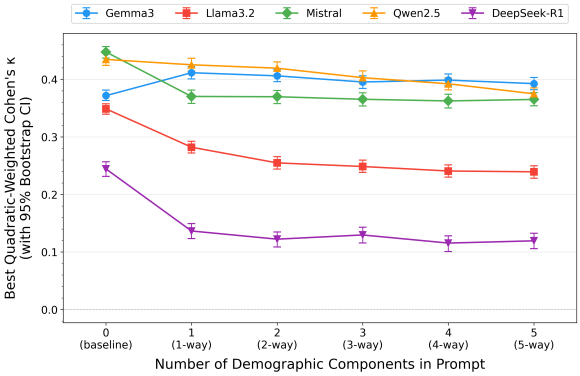

Figure 3: Best quadratic-weighted Cohen's κ on the **Politeness task** by the number of demographic prompt components across five LLMs.

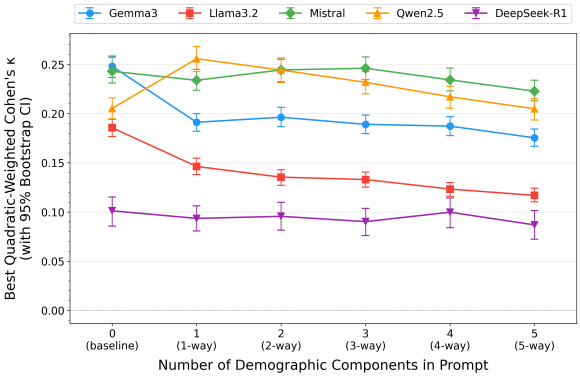

Figure 4: Best quadratic-weighted Cohen's κ on the **Offensiveness task** by the number of demographic prompt components across five LLMs.

or declining from the single-component level onward, **indicating a weak and model-dependent demographic signal.**

Politeness and Offensiveness. Both tasks show broad resistance to demographic prompting. On Politeness, four of the five models (all except Gemma) achieve their best results at baseline, with degradations that are not marginal: adding `race` drops Llama's κ from 0.349 to 0.252 and DeepSeek's from 0.244 to 0.101 (see Figure 3; Table 14). Gemma is the sole exception, with every single-attribute prompt exceeding its baseline, though combining attributes degrades these gains. On **Offensiveness, `race` is the only attribute that can selectively improve alignment**: Qwen peaks with `race` alone, and `race` appears in every top Mistral configuration (see Figure 4; Table 15). For the remaining models, no combination surpasses the baseline, and introducing attributes such as `age` or `gender` produces consistent, often significant degradations.

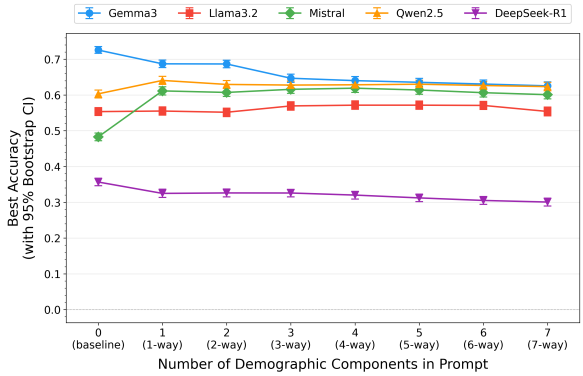

Figure 5: Best accuracy on the **Emotion task** by the number of demographic prompt components across five LLMs.

Emotion Attribution. Emotion Attribution produces the sharpest polarization across models (see Figure 5; Table 16). **age dominates for the models that benefit**: Mistral improves by 12.9 pp and Qwen by 3.8 pp, while Llama gains only through specific 4- and 5-way combinations anchored on `age`, `gender`, and parental occupations (up to 0.572 vs. baseline 0.560). Conversely, Gemma and DeepSeek are degraded by every demographic prompt without exception. That `age` cannot help these two models, whose strong baselines appear robust to persona prompting, highlights that **the utility of any demographic attribute is conditioned on model architecture as much as on dataset signal.**

4.2 The Role of Directional Coherence in Demographic Prompting (RQ2)

Model	Toxicity	Sentiment	Politeness	Offensiveness	Emotion
Gemma	0.261	-0.234	-0.300	-0.100	0.178
Llama	-0.333	-0.301	-0.400	0.400	-0.214
Qwen	0.023	-0.283	-0.100	-0.500	0.250
DeepSeek	0.571	-0.108	-0.400	0.100	0.285
Mistral	-0.595	0.267	-0.200	-0.100	-0.285

Table 1: Spearman rank correlation results (ρ) for **SHAP importance vs. LLM performance**. We compare Cohen's kappa (κ) for all tasks except Emotion, for which we use accuracy. None of the settings are statistically significant.

Using the setup described in Section 3.5, we organize the RQ2 findings around two contrasts: whether the *magnitude* of demographic influence (SHAP, Table 21) or the *learnability* of demographic signal (LSVC κ, Table 22) predicts LLM alignment, and connect both to the *directional coherence* of each attribute's annotation signal

(Fisher-averaged $\bar{\rho}$, in Table 3 (full results in Tables 23–27)).

SHAP importance does not predict LLM alignment. Table 1 reports Spearman rank correlations between the normalized Mean |SHAP| importance of each single-attribute demographic feature and the corresponding single-component Cohen's κ (or accuracy for Emotion) achieved by each model. None of the 25 (model $\times$ task) correlations reaches statistical significance ($p < 0.05$). The ρ values scatter widely from -0.595 (Mistral, Toxicity) to +0.571 (DeepSeek, Toxicity) with no consistent pattern. The implication is that **knowing which demographic attributes most strongly predict variation in human annotations provides no reliable information about which attributes will improve a given LLM's alignment when used as a prompt cue**.

Model	Toxicity	Sentiment	Politeness	Offensiveness	Emotion
Gemma	0.523	0.426	0.400	-0.700	-0.522
Llama	0.071	0.376	0.200	-0.700	-0.666
Qwen	**0.714**	0.283	0.700	-0.700	0.216
DeepSeek	-0.190	0.502	-0.500	0.700	0.342
Mistral	-0.047	0.175	-0.100	-0.800	-0.540

Table 2: Spearman rank correlation results (ρ) for **LSVC performance vs. LLM performance**. We compare Cohen's kappa (κ) for all tasks except Emotion, for which we use accuracy. Statistically significant values ($p < 0.05$) are indicated in **bold**.

LSVC learnability appears to be a more informative (though task-dependent) predictor. Table 2 reports the Spearman rank correlation between LSVC Cohen's κ (or accuracy for Emotion) and the corresponding single-component LLM κ (or accuracy). On Toxicity, the pattern is predominantly positive: Gemma ($\rho = 0.523$), **Qwen** ($\rho = 0.714$, $p < 0.05$), and Mistral (-0.047) near zero. Concretely, the three attributes with the highest LSVC κ on Toxicity (Table 22) `education` (0.201), `lgbtq_status` (0.200), and `race` (0.186) are the same attributes yielding Qwen's strongest single-component alignment gains ($\kappa = 0.198$, 0.246, and 0.172, vs. baseline 0.157). On Sentiment, all five correlations are positive (0.176–0.502), suggesting a weak-to-moderate tendency for more learnable attributes to benefit alignment, though none reaches significance. Two systematic negative patterns emerge. On Offensiveness, four of the five models produce $\rho \leq -0.700$ (Gemma, Llama, Qwen: -0.700; Mistral: -0.800), meaning that the most learnable attributes tend to yield the worst LLM alignment. This inversion is concrete: `age`, the highest-LSVC attribute (0.316, Table 22), produces some of the largest alignment degradations (e.g., Gemma drops from 0.248 to 0.152; Llama from 0.186 to 0.131, Table 15), while `race` ranked second-lowest in LSVC κ (0.302) is the only attribute that significantly improves any model's alignment (Qwen: 0.256 vs. baseline 0.205).

Attribute	Toxicity	Sentiment	Politeness	Offensiveness	Emotion
age	+0.125	+0.032	+0.048	+0.139	+0.280
race	+0.250	+0.207	+0.144	+0.058	—
gender	-0.087	-0.633	-0.046	-0.182	-0.087
lgbtq_status	+0.208	—	—	—	—

Table 3: Fisher-averaged $\bar{\rho}$ for selected demographic attributes across tasks. All values are significant. **This table reports only the attributes referenced in the main discussion; complete per-attribute results are in Tables 23–27 (Appendix G).**

Directional coherence explains when learnability helps versus hurts. The Fisher-averaged $\bar{\rho}$ values (Table 3 and Tables 23–27) resolve the apparent contradiction between positive and negative LSVC correlations across tasks. On Toxicity, the attributes with the most coherent directional signals, `race` ($\bar{\rho} = +0.250$) and `lgbtq_status` ($\bar{\rho} = +0.208$) (Table 3) are precisely those driving alignment gains in Table 12, and their high LSVC κ values contribute to the positive Spearman correlations in Table 2. On Offensiveness, however, the pattern inverts. Although `age` is the most learnable attribute (LSVC $\kappa = 0.316$), its moderate Fisher $\bar{\rho}$ (+0.139) (Table 3) reflects lexical patterns a classifier can exploit but a simple persona prompt cannot, learnability here reflects classifier capacity, not prompt-exploitable structure. `Gender`'s negative $\bar{\rho}$ (-0.182; Man vs. Woman $\rho = -0.319$) means its subgroups disagree on which words signal offensiveness, making persona-based alignment structurally challenging. Together, these produce the strong negative Spearman correlations in Table 2. The same mechanism applies to Sentiment, where `gender` carries the most severely opposed signal in the entire study (Fisher $\bar{\rho} = -0.633$; Female vs. Male $\rho = -0.633$), explaining why gender-based prompts fail to produce alignment gains despite moderate SHAP importance (0.107, Table 21) and LSVC learnability ($\kappa = 0.203$, Table 22). Detailed per-attribute breakdowns are in Appendix G.

4.3 Neuron Probing (RQ3)

Rationale and Interpretation Framework. Our neuron probing analysis is grounded in interpretability findings showing that neurons with higher activation values during inference indicate knowledge usage relevant to the input (Hong et al., 2025; Cao et al., 2025; Zhao et al., 2024; Tang et al., 2024). Building on this, Ying et al. (2025) propose that models activating a larger proportion of specialized neurons for a given context demonstrate stronger knowledge utilization and better task performance. Following this reasoning, we test the **hypothesis** that *a higher proportion of specialized neurons triggered by a demographic cue reflects deeper internal engagement with that attribute, potentially steering the model toward the corresponding annotation norms and improving alignment scores (κ or accuracy)*. We treat this as a testable hypothesis rather than an assumption. To our knowledge, neuron probing has not previously been applied to demographic alignment in subjective annotation tasks. Using the setup described in Section 3.6, we compute specialized neurons for single-attribute prompts and report full results in Tables 28–32 (Appendix H). We then measure the Pearson correlation (following Ying et al. (2025)) between each attribute's mean specialization proportion and its alignment score (κ or accuracy), reported in Table 4, to examine whether attributes activating more specialized neurons also achieve stronger alignment with human annotations.

Model	Toxicity	Sentiment	Politeness	Offensiveness	Emotion
Gemma	-0.222	**-0.888**	-0.823	-0.782	**-0.982**
Llama	**0.741**	-0.657	0.052	-0.120	**-0.772**
Qwen	**0.769**	-0.608	0.118	0.432	-0.318
DeepSeek	0.706	-0.049	-0.491	0.682	0.033
Mistral	0.660	**-0.727**	0.731	0.345	0.678

Table 4: Pearson correlation between the number of neurons and LLM performance/alignment metrics. We compare Cohen's kappa (κ) for all tasks except Emotion, for which we use accuracy. Values in **bold** are statistically significant ($p < 0.05$).

Selective support and the high-volume paradox. The specialized-neuron hypothesis receives partial support on Toxicity. Table 4 shows strong positive correlations for Llama ($r = 0.742$) and Qwen ($r = 0.769$), indicating that demographic attributes that activate more specialized neurons tend to produce better alignment with human toxicity annotations. Attributes such as `lgbtq_status` and `religion_important` illustrate this pattern for Qwen: they activate the highest proportions of specialized neurons (see Table 28) and also yield the largest alignment gains (see Table 12). Importantly, these attributes also exhibit directionally coherent lexical signals (Section 4.2), suggesting that specialized neuron activation correlates with alignment only when the underlying annotation signal is structurally exploitable. However, this relationship does not hold universally. DeepSeek activates the largest number of specialized neurons across tasks (see Tables 28 to 32), yet its no-demographic baseline consistently performs best and no demographic attribute improves performance. **This high-volume paradox suggests that neuron quantity alone does not reflect useful knowledge engagement; instead, widespread activation may indicate broad representational perturbation rather than targeted alignment-relevant processing.**

Task-dependent breakdown and signal coherence. In several model–task combinations, the neuron–performance relationship reverses. Significant negative correlations appear on Sentiment (Gemma r=-0.888, Mistral r=-0.727) and Emotion (Gemma r=-0.983, Llama r=-0.772), where attributes activating more specialized neurons often degrade performance, suggesting interference with already well-calibrated representations. For instance, on Sentiment, Gemma's most activating attributes (`political_id`, `hispanic_latino`) produce the largest alignment drops, while `gender` activates the fewest neurons yet causes the smallest degradation despite carrying the most conflicting annotation signal. Similarly, on Politeness, Gemma's best-performing attribute (`occupation`) activates the fewest specialized neurons. **Across tasks, specialized-neuron counts alone are an unreliable proxy for alignment utility; the directional coherence of the underlying annotation signal is the critical factor determining whether demographic prompting helps or harms performance.** See Appendix H for detailed discussion.

5 Conclusion

We presented a systematic study of demographic prompting across five tasks, enumerating all possible attribute combinations. Three findings stand out. First, the optimal amount of demographic information is model- and task-dependent. The baseline is best in 12 of 25 model–task pairs; among the 13 pairs benefiting numerically from demographic prompting, 11 peak with 1–3 attributes and

two with higher-order combinations. Across all pairs, however, the full-attribute configuration is never best. Second, predicting which attributes will improve alignment requires jointly considering both learnability and directional coherence of each attribute's annotation signal. Third, specialized neuron activation correlates with alignment only when the underlying signal is structurally coherent, and activation volume alone does not index steerability. Collectively, these results demonstrate that demographic prompting is not a monolithic intervention, and that principled attribute selection grounded in signal quality is essential for effective use. Based on these findings, we provide task- and model-specific recommendations for practitioners in Appendix J.

6 Limitations

English-only scope. All five datasets and our experimental setups are in English, so whether our findings on directional coherence and over-specification transfer to other languages or cultural settings remains an open question.

LLMs. All five LLMs range from 3B to 12B parameters. Larger models may exhibit different sensitivity to demographic prompts, and our findings may not generalize to proprietary models such as GPT-4 or Claude, which differ in training data, alignment procedures, and scale. We limited our tests to five models due to resource constraints and balancing the research budget.

Demographic category granularity. We treat demographic attributes as fixed categorical variables inherited from each dataset. This does not capture intersectional identities, within-category heterogeneity, or the fluid nature of certain demographic dimensions. Additionally, annotator pools in several datasets skew toward particular groups (e.g., predominantly White participants), which may limit the representativeness of subgroup-level analyses.

Prompt template sensitivity. Although we average results across three paraphrases of each prompting template to improve robustness, our findings remain conditioned on the general structure of persona-style prompts. Alternative prompting strategies (e.g., chain-of-thought or role-play framing) could yield different alignment patterns.

Correlational neuron analysis. Our neuron probing analysis is correlational rather than causal. We identify associations between specialized neuron activation and alignment outcomes, but *we do not intervene on specific neurons to confirm a causal mechanism*.

Inter-annotator agreement in source datasets. We rely on original annotation labels as ground truth but do not report inter-annotator agreement (IAA) within the source datasets. Since subjective tasks are known to exhibit substantial annotator disagreement, low IAA would place an inherent ceiling on achievable alignment. Our κ values should therefore be interpreted relative to this ceiling rather than as absolute measures of model quality.

Small-n rank correlations and correlational scope. The Spearman correlations central to RQ2 (Tables 1 and 2) are computed over as few as five matched attributes for Politeness and Offensiveness. At this sample size, individual rank swaps can substantially shift ρ, so the per-task correlations should be interpreted as indicative trends rather than precise effect-size estimates, though the consistent directionality across tasks and models lends support to our conclusions. Additionally, our RQ2 analysis is entirely correlational: we identify associations between dataset-side signal properties and LLM alignment but *do not* manipulate these properties directly. Unmeasured confounds such as label distribution skew or text-level difficulty may partially account for the observed patterns, and we frame directional coherence as a diagnostic indicator rather than a confirmed causal mechanism.

Distillation effects on persona sensitivity DeepSeek-R1-Distill-Qwen-7B was fine-tuned on chain-of-thought reasoning traces from the larger DeepSeek-R1 model. This distillation process may compress or discard the representational capacity needed to respond differentially to persona cues. Our high-volume paradox finding may therefore reflect a distillation artifact rather than a general property of demographic prompting, and we do not disentangle distillation effects from model scale or architecture in the current study.

7 Ethical Statement

This work studies how demographic information in prompts affects LLM–human alignment across subjective annotation tasks. A potential positive

impact of this research is that it helps clarify when demographic prompting is useful and when it may instead degrade alignment. In particular, our results highlight that demographic prompting is highly model- and task-dependent and that providing more demographic information does not necessarily improve alignment. These findings may help practitioners avoid unnecessary or ineffective demographic conditioning and encourage more careful evaluation of persona-based prompting systems.

At the same time, demographic prompting raises important ethical concerns. Representing individuals through demographic attributes may oversimplify heterogeneous groups, reinforce stereotypes, or encourage the assumption that members of the same demographic group share a common perspective. Such systems could also be misused for demographic profiling, targeted persuasion, or discriminatory decision making. Furthermore, demographic attributes can constitute sensitive personal information, creating privacy concerns if such information is collected or inferred without appropriate consent. Our study does not argue that greater demographic alignment is inherently desirable, nor that demographic categories should be treated as proxies for individual beliefs or preferences. Rather, we use these attributes to study how existing LLMs respond to demographic cues. The demographic categories used in our experiments are inherited from the original datasets and may not fully represent the diversity or intersectionality of real-world populations. Therefore, our findings should be interpreted as diagnostic evidence about model behavior rather than as justification for demographic-based personalization or decision making.

Acknowledgements

This project was fully supported by the University of South Florida. We would like to thank all the anonymous reviewers for their valuable feedback.

References

Shayan Alipour, Indira Sen, Mattia Samory, and Tanu Mitra. 2025. Robustness and confounders in the demographic alignment of LLMs with human perceptions of offensiveness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22025–22047, Vienna, Austria. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. 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.

Yixin Cao, Jiahao Ying, Yaoning Wang, Xipeng Qiu, Xuanjing Huang, and Yugang Jiang. 2025. Model utility law: Evaluating llms beyond performance through mechanism interpretable metric. *arXiv preprint arXiv:2504.07440*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717, Online. Association for Computational Linguistics.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. 2025. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19513–19535, Suzhou, China. Association for Computational Linguistics.

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

Mahammed Kamruzzaman, Abdullah Al Monsur, Gene Louis Kim, and Anshuman Chhabra. 2025. From anger to joy: How nationality personas shape emotion attribution in large language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 48–68, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. 2024. " a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms. *arXiv preprint arXiv:2409.11636*.

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.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in pre-trained language models. *arXiv preprint arXiv:2406.10130*.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

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.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

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

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Katia Schwerzmann and Alexander Campolo. 2025. "desired behaviors": alignment and the emergence of a machine learning ethics. *AI Soc.*, 40(7):5181–5194.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

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

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language

models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *arXiv preprint arXiv:2501.14457*.

Jie Zeng, Qianyu He, Qingyu Ren, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. 2025. Order matters: Investigate the position bias in multi-constraint instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12479–12492.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

A Dataset Details

Here, we provide a comprehensive overview of each of the datasets. Table 5 details the label distributions for each of the five tasks, capturing the class frequencies for Toxicity, Sentiment Analysis, Politeness, Offensiveness, and Emotion Attribution. We also provide specific demographic breakdowns including percentage shares for attributes such as age, race, gender, and education, across individual tables for Toxicity (Table 6), Sentiment Analysis (Table 7), Politeness (Table 8), Offensiveness (Table 9), and Emotion Attribution (Table 10). These summaries highlight the diversity of the rater pools whose judgments form the basis of our alignment evaluation.

1. **Toxicity:** The toxicity dataset was constructed from an initial corpus of 549,058 comments collected from Twitter, Reddit, and 4chan between December 2019 and August 2020. To address the natural class imbalance between benign and toxic content, Kumar et al. (2021) employed a stratified sampling approach using scores from the Perspective API TOXICITY model. This method deliberately oversampled comments within score ranges that typically generate the highest rater disagreement. The resulting final dataset contains 107,620 comments, distributed across Twitter, 4chan, and Reddit. These comments were evaluated by a diverse group of 17,280 crowdsourced participants.

2. **Sentiment:** The sentiment analysis dataset was constructed to evaluate age-related bias by isolating and manipulating age-identifying terms in naturalistic text. Díaz et al. (2018) initially collected a corpus by scraping 4,151 blog posts and 64,283 comments from a prominent “elderblogger" community. From this text, they extracted sentences containing the word “old" and applied strict exclusion criteria to isolate instances where the term specifically described people rather than objects. This filtering yielded a curated set of 121 base sentences. To establish a comparative baseline, the researchers duplicated these sentences and systematically replaced “old" (and its variants) with “young" (and its variants), resulting in a final standardized dataset of 242 paired sentences. Additionally, for broader synthetic testing and model retraining, the study generated a template-based dataset of around 135k sentences and isolated a subset of 13,781 age-related tweets filtered from the 1-million-tweet Sentiment140 corpus.

3. **Politeness and Offensiveness:** The POPQUORN dataset is derived from the Ruddit corpus (Hada et al., 2021), but it departs from the original Best–Worst Scaling (BWS) annotation method. Because BWS can struggle with skewed data and often produces label distributions that differ from standard rating scales (Louviere et al., 2015), the data was reannotated to better suit our experimental design. Specifically, instances were rescored on a 1-to-5 scale, ranging from 1 (not at all polite/offensive) to 5 (very polite/offensive). To ensure data quality and reduce noise, the MACE framework (Hovy et al., 2013) was applied to filter out unreliable annotators. The refined dataset ultimately features responses from 1,484 individuals (complete with demographic metadata detailing their age, gender, race, education, and occupation), yielding 25,043 instances for the politeness task and 13,037 instances for the offensiveness task.

4. **Emotion:** ISEAR dataset includes 7,665 events of 7 emotion categories. They utilize information from 3000 respondents in the dataset covering 16 countries.

Task	Dataset	Labels
Toxicity	Kumar et al. (2021)	Not at all toxic (52.02%), Slightly toxic (19.15%), Moderately toxic (12.95%), Very toxic (9.48%), Extremely toxic (6.40%)
Sentiment	Díaz et al. (2018)	Neutral (42.88%), Somewhat positive (23.47%), Somewhat negative (19.22%), Very positive (9.54%), Very negative (4.89%)
Politeness	Pei and Jurgens (2023)	somewhat polite (29.03%), moderately polite (24.90%), very polite (21.02%), barely polite (13.88%), not polite at all (11.17%)
Offensiveness	Pei and Jurgens (2023)	not offensive at all (57.66%), barely offensive (16.80%), somewhat offensive (12.20%), moderately offensive (7.97%), very offensive (5.38%)
Emotion	Scherer and Wallbott (1994)	anger (14.29%), sadness (14.29%), disgust (14.29%), shame (14.29%), fear (14.29%), joy (14.29%), guilt (14.29%)

Table 5: Label distribution for each task.

Attribute	Values (Percentage share)
age_range	25 - 34 (39.51%), 35 - 44 (25.00%), 45 - 54 (13.11%), 18 - 24 (11.77%), 55 - 64 (7.33%), 65 or older (3.10%), Prefer not to say (0.16%), Under 18 (0.02%)
education	Bachelor's degree (40.86%), Some college (20.14%), Master's degree (15.01%), Associate degree (11.08%), High school graduate (8.97%), Professional degree (1.59%), Doctoral degree (1.19%), Less than high school (0.54%), Prefer not to say (0.39%), Other (0.18%)
gender	Female (51.64%), Male (47.03%), Unknown (0.78%), Nonbinary (0.54%)
parental_status	Yes (51.23%), No (47.79%), Prefer not to say (0.99%)
lgbtq_status	Heterosexual (82.42%), Bisexual (10.86%), Homosexual (3.49%), Prefer not to say (1.99%), Other (0.86%)
political_affiliation	Liberal (40.83%), Conservative (26.69%), Independent (26.30%), Prefer not to say (4.04%), Other (2.14%)
race	White (71.86%), Black or African American (12.29%), Asian (5.78%), Hispanic (2.72%), Prefer not to say (1.04%), White/Hispanic (0.89%), American Indian or Alaska Native (0.88%), White/Black (0.81%), Other (0.78%), White/Asian (0.69%), White/American Indian (0.59%), White/Other (0.31%), Missing (0.24%), Native Hawaiian/Pacific Islander (0.18%), Various mixed backgrounds (<0.12% each)
religion_importance	Not important (31.79%), Very important (31.44%), Somewhat important (23.42%), Not too important (12.00%), Prefer not to say (1.36%)

Table 6: Demographic characteristics and percentage shares for the **Toxicity** task dataset

B LLM

For our analysis, we utilize five instruction-tuned language models sourced directly from the Hugging Face[3]: (1) **Llama-3.2-3B-Instruct** (*meta-llama/Llama-3.2-3B-Instruct*); (2) **Mistral-7B-Instruct-v0.3** (*mistralai/Mistral-7B-Instruct-v0.3*); (3) **Gemma-3-12B-it** (*google/gemma-3-12b-it*); (4) **Qwen2.5-7B-Instruct** (*Qwen/Qwen2.5-7B-Instruct*); and (5) **DeepSeek-R1-Distill-Qwen-7B** (*deepseek-ai/DeepSeek-R1-Distill-Qwen-7B*). We intentionally include a diverse selection of architectures, featuring both Western-trained models, such as Llama (Meta), Gemma (Google), and Mistral (Mistral AI), alongside Chinese-trained models, such as Qwen (Alibaba Cloud) and DeepSeek. This selection allows for a balanced assessment across different training methodologies and cultural data influences. To accommodate budget constraints and facilitate our available computational capabilities, we focused on these specific model scales. Unless otherwise specified, we used the models' default configuration settings. Generation was performed using the decoding configuration described below. For all models, responses were generated via HuggingFace using greedy decoding (`do_sample=False`), with `max_new_tokens` set to each model's default as specified in its generation configuration. We chose greedy decoding (`temperature=0.0`) because it is fully deterministic and appropriate for classification tasks, where the goal is to obtain the model's single best prediction rather than diverse sampled outputs. We note that the neuron activation extraction step does not involve generation; instead, it performs a single deterministic forward pass over the pre-generated responses to record FFN activations. Therefore, no additional generation hyperparameters apply to the neuron extraction stage.

[3] https://huggingface.co/

C Prompt Details

In Table 11, we present the specific prompting templates used for both the baseline evaluation and the with-demographic assessment, including two rephrased variants for each to ensure results are not artifact-dependent. In these templates, the place-

Attribute	Values (Percentage share)
age	60-69 (40.73%), 50-59 (36.17%), 70-79 (20.22%), 80-89 (2.68%), 90-99 (0.14%), 100+ (0.06%)
race	White (75.98%), Black or African American (13.90%), Asian (6.44%), Other (2.10%), American Indian or Alaska Native (1.22%), Native Hawaiian or Pacific Islander (0.22%), Middle Eastern (0.14%)
hispanic_latino	No (83.39%), Yes (16.61%)
income	$50,000 - $74,999 (20.80%), $35,000 - $49,999 (15.83%), $25,000 - $34,999 (13.77%), $75,000 - $99,999 (13.68%), $15,000 - $24,999 (10.87%), $100,000 - $149,999 (9.26%), Less than $10,000 (5.54%), $10,000 - $14,999 (5.18%), More than $200,000 (2.62%), $150,000 - $199,999 (2.45%)
education	Some college or associate's degree (37.94%), Bachelor's degree (25.34%), High school graduate/GED (19.70%), Graduate or professional degree (15.08%), Less than high school (1.94%)
employment_status	Retired (47.75%), Working full-time (24.90%), Working part-time (11.85%), Unemployed (7.93%), On disability (7.57%)
living_situation	Live with spouse/partner (51.98%), Live alone (29.44%), Live with family (16.84%), Other (1.11%), Nursing home (0.36%), Assisted living (0.21%), Retirement community (0.06%)
political_id	Moderate (38.80%), Somewhat conservative (20.36%), Somewhat liberal (17.09%), Very conservative (14.90%), Very liberal (8.85%)
gender	Male (50.40%), Female (49.54%), Nonbinary (0.06%)

Table 7: Demographic characteristics and percentage shares for the **Sentiment** task dataset.

Attribute	Values (Percentage share)
race	White (72.55%), Black or African American (12.65%), Hispanic or Latino (6.88%), Asian (6.32%), Prefer not to disclose (0.79%), Native Hawaiian or Pacific Islander (0.40%), American India or Alaska Native (0.20%), Hebrew (0.20%)
age	>65 (13.48%), 60-64 (11.06%), 18-24 (10.94%), 30-34 (10.76%), 40-44 (10.29%), 50-54 (9.29%), 54-59 (9.22%), 25-29 (8.71%), 45-49 (8.31%), 35-39 (7.55%), Prefer not to disclose (0.40%)
occupation	Employed (43.89%), Retired (15.19%), Self-employed (12.78%), Unemployed (12.12%), Homemaker (6.55%), Student (5.37%), Other (2.33%), Prefer not to disclose (1.78%)
education	College degree (46.95%), High school diploma or equivalent (29.88%), Graduate degree (19.00%), Other (2.19%), Prefer not to disclose (1.19%), Less than a high school diploma (0.78%)
gender	Woman (50.07%), Man (46.74%), Non-binary (2.60%), Prefer not to disclose (0.60%)

Table 8: Demographic attributes and percentage shares for the **Politeness** task dataset.

holder **`text`** refers to the specific content the model is tasked to classify, while **`task`** defines the classification dimension, such as toxicity levels or sentiment polarity. The **`options_string`** placeholder represents the set of labels corresponding to the specific task; for instance, a toxicity task include a 5-point scale ranging from "Not at all toxic" to "Extremely toxic." Crucially, the order of these options is randomized for every prompt. As prior studies (Zeng et al., 2025; Shi et al., 2025) have demonstrated that large language models are susceptible to ordering bias, often showing a preference for labels appearing earlier in a list, this randomization ensures that our results reflect the model's actual judgment rather than a structural artifact of the prompt. For example, a baseline prompt for a toxicity task might appear as: *"Given the following text: 'He's lucky get didn't get a bullet put in his head', rate its overall toxicity. Respond using one of the following options: Not at all toxic, Slightly toxic, Moderately toxic, Very toxic, Extremely toxic. Respond with an answer in English, without any explanation"* By shuffling these options in subsequent iterations, we mitigate positional preference and increase the robustness of our findings.

D Detailed Model Results for Each Task

In this section, we provide the full performance breakdown for all five LLMs across the five experimental tasks. Tables 12 through 16 present the results for the baseline, single-component, and multi-component configurations.

Specifically, Table 12 and Table 13 report the quadratic-weighted Cohen's κ for the Toxicity and Sentiment tasks, respectively. For the Politeness and Offensiveness tasks, Cohen's κ results are detailed in Table 14 and Table 15. Table 16 presents the classification accuracy for the Emotion task, as this dataset lacks natural ordinality.

Furthermore, we provide comprehensive alignment metrics including Cohen's κ, Micro Accuracy, and Macro Accuracy for each task to facilitate a more granular analysis: Toxicity (Table 17), Sentiment (Table 18), Politeness (Table 19), and Offensiveness (Table 15). In these tables, significant improvements or degradations relative to the no-demographic baseline are indicated based on a two-sided paired bootstrap test ($B = 10,000$).

Attribute	Values (Percentage share)
race	White (75.21%), Black or African American (12.29%), Asian (7.28%), Native American (2.29%), Hispanic or Latino (1.92%), Arab American (0.38%)
age	54-59 (13.75%), 35-39 (11.87%), 18-24 (11.85%), 30-34 (11.46%), >65 (11.08%), 25-29 (9.48%), 40-44 (8.76%), 45-49 (8.37%), 50-54 (7.65%), 60-64 (5.73%)
occupation	Employed (48.52%), Self-employed (12.98%), Unemployed (12.92%), Retired (11.09%), Student (7.63%), Homemaker (4.56%), Other (1.92%), Prefer not to disclose (0.38%)
education	College degree (46.30%), High school diploma or equivalent (30.56%), Graduate degree (19.03%), Other (2.29%), Less than a high school diploma (1.82%)
gender	Woman (50.02%), Man (47.30%), Non-binary (2.68%)

Table 9: Demographic attributes and percentage shares for the **Offensiveness** task dataset.

Attribute	Values (Percentage share)
gender	female (54.78%), male (45.09%), Other (0.04%)
father's_occupation.	white collar academic (26.84%), white collar nonacademic (19.31%), self-employed nonacademic (15.13%), blue collar untrained (11.32%), blue collar trained (11.23%), unemployed (7.39%), self-employed academic (4.56%), housewife (1.00%), student (0.09%)
mother's_occupation.	housewife (43.96%), white collar nonacademic (20.05%), white collar academic (14.68%), blue collar untrained (6.85%), self-employed nonacademic (5.57%), blue collar trained (4.29%), unemployed (1.46%), self-employed academic (0.64%), student (0.27%)
field_of_study	psychology (40.94%), medical (16.98%), social sciences (14.58%), other (9.04%), languages (5.93%), law (4.47%), natural science (4.38%), engineering (1.46%), fine arts (1.10%)
country	zambia (10.04%), sweden (7.39%), china mainland (7.21%), spain (7.12%), bulgaria (6.66%), malawi (6.57%), finland (6.38%), netherlands (6.30%), austria (6.26%), india (6.21%), usa (5.55%), new zealand (5.39%), brazil (5.29%), australia (5.29%), honduras (5.01%), norway (3.29%)
age	20.0 (16.69%), 21.0 (15.24%), 22.0 (12.92%), 19.0 (12.14%), 18.0 (8.66%), 23.0 (7.85%), 24.0 (5.84%), 25.0 (3.82%), 26.0 (3.19%), 27.0 (2.74%), 28.0 (2.46%), 30.0 (2.01%), 33.0 (1.28%), 29.0 (1.19%), 35.0 (1.19%), 32.0 (1.10%), 34.0 (0.82%), 31.0 (0.82%)
religion	catholic (32.74%), protestant (26.80%), areligious (24.28%), hindu (5.66%), others (3.65%), native (3.01%), jewish (0.53%), buddhist (0.37%)

Table 10: Demographic characteristics and percentage shares for the **Emotion** task dataset.

Category	Version	Prompt Template
Baseline	Original	Given the following {text}, rate its overall {task}. Respond using one of the following options: {options_string}. Respond with an answer in English, without any explanation.
	Rephrased 1	Analyze the {task} of the {text} provided below. Select the most appropriate rating from this list: {options_string}. Give your response in English. Do not provide any context, reasoning, or additional text.
	Rephrased 2	Your task is to provide a {task} rating for the following {text}. You must use one of these predefined labels: {options_string}. Return only the English label as your final answer, skipping all introductory or explanatory remarks.
Demographic	Original	Given the following text: {text}, how would a person of {demographic_attributes} rate this {task}? Respond using one of the following options: {options_string}. Respond with an answer in English, without any explanation.
	Rephrased 1	Adopt the perspective of an individual with the following background: {demographic_attributes}. Based on that viewpoint, evaluate the {task} of this text: {text}. Choose from: {options_string}. Response must be in English and contain only the selected option.
	Rephrased 2	If a person identifying as {demographic_attributes} were to read the text below, how would they likely rate its {task}? Choose from these given options: {options_string}. Provide a one-word/phrase answer in English with no explanation.

Table 11: Baseline and with demographic Prompting Templates

E SHAP (SHapley Additive exPlanations)

To interpret the model's decision-making process and evaluate the influence of demographic attributes on predictions, we utilized SHAP (SHapley Additive exPlanations). After training a *Logistic Regression* model on One-Hot Encoded demographic data, we applied a LinearExplainer to compute the SHAP values for each instance. To assess feature importance at the categorical level, we ag-

no.	demographic in the prompt	Gemma κ	Llama κ	Qwen κ	DeepSeek κ	Mistral κ
Baseline, single-component and all components together configurations						
0	baseline (no demographic)	0.230	0.275	0.157	0.100	0.276
1	age_range	0.213	$0.226^{\dagger}$	0.151	$0.057^{\dagger}$	$0.237^{\dagger}$
1	education	0.235	$0.224^{\dagger}$	0.198^{*}	$0.060^{\dagger}$	$0.234^{\dagger}$
1	gender	$0.191^{\dagger}$	$0.249^{\dagger}$	0.141	$0.068^{\dagger}$	0.263
1	is_parent	$0.208^{\dagger}$	$0.235^{\dagger}$	0.168	$0.078^{\dagger}$	$0.242^{\dagger}$
1	lgbtq_status	0.236	0.264	0.246^{*}	$0.055^{\dagger}$	0.316^{*}
1	political_affilation	0.234	$0.251^{\dagger}$	0.192^{*}	$0.059^{\dagger}$	0.299^{*}
1	race	$0.195^{\dagger}$	0.289	0.172	$0.064^{\dagger}$	0.267
1	religion_important	0.271^{*}	0.270	0.242^{*}	$0.072^{\dagger}$	0.287
8	all together	0.231	0.258	0.240^{*}	$0.070^{\dagger}$	0.271
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Gemma						
3	education, is_parent, religion_important	0.266^{*}	–	–	–	–
2	education, religion_important	0.266^{*}	–	–	–	–
3	education, political_affilation, religion_important	0.260^{*}	–	–	–	–
3	education, lgbtq_status, religion_important	0.258^{*}	–	–	–	–
2	political_affilation, religion_important	0.257^{*}	–	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Llama						
2	lgbtq_status, race	–	0.299^{*}	–	–	–
4	is_parent, lgbtq_status, political_affilation, race	–	0.290^{*}	–	–	–
3	lgbtq_status, political_affilation, race	–	0.289	–	–	–
3	education, gender, lgbtq_status	–	0.286	–	–	–
3	is_parent, lgbtq_status, race	–	0.285	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Qwen						
3	age_range, lgbtq_status, religion_important	–	–	0.294^{*}	–	–
2	lgbtq_status, race	–	–	0.289^{*}	–	–
4	gender, lgbtq_status, political_affilation, religion_important	–	–	0.288^{*}	–	–
3	education, lgbtq_status, race	–	–	0.286^{*}	–	–
4	age_range, education, lgbtq_status, race	–	–	0.284^{*}	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for DeepSeek						
–	[No entries]	–	–	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Mistral						
2	lgbtq_status, political_affiliation	–	–	–	–	0.323^{*}
2	lgbtq_status, religion_important	–	–	–	–	0.316^{*}
3	lgbtq_status, political_affilation, religion_important	–	–	–	–	0.313^{*}
3	is_parent, lgbtq_status, political_affiliation	–	–	–	–	0.304^{*}
4	gender, lgbtq_status, political_affiliation, race	–	–	–	–	0.303^{*}

Table 12: **Cohen's κ on the Toxicity dataset** for Gemma, Llama, Qwen, DeepSeek, and Mistral. Statistical significance is assessed via a two-sided paired bootstrap test ($B = 10{,}000$ resamples); we report the 95% confidence interval on $\Delta\kappa = \kappa_{\text{config}} - \kappa_{\text{baseline}}$ and consider a result significant when the CI excludes zero ($p < 0.05$). * significant improvement; † significant degradation relative to baseline.

gregated the absolute SHAP values of the encoded binary features back to their parent demographic groups (e.g., combining gender Male and gender Female into gender). This approach allowed us to extract both global feature importance rankings and local, instance-level explanations, revealing the magnitude and directionality of how specific demographic groups bias the model toward or away from the original label.

F Word–Demographic Interaction Features and Spearman Rank Correlation Analysis

F.1 Measuring Annotator Demographic Influence via Learned Interaction Features

To quantify how annotator demographic characteristics shape the target label at the lexical level, we trained a *Linear Support Vector Classifier (LinearSVC)* with explicitly constructed *word–demographic interaction features*. The motivation for this design is to move beyond post-hoc analysis: rather than training a model on text and demographics separately and then examining correlations after the fact, we encode the relationship

no.	demographic in the prompt	Gemma κ	Llama κ	Qwen κ	DeepSeek κ	Mistral κ
Baseline, single-component and all components together configurations						
0	baseline (no demographic)	0.388	0.347	0.337	0.265	0.419
1	age	0.355†	0.335	0.315†	0.241†	0.405
1	education	0.354†	0.330	0.346	0.240†	0.415
1	employment_status	0.339†	0.321†	0.310†	0.254	0.406
1	gender	0.366†	0.350	0.349	0.262	0.414
1	hispanic_latino	0.339†	0.335	0.342	0.262	0.405
1	income	0.338†	0.326†	0.301†	0.251	0.409
1	living_situation	0.352†	0.333	0.326	0.260	0.417
1	political_id	0.300†	0.303†	0.324	0.239†	0.396†
1	race	0.362†	0.343	0.345	0.246	0.410
9	all together	0.305†	0.281†	0.267†	0.227†	0.401
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Gemma, Llama, Mistral						
–	[No entries]	–	–	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Qwen						
2	education, hispanic_latino	–	–	0.357*	–	–
2	education, race	–	–	0.354	–	–
3	education, gender, hispanic_latino	–	–	0.353	–	–
3	employment_status, gender, hispanic_latino	–	–	0.352	–	–
3	age, gender, hispanic_latino	–	–	0.350	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for DeepSeek						
5	education, gender, income, living_situation, race	–	–	–	0.291*	–
5	age, employment_status, hispanic_latino, income, race	–	–	–	0.283	–
5	age, employment_status, hispanic_latino, living_situation, race	–	–	–	0.282	–
5	employment_status, gender, living_situation, political_id, race	–	–	–	0.281	–
5	age, gender, hispanic_latino, living_situation, race	–	–	–	0.278	–

Table 13: **Cohen's κ on the Sentiment dataset** for Gemma, Llama, Qwen, DeepSeek, and Mistral. * significant improvement; † significant degradation relative to baseline, calculated in the similar way as in Table 12.

directly into the feature space so that the model learns differential word weights during training.

F.1.1 Feature Construction

For each comment in the dataset, we first constructed two independent feature representations. **Text features were extracted using TF-IDF vectorization** (maximum vocabulary of 30,000 unigrams, minimum document frequency of 2), producing a sparse matrix $\mathbf{X}_{\text{words}}$ of shape $(n_{\text{samples}} \times n_{\text{text_features}})$. Annotator demographic attributes such as race, gender, religion, education, LGBTQ status, political affiliation, parental status, and age range were encoded using one-hot encoding, producing a sparse matrix $\mathbf{X}_{\text{demo}}$ of shape $(n_{\text{samples}} \times n_{\text{dem_categories}})$, where each column corresponds to one unique value within a demographic attribute (e.g., `gender_Female`, `gender_Male`, `gender_Nonbinary`).

The core of our approach is the explicit construction of word–demographic interaction features before model fitting. For each demographic indicator column d_j in $\mathbf{X}_{\text{demo}}$, we compute the element-wise product of the entire TF-IDF word matrix with that column:

$$\mathbf{X}_{\text{inter},j} = \mathbf{X}_{\text{words}} \odot d_j \tag{4}$$

where $\odot$ denotes column-wise multiplication (each word's TF-IDF value is multiplied by the annotator's demographic indicator for category j). This produces one interaction block per demographic category, yielding an interaction matrix of shape $(n_{\text{samples}} \times n_{\text{text_features}} \times n_{\text{dem_categories}})$. The final feature matrix passed to the classifier is the horizontal concatenation:

$$\mathbf{X}_{\text{full}} = [\mathbf{X}_{\text{words}} \mid \mathbf{X}_{\text{demo}} \mid \mathbf{X}_{\text{inter},1} \mid \mathbf{X}_{\text{inter},2} \mid \cdots \mid \mathbf{X}_{\text{inter},k}] \tag{5}$$

This design means the classifier learns a separate weight for every (word, demographic category) pair for example, $w[\texttt{gender_Female} \times \text{“threat”}]$ is learned independently from $w[\texttt{gender_Male} \times \text{“threat”}]$. Crucially, these weights are estimated jointly during a single model fit rather than derived post-hoc, so they reflect the interaction as the model actually uses it in prediction. We deliberately exclude word–word interactions to avoid feature explosion while preserving full interpretability of the learned weights. Each model was trained separately per demographic attribute on an 80/20 stratified split, with a text-only baseline trained under identical conditions for comparison.

no.	demographic in the prompt	Gemma κ	Llama κ	Qwen κ	DeepSeek κ	Mistral κ
Baseline, single-component and all components together configurations						
0	baseline (no demographic)	0.372	0.349	0.435	0.244	0.448
1	age	0.376	$0.253^{\dagger}$	$0.376^{\dagger}$	$0.109^{\dagger}$	$0.354^{\dagger}$
1	education	0.383	$0.237^{\dagger}$	$0.361^{\dagger}$	$0.137^{\dagger}$	$0.349^{\dagger}$
1	gender	0.408^{*}	$0.282^{\dagger}$	0.425	$0.105^{\dagger}$	$0.368^{\dagger}$
1	occupation	0.411^{*}	$0.267^{\dagger}$	$0.389^{\dagger}$	$0.117^{\dagger}$	$0.370^{\dagger}$
1	race	0.407^{*}	$0.252^{\dagger}$	$0.409^{\dagger}$	$0.101^{\dagger}$	$0.353^{\dagger}$
5	all together	0.393^{*}	$0.239^{\dagger}$	$0.375^{\dagger}$	$0.119^{\dagger}$	$0.365^{\dagger}$
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Gemma						
2	education, occupation	0.406^{*}	–	–	–	–
2	gender, race	0.404^{*}	–	–	–	–
2	gender, occupation	0.401^{*}	–	–	–	–
4	education, gender, occupation, race	0.399^{*}	–	–	–	–
2	education, race	0.398^{*}	–	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Llama, Qwen, DeepSeek, and Mistral						
–	[No entries]	–	–	–	–	–

Table 14: **Cohen's κ on the Politeness dataset** for Gemma, Llama, Qwen, DeepSeek, and Mistral. * significant improvement; † significant degradation relative to baseline, calculated in the similar way as in Table 12.

no.	demographic in the prompt	Gemma κ	Llama κ	Qwen κ	DeepSeek κ	Mistral κ
Baseline, single-component and all components together configurations						
0	baseline (no demographic)	0.248	0.186	0.205	0.101	0.243
1	age	$0.152^{\dagger}$	$0.131^{\dagger}$	$0.158^{\dagger}$	0.094	$0.179^{\dagger}$
1	education	$0.182^{\dagger}$	$0.137^{\dagger}$	0.198	0.084	$0.223^{\dagger}$
1	gender	$0.160^{\dagger}$	$0.128^{\dagger}$	0.209	0.088	$0.219^{\dagger}$
1	occupation	$0.191^{\dagger}$	$0.146^{\dagger}$	0.204	$0.078^{\dagger}$	0.227
1	race	$0.167^{\dagger}$	$0.139^{\dagger}$	0.256^{*}	0.085	0.234
5	all together	$0.175^{\dagger}$	$0.117^{\dagger}$	0.205	0.087	$0.223^{\dagger}$
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Gemma, Llama, and DeepSeek						
–	[No entries]	–	–	–	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Qwen						
2	gender, race	–	–	0.244^{*}	–	–
2	occupation, race	–	–	0.239^{*}	–	–
2	education, race	–	–	0.234^{*}	–	–
3	education, occupation, race	–	–	0.232^{*}	–	–
3	gender, occupation, race	–	–	0.220	–	–
Top 5 configurations (no. of components $\geq$ 2 and not all together) exceeding baseline κ for Mistral						
3	education, occupation, race	–	–	–	–	0.246
2	occupation, race	–	–	–	–	0.244

Table 15: **Cohen's κ on the Offensiveness dataset** for Gemma, Llama, Qwen, DeepSeek, and Mistral. * significant improvement; † significant degradation relative to baseline, calculated in the similar way as in Table 12.

F.1.2 Extracting Demographic-Specific Word Weights

After fitting, we extracted the learned interaction weights from the model's coefficient vector. For each demographic attribute (e.g., gender), we identified the top-200 most influential words by selecting the words with the highest maximum absolute interaction weight across all demographic categories within that attribute. This max-pooling selection ensures that the vocabulary of 200 words is not dominated by a single category, and that the resulting word set is representative of lexical signals that matter for at least one group.

For each word in this vocabulary, we then retrieved the learned interaction weight for every demographic category within the attribute—regardless of whether that (category, word) pair ranked in the top 200. This produces a wide-format weight table $\mathbf{W}$ of shape $(n_{\text{words}} \times n_{\text{categories}})$, where entry $W[\text{word}, \text{category}]$ is the model's learned weight for how strongly that word pushes toward a toxic classification when the annotator belongs to that demographic category. A weight of zero indicates that the word was either absent from the training vocabulary for that category or had no discriminative value in that context.

F.1.3 Spearman Rank Correlation Across Demographic Categories

To quantify the degree to which two demographic categories produce similar lexical toxicity patterns,

no.	demographic in the prompt	Gemma Accuracy	Llama Accuracy	Qwen Accuracy	DeepSeek Accuracy	Mistral Accuracy
Baseline, single-component and all components together configurations						
0	baseline (no demographic)	0.7256	0.5601	0.6032	0.3566	0.4829
1	age	0.6871	0.5548	0.6407	0.3177	0.6114
1	country	0.4741	0.4265	0.6250	0.3248	0.5710
1	father_occupation	0.4633	0.4567	0.5911	0.2986	0.5911
1	field_of_study	0.4176	0.4720	0.6081	0.3099	0.5817
1	gender	0.6538	0.5348	0.6114	0.3174	0.6066
1	mother_occupation	0.5229	0.4969	0.5768	0.2876	0.5713
1	religion	0.5676	0.4682	0.5791	0.2962	0.5620
7	all together	0.6255	0.5537	0.6234	0.3006	0.6008
Top 5 configurations (no. of components ≥ 2 and not all together) exceeding baseline Accuracy for Gemma, DeepSeek						
–	[No entries]	–	–	–	–	–
Top 5 configurations (no. of components ≥ 2 and not all together) exceeding baseline Accuracy for Llama						
4	age, father_occupation, gender, mother_occupation	–	0.5715	–	–	–
5	age, country, father_occupation, gender, mother_occupation	–	0.5715	–	–	–
6	age, country, father_occupation, field_of_study, gender, mother_occupation	–	0.5708	–	–	–
3	age, gender, mother_occupation	–	0.5693	–	–	–
4	age, country, father_occupation, mother_occupation	–	0.5691	–	–	–
Top 5 configurations (no. of components ≥ 2 and not all together) exceeding baseline Accuracy for Qwen						
5	age, country, father_occupation, gender, mother_occupation	–	–	0.6301	–	–
2	age, country	–	–	0.6295	–	–
5	age, father_occupation, field_of_study, gender, mother_occupation	–	–	0.6285	–	–
4	age, country, father_occupation, mother_occupation	–	–	0.6285	–	–
3	age, father_occupation, gender	–	–	0.6277	–	–
Top 5 configurations (no. of components ≥ 2 and not all together) exceeding baseline Accuracy for Mistral						
6	country, father_occupation, field_of_study, gender, mother_occupation, religion	–	–	–	–	0.5998
5	father_occupation, field_of_study, gender, mother_occupation, religion	–	–	–	–	0.5995
4	father_occupation, gender, mother_occupation, religion	–	–	–	–	0.5969
3	father_occupation, gender, religion	–	–	–	–	0.5923
2	gender, religion	–	–	–	–	0.5872

Table 16: **Accuracy on the Emotion dataset** for Gemma, Llama, Qwen, DeepSeek, and Mistral.

		Gemma			Llama			Qwen			DeepSeek			Mistral		
no.	demographic in the prompt	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro
Baseline, single-component, and all components together configurations																
0	baseline (no demographic)	0.230	0.222	0.235	0.275	0.300	0.268	0.157	0.214	0.243	0.100	0.286	0.228	0.276	0.244	0.271
1	age_range	0.213	0.139	0.236	0.226	0.287	0.256	0.151	0.171	0.224	0.057	0.196	0.210	0.237	0.197	0.263
1	education	0.235	0.161	0.245	0.224	0.305	0.256	0.198	0.226	0.253	0.060	0.204	0.212	0.234	0.201	0.260
1	gender	0.191	0.137	0.235	0.249	0.322	0.262	0.141	0.178	0.233	0.068	0.203	0.212	0.263	0.230	0.271
1	is_parent	0.208	0.155	0.242	0.235	0.303	0.263	0.168	0.228	0.248	0.078	0.206	0.214	0.242	0.211	0.263
1	lgbtq_status	0.236	0.192	0.255	0.264	0.456	0.262	0.246	0.393	0.263	0.055	0.200	0.208	0.316	0.334	0.299
1	political_affilation	0.234	0.170	0.251	0.251	0.327	0.267	0.192	0.222	0.242	0.059	0.206	0.213	0.299	0.278	0.293
1	race	0.195	0.163	0.251	0.289	0.397	0.280	0.172	0.222	0.244	0.064	0.206	0.208	0.267	0.249	0.274
1	religion_important	0.271	0.213	0.268	0.270	0.354	0.270	0.242	0.295	0.266	0.072	0.194	0.209	0.287	0.244	0.275
8	all components together	0.231	0.167	0.251	0.258	0.319	0.270	0.240	0.307	0.273	0.070	0.199	0.215	0.271	0.237	0.274

Table 17: **Cohen κ, Micro Accuracy, and Macro Accuracy for Toxicity Dataset**. This table lists the baseline, all single-component, and all components together configurations.

we computed Spearman rank correlation coefficients between the weight vectors of every pair of categories within each demographic attribute. For two categories c_i and c_j, the Spearman correlation $\rho(c_i, c_j)$ measures the rank-order agreement between the vectors $\mathbf{W}[:, c_i]$ and $\mathbf{W}[:, c_j]$ across the shared 200-word vocabulary.

A high positive ρ indicates that the two demographic groups assign similar relative importance to the same words when judging toxicity. A negative ρ indicates that words strongly associated with toxicity for one group tend to be strongly associated with non-toxicity for the other a pattern that directly reflects divergent annotation behavior. A near-zero ρ indicates no systematic lexical alignment between the two groups.

For each demographic attribute with k unique categories, this procedure yields $\binom{k}{2}$ pairwise correlations. To summarize the overall degree of within-attribute divergence as a single comparable value, we aggregated pairwise correlations using the Fisher z-transformation:

$$z_i = \operatorname{arctanh}(\rho_i) \tag{6}$$

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \tag{7}$$

$$\bar{\rho} = \tanh(\bar{z}) \tag{8}$$

where the sum runs over all $\binom{k}{2}$ pairs. Fisher z-averaging is preferred over the arithmetic mean of ρ values because ρ is bounded in $[-1, 1]$ and compressed near the extremes; averaging in z-space and back-transforming avoids this distortion and produces a more reliable estimate of the central tendency across pairs.

no.	demographic in the prompt	Gemma			Llama			Qwen			DeepSeek			Mistral		
		κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro
Baseline and single-component configurations (shared across models)																
0	baseline (no demographic)	0.388	0.269	0.360	0.347	0.237	0.343	0.337	0.279	0.342	0.265	0.290	0.290	0.419	0.318	0.358
1	age	0.355	0.270	0.330	0.335	0.253	0.327	0.315	0.233	0.336	0.241	0.267	0.272	0.405	0.301	0.359
1	education	0.354	0.278	0.329	0.330	0.274	0.311	0.346	0.280	0.353	0.240	0.297	0.276	0.415	0.321	0.353
1	employment_status	0.339	0.267	0.322	0.321	0.249	0.313	0.310	0.255	0.329	0.254	0.287	0.270	0.406	0.316	0.358
1	gender	0.366	0.262	0.344	0.350	0.256	0.333	0.349	0.269	0.350	0.262	0.283	0.283	0.414	0.313	0.347
1	hispanic_latino	0.339	0.274	0.323	0.335	0.265	0.318	0.342	0.312	0.341	0.262	0.303	0.278	0.405	0.334	0.348
1	income	0.338	0.276	0.318	0.326	0.275	0.324	0.301	0.273	0.327	0.251	0.290	0.272	0.409	0.332	0.350
1	living_situation	0.352	0.267	0.332	0.333	0.257	0.317	0.326	0.266	0.347	0.260	0.288	0.286	0.417	0.322	0.360
1	political_id	0.300	0.296	0.307	0.303	0.263	0.298	0.324	0.290	0.342	0.239	0.312	0.266	0.396	0.341	0.340
1	race	0.362	0.285	0.337	0.343	0.269	0.319	0.345	0.304	0.346	0.246	0.301	0.282	0.410	0.341	0.348
9	all together	0.305	0.289	0.312	0.281	0.262	0.300	0.267	0.252	0.330	0.227	0.278	0.266	0.401	0.331	0.346

Table 18: **Cohen κ, Micro Accuracy, and Macro Accuracy for Sentiment Dataset**. This table lists the baseline, all single-component, and all components together configurations.

no.	demographic in the prompt	Gemma			Llama			Qwen			DeepSeek			Mistral		
		κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro
Baseline and single-component configurations (shared across models)																
0	baseline (no demographic)	0.372	0.302	0.317	0.349	0.263	0.327	0.435	0.295	0.337	0.244	0.276	0.266	0.448	0.314	0.329
1	age	0.376	0.307	0.293	0.253	0.298	0.271	0.376	0.282	0.315	0.109	0.265	0.226	0.354	0.284	0.291
1	education	0.383	0.307	0.300	0.237	0.291	0.269	0.361	0.277	0.314	0.137	0.271	0.235	0.349	0.285	0.292
1	gender	0.408	0.315	0.315	0.282	0.305	0.283	0.425	0.303	0.334	0.105	0.262	0.224	0.368	0.287	0.298
1	occupation	0.411	0.307	0.314	0.267	0.289	0.275	0.389	0.285	0.322	0.117	0.266	0.227	0.370	0.287	0.297
1	race	0.407	0.308	0.314	0.252	0.305	0.272	0.409	0.297	0.330	0.101	0.256	0.223	0.353	0.285	0.294
5	all together	0.393	0.307	0.300	0.239	0.293	0.267	0.375	0.279	0.319	0.119	0.262	0.227	0.365	0.283	0.292

Table 19: **Cohen κ, Micro Accuracy, and Macro Accuracy for Politeness Dataset**. This table lists the baseline, all single-component, and all components together configurations.

F.1.4 Reporting Strategy

Because the number of pairwise comparisons grows quadratically with the number of demographic categories (reaching $\binom{48}{2} = 1{,}128$ pairs for race alone) it is neither practical nor informative to report all individual correlations in the main text. We therefore adopt a two-level reporting structure. In Tables 23–27, we report the Fisher-averaged $\bar{\rho}$ as a compact cross-attribute divergence score alongside the single most divergent pair (highest $|\rho|$) within each attribute. This dominant pair captures the most salient signal within the attribute and flags cases where the summary $\bar{\rho}$ may be masking a strong but localized divergence. For gender, for example, the Fisher-averaged $\bar{\rho} = -0.087$ (non-significant at the aggregate level), but this summary conceals a strongly negative Female/Male correlation of $\rho = -0.629$ ($p < 0.001$), which indicates that words driving toxicity predictions for Female annotators systematically suppress them for Male annotators and vice versa.

G Per-Task Analysis of RQ2

Below we provide per-task commentary connecting the three dataset-side analyses (SHAP importance from Table 21, LSVC learnability from Table 22, and Fisher-averaged directional coherence from Tables 23–27) to the LLM alignment outcomes.

Toxicity. Toxicity provides the strongest evidence for dataset-side signal predicting alignment. SHAP ranks `religion_important` highest (0.189), yet its Fisher $\bar{\rho}$ is near zero (-0.001), indicating that its subgroups do not agree on which words indicate toxicity. By contrast, `race` (SHAP: 0.106, rank 6/8) and `lgbtq_status` (SHAP: 0.100, rank 7/8) carry the highest positive Fisher $\bar{\rho}$ ($+0.250$ and $+0.208$) and also rank among the top three LSVC κ values (0.186 and 0.200, Table 22). These are precisely the attributes anchoring the best-performing multi-component configurations in Table 12: `lgbtq_status` appears in every top-5 configuration for Llama, Qwen, and Mistral. The LSVC Spearman correlation is positive for Gemma (0.524) and significant for Qwen (**0.714**, $p < 0.05$). This task illustrates the key principle: an attribute can be a strong predictor of human label variation (high SHAP) without offering a coherent signal that a prompted LLM can exploit; conversely, attributes with moderate SHAP importance but high directional coherence enable the strongest alignment gains.

Sentiment. All five LSVC Spearman correlations are positive (0.176–0.502), but none reaches significance. The Fisher $\bar{\rho}$ analysis explains the limited alignment gains on this task. `Gender`—one of the few attributes occasionally beating baseline (Llama: $\kappa = 0.350$ vs. 0.347; Qwen: 0.349 vs. 0.337)—carries the most severely opposed signal in the entire study ($\bar{\rho} = -0.633$). Even attributes with positive Fisher $\bar{\rho}$ (e.g., `race`: $+0.207$; `income`: $+0.148$) produce only marginal alignment improvements, and the "all together" configuration

no.	demographic in the prompt	Gemma			Llama			Qwen			DeepSeek			Mistral		
		κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro	κ	Micro	Macro
Baseline and single-component configurations (shared across models)																
0	baseline (no demographic)	0.248	0.218	0.288	0.186	0.260	0.282	0.205	0.291	0.288	0.101	0.313	0.222	0.284	0.329	0.308
1	age	0.152	0.118	0.252	0.131	0.139	0.253	0.158	0.232	0.277	0.094	0.261	0.226	0.179	0.209	0.280
1	education	0.182	0.137	0.265	0.137	0.140	0.253	0.198	0.271	0.287	0.084	0.262	0.221	0.223	0.250	0.289
1	gender	0.160	0.123	0.259	0.128	0.145	0.253	0.209	0.311	0.291	0.088	0.273	0.227	0.219	0.246	0.285
1	occupation	0.191	0.138	0.258	0.146	0.152	0.250	0.204	0.295	0.288	0.078	0.271	0.233	0.227	0.261	0.292
1	race	0.167	0.138	0.270	0.139	0.179	0.260	0.256	0.384	0.308	0.085	0.280	0.225	0.234	0.263	0.300
5	all together	0.175	0.134	0.261	0.117	0.125	0.247	0.205	0.299	0.297	0.087	0.274	0.223	0.223	0.256	0.294

Table 20: **Cohen κ, Micro Accuracy, and Macro Accuracy for Offensiveness Dataset**. This table lists the baseline, all single-component, and all components together configurations.

Toxicity		Sentiment		Politeness		Offensiveness		Emotion	
Attribute	**SHAP**	**Attribute**	**SHAP**	**Attribute**	**SHAP**	**Attribute**	**SHAP**	**Attribute**	**SHAP**
Religion Importance	0.1889	Living situation	0.1889	Race	0.2255	Age	0.3661	Country	0.2151
Education	0.1495	Household income	0.1459	Age	0.2200	Occupation	0.2165	Age	0.1731
Is Parent	0.1272	Political identification	0.1456	Occupation	0.2186	Race	0.1568	Mother's Occupation	0.1535
Age Range	0.1254	Gender	0.1067	Education	0.1799	Education	0.1546	Religion	0.1421
Gender	0.1250	Age	0.1039	Gender	0.1560	Gender	0.1060	Father's Occupation	0.1267
Race	0.1063	Education	0.0980					Field of Study	0.1047
LGBTQ Status	0.0997	Employment status	0.0833					Sex (Gender)	0.0848
Political Affiliation	0.0779	Race	0.0825						
		Hispanic/Latino	0.0452						

Table 21: SHAP Feature Importance Across All Tasks

degrades every model substantially (Table 13). The positive but non-significant LSVC correlations suggest that learnability and alignment point in the same direction on this task, but the overall weakness of directionally coherent signal limits the practical utility of demographic prompting for Sentiment.

Politeness. Politeness exhibits a split: Gemma (0.400) and Qwen (0.700) show positive LSVC correlations, while DeepSeek (-0.500) is negative. Gemma is the sole model benefiting from demographic prompting, peaking with `occupation` ($\kappa = 0.411$) and `gender` (0.408). `Occupation` has the fourth-highest LSVC κ on Politeness (0.472, Table 22) and a modestly positive Fisher $\bar{\rho}$ ($+0.053$) with a strong dominant-pair correlation of $+0.474$. By contrast, `gender`—the highest-LSVC attribute (0.507)—carries a negative Fisher $\bar{\rho}$ (-0.046; Man vs. Woman $\rho = -0.459$), meaning its high LSVC learnability reflects classifier capacity to exploit the *divergence* between subgroups rather than a coherent shared signal. That Gemma nonetheless benefits from `gender` prompting suggests model-specific factors beyond dataset-side signal also play a role.

Offensiveness. The systematic negative LSVC correlations ($\rho \leq -0.700$ for four of five models) are the most striking result in Table 2. The LSVC ranks `age` highest ($\kappa = 0.316$), followed by `education` (0.310) and `gender` (0.306), yet these are among the attributes that most degrade LLM alignment (e.g., `age` drops Gemma from 0.248 to 0.152). Meanwhile, `race`—ranked lowest by LSVC (0.302)—is the only attribute that significantly improves any model (Qwen: 0.256 vs. 0.205). The Fisher analysis clarifies: `gender` ($\bar{\rho} = -0.182$) is directionally opposed, and `age` ($\bar{\rho} = +0.139$) has moderate coherence that does not translate to persona-exploitable structure. `Race` ($\bar{\rho} = +0.058$) has low overall coherence but may contain specific subgroup pairings that Qwen can exploit. DeepSeek is the exception ($\rho = +0.700$), though this does not correspond to actual alignment gains—DeepSeek's baseline remains its best performance on every task.

Emotion. The dominant alignment attribute (`age`, which lifts Mistral from 0.483 to 0.611 and Qwen from 0.603 to 0.641) ranks only fourth in LSVC accuracy (0.539, Table 22), behind `country` (0.577), `field_of_study` (0.559), and `religion` (0.546). Despite their higher learnability, these top-LSVC attributes produce worse alignment: `country` drops Gemma from 0.726 to 0.474; `field_of_study` drops it to 0.418. Fisher coherence does not resolve this: `country` ($\bar{\rho} = +0.356$) and `field_of_study` ($\bar{\rho} = +0.307$) both exceed `age` ($\bar{\rho} = +0.280$). This discrepancy reinforces the conclusion from the main text: dataset-side signal—even when both learnable and directionally coherent—does not deterministically predict which attributes a specific model architecture can successfully leverage. The negative LSVC correlations for Gemma (-0.523), Llama (-0.667), and Mistral (-0.541) reflect the fact that `age` drives alignment for these models despite not ranking highest on

Toxicity		Sentiment		Politeness		Offensiveness		Emotion	
Attribute	**LSVC κ**	**Attribute**	**LSVC κ**	**Attribute**	**LSVC κ**	**Attribute**	**LSVC κ**	**Attribute**	**LSVC Acc**
Education	0.2007	Employment status	0.2048	Gender	0.5067	Age	0.3164	Country	0.5769
LGBTQ status	0.1999	Gender	0.2027	Race	0.4993	Education	0.3098	Field of study	0.5587
Race	0.1857	Race	0.2019	Education	0.4845	Gender	0.3055	Religion	0.5463
Religion important	0.1773	Hispanic/Latino	0.2005	Occupation	0.4724	Race	0.3023	Age	0.5385
Is parent	0.1764	Education	0.1947	Age	0.4507	Occupation	0.2971	Father occupation	0.5385
Age range	0.1739	Income	0.1946					Mother occupation	0.5372
Political affiliation	0.1608	Age	0.1873					Gender	0.5339
Gender	0.1607	Living situation	0.1755						
		Political id	0.1614						

Table 22: LSVC Performance Across All Tasks

Demographic	Fisher $\bar{\rho}$	Dominant Pair	Dominant ρ
race	0.2503***	Black or African American, Hispanic vs White, Black or African American, American Indian or Alaska Native	0.521***
gender	-0.0869***	Female vs Male	-0.629***
religion_important	-0.0009***	important_Not important vs important_Very important	-0.365***
education	0.1589***	Associate degree in college (2-year) vs Professional degree (JD, MD)	0.489***
lgbtq_status	0.2082***	status_Homosexual vs status_Other	0.409***
political_affilation	0.0019***	affilation_Independent vs affilation_Liberal	-0.399***
is_parent	-0.2028***	parent_No vs parent_Yes	-0.636***
age_range	0.1245***	range_25 - 34 vs range_35 - 44	-0.489***

Table 23: Summary of Spearman Correlations by Demographic Category for **Toxicity dataset**. *Note:* *** $p < .001$, ** $p < .01$, * $p < .05$, ns = not significant.

either LSVC learnability or Fisher coherence.

H Detailed Results for Neuron Probing

H.1 Ablation Study

To determine the optimal threshold for Key Neuron selection, we evaluate the effect of masking neurons at varying top-k values ($k \in 5, 10, 15, 20, 30, 40, 50, 100$) on both in-task performance (Cohen's κ) and out-of-distribution performance (WikiText-2 next-token accuracy (Merity et al., 2016)). Figure 6 shows results for Llama-3.2-3B-Instruct on the Sentiment task, evaluated on all samples. At each threshold, we compare masking the specified top-k neurons (selected by activation magnitude) against masking the same number of randomly chosen neurons. At $k = 10$, masking the specified neurons lowers task Cohen's κ from 0.370 to 0.288, a 22.1% drop. Masking the same number of randomly chosen neurons costs only 1.7%, a separation of 7.1σ. General capability is largely preserved: WikiText-2 next-token accuracy falls by just 4.1%. Random masking at the same scale causes minimal degradation on either metric, verifying that the selected neurons are not arbitrary. At larger k the specified mask also begins to lose general capability, costing 30.0% of WikiText-2 accuracy by $k = 100$. This suggests the mask has grown beyond the task-specific neurons and started removing general-purpose ones. Random masks of the same size remain far less damaging, costing at most 9.5%. We therefore set $k = 10$ for all subsequent experiments, as it captures task-relevant neurons while preserving the model's broader representational integrity.

H.2 Selective Support on Toxicity

The hypothesis for RQ3 receives its clearest support on Toxicity. Table 4 shows statistically significant positive Pearson correlations for Llama ($r = 0.742$) and Qwen ($r = 0.769$): attributes that activate a larger proportion of new/specialized neurons tend to produce better alignment with human toxicity annotations. The per-attribute data (Table 28) make the pattern concrete: for Qwen, `lgbtq_status` and `religion_important` activate the highest specialised-neuron proportions (0.436 and 0.387) and also yield the strongest single-component alignment gains ($\kappa = 0.246$ and 0.242 vs. baseline 0.157, Table 12). For Mistral, `lgbtq_status` similarly activates the most specialised neurons (0.422) and achieves the highest single-component κ (0.316). Crucially, these are precisely the attributes identified in Section 4.2 as carrying directionally coherent lexical signals (Fisher $\bar{\rho} = +0.208$ for `lgbtq_status` and $+0.250$ for `race`, Table 23), suggesting that specialised neuron activation is a meaningful correlate of alignment only when the underlying annotation signal is structurally exploitable.

H.3 The DeepSeek High-Volume Paradox

A striking counter-example emerges from DeepSeek. Across all five tasks, DeepSeek activates substantially more specialised neurons than any other model (see Tables 28–32, Ap-

Demographic	Fisher $\bar{\rho}$	Dominant Pair	Dominant ρ
age	0.0319***	50-59 vs 60-69	-0.475***
race	0.2074***	Black or African American vs Other	0.452***
hispanic_latino	-0.2547***	No vs Yes	-0.255***
income	0.1481***	Less than $10,000 vs More than $200,000	0.386***
education	-0.0216***	Bachelor's degree vs Some college or associate's degree	-0.306***
employment_status	0.0702***	On disability vs Unemployed	0.422***
living_situation	0.0610***	I live alone vs I live with a spouse or romantic partner	-0.390***
political_id	0.0365***	Moderate vs Very conservative	-0.275***
gender	-0.6329***	Female vs Male	-0.633***

Table 24: Summary of Spearman Correlations by Demographic Category for **Sentiment Analysis dataset**. *Note:* *** $p < .001$, ** $p < .01$, * $p < .05$, ns = not significant.

Demographic	Fisher $\bar{\rho}$	Dominant Pair	Dominant ρ
race	0.1440***	Black or African American vs White	-0.502***
gender	-0.0462***	Man vs Woman	-0.459***
age	0.0483***	50-54 vs 54-59	0.243***
education	0.0275***	College degree vs High school diploma or equivalent	-0.429***
occupation	0.0534***	Prefer not to disclose vs Student	0.474***

Table 25: Summary of Spearman Correlations by Demographic Category for **Politeness dataset**. *Note:* *** $p < .001$, ** $p < .01$, * $p < .05$, ns = not significant.

Demographic	Fisher $\bar{\rho}$	Dominant Pair	Dominant ρ
race	0.0580***	Arab American vs Native American	0.439***
gender	-0.1823***	Man vs Woman	-0.319***
age	0.1385***	25-29 vs 60-64	0.305***
education	-0.0534***	Less than a high school diploma vs Other	0.212**
occupation	0.0967***	Other vs Prefer not to disclose	0.417***

Table 26: Summary of Spearman Correlations by Demographic Category for **Offensiveness dataset**. *Note:* *** $p < .001$, ** $p < .01$, * $p < .05$, ns = not significant.

Demographic	Fisher $\bar{\rho}$	Dominant Pair	Dominant ρ
AGE	0.2804***	28.0 vs 31.0	0.506***
sex_map	-0.0874***	female vs male	-0.327***
reli_map	0.2178***	native vs others	0.544***
focc_map	0.2647***	Unknown vs unemployed	0.511***
mocc_map	0.2647***	Unknown vs self-employed nonacademic	0.444***
fiel_map	0.3071***	natural science vs other	0.517***
coun_map	0.3557***	australia vs usa	0.548***

Table 27: Summary of Spearman Correlations by Demographic Category for **Emotion dataset**. *Note:* *** $p < .001$, ** $p < .01$, * $p < .05$, ns = not significant.

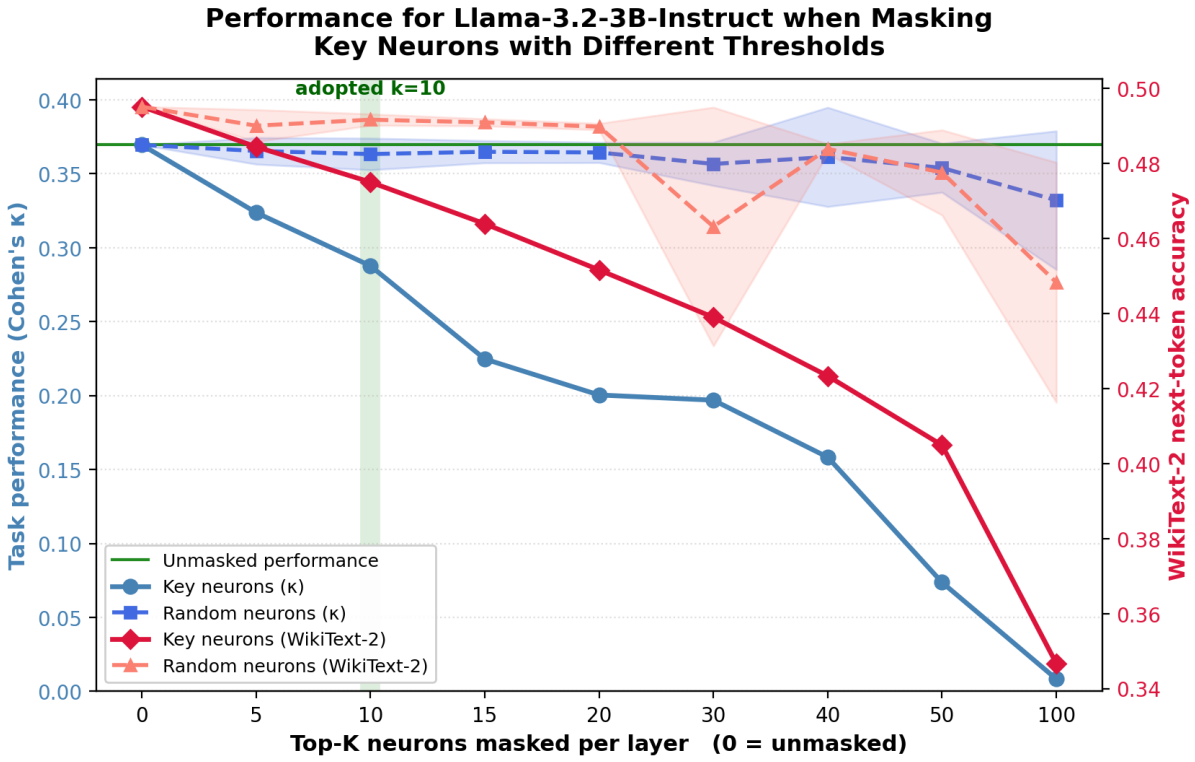

Figure 6: Ablation study for the Sentiment Task

pendix). If the specialised-neuron hypothesis held universally, DeepSeek should be the most demographically steerable model. The alignment results show precisely the opposite: DeepSeek is the only model whose no-demographic baseline is its best configuration on every task, and no single-component prompt produces a statistically significant improvement over baseline on any task.

We term this the *high-volume paradox*. One interpretation, consistent with the mechanistic interpretability literature (Cao et al., 2025), is that activation quantity does not index the quality or relevance of knowledge engaged: DeepSeek's high neuron volumes may reflect broad, undifferentiated representational perturbation in response to persona injection rather than targeted engagement with annotation-relevant knowledge.

H.4 Negative Correlations and Task-Dependent Breakdown

Table 4 reveals that the neuron–performance correlation is significantly *negative* for several model–task combinations. On Sentiment, Gemma ($r = -0.888$) and Mistral ($r = -0.727$) both show strong, statistically significant negative correlations; on Emotion, Gemma reaches $r = -0.983$, the largest-magnitude value in the table and Llama shows $r = -0.772$. The structural explanation connects directly to the dataset-level analysis in Section 4.2. For Gemma on Sentiment, the attributes activating the *most* specialised neurons `political_id` (0.288) and `hispanic_latino` (0.262) also produce the largest alignment degradations: `political_id` drops Gemma's κ from 0.388 to 0.300. Conversely, `gender`, which activates the *fewest* specialised neurons for Gemma on this task (0.183), causes the smallest degradation ($\kappa = 0.366$). The key factor is signal coherence: `gender` on Sentiment carries the most severely opposed annotation signal in the study (Fisher $\bar{\rho} = -0.633$), meaning additional neuron activation engages a structurally irreconcilable subgroup divergence that no single persona can resolve. The Emotion results follow the same logic: Gemma's strong baseline (0.726, the highest on that task) is disrupted by the demographic attributes that activate the most new neurons `field_of_study` (0.696) and `country` (0.671) produce the steepest accuracy drops, to 0.418 and 0.474 respectively. High specialised-neuron activation here reflects interference with an already well-calibrated representation, not useful knowledge engagement.

H.5 Gemma's Politeness Exception

Gemma is the only model that benefits from demographic prompting on Politeness, with `occupation` ($\kappa = 0.411$), `gender` (0.408), and `race` (0.407) all exceeding the baseline (0.372). Yet the Pearson correlation for Gemma on Politeness is negative ($r = -0.824$), and Table 30 (Appendix) shows that `occupation` (the best-performing attribute) activates the *fewest* specialised neurons for Gemma (0.595), while `age` activates the most (0.633). The resolution again lies in signal quality: `occupation` carries the most directionally coherent Fisher-averaged signal on this task ($\bar{\rho} = +0.053$, dominant-pair $\rho = +0.474$), while `age` ($\bar{\rho} = +0.048$) and `race` ($\bar{\rho} = +0.144$ but dominant Black vs. White pair $\rho = -0.502$) carry more structurally conflicted signals. Specialised neuron count is once more a poor proxy for alignment utility when signal coherence is not accounted for.

Component	Gemma	Llama	Qwen	DeepSeek	Mistral
age_range	0.4413	0.4361	0.3647	0.7450	0.3759
education	0.4163	0.4450	0.3737	0.7550	0.3820
gender	0.4471	0.4385	0.3691	0.7508	0.3725
is_parent	0.4270	0.4279	0.3634	0.7820	0.3788
lgbtq_status	0.4317	0.4834	0.4360	0.7312	0.4220
political_affilation	0.4273	0.4351	0.3678	0.7003	0.3814
race	0.4384	0.4908	0.3764	0.7166	0.3836
religion_important	0.4425	0.4456	0.3879	0.7953	0.3801

Table 28: Specialized Neurons across Components and Models for **Toxicity**

Component	Gemma	Llama	Qwen	DeepSeek	Mistral
age	0.2127	0.3072	0.3075	0.7402	0.3384
education	0.2157	0.3115	0.3006	0.7486	0.3317
employment_status	0.2190	0.3171	0.3099	0.7607	0.3413
gender	0.1827	0.3053	0.2897	0.7211	0.3312
hispanic_latino	0.2617	0.3159	0.3100	0.6928	0.3484
income	0.2560	0.3153	0.3202	0.7944	0.3389
living_situation	0.2118	0.3142	0.3070	0.7563	0.3396
political_id	0.2880	0.3358	0.3148	0.6979	0.3528
race	0.2191	0.3266	0.3144	0.7414	0.3486

Table 29: Specialized Neurons across Components and Models for **Sentiment**

Component	Gemma	Llama	Qwen	DeepSeek	Mistral
age	0.6328	0.5247	0.4606	0.6955	0.4504
education	0.6243	0.5169	0.4506	0.6946	0.4476
gender	0.6079	0.5327	0.4562	0.6993	0.4498
occupation	0.5952	0.5108	0.4439	0.6916	0.4558
race	0.6216	0.5529	0.4524	0.6964	0.4500

Table 30: Specialized Neurons across Components and Models for **Politeness**

I Refusal

Table 33 reports task-wise refusal rates averaged across all 3 prompting templates. Refusal patterns

Component	Gemma	Llama	Qwen	DeepSeek	Mistral
age	0.4722	0.4131	0.3765	0.7188	0.4242
education	0.4468	0.4044	0.3849	0.7175	0.4183
gender	0.4546	0.4108	0.3837	0.7232	0.4216
occupation	0.4458	0.3967	0.3706	0.7048	0.4315
race	0.4723	0.4442	0.3844	0.7058	0.4333

Table 31: Specialized Neurons across Components and Models for **Offensiveness**

Component	Gemma	Llama	Qwen	DeepSeek	Mistral
age	0.5734	0.5384	0.3666	0.7031	0.4510
country	0.6714	0.5800	0.3382	0.7092	0.4585
father_occupation	0.6665	0.5577	0.3831	0.6953	0.4613
field_of_study	0.6959	0.5652	0.3596	0.6888	0.4580
gender	0.5828	0.5557	0.3638	0.6837	0.4524
mother_occupation	0.6251	0.5366	0.3661	0.6996	0.4544
religion	0.6163	0.5664	0.3578	0.7024	0.4612

Table 32: Specialized Neurons across Components and Models for **Emotion**

vary substantially across models and tasks. Llama-3.2-3B exhibits the highest refusal rates across all five tasks, peaking at 4.01% on Toxicity, likely due to its safety-tuning behavior/conservative guardrails at smaller model scale. In contrast, Qwen2.5-7B produces the fewest refusals, remaining below 0.11% on every task, suggesting more permissive output behavior under persona prompting. Across tasks, Toxicity and Offensiveness elicit the most refusals, consistent with the sensitive nature of these classification targets models are more likely to decline rating content that contains potentially harmful language. Sentiment and Emotion produce the lowest refusal rates, reflecting the comparatively neutral framing of these tasks. These refusals were excluded from all alignment calculations reported in the main text; including them as incorrect predictions does not change the relative ordering of model performance.

J Practical Suggestions for Demographic Prompting

Our findings reveal that demographic prompting is not a one-size-fits-all strategy: its effectiveness depends on the interaction between model architecture, task characteristics, and the structural quality of the demographic signal being prompted. Below, we distill actionable guidance organized by task and model, grounded in the empirical patterns from Sections 4.1 through 4.3.

Toxicity detection is the most favorable setting for demographic prompting, but attribute selection is critical. Across four of five models, compact demographic prompts (one to three attributes)

Model	Toxicity	Sentiment	Politeness	Offensiveness	Emotion
Gemma3-12B	1,204 (1.12%)	142 (0.58%)	387 (1.55%)	892 (0.83%)	98 (0.41%)
Llama-3.2-3B	4,312 (4.01%)	876 (3.61%)	621 (2.48%)	2,147 (3.29%)	412 (1.72%)
Qwen2.5-7B	83 (0.08%)	12 (0.05%)	27 (0.11%)	54 (0.08%)	8 (0.03%)
DeepSeek-R1-7B	1,847 (1.72%)	324 (1.34%)	498 (1.99%)	1,103 (1.69%)	215 (0.90%)
Mistral-7B	956 (0.89%)	187 (0.77%)	312 (1.25%)	645 (0.99%)	104 (0.43%)

Table 33: Task-wise refusal rates for LLMs, averaged across all prompting templates.

improve alignment over the no-demographic baseline on toxicity (Table 12). Users seeking to improve LLM–human alignment on toxicity tasks should prioritize attributes with high directional coherence, such as `lgbtq_status` ($\bar{\rho} = +0.208$) and `race` ($\bar{\rho} = +0.250$), which anchor the top-performing configurations for Llama, Qwen, and Mistral. Conversely, attributes with high SHAP importance but low coherence, such as `religion_important` (Fisher $\bar{\rho} \approx 0$), should be used cautiously despite their strong influence on human label variation. Users should avoid prompting with the full attribute set, as the 8-component configuration never outperforms the best compact prompt and often degrades performance significantly.

For sentiment analysis, demographic prompting offers limited and model-specific benefits; a conservative approach is advisable. Sentiment presents weak and inconsistent gains from demographic prompting (Figure 2). For Gemma and Mistral, no demographic prompt surpasses baseline, so users of these models should default to unprompted classification. Qwen users may see modest improvement with `education` + `hispanic_latino`, and Llama users with `gender` alone, but these gains are small. The severely opposed signal carried by `gender` on this task ($\bar{\rho} = -0.633$) means that gender-based persona prompts are structurally unlikely to help, despite the intuitive appeal of gendered perspectives on sentiment. Users working on sentiment tasks should generally avoid multi-attribute prompts, as the "all together" configuration degrades every model substantially.

Politeness and offensiveness tasks are broadly resistant to demographic prompting; users should prefer unprompted baselines unless model-specific exceptions apply. On politeness, four of five models perform best at baseline, with demographic attributes often causing significant degradation (e.g., `race` drops Llama's κ from 0.349 to 0.252). The sole exception is Gemma, where single-attribute prompts such as `occupation` and `gender` improve alignment. Users of Gemma on politeness tasks may therefore experiment with these attributes individually, but should avoid combining them, as multi-attribute configurations erode the gains. On offensiveness, `race` is the only attribute that selectively improves alignment (for Qwen and Mistral), while `age` and `gender`, despite being the most learnable attributes by LSVC κ, consistently degrade performance due to their opposed subgroup signals. Users should treat high learnability as a necessary but insufficient condition: an attribute that a classifier can exploit may still be unusable by a persona prompt if its subgroups disagree directionally.

On emotion attribution, `age` is the dominant beneficial attribute, but only for select models. Mistral and Qwen benefit substantially from `age`-anchored prompts on the emotion task (Mistral improves by 12.9 percentage points; Qwen by 3.8 pp), making `age` the recommended single attribute for users of these models on emotion-related tasks. Llama users may achieve smaller gains through specific 4- and 5-way combinations anchored on `age`, `gender`, and parental occupations, though the improvement over baseline is marginal. Users of Gemma and DeepSeek should avoid demographic prompting entirely on this task, as every demographic configuration degrades their strong baselines without exception.

DeepSeek shows limited demographic steerability across most tasks. DeepSeek's no-demographic baseline is the best-performing configuration on four of the five tasks examined. Sentiment is the exception, where a specific five-attribute combination significantly improves alignment over baseline ($\kappa = 0.291$ vs. 0.265). The high-volume paradox identified in our neuron probing analysis (Section 4.3) nevertheless suggests that DeepSeek's broad neuron activation in response to persona cues does not reliably translate into improved demographic alignment. This may partly reflect the distillation process (see Limitations). Thus, demographic prompting with DeepSeek-R1-Distill-

Qwen-7B should be used cautiously and validated on a task-specific basis rather than assumed to improve alignment.

When in doubt, prefer fewer attributes and verify signal quality before prompting. Our RQ1 results suggest that compact prompts are often preferable when demographic prompting is beneficial, although the optimal number of attributes varies across models and tasks. Across all model–task pairs, the full-attribute configuration is never the best-performing configuration. Users who must decide which attributes to include should, where feasible, assess (1) whether the attribute's subgroups exhibit directionally coherent annotation signals (positive Fisher $\bar{\rho}$), and (2) whether the attribute contributes learnable lexical signal (high LSVC κ). Attributes satisfying both conditions (e.g., `lgbtq_status` and `race` on toxicity) are the strongest candidates. Attributes with high learnability but opposed signals (e.g., `gender` on sentiment or offensiveness) should be excluded, as they create structurally irreconcilable demands that no single persona prompt can satisfy. In the absence of dataset-level diagnostics, defaulting to the unprompted baseline is the safest strategy.