Robustness and Confounders in the Demographic Alignment of LLMs with Human Perceptions of Offensiveness

Shayan Alipour[1], Indira Sen[2], Mattia Samory[1], and Tanushree Mitra[3]
[1] Sapienza University of Rome, [2] University of Mannheim, [3] University of Washington
shayan.alipour@uniroma1.it, indira.sen@uni-mannheim.de,
mattia.samory@uniroma1.it, tmitra@uw.edu

Abstract

Despite a growing literature finding that large language models (LLMs) exhibit demographic biases, reports with whom they align best are hard to generalize or even contradictory. In this work, we examine the alignment of LLMs with human annotations in five offensive language datasets, comprising approximately 220K annotations. While demographic traits, particularly race, influence alignment, these effects vary across datasets and are often entangled with other factors. Confounders introduced in the annotation process—such as document difficulty, annotator sensitivity, and within-group agreement—account for more variation in alignment patterns than demographic traits. Alignment increases with annotator sensitivity and group agreement, and decreases with document difficulty. Our results underscore the importance of multi-dataset analyses and confounder-aware methodologies in developing robust measures of demographic bias.

1 Introduction

A growing body of literature explores LLMs as a quick, inexpensive, and reliable alternative to human annotators (Chiang and Lee, 2023; Törnberg, 2023; Zhu et al., 2023; Gilardi et al., 2023). The eventuality of annotating data with LLMs is more than speculation: crowdworkers, who are often key to data annotation, already rely on LLMs for efficiency (Veselovsky et al., 2023). Since data annotations are a primary concern for any machine learning application, it is essential to assess the quality of LLM-generated annotations.

In particular, tasks that reflect annotators' subjectivity, such as the perceived offensiveness of a message (Davani et al., 2023), raise the question of which subjectivities are reflected in LLM-generated annotations. Recent research finds evidence of demographic bias, that is, systematic alignment between LLMs' annotations and those from select demographic groups of human annotators. If LLMs replicate the views of one demographic group over others, downstream applications risk perpetuating structural harms like marginalizing minority views.

Although LLMs' demographic bias has been identified for a variety of subjective constructs, including resumé screening (Wilson and Caliskan, 2024; Dammu et al., 2024), healthcare (Jiang et al., 2024; Zack et al., 2023), political opinions (Motoki et al., 2024), and offensiveness (Sun et al., 2023; Santy et al., 2023), we know little about whether such bias is consistent, since many of these studies focus on different NLP tasks and datasets. Even within the same dataset and task, the identified biases are unclear or even seemingly contradictory. For example, for offensiveness annotations in the POPQUORN dataset, Sun et al. find that LLMs align most with white and female annotators, while Schäfer et al. do not find such alignment. Using a different dataset, Santy et al. find alignment is highest with Asian Americans. Understanding which demographic biases are consistent in LLM annotations is fundamental to tackling them. Our work fills this gap with a systematic study of LLM alignment on offensiveness labeling with different genders and ethnicities across five datasets. We further investigate factors beyond annotator demographics that might drive human-LLM misalignment.

First, to exclude that LLM bias may simply be attributed to low performance, we verify *RQ1: to what extent LLMs can substitute human annotators in detecting offensive language*. Indeed, LLMs are strong performers: correlations with aggregate human labels are positive and significant—ranging from 0.2 to 0.8. Through permutation and bootstrapping tests, we show that LLMs match human performance in all datasets and surpass it in three.

Next, we test *RQ2: which demographic biases are consistently reproduced in LLM-generated annotations across datasets.* Demographic biases exist within each dataset; however, most of these

Findings of the Association for Computational Linguistics: ACL 2025, pages 22025–22047
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

biases lack consistency. We find that only the difference in bias between the White and Black demographics remained consistent across all five datasets. Other biases appeared inconsistently or even showed opposite results across datasets.

To unpack such differences between datasets, we therefore explore *RQ3: to what extent confounding factors explain demographic bias*. We consider three hypotheses about the dataset annotation process: HPa) documents that are difficult to annotate may be assigned unevenly across demographics; HPb) annotators may show strong individual rather than demographic preferences in their annotation; and HPc) the disagreement between annotators of a same demographic group may affect the measures of alignment. We find that confounders explain a large fraction of the variance of LLM–human alignment, and thus can help us unpack cases when demographic bias is inconsistent.

Overall Contributions and Novelty. Prior studies on human-LLM alignment rely on single datasets, leading to non-generalizable findings. We conduct the first large-scale, multi-dataset analysis across five NLP datasets and uncover inconsistent alignments, despite prior claims (Sun et al., 2023; Santy et al., 2023). However, we show that demographic factors alone do not explain misalignment. Through a robust regression framework, we identify key confounders that drive alignment patterns, a crucial aspect overlooked in past research. Finally, we provide actionable recommendations for bias estimation and contribute a harmonized dataset of 220k offensive language annotations, enabling more rigorous bias assessment in future studies.[1]

2 Related Work

Our work assesses one aspect of alignment—demographics—within the context of LLM labeling. Therefore, it lies at the intersection of using LLMs for annotation, annotation subjectivity, and demographic bias in LLMs.

2.1 Using LLMs for Data Annotation

Recent work has focused on using generative LLMs, such as Flan-T5 and GPT for data labeling for various social constructs like offensive language (Zampieri et al., 2023), stance detection (Aiyappa et al., 2024), hate speech (Huang et al., 2023), and framing (Gilardi et al., 2023). However, there is contention about the quality of LLM-generated annotations. Since ChatGPT's release, some studies have claimed it outperforms human annotators (Gilardi et al., 2023; Wu et al., 2023; Chiang and Lee, 2023; Törnberg, 2023; Zhu et al., 2023), while others find LLMs do not reach human-level performance (Kristensen-McLachlan et al., 2023).

2.2 Subjectivity and Demographic Factors in Human Annotations

Ground-truth labels for an instance are often aggregated from the ratings from multiple annotators. Yet, annotation is often an interpretive task that depends on the annotator's positionality, social situation, and lived experiences (Paullada et al., 2021; Santy et al., 2023), which challenges the assumption of the existence of such a single true label. Recent work in NLP has recognized that such aggregation can squash the opinions and views of marginalized populations (Davani et al., 2022, 2023). This is particularly pertinent for subjective tasks such as detecting offensive, abusive, or toxic content. Here, an annotator's demographic identity (Al Kuwatly et al., 2020), attitudes (Sap et al., 2021), and personal experiences (Sang and Stanton, 2022) affect the perception of toxicity—together with artifacts of the annotation process such as the annotation instructions and interface (Kern et al., 2023). To account for this variance in annotation distributions, researchers recommend explicitly factoring in dimensions that would lead to disagreement, even *before* the annotation task (Fleisig et al., 2024).

2.3 Demographic Alignment of LLMs for Annotation

A growing body of work examines LLM alignment with human annotators in content analysis (Sun et al. 2023; Santy et al. 2023, inter alia; see Table 3). Default prompting is particularly relevant for real-world applications, where demographic information is rarely available when deploying LLMs for content analysis. Unlike sociodemographic prompting, which artificially steers model behavior, default prompting reveals biases that emerge naturally. Prior work has explored the effects of explicit demographic cues (Beck et al., 2023; Sun et al., 2023; Schäfer et al., 2024), but most LLM benchmarking studies use default prompting (Gilardi et al., 2023; Ziems et al., 2024). By focusing on this setup, we ensure our findings reflect how LLMs align with human annotators in practice

[1] https://github.com/shayanalipour/llm-alignment-bias

while validating patterns across datasets.

Most work assesses default prompting on one dataset, often different from those of comparable works, leading to unclear overall findings. For example, for offensiveness and politeness, Sun et al. find that LLMs align best with White women. Schäfer et al. use the same dataset and find that LLMs align better with White people, but not women. Differing from both of these, Santy et al. find that for hate speech detection, models align best with Asian Americans. Our work unpacks these seemingly contradictory findings by conducting a systematic study across several datasets.

Closest to our work, Hu and Collier tests the demographic alignment of multiple LLMs across different datasets and tasks. The present work takes this direction further and tests the *robustness* of such alignment. Like Hu and Collier, our work shows that beyond demographic variables, the characteristics of the annotation process itself are correlated with model alignment: we lay out confounding factors in the annotation process that may explain inconsistencies in existing literature.

3 Data

To fully understand how LLM annotations align with human opinions, we leverage five datasets encoding annotators' perceptions of offensiveness—a construct that has been proven to vary according to annotators' sociodemographic characteristics. Specifically, we use the following datasets: Annotator with Attitudes (AwA) dataset (Sap et al., 2021), UC Berkeley's Measuring Hate Speech Corpus (MHSC) (Kennedy et al., 2020), NLPositionality (NLPos) dataset (Santy et al., 2023), POPQUORN dataset (POPQ) (Pei and Jurgens, 2023), and Social Bias Inference Corpus (SBIC) (Sap et al., 2020).[2] For AwA, MHSC, and SBIC, participants were recruited from Amazon Mechanical Turk, while for POPQ, they were recruited through Prolific. Only Sap et al. (SBIC) mention that they restricted the annotator pool to people from the US and Canada, while Santy et al. (NLPos) explicitly recruited a diverse pool of annotators from over 80 countries.

We use the annotator backgrounds in these datasets to consider the potential alignment between LLMs' responses and those of annotators in specific demographic groups (Table 1). Following past research (Sap et al., 2021, 2019), we focus on annotators' ethnicity and gender as major factors in demographic alignment. Although age is a sociodemographic trait present in all of the datasets included in this study, gender and ethnicity have the advantages that they are coded harmonically across datasets, and that their empirical distributions are such that they are likely to provide sufficient statistical power for analyses. More details about the datasets can be found in the appendix A.

Demographic	AwA	MHSC	NLPos	POPQ	SBIC
Man	56.32	43.07	43.02	48.53	48.08
Woman	43.68	56.93	56.98	51.47	51.92
Asian	-	5.95	20.92	7.85	6.73
Black	33.83	8.64	7.34	13.11	4.10
Hispanic	-	4.01	9.40	-	6.54
White	66.17	81.40	62.35	79.04	82.63

Table 1: Demographic distributions across five datasets shown as percentages.

4 Methods

Next, we outline the models used for annotation, explain our approach to prompting, and describe how we operationalize our research questions.

4.1 Models

We conduct our experiments with three state-of-the-art models — GPT4o mini (Achiam et al., 2023), Gemini 1.5 Flash (Team et al., 2024), and Solar-10.7B-Instruct (Kim et al., 2023) — to identify consistent trends in detecting offensive and hateful language and explore any inter-model variations. Recent studies show that while GPT models perform well on hate speech detection (Huang et al., 2023), they may exhibit inherent demographic biases (Zack et al., 2024; Wang et al., 2023; Tao et al., 2023). Similarly, Gemini 1.5 Flash, a distilled version of Gemini 1.5 Pro, has been reported to achieve near parity with OpenAI's models across many benchmarks (Team et al., 2024). However, extensive studies, similar to those conducted for GPT models, have not been carried out for open-source models to evaluate their suitability for text annotation tasks. Therefore, we include in our analysis Solar-10.7B-Instruct, an open-source model

[2] Three of the five datasets (AwA, POPQ, and SBIC) focus on offensiveness, while the remaining two datasets (NLPos and MHSC) focus on hate speech. Past research has looked into the association between offensiveness language and hate speech, concluding that both constructs, while not identical, are often similarly perceived (Davidson et al., 2017; Founta et al., 2018; Fortuna and Nunes, 2018). Offensive language can be considered a superset of hate speech, where the latter is offensive language targeting protected groups or minorities.

praised for its performance and reportedly capable of outperforming larger models, including Mistral 8X7B (Kim et al., 2023).

4.2 Prompting Strategies

We designed a scoring system that matches the questions asked to human annotators across various datasets for evaluating offensive or toxic content. Following the approach in Wei et al., we required the model to not only rate the comments but also justify its scores. The wording of the prompts was the same as that used for human annotators, with an additional instruction: “Begin your response by mentioning one of the valid options, then provide a concise explanation for your rating.” Our prompts are included in Table 10 in the appendix. We used regular expressions to extract final labels from the models’ responses.

4.3 RQ1: LLMs as Annotators

To evaluate LLM performance in offensive language detection, we compared their labels with human annotators’ ground-truth labels, calculated using two methods: rounded average and majority. We report results using rounded average, while the majority-label results are detailed in the appendix. We calculated the Pearson correlation coefficients between the model and human labels, performed t-tests to assess statistical significance, and calculated 95% confidence intervals (CI) using bootstrapping with 1,000 samples. To further test the robustness of the correlations, we ran permutation tests by shuffling demographic information and recalculating correlations for 1,000 iterations.

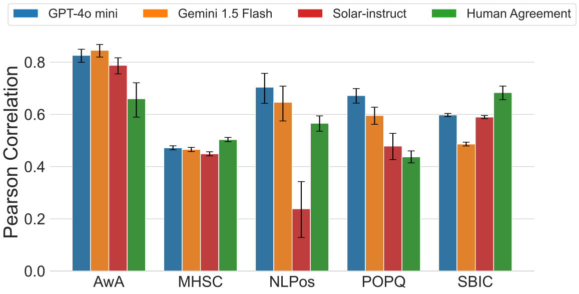

Figure 1: Comparison of model correlations with human annotators against human agreement

To measure agreement among human annotators, we used a leave-one-out approach. For each annotator, we excluded their label from each post and recalculated the ground truth from the remaining labels. This was repeated for every post that the annotator labeled. We then measured the correlation between the annotator’s labels and the recalculated ground truth. By averaging these correlations across all annotators, we determined overall human agreement. To evaluate the robustness, we applied bootstrapping (1,000 samples) to this correlation distribution to estimate 95% confidence intervals.

4.4 RQ2: Demographic Bias Robustness

To analyze demographic biases in LLM-generated annotations, we calculate the ground truth for each demographic by filtering the annotations for that group and aggregating them per post. We then compute the Pearson correlation r between the model’s predictions and the demographic-specific ground truth. We also apply the aforementioned robustness checks, including t-tests, confidence interval estimation using bootstrapping, and permutation tests on the demographic labels.

To assess whether the model consistently aligns better with one demographic than another, we use Steiger’s Z test (Steiger, 1980; Hoerger, 2013) to determine if the difference between correlations Δr is statistically significant. To account for multiple comparisons, we adjust p-values using the Holm-Bonferroni correction. Additionally, we compute confidence intervals for the difference in correlations using bootstrapping. In particular, we measured $\Delta r = r(P, D_1) - r(P, D_2)$ where P represents the model’s predictions, D_1 and D_2 are two demographic groups. This involves resampling the annotations for each demographic pair and recalculating the difference over 1,000 iterations. If the 95% CI for the bootstrapped distribution includes zero, this suggests that the observed difference in correlations may be due to random variation in the sample distribution.

4.5 RQ3: Demographic Bias Confounders

We consider alternative hypotheses for the alignment between the LLMs and humans, beyond demographic bias. We use individual annotations as observations and operationalize **alignment** as an indicator variable set to 1 when LLM and human annotations coincide. While considering exact alignment does not capture the direction and magnitude of the differences in annotations, this operalization is practically useful since alignment is high overall (~60% of annotations align perfectly). We model alignment via logistic regression with annotators’ demographic traits as independent variables. Next, we develop additional hypotheses for factors

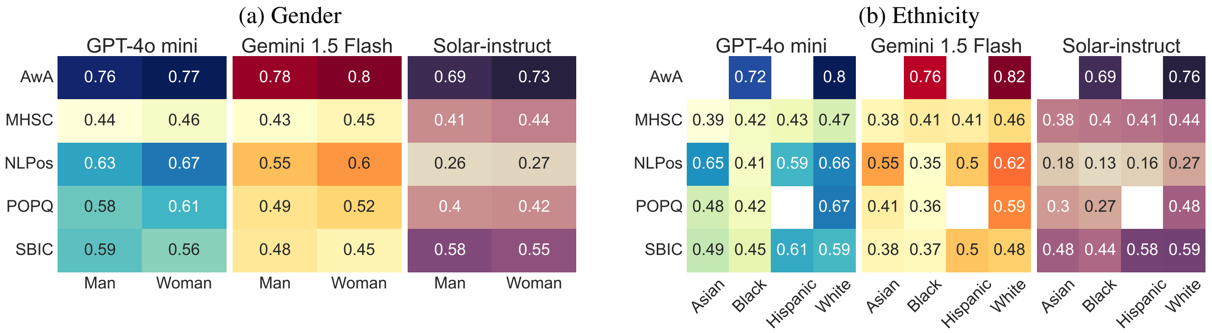

Figure 2: Pearson correlation coefficients between model outputs and human annotator labels, broken down by gender (a) and ethnicity (b) across five datasets. The ground truth for each post is determined by averaging the labels from annotators belonging to the target demographic. Darker shades indicate stronger correlations. Confidence intervals and p-values for statistical significance are reported in Table 6 in the appendix.

that may confound demographic alignment, which we include as additional independent variables.

HPa: difficulty. Documents that are difficult to annotate may be assigned unevenly across demographics, which may in turn negatively affect LLM's alignment. We measure difficulty as the negative Kullback-Leibler divergence between a document's labels and the uniform distribution. Intuitively, the more diverse the annotators' labels, the more difficult the document is to annotate.

HPb: sensitivity. Since few annotators partake in most annotation tasks, some annotators contribute more labels than others, and the representation of demographic traits is unequal, individual annotator factors may dominate the apparent demographic biases. At its simplest, some annotators may systematically label documents as more offensive than other annotators, irrespective of their demographic or a document's aggregate label. To measure sensitivity, we rank annotators of a document based on their labels. The higher the annotators' rank, the more likely they align with LLMs tuned to discourage offensive content.

HPc: agreement. Although alignment is typically measured with a whole demographic group of annotators, different groups may have varying levels of internal agreement. When a group internally disagrees, alignment with it is more complex (and arguably, less meaningful). We operationalize agreement as the negative absolute difference between the individual annotators' labels and the average label of their demographic groups, therefore computing distinct agreement values for gender and ethnicity. The more annotators behave similarly to their reference group, the higher the agreement.

In addition to confounders, we include the document's offensiveness **label** as a control variable to account for skews in the LLMs' annotations, since we aim at measuring whether LLMs replicate demographics' annotations of individual documents rather than generic similarity in label distributions. LLMs may prefer to label documents as offensive, given their terms of service and training for general, safe-for-work applications. We also control for dataset-specific levels of the dependent and independent variables by including corresponding intercepts. We center all binary variables, and standardize confounders and labels within each dataset by centering them and dividing their values by two standard deviations to make their scales comparable and interpretable (Gelman, 2008).

5 Results

In this section, we address three questions about the capabilities of LLMs in simulating human judgment when annotating offensive language. First (**RQ1**), we examine the extent to which LLMs can accurately replicate human annotations overall. Second (**RQ2**), we investigate the alignment between these models and the annotations of sociodemographic subgroups of annotators and whether these alignments are consistent across datasets. Third (**RQ3**), we model the alignment behavior by considering potential confounding factors.

5.1 RQ1: Viability of LLMs as Annotators

Our results demonstrate that LLMs closely mirror human annotations. Figure 1 shows that the correlations between LLM labels and ground truth labels are strong, positive, and significant, measured using both rounded averages and majority votes.

To put the models' performance into perspective, we compare it with the performance that individual human annotators achieved compared to the remaining annotators. In three datasets—AwA, POPQ, and NLPos—LLMs surpass individual human annotators. In the other two datasets (SBIC and MHSC), LLMs perform competitively. The lower performance of the SOLAR model on the NLPos dataset can be attributed to the dataset's inherent difficulty (e.g., sarcam, inclusion of controversial and implicit hate examples) and the significantly smaller size of the SOLAR model compared to other models, which arguably limits its ability to handle such nuanced tasks.

5.2 RQ2: Robustness of Demographic Bias

Figure 2 shows the correlation r between LLM labels and the labels of annotators from each demographic group. In Figure 2.a, all three models show similar patterns as they align more closely with women's annotations in the AwA, MHSC, NLPos, and POPQ datasets, while aligning better with men's annotations in SBIC. For ethnicity, Figure 2.b shows that models align better with the White demographic, except in SBIC, where GPT-4o and Gemini achieve a higher r with the Hispanic demographic. Across datasets, the Black demographic generally shows the lowest correlations, except in MHSC, where it surpasses the Asian demographic. The correlations for high performing models, as summarized in Table 6, are significant and not due to chance, as confirmed by permutation tests (see Figures 6 and 9 in the appendix).

However, echoing Movva et al., statistical significance alone does not guarantee consistent alignment across demographic groups. To assess whether these differences are robust, we examine the correlation differences using Steiger's Z test and bootstrapping. Without this additional analysis, there is a risk of over-interpreting the correlations, which might reflect dataset-specific variations rather than true alignment. Figure 3 displays the results for demographic pair comparisons when using the rounded average aggregation method. Some comparisons reveal no significant or robust differences. For example, in the POPQ dataset, the correlation difference between men and women changes sign depending on the sample. We observe that the models align slightly better with the Hispanic demographic than with the White demographic in SBIC, but this trend reverses in MHSC and NLPos. The only consistent finding across all datasets is the Black-White pair, where models show lower correlations with the Black people.

Results from Figure 3 highlight a general lack of robustness indicating that demographic annotations are influenced by many factors beyond demographic identity, such as individual interpretation or dataset composition. While the correlation values r between models and annotator groups are statistically significant, testing the robustness of correlation differences shows that demographic identity alone does not consistently explain the observed variance in model alignment.

We also examine intersectional cases, considering race and gender together. For brevity, we present these findings in the appendix (Figures 13 and 14). The overall patterns remain consistent with our main conclusions. Specifically, in RQ2, we observe higher variability when considering race and gender jointly. This variability suggests that LLMs do not align consistently with a single race-gender demographic across all datasets.

5.3 RQ3: Significant Confounders of Demographic Alignment

We now explore potential confounders that may account for inconsistencies in RQ2. Table 2 shows the summary statistics of the logistic regressions of the alignment between LLM and human labels (1 only if they use the same label). The first regression (on the left in the table) explains the relationship between LLM alignment and the annotator's gender and ethnicity, controlling for per-dataset differences. As for our previous analyses, alignment is higher for White than for Black annotators, with marginally significant differences between women and men. This demographic-only model explains little variance in the LLM-human alignment (pseudo-R^2 = 0.015). The second model (on the right in the table) attempts to explain the remaining variance by accounting for confounding factors. Indeed, modeling confounders substantially improves model fit (pseudo-R^2 = 0.213). While most confounders show statistically significant coefficients, the Hispanic demographic stands as an exception. This insignificance may partially stem from the Hispanic demographic not being present in all datasets, which could dilute the coefficient's impact. This behavior aligns with our broader findings that sociodemographic strata, including ethnicity, may not be the primary sources of bias once confounders and between-dataset variations are accounted for.

First, the more difficult it is to label a document,

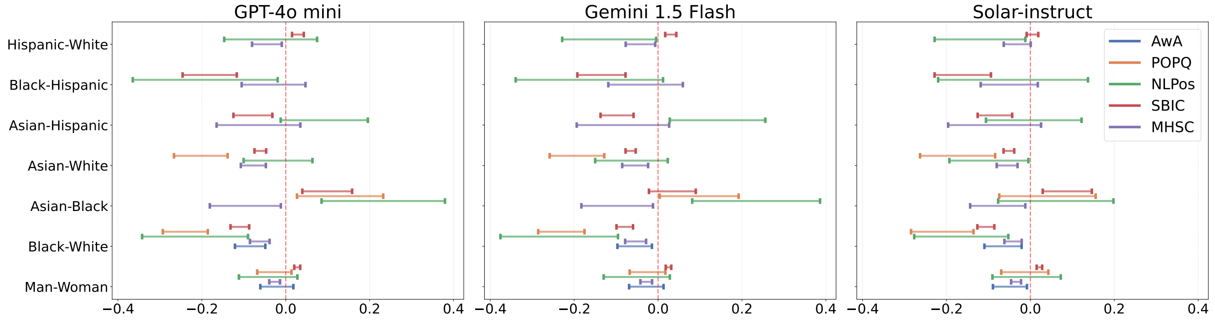

Figure 3: The 95% confidence intervals (CI) for the difference in correlation between the model's predictions and two demographic groups, computed as: $\Delta r = r(P, D_1) - r(P, D_2)$, where P represents the model's predictions, and D_1 and D_2 are two demographic groups. The intervals are derived from 1,000 bootstrap samples. If the CI includes zero, the difference is not statistically significant. See Table 8 in the appendix for further details.

i.e., the more diversity in the annotators' labels regardless of their demographics, the lower the alignment with the LLM. Conversely, the higher the agreement between annotators of the same demographic on a document, the higher the likelihood of alignment. Additionally, the more sensitive one annotator is to offensiveness compared to other annotators of the same document, the higher the alignment, which remains true even when controlling for the overall label of the document, irrespective of the annotators' demographics. In fact, the largest coefficients are associated with confounders at the level of the document—its label and annotation difficulty—which do not directly model the preference of individuals or groups of annotators.

Yet, these confounders do not fully mediate demographic alignment: when explicitly modeled, we see increased rather than decreased significance of demographic coefficients. Additionally, even with the confounders, the overall explanation for the variance in annotations is moderate (pseudo-R^2 = 0.213), indicating that other hidden confounders need to be accounted for to explain (mis)alignment, e.g., social media usage or attitudes towards free speech (Fleisig et al., 2023; Sap et al., 2021).

In summary, we find that **LLMs' demographic bias is at least partially explained by confounding factors**. Especially, LLMs' overall tendency to rate documents as offensive matches demographic s that are assigned more, more clearly offensive documents and/or include more sensitive, mutually agreeing individual annotators. Since the procedures of annotator recruitment and document assignment for a demographic vary between annotation tasks, LLMs may appear biased toward different demographics in the resulting datasets. Thus, by resorting to factors at the document, individual annotator, and annotator sample levels, **we can reconcile inconsistencies in the demographic bias observed across datasets**—that would otherwise contradict the assumption that LLMs replicate certain demographics' annotations.

D.V.: alignment	model 1	model 2
dataset=popq	-0.320^{***}	-0.410^{***}
dataset=nlpos	1.075^{***}	1.547^{***}
dataset=sbic	0.802^{***}	1.018^{***}
dataset=mhsc	0.843^{***}	1.055^{***}
gender=woman	-0.021^{*}	-0.036^{***}
ethnicity=asian	-0.175^{***}	-0.127^{***}
ethnicity=black	-0.071^{***}	-0.118^{***}
ethnicity=hispanic	0.049^{*}	-0.045
difficulty		-1.747^{***}
sensitivity		0.499^{***}
$\text{agreement}_{\text{ethnicity}}$		0.251^{***}
$\text{agreement}_{\text{gender}}$		0.208^{***}
label		1.788^{***}
intercept	0.439^{***}	0.545^{***}
observations	219359	219359
pseudo R^2	0.015	0.213

*p<0.05; **p<0.01; ***p<0.001

Table 2: Logistic regression of LLM–human alignment. Model 1 (left) explains whether an LLM chooses the same label as a human annotator by regressing over the annotator's gender (vs. man as the reference level), ethnicity (vs. White), and the annotated document's dataset (vs. AwA), encoded as indicator variables. Model 2 (right) additionally accounts for potential confounders: the document's difficulty, the annotator's sensitivity, and the agreement of the annotator with other annotators of the same gender and ethnicity, as well as the annotator's label as a control variable to account for the LLM's overall label skew.

6 Discussion and Conclusions

Our findings corroborate some of the previous results on demographic biases in LLMs' offensiveness ratings. In line with Sun et al. and Schäfer et al., LLMs consistently align better with White than Black annotators. This bias replicates across datasets and is measurable even when accounting for several confounders. However, our systematic analysis offer a nuanced picture compared to past single-dataset studies. Apparent demographic biases such as those based on annotators' gender (Sun et al., 2023) or the Asian American demographic (Santy et al., 2023), are statistically significant but contradictory in different datasets. We show that LLM–human alignment may appear associated with annotators' demographics, when in fact it may be due to factors that are distributed inconsistently across demographics and datasets.

Implications for Demographic Alignment in Data Annotation. LLMs showed consistently worse alignment with Black people's annotations of offensiveness. Therefore, we echo the warnings about the potential negative consequences of using LLM-generated annotations without understanding whose points of view they reinforce or neglect (Santy et al., 2023). However, given inconsistent alignment with any other demographic factor, we caution against narratives that anthropomorphize LLMs and essentialize annotators.

Recommendations for Measuring Misalignment. Our results show that measuring demographic biases in a single dataset can produce unreliable results. Systematic benchmarking across multiple datasets, coupled with replication and meta-analytical studies, is essential to support general claims about demographic bias. Moreover, including confounding variables in analyses is crucial to distinguish between demographic bias and other sources of LLM-human alignment.

The heterogeneity of results also underscores the need for datasets with greater demographic representation, more redundant labeling, and increased data diversity (Fleisig et al., 2024). While acknowledging the practical and financial constraints of building such datasets, our research provides insights into strategic approaches for enhancing data quality. Specifically, we emphasize the importance of accounting for confounders, such as document difficulty and the sensitivity of individual annotators, in their interplay with annotator demographics. Since, at present, such confounders only emerge at the end of the annotation process, we see an opportunity for developing annotation solutions that dynamically adjust annotator recruitment and document assignment in response to emerging patterns.

Finally, while sociodemographic prompting has had some success in simulating human samples (Argyle et al., 2023), it might not have the same success in simulating human annotators. But, this is not necessarily an indication of unrepresentativity of LLMs (though it is possible LLMs are, in fact, unrepresentative). In line with past research on demographic variation in annotation (Orlikowski et al., 2023), it is crucial to explore whether different demographics correlate with confounding factors, as humans do not conduct content analysis solely driven by their demographic identity. This leads to unexplained variation and misalignment with LLM labels when sociodemographic variables are the only ones considered.

7 Limitations

Our study has several limitations that point to opportunities for future work. The analysis is restricted to the English language, which limits the applicability of our findings across languages with different cultural nuances and linguistic structures. Additionally, our focus on demographic factors like gender and race provides only a partial view of potential biases, possibly overlooking other demographic characteristics that may influence alignment patterns. Our analysis is limited to three models consisting of two proprietary one (GPT-4o mini and Gemini 1.5 Flash) and one open-source model (Solar 10.7b-instruct), due to the budget and resource constraints in an academic setting. To address the need for broader model coverage, we extended the analysis by testing three additional models—Claude 3.5 Sonnet, Mistral 7b-instruct, and Llama 3.1 8b-instruct—on two of the five datasets, AwA and POPQ. The results, presented in the appendix, align closely with the patterns observed in our main analysis. Further exploration of additional models or alternative configurations of the same models, such as different hyperparameter settings (e.g., temperature), could provide deeper insights into alignment patterns. Regarding confounders, although we accounted for factors such as document difficulty and annotator sensitivity, other factors like the target of offensive language, could further enhance our model. Finally, we only assess the biases associated with default prompting, i.e., prompting without sociodemographic signals. While other researchers have looked into sociodemographic prompting (Beck et al., 2023; Sun et al., 2023; Schäfer et al., 2024), it is important to consider the default case in detail since in real-world settings, all relevant demographic variables may not be known a priori. Indeed, most recent work benchmarking the use of LLMs for content labeling do so without sociodemographic prompting (Gilardi et al., 2023; Ziems et al., 2024).

8 Ethics Statement

As language technologies become widely used for algorithmic decision-making, such as using NLP techniques for detecting offensive content as a type of content moderation tool, there are growing concerns about these technology's biases against marginalized populations; populations who are themselves most susceptible to receiving offensive attacks on platforms. In this work, we assess one aspect of such bias in the latest generation of language technology — demographic misalignment in prompt-based Large Language Models. Our findings across multiple datasets show that current LLMs have varying and inconsistent alignment with different demographics, but have especially lower alignment with Black people, and for offensive and potentially offensive content.

While current work has assessed the utility of including demographic information in prompts to induce personas ("demographic steering") (Santurkar et al., 2023), also in the context of data annotation (Sun et al., 2023; Beck et al., 2023), the preliminary results indicate that this type of steering does *not* improve alignment. Therefore, we need to assess further strategies such as fine-tuning (instruction or otherwise), Retrieval Augmented Generation, or even pre-training to address this misalignment.

We use five openly available datasets in our experiments, where the creators of these datasets made annotator demographics available along with the distribution of annotator labels (Sun et al., 2023; Sap et al., 2020; Kennedy et al., 2020). All of the annotator data released by these authors are anonymized and we do not attempt to deanonymize any of the annotators. While we attempted to include understudied demographic identities in this work, we only consider men and women within our gender variables. This is because the annotations of other genders (non-binary people) were significantly fewer and could not be quantitatively modeled. However, it is important to represent gender minorities when assessing the alignment of LLMs, especially when they are used to label offensive content targeting these groups. We hope to address this in future work by having a larger and more diverse pool of annotators.

While the results of our work indicate the need for strategies to improve alignment, there are also concerns of demographic essentialization and ecological fallacies (Orlikowski et al., 2023); their demographic identity could be *one* of the many factors affecting an annotator's perception of offensiveness. Other important factors to consider could be lived experiences, particularly past experiences with harassment. In future work, we hope to disentangle demographic and individual patterns when annotating content and devise ways of incorporating these into LLMs.

Not least, the premise for demographic analyses is the categorization of humans into demographic groups. These groups afford comparisons between individuals and data collections—and indeed, the harmonization of demographic descriptors of annotators across multiple datasets is one of the contributions of the present work. Yet, we stress that the choice of how to categorize humans into demographic groups is subjective, culturally dependent, and ultimately the outcome of a construction process; demographic grouping necessarily overlooks an irreducible variety of identities and experiences. The outlook of this work is to not to reinforce the validity of the demographic categories included in this study, but to problematize general claims made about them.

8.1 Reproducibility

Our analysis relies on five distinct datasets, with four freely accessible through public repositories. The fifth dataset AwA (Annotators with Attitude) can be obtained through a formal request to the original authors. We used three proprietary models *gpt-4o-mini-2024-07-18* (temperature 1.0), *gemini-1.5-flash-002* (temperature 1.0), and *claude-3-5-sonnet-20240620* (temperature 1.0), all of which can be accessed from OpenAI, Google and Anthropic's APIs. Regarding the open-source models, we used *solar:10.7b instruct-v1-q8_0*, *mistral:7b-instruct-fp16*, and *llama3.1:8b-instruct-q8_0* (temperature 0.8) through Ollama and then run the inference on a Linux 64-bit system equipped with an NVIDIA GeForce RTX 3090 GPU. The exact prompts that we used for all LLMs are included in the appendix (Table 10). The code to run the experiments is available at *llm-demographic-bias* repository on Github.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rachith Aiyappa, Shruthi Senthilmani, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2024. Benchmarking zero-shot stance detection with flant5-xxl: Insights from training data, prompting, and decoding strategies into its near-sota performance. *arXiv preprint arXiv:2403.00236*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in LLM generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2023. Disentangling disagreements on offensiveness: A cross-cultural study. In *The 61st Annual Meeting of the Association for Computational Linguistics*.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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

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

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Andrew Gelman. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873.

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.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

M Hoerger. 2013. Zh: An updated version of steiger's z and web-based calculator for testing the statistical significance of the difference between dependent correlations.

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

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Yixing Jiang, Jesutofunmi A Omiye, Cyril Zakka, Michael Moor, Haiwen Gui, Shayan Alipour, Seyed Shahabeddin Mousavi, Jonathan H Chen, Pranav Rajpurkar, and Roxana Daneshjou. 2024. Evaluating general vision-language models for clinical medicine. *medRxiv*, pages 2024–04.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

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

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kardos, Mia Jacobsen, and Lene Aarøe. 2023. Chatbots are not reliable text annotators. *arXiv preprint arXiv:2311.05769*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing llm and human annotations of conversational safety. *arXiv preprint arXiv:2406.06369*.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

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

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

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

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

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, Sabine Weber, and Roman Klinger. 2024. Which demographics do llms default to during annotation? *Preprint*, arXiv:2410.08820.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. *arXiv preprint arXiv:2310.15683*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1578–1590.

Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2023. Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *medRxiv*, pages 2023–07.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Dataset Details and Demographic Distribution

Here, we provided a more detailed summary of the 5 different datasets used in this work.

The **"Annotator with Attitudes"** (AwA) dataset (Sap et al., 2021) curates a dataset on potentially offensive content targeting Black people. We use their Breadth-of-Posts dataset, which contains 626 posts annotated by 177 annotators, totaling 3,349 annotations from different genders, ethnicities, and political backgrounds. Annotators were asked to rate how much they perceived each post as toxic, hateful, disrespectful, or offensive on a 5-point Likert scale, ranging from 1 (not at all) to 5 (very much so). To remain consistent across datasets, we use the annotators' gender and ethnicity as demographic variables.

UC Berkeley's **Measuring Hate Speech Corpus** (MHSC) (Kennedy et al., 2020) contains 90,174 annotations from 7,725 annotators on 39,263 online comments. The metric used for comparison was the "hatespeech" ordinal label of each comment measuring the identified severity on a three-level scale: yes, no, and unclear.

We used the **NLPositionality** (NLPos) dataset (Santy et al., 2023), which was originally used for the hate speech detection task in their paper. This dataset contains annotations from 412 annotators on 299 posts, totaling 4,417 annotations. Annotators were asked to evaluate an instance using a 3-point scale.

The **POPQUORN dataset** (Pei and Jurgens, 2023) (POPQ) contains 12,088 annotations on 1,500 online comments from 243 annotators from a sample of the US adult population that was representative based on age, gender, and ethnicity. Annotators were asked to provide an offensiveness score of each text sample on a 1-5 scale, from "Not offensive at all" to "Very offensive", gathered from annotators through a multiple-choice task.

The **Social Bias Inference Corpus** (Sap et al., 2020), referred to as SBIC, contains 109,349 annotations on 44,232 online posts from 280 annotators. The dataset is acknowledged by its creators to be racially skewed, with a vast majority of annotators being White and nearly none being both Black and male. Annotators were asked to evaluate whether each post could be considered offensive, disrespectful, or toxic to anyone/someone, with the following valid response options: 1 (Yes, this could be offensive), 2 (Maybe, I'm not sure), 3 (No, this is harmless), and 4 (I don't understand the post).

Naturally, a major factor determining the subjective offensiveness of a particular statement is the group or individual targeted by said statement. Therefore, we reasoned that demographic factors most frequently targeted by offensive statements were likely to have major effects on the perception of offensiveness, providing valuable insights in determining potential alignment. Considering available annotator data regarding targeted groups, it became clear that ethnicity and gender were the two most significant factors represented in targeted language. 11.7% and 7.6% of SBIC annotations noted targeted language towards Black folks and women respectively - figures around twice those corresponding to any other group. In MHSC, 35.7% of annotations indicated offensive language targeting a racial group and 29.8% targeted a gender, with 20.6% targeting women and 16.9% targeting Black people specifically. In contrast, age groups were targeted to a much lesser degree (1.5% of Hate Speech annotations), displaying less evidence supporting its status as a largely influential factor. Overall, offensive language was shown to target racial and gender demographic groups, specifically Black people and women, indicating a high likelihood that annotator ethnicity and gender would have significant influences on the variable perception of offensive statements.

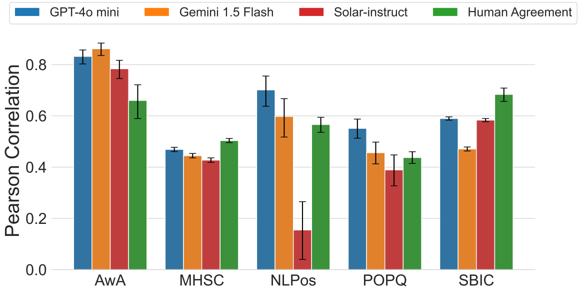

Figure 4: (Majority vote) Comparison of model correlations with human annotators against human agreement (individual annotators with their peers). The ground truth for each post is determined by the majority vote of annotators' labels. For human agreement, correlations are measured by leaving out one annotator and comparing their labels to the ground truth from the remaining annotators. Error bars represent 95% CIs.

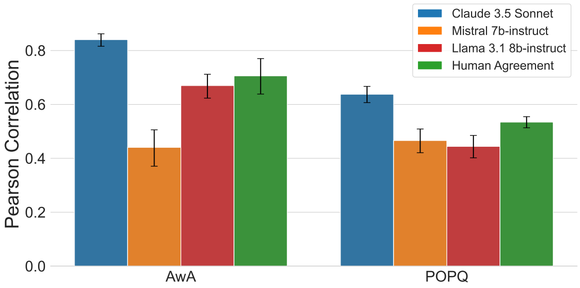

Figure 5: (Average vote) Comparison of model correlations with human annotators against human agreement (individual annotators with their peers). The ground truth for each post is determined by the average vote of annotators' labels. For human agreement, correlations are measured by leaving out one annotator and comparing their labels to the ground truth from the remaining annotators. Error bars represent 95% CIs.

Paper	Models	Datasets	Tasks	Demog.	Prompt	Findings for Default Prompting
Beck et al. (2023)	G3, T5, O, P	several	se, st, ol, hs	g, r, a, e, p	def+dem	does not test default alignment
Giorgi et al. (2024)	L3, P, S10, St	MHSC	hs	g, r, a, e, re, s, i	dem	N/A
Hu and Collier (2024)	G4, G35, L2, T	several	se, st, ol, hs, sa	g, r, l, e, n, a, re, p	def+dem	does not test default alignment
Movva et al. (2024)	G4	D	st	r, g	def	no clear alignment
Santy et al. (2023)	G4	NLPos	hs, sa	g, r, l, e, n, a, re	def	for hate speech: better alignment with Asian-Americans than White
Schäfer et al. (2024)	G4o, C	POPQ	ol, po	r, g, a, e, o	def+dem	better alignment with White than Black people
Sun et al. (2023)	FT, FU, G35, G4	POPQ	ol, po	r, g	def+dem	best alignment with White people and women
Ours	G4o, G, SI	several	ol, hs	r, g	def	better alignment with White than Black people

Table 3: Summary of recent research on LLMs' annotations and their demographic alignment with humans. Model abbreviations: L2=LlaMa-2, L3=Llama-3, P=Phi-3, St=Starling-LM, T=Tulu, FU=FLAN-UL2, FT=FLAN-T5-XXL, G3=GPT-3, G35=GPT-3.5, G4=GPT-4, G4o=GPT4o, C=Claude, T5=T5, O=OPT, P=Pythia, S10=SOLAR-10, SI=Solar-Instruct. Dataset abbreviation: MHSC=Measuring Hate Speech, POPQ=POPQUORN, NLPos=NLPositionality, D=DICES. Task abbreviations: se=sentiment, st=stance, ol=offensive language, hs=hate speech, sa=social acceptability, po=politeness, st=safety. Demographic variable abbreviations: g=gender, r=race, l=location, e=education, n=native language, a=age, re=religion, p=political leaning, s=sexuality, i=ideology, o=occupation. Prompt abbreviations: def=default, dem=demographic prompting.

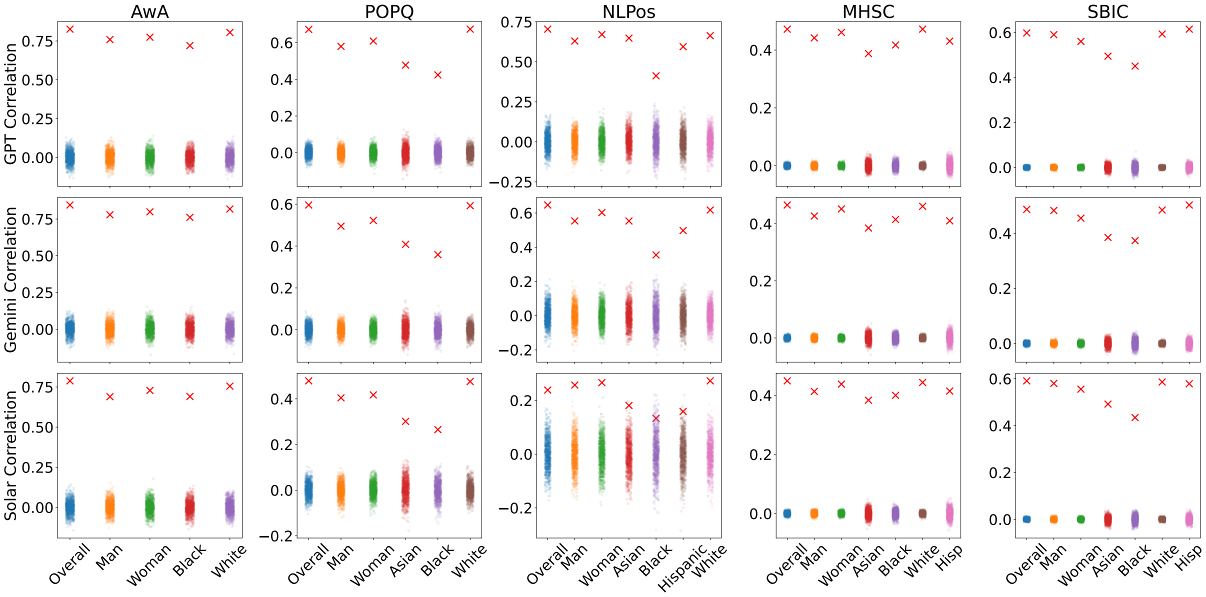

Figure 6: (Average vote) Results of a permutation test comparing Pearson correlation coefficients between model outputs and human annotator labels across demographic groups and datasets. Ground truth labels are based on the average of annotations from people in the target demographic. Each row shows results for a model, with observed correlations marked as red crosses and the null distribution from 1,000 random label permutations shown as scatter points. For the Solar model in the NLPos dataset, a few cases in the Asian, Black, and Hispanic demographics have higher correlations in the null distribution than the observed ones. In all other cases, observed correlations are consistently higher than shuffled ones.

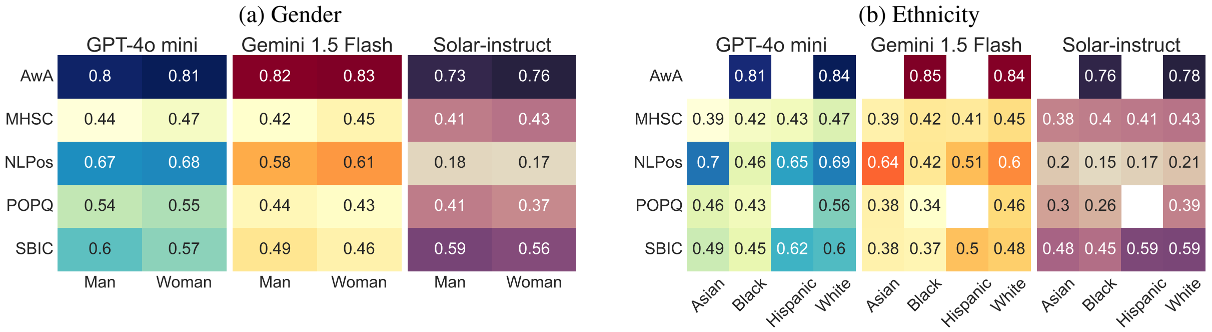

Figure 7: (Majority vote) Pearson correlation coefficients between model outputs and human annotator labels, broken down by gender (a) and ethnicity (b) across five datasets.The ground truth for each post is determined by the majority vote of annotators from the target demographic. Darker shades indicate stronger correlations. Confidence intervals and p-values for statistical significance are reported in Table 7 in the appendix.

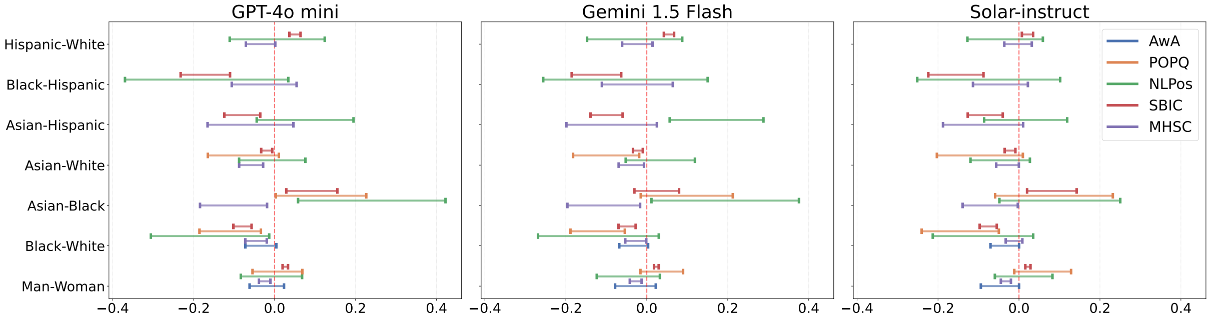

Figure 8: (Majority vote) The 95% confidence intervals (CI) for the difference in correlation between the model's predictions and two demographic groups, computed as: $\Delta r = r(P, D_1) - r(P, D_2)$, where P represents the model's predictions, and D_1 and D_2 are two demographic groups. Ground truth for each post is determined by the majority vote of annotators' labels. The intervals are derived from 1,000 bootstrap samples. If the CI includes zero, the difference is not statistically significant. See table 9 in the appendix for further details.

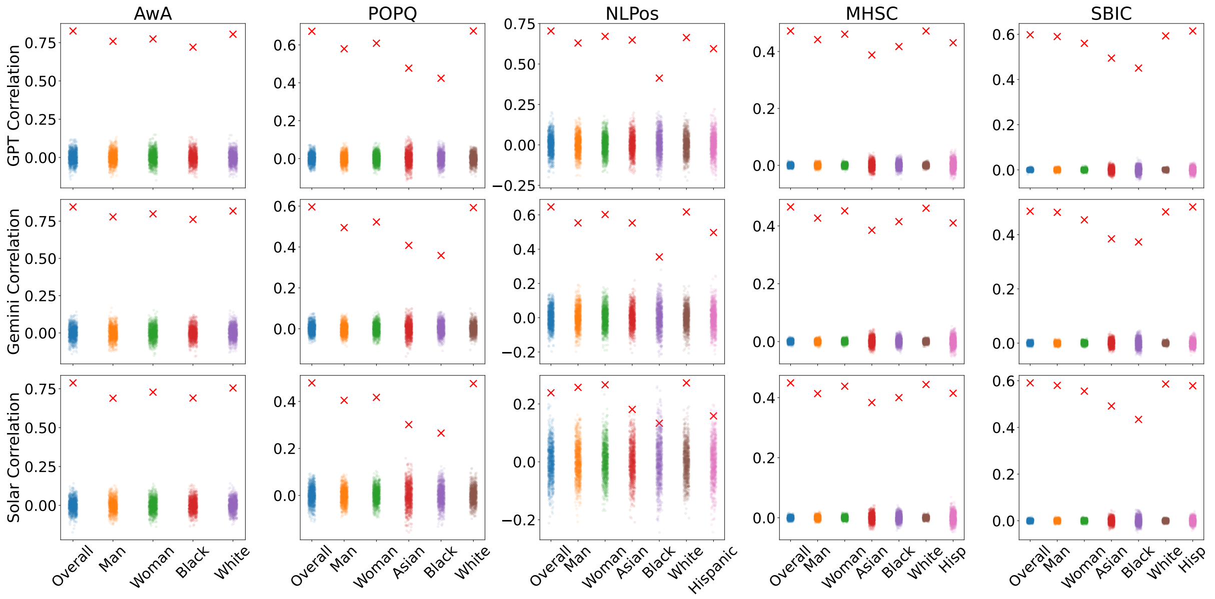

Figure 9: (Majority vote) Results of a permutation test comparing Pearson correlation coefficients between model outputs and human annotator labels across demographic groups and datasets. Ground truth is determined by the majority vote of annotators' labels. Each row shows results for a model, with observed correlations marked as red crosses and the null distribution from 1,000 random label permutations shown as scatter points. For the Solar model in the NLPos dataset, a few cases in the Asian, Black, and Hispanic demographics have higher correlations in the null distribution than the observed ones. In all other cases, observed correlations are consistently higher than shuffled ones.

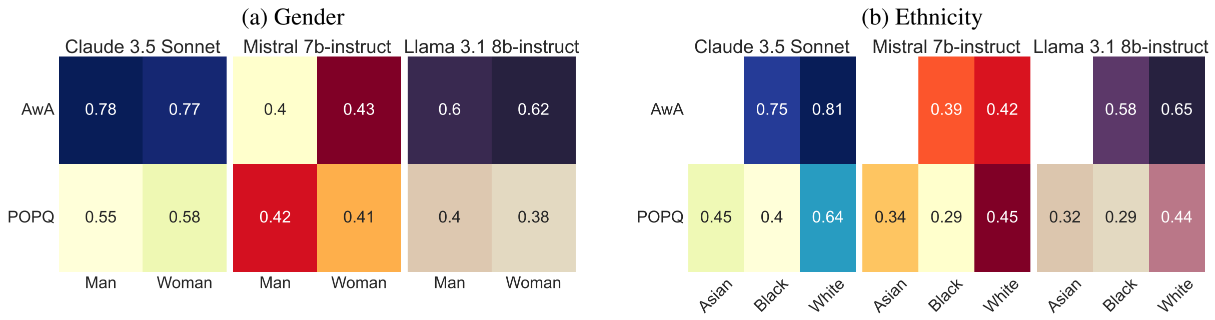

Figure 10: Pearson correlation coefficients between model outputs and human annotator labels, broken down by gender (a) and ethnicity (b) across two datasets.The ground truth for each post is determined by the average vote of annotators from the target demographic. Darker shades indicate stronger correlations. Confidence intervals and p-values for statistical significance are reported in Table 11 in the appendix.

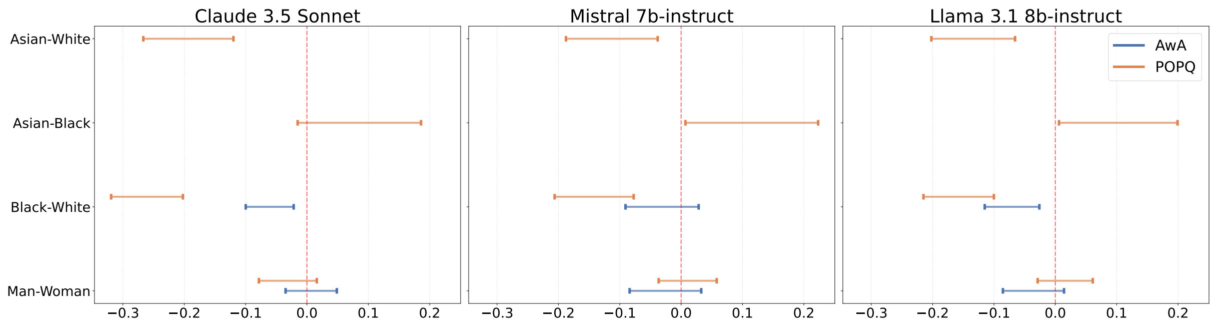

Figure 11: The 95% confidence intervals (CI) for the difference in correlation between the model's predictions and two demographic groups, computed as: $\Delta r = r(P, D_1) - r(P, D_2)$, where P represents the model's predictions, and D_1 and D_2 are two demographic groups. The intervals are derived from 1,000 bootstrap samples. If the CI includes zero, the difference is not statistically significant. See table 12 for further details.

GPT	model 1	model 2
dataset=popq	–0.320***	–0.407***
dataset=nlpos	1.075***	1.560***
dataset=sbic	0.802***	1.025***
dataset=mhsc	0.843***	1.062***
gender=woman	–0.021*	–0.042***
ethnicity=asian	–0.175***	–0.111***
ethnicity=black	–0.071***	–0.104***
ethnicity=hispanic	0.049*	–0.028
difficulty		–1.870***
sensitivity		0.566***
agreement$_{\text{ethnicity}}$		0.276***
agreement$_{\text{gender}}$		0.136***
label		1.851***
intercept	0.439***	0.559***
observations	219359	219359
pseudo R^2	0.015	0.213
Gemini		
dataset=popq	–0.220***	–0.269***
dataset=nlpos	1.062***	1.418***
dataset=sbic	0.473***	0.545***
dataset=mhsc	0.711***	0.847***
gender=woman	–0.032***	–0.053***
ethnicity=asian	–0.183***	–0.106***
ethnicity=black	–0.035*	–0.030
ethnicity=hispanic	0.031	–0.001
difficulty		–1.751***
sensitivity		0.522***
agreement$_{\text{ethnicity}}$		0.099***
agreement$_{\text{gender}}$		–0.002
label		1.400***
intercept	0.268***	0.297***
observations	219359	219359
pseudo R^2	0.010	0.166
Solar		
dataset=popq	–0.956***	–1.175***
dataset=nlpos	0.436***	0.414***
dataset=sbic	0.532***	0.810***
dataset=mhsc	0.512***	0.890***
gender=woman	0.026**	–0.014
ethnicity=asian	–0.215***	–0.106***
ethnicity=black	–0.073***	–0.135***
ethnicity=hispanic	0.102***	0.046
difficulty		–1.211***
sensitivity		0.454***
agreement$_{\text{ethnicity}}$		0.206***
agreement$_{\text{gender}}$		0.101***
label		2.853***
intercept	–0.308***	–0.370***
observations	219359	219359
pseudo R^2	0.016	0.286

*p<0.05; **p<0.01; ***p<0.001

Table 4: (Majority vote) Logistic regression of alignment for all three models using the methodology described in 4.5. The results indicate a similar trend, where confounders explain more variation in alignment patterns compared to demographic traits alone.

Gemini	model 1	model 2
dataset=popq	–0.220***	–0.269***
dataset=nlpos	1.062***	1.418***
dataset=sbic	0.473***	0.546***
dataset=mhsc	0.711***	0.848***
gender=woman	–0.032***	–0.048***
ethnicity=asian	–0.183***	–0.103***
ethnicity=black	–0.035*	–0.025
ethnicity=hispanic	0.031	0.002
difficulty		–1.692***
sensitivity		0.513***
agreement$_{\text{ethnicity}}$		0.058***
agreement$_{\text{gender}}$		0.105***
label		1.390***
intercept	0.268***	0.295***
observations	219359	219359
pseudo R^2	0.010	0.166
Solar		
dataset=popq	–0.956***	–1.164***
dataset=nlpos	0.436***	0.427***
dataset=sbic	0.532***	0.807***
dataset=mhsc	0.512***	0.885***
gender=woman	0.026**	–0.010
ethnicity=asian	–0.215***	–0.149***
ethnicity=black	–0.073***	–0.177***
ethnicity=hispanic	0.102***	–0.006
difficulty		–0.999***
sensitivity		0.448***
agreement$_{\text{ethnicity}}$		0.351***
agreement$_{\text{gender}}$		0.256***
label		2.848***
intercept	–0.308***	–0.386***
observations	219359	219359
pseudo R^2	0.016	0.288

*p<0.05; **p<0.01; ***p<0.001

Table 5: (Average vote) Logistic regression of LLM–human alignment for Gemini (top) and Solar (bottom). For each LLM, model 1 (left) explains whether the LLM chooses the same label as a human annotator by regressing over the annotator's gender (vs. man as the reference level), ethnicity (vs. White), and the annotated document's dataset (vs. AwA), encoded as indicator variables. Model 2 (right) additionally accounts for potential confounders: the document's difficulty, the annotator's sensitivity, and the agreement of the annotator with other annotators of the same gender and ethnicity, as well as the annotator's label as a control variable to account for the LLM's overall label skew.

Demo.	Model	AwA		MHSC		NLPos		POPQ		SBIC	
		Corr.	95% CI	Corr.	95% CI	Corr.	95% CI	Corr.	95% CI	Corr.	95% CI
Overall	GPT	0.83^*	(0.80, 0.85)	0.47^*	(0.46, 0.48)	0.70^*	(0.64, 0.76)	0.67^*	(0.64, 0.70)	0.60^*	(0.59, 0.60)
	Gemini	0.85^*	(0.82, 0.87)	0.47^*	(0.46, 0.47)	0.65^*	(0.57, 0.71)	0.60^*	(0.56, 0.63)	0.49^*	(0.48, 0.49)
	Solar	0.79^*	(0.76, 0.82)	0.45^*	(0.44, 0.46)	0.24^*	(0.13, 0.34)	0.48^*	(0.43, 0.53)	0.59^*	(0.58, 0.60)
Man	GPT	0.76^*	(0.72, 0.79)	0.44^*	(0.43, 0.45)	0.63^*	(0.56, 0.69)	0.58^*	(0.55, 0.61)	0.59^*	(0.58, 0.60)
	Gemini	0.78^*	(0.74, 0.81)	0.43^*	(0.42, 0.44)	0.55^*	(0.47, 0.63)	0.49^*	(0.46, 0.53)	0.48^*	(0.47, 0.49)
	Solar	0.69^*	(0.64, 0.73)	0.41^*	(0.40, 0.42)	0.26^*	(0.15, 0.36)	0.40^*	(0.35, 0.46)	0.58^*	(0.57, 0.59)
Woman	GPT	0.77^*	(0.74, 0.80)	0.46^*	(0.45, 0.47)	0.67^*	(0.60, 0.73)	0.61^*	(0.58, 0.64)	0.56^*	(0.55, 0.57)
	Gemini	0.80^*	(0.77, 0.83)	0.45^*	(0.44, 0.46)	0.60^*	(0.52, 0.67)	0.52^*	(0.48, 0.56)	0.45^*	(0.45, 0.46)
	Solar	0.73^*	(0.69, 0.77)	0.44^*	(0.43, 0.45)	0.27^*	(0.16, 0.37)	0.42^*	(0.36, 0.47)	0.55^*	(0.55, 0.56)
Asian	GPT	–	–	0.39^*	(0.36, 0.41)	0.65^*	(0.57, 0.71)	0.48^*	(0.42, 0.53)	0.49^*	(0.48, 0.51)
	Gemini	–	–	0.38^*	(0.36, 0.41)	0.55^*	(0.46, 0.63)	0.41^*	(0.35, 0.47)	0.38^*	(0.36, 0.41)
	Solar	–	–	0.38^*	(0.36, 0.41)	0.18^*	(0.06, 0.29)	0.30^*	(0.21, 0.38)	0.48^*	(0.46, 0.50)
Black	GPT	0.72^*	(0.68, 0.76)	0.42^*	(0.40, 0.44)	0.41^*	(0.28, 0.53)	0.42^*	(0.37, 0.47)	0.45^*	(0.43, 0.47)
	Gemini	0.76^*	(0.72, 0.79)	0.41^*	(0.40, 0.43)	0.35^*	(0.22, 0.48)	0.36^*	(0.30, 0.41)	0.37^*	(0.35, 0.40)
	Solar	0.69^*	(0.65, 0.73)	0.40^*	(0.38, 0.42)	0.13	(−0.01, 0.27)	0.27^*	(0.19, 0.34)	0.44^*	(0.42, 0.47)
Hispanic	GPT	–	–	0.43^*	(0.40, 0.46)	0.59^*	(0.50, 0.67)	–	–	0.61^*	(0.60, 0.63)
	Gemini	–	–	0.41^*	(0.38, 0.44)	0.50^*	(0.39, 0.59)	–	–	0.50^*	(0.48, 0.52)
	Solar	–	–	0.41^*	(0.39, 0.44)	0.16^*	(0.03, 0.28)	–	–	0.58^*	(0.57, 0.60)
White	GPT	0.80^*	(0.78, 0.83)	0.47^*	(0.46, 0.48)	0.66^*	(0.59, 0.72)	0.67^*	(0.65, 0.70)	0.59^*	(0.59, 0.60)
	Gemini	0.82^*	(0.79, 0.84)	0.46^*	(0.45, 0.47)	0.62^*	(0.54, 0.68)	0.59^*	(0.56, 0.62)	0.48^*	(0.48, 0.49)
	Solar	0.76^*	(0.72, 0.79)	0.44^*	(0.44, 0.45)	0.27^*	(0.16, 0.37)	0.48^*	(0.42, 0.52)	0.59^*	(0.58, 0.59)

Table 6: (Average vote) Correlation results across datasets, models, and demographics, with ground truth determined by averaging labels from annotators within the target demographic. * indicates statistical significance at p-value < 0.05 (corrected for multiple comparisons).

Demo.	Model	AwA		MHSC		NLPos		POPQ		SBIC	
		Corr.	95% CI	Corr.	95% CI	Corr.	95% CI	Corr.	95% CI	Corr.	95% CI
Overall	GPT	0.83^*	(0.80, 0.86)	0.47^*	(0.46, 0.48)	0.70^*	(0.64, 0.76)	0.55^*	(0.51, 0.59)	0.59^*	(0.58, 0.60)
	Gemini	0.86^*	(0.84, 0.88)	0.44^*	(0.44, 0.45)	0.60^*	(0.52, 0.67)	0.46^*	(0.41, 0.50)	0.47^*	(0.46, 0.48)
	Solar	0.78^*	(0.75, 0.82)	0.43^*	(0.42, 0.44)	0.15^*	(0.04, 0.27)	0.39^*	(0.33, 0.45)	0.58^*	(0.58, 0.59)
Man	GPT	0.80^*	(0.76, 0.83)	0.44^*	(0.43, 0.45)	0.67^*	(0.60, 0.73)	0.54^*	(0.50, 0.58)	0.60^*	(0.60, 0.61)
	Gemini	0.82^*	(0.79, 0.85)	0.42^*	(0.41, 0.43)	0.58^*	(0.50, 0.65)	0.44^*	(0.40, 0.49)	0.49^*	(0.48, 0.50)
	Solar	0.73^*	(0.68, 0.77)	0.41^*	(0.40, 0.42)	0.18^*	(0.06, 0.29)	0.41^*	(0.34, 0.47)	0.59^*	(0.58, 0.60)
Woman	GPT	0.81^*	(0.77, 0.84)	0.47^*	(0.46, 0.48)	0.68^*	(0.61, 0.74)	0.55^*	(0.51, 0.59)	0.57^*	(0.56, 0.58)
	Gemini	0.83^*	(0.80, 0.86)	0.45^*	(0.44, 0.46)	0.61^*	(0.53, 0.68)	0.43^*	(0.38, 0.47)	0.46^*	(0.45, 0.47)
	Solar	0.76^*	(0.72, 0.80)	0.43^*	(0.42, 0.44)	0.17^*	(0.05, 0.28)	0.37^*	(0.30, 0.44)	0.56^*	(0.56, 0.57)
Asian	GPT	–	–	0.39^*	(0.37, 0.41)	0.70^*	(0.63, 0.76)	0.46^*	(0.40, 0.52)	0.49^*	(0.48, 0.51)
	Gemini	–	–	0.39^*	(0.36, 0.41)	0.64^*	(0.55, 0.71)	0.38^*	(0.32, 0.45)	0.38^*	(0.36, 0.40)
	Solar	–	–	0.38^*	(0.36, 0.41)	0.20^*	(0.07, 0.32)	0.30^*	(0.21, 0.39)	0.48^*	(0.46, 0.50)
Black	GPT	0.81^*	(0.77, 0.84)	0.42^*	(0.40, 0.44)	0.46^*	(0.32, 0.58)	0.43^*	(0.37, 0.49)	0.45^*	(0.43, 0.48)
	Gemini	0.85^*	(0.81, 0.88)	0.42^*	(0.40, 0.44)	0.42^*	(0.28, 0.54)	0.34^*	(0.27, 0.40)	0.37^*	(0.35, 0.40)
	Solar	0.76^*	(0.71, 0.80)	0.40^*	(0.38, 0.42)	0.15	(−0.01, 0.30)	0.26^*	(0.17, 0.34)	0.45^*	(0.42, 0.47)
Hispanic	GPT	–	–	0.43^*	(0.40, 0.46)	0.65^*	(0.56, 0.72)	–	–	0.62^*	(0.60, 0.63)
	Gemini	–	–	0.41^*	(0.38, 0.44)	0.51^*	(0.40, 0.61)	–	–	0.50^*	(0.49, 0.52)
	Solar	–	–	0.41^*	(0.39, 0.44)	0.17^*	(0.03, 0.30)	–	–	0.59^*	(0.57, 0.60)
White	GPT	0.84^*	(0.81, 0.86)	0.47^*	(0.46, 0.48)	0.69^*	(0.62, 0.75)	0.56^*	(0.52, 0.59)	0.60^*	(0.59, 0.61)
	Gemini	0.84^*	(0.81, 0.87)	0.45^*	(0.44, 0.46)	0.60^*	(0.51, 0.67)	0.46^*	(0.42, 0.50)	0.48^*	(0.47, 0.49)
	Solar	0.78^*	(0.74, 0.82)	0.43^*	(0.42, 0.44)	0.21^*	(0.09, 0.32)	0.39^*	(0.33, 0.45)	0.59^*	(0.59, 0.60)

Table 7: (Majority vote) Correlation results across datasets, models, and demographics, with ground truth determined by the majority vote of annotators' labels. * indicates statistical significance at p-value < 0.05 (corrected for multiple comparisons).

Demo. Pair	Model	AwA	MHSC	NLPos	POPQ	SBIC
Man - Woman	GPT	(-0.06, 0.02)	(-0.04, -0.01)**	(-0.11, 0.03)	(-0.07, 0.01)	(0.02, 0.03)**
	Gemini	(-0.07, 0.01)	(-0.04, -0.01)**	(-0.13, 0.03)	(-0.07, 0.02)	(0.02, 0.03)**
	Solar	(-0.09, -0.01)*	(-0.05, -0.02)**	(-0.09, 0.07)	(-0.07, 0.04)	(0.01, 0.03)**
Black - White	GPT	(-0.12, -0.05)**	(-0.09, -0.04)**	(-0.34, -0.09)**	(-0.29, -0.19)**	(-0.13, -0.09)**
	Gemini	(-0.10, -0.01)**	(-0.08, -0.03)**	(-0.38, -0.10)**	(-0.29, -0.18)**	(-0.10, -0.06)**
	Solar	(-0.11, -0.02)**	(-0.06, -0.02)**	(-0.28, -0.05)	(-0.28, -0.14)**	(-0.13, -0.09)**
Asian - Black	GPT	–	(-0.18, -0.01)*	(0.09, 0.38)**	(0.03, 0.23)*	(0.04, 0.16)*
	Gemini	–	(-0.18, -0.01)	(0.08, 0.39)**	(0.00, 0.19)	(-0.02, 0.09)
	Solar	–	(-0.14, -0.01)	(-0.08, 0.20)	(-0.07, 0.16)	(0.03, 0.15)
Asian - Hispanic	GPT	–	(-0.16, 0.03)	(-0.01, 0.20)	–	(-0.12, -0.03)**
	Gemini	–	(-0.19, 0.03)	(0.03, 0.26)*	–	(-0.14, -0.06)**
	Solar	–	(-0.20, 0.03)	(-0.11, 0.12)	–	(-0.13, -0.04)**
Asian - White	GPT	–	(-0.11, -0.05)**	(-0.10, 0.06)	(-0.27, -0.14)**	(-0.07, -0.05)**
	Gemini	–	(-0.09, -0.02)**	(-0.15, 0.02)	(-0.26, -0.13)**	(-0.08, -0.05)**
	Solar	–	(-0.08, -0.03)**	(-0.19, -0.00)	(-0.26, -0.08)**	(-0.06, -0.04)**
Black - Hispanic	GPT	–	(-0.11, 0.05)	(-0.36, -0.02)	–	(-0.25, -0.12)**
	Gemini	–	(-0.12, 0.06)	(-0.34, 0.01)	–	(-0.19, -0.08)**
	Solar	–	(-0.12, 0.02)	(-0.22, 0.14)	–	(-0.23, -0.09)**
Hispanic - White	GPT	–	(-0.08, -0.01)*	(-0.15, 0.07)	–	(0.02, 0.04)**
	Gemini	–	(-0.08, -0.01)	(-0.23, -0.00)	–	(0.02, 0.04)**
	Solar	–	(-0.06, 0.00)	(-0.23, -0.01)	–	(-0.01, 0.02)

Table 8: (Average vote) The 95% confidence intervals (CI) for the difference in correlation between the model's predictions and two demographic groups, computed as: $\Delta r = r(P, D_1) - r(P, D_2)$, where P represents the model's predictions, and D_1 and D_2 are two demographic groups. Ground truth for each post is determined by averaging the labels from annotators in the target demographic. The CIs are based on 1000 bootstrap samples. p-values were computed using Steiger's Z test and corrected for multiple comparisons with Holm's method. Significance levels are indicated by asterisks: $*p < 0.1$, $**p < 0.05$

Demo. Pair	Model	AwA	MHSC	NLPos	POPQ	SBIC
Man - Woman	GPT	(-0.06, 0.02)	(-0.04, -0.01)**	(-0.08, 0.07)	(-0.05, 0.07)	(0.02, 0.03)**
	Gemini	(-0.08, 0.02)	(-0.04, -0.01)**	(-0.12, 0.03)	(-0.02, 0.09)	(0.02, 0.03)**
	Solar	(-0.09, 0.00)	(-0.04, -0.02)**	(-0.06, 0.08)	(-0.01, 0.13)	(0.02, 0.03)**
Black - White	GPT	(-0.07, 0.00)*	(-0.07, -0.02)**	(-0.31, -0.01)*	(-0.19, -0.03)**	(-0.10, -0.06)**
	Gemini	(-0.07, 0.00)*	(-0.05, -0.00)	(-0.27, 0.03)	(-0.19, -0.05)**	(-0.07, -0.03)**
	Solar	(-0.07, 0.00)	(-0.03, 0.01)	(-0.21, 0.04)	(-0.24, -0.05)*	(-0.10, -0.05)**
Asian - Black	GPT	–	(-0.18, -0.02)*	(0.06, 0.42)**	(0.00, 0.23)	(0.03, 0.16)*
	Gemini	–	(-0.20, -0.02)	(0.01, 0.38)	(-0.01, 0.21)	(-0.03, 0.08)
	Solar	–	(-0.14, -0.00)	(-0.05, 0.25)	(-0.06, 0.23)	(0.02, 0.14)*
Asian - Hispanic	GPT	–	(-0.17, 0.05)	(-0.04, 0.20)	–	(-0.12, -0.04)**
	Gemini	–	(-0.20, 0.02)	(0.06, 0.29)**	–	(-0.14, -0.06)**
	Solar	–	(-0.19, 0.01)	(-0.09, 0.12)	–	(-0.13, -0.04)**
Asian - White	GPT	–	(-0.09, -0.03)**	(-0.09, 0.08)	(-0.16, 0.01)	(-0.03, -0.01)*
	Gemini	–	(-0.07, -0.01)	(-0.05, 0.12)	(-0.18, -0.02)	(-0.03, -0.01)*
	Solar	–	(-0.06, -0.00)	(-0.12, 0.03)	(-0.20, 0.01)	(-0.04, -0.01)*
Black - Hispanic	GPT	–	(-0.11, 0.05)	(-0.37, 0.03)	–	(-0.23, -0.11)**
	Gemini	–	(-0.11, 0.06)	(-0.26, 0.15)	–	(-0.19, -0.06)**
	Solar	–	(-0.11, 0.02)	(-0.25, 0.10)	–	(-0.22, -0.09)**
Hispanic - White	GPT	–	(-0.07, 0.00)	(-0.11, 0.12)	–	(0.04, 0.06)**
	Gemini	–	(-0.06, 0.01)	(-0.15, 0.09)	–	(0.04, 0.07)**
	Solar	–	(-0.04, 0.03)	(-0.13, 0.06)	–	(0.01, 0.04)**

Table 9: (Majority vote) The 95% confidence intervals (CI) for the difference in correlation between the model's predictions and two demographic groups, computed as: $\Delta r = r(P, D_1) - r(P, D_2)$, where P represents the model's predictions, and D_1 and D_2 are two demographic groups. Ground truth for each post is determined by the majority vote of annotators' labels. The CIs are based on 1000 bootstrap samples. p-values were computed using Steiger's Z test and corrected for multiple comparisons with Holm's method. Significance levels are indicated by asterisks: $*p$ < 0.1, $**p$ < 0.05

Dataset	Prompt
AwA	How much could this post be seen as toxic/hateful/disrespectful or offensive to anyone, on a scale from 1 (not at all) to 5 (very much so)? Valid options: 1, 2, 3, 4, 5. Begin your response by selecting one of the valid options, then provide a concise explanation for your rating.
MHSC	Does this comment contain hate speech, defined as "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled (e.g., "jew bankers", "nâĂŞ hooligans")." The comment contains hate speech if slurs or derogatory language are used toward any person or group regardless of whether or not the target is a member of the slur-target group. (e.g. a comment calling Trump a faggot or a n***** still contains hate speech). Valid options are 1 (Yes), 2 (Unclear), 3 (No). Begin your response by mentioning one of the valid options, then provide a concise explanation for your rating.
NLPos	Do you think this post is hateful speech? Valid options are 1 (Yes), 2 (Not sure), 3 (No). Begin your response by mentioning one of the valid options, then provide a concise explanation for your rating.
POPQ	Consider you read the above comment on Reddit. How offensive do you think it is, on a scale from 1 (not offensive at all) to 5 (very offensive)? Valid options: 1, 2, 3, 4, 5. Begin your response by selecting one of the valid options, then provide a concise explanation for your rating.
SBIC	Could this post be considered offensive, disrespectful, or toxic to anyone/someone? Valid options are: 1 (Yes, this could be offensive), 2 (Maybe, I'm not sure), 3 (No, this is harmless), 4 (I don't understand the post). Begin your response by mentioning one of the valid options, then provide a concise explanation for your rating.

Table 10: Prompts used for inference to annotate comments and posts, based on the original questions and wording provided to human annotators in each dataset.

Demo.	**Model**	**AwA**		**POPQ**	
		Corr.	95% CI	Corr.	95% CI
Overall	Claude	0.84^*	$(0.82, 0.86)$	0.64^*	$(0.61, 0.67)$
	Mistral	0.44^*	$(0.37, 0.51)$	0.47^*	$(0.42, 0.51)$
	Llama	0.67^*	$(0.62, 0.71)$	0.44^*	$(0.40, 0.48)$
Man	Claude	0.78^*	$(0.75, 0.81)$	0.55^*	$(0.51, 0.58)$
	Mistral	0.40^*	$(0.33, 0.47)$	0.42^*	$(0.37, 0.47)$
	Llama	0.60^*	$(0.55, 0.65)$	0.40^*	$(0.36, 0.44)$
Woman	Claude	0.77^*	$(0.74, 0.80)$	0.58^*	$(0.54, 0.61)$
	Mistral	0.43^*	$(0.35, 0.49)$	0.41^*	$(0.36, 0.46)$
	Llama	0.62^*	$(0.57, 0.67)$	0.38^*	$(0.34, 0.43)$
Asian	Claude	–	–	0.45^*	$(0.39, 0.50)$
	Mistral	–	–	0.34^*	$(0.27, 0.41)$
	Llama	–	–	0.32^*	$(0.25, 0.39)$
Black	Claude	0.75^*	$(0.72, 0.78)$	0.40^*	$(0.35, 0.45)$
	Mistral	0.39^*	$(0.32, 0.46)$	0.29^*	$(0.23, 0.35)$
	Llama	0.58^*	$(0.53, 0.63)$	0.29^*	$(0.24, 0.35)$
White	Claude	0.81^*	$(0.78, 0.84)$	0.64^*	$(0.61, 0.67)$
	Mistral	0.42^*	$(0.34, 0.48)$	0.45^*	$(0.41, 0.50)$
	Llama	0.65^*	$(0.60, 0.69)$	0.44^*	$(0.40, 0.48)$

Table 11: Correlation results between model predictions and average votes within each target demographic. * denotes p-values < 0.05 (corrected for multiple comparisons).

Demo. Pair	**Model**	**AWA**	**POPQ**
Man - Woman	Claude	(-0.03, 0.05)	(-0.08, 0.02)
	Llama	(-0.09, 0.01)	(-0.03, 0.06)
	Mistral	(-0.08, 0.03)	(-0.04, 0.06)
Black - White	Claude	(-0.10, -0.02)**	(-0.32, -0.20)**
	Llama	(-0.12, -0.03)**	(-0.21, -0.10)**
	Mistral	(-0.09, 0.03)	(-0.21, -0.08)**
Asian - Black	Claude	–	(-0.02, 0.19)
	Llama	–	(0.01, 0.20)
	Mistral	–	(0.01, 0.22)
Asian - White	Claude	–	(-0.27, -0.12)**
	Llama	–	(-0.20, -0.07)**
	Mistral	–	(-0.19, -0.04)*

Table 12: The 95% confidence intervals (1000 bootstrap samples) for the difference in correlation between the model's predictions and two demographic groups. p-values were computed using Steiger's Z test and corrected for multiple comparisons with Holm's method. Significance levels are indicated by asterisks: $^*p < 0.1$, $^{**}p < 0.05$

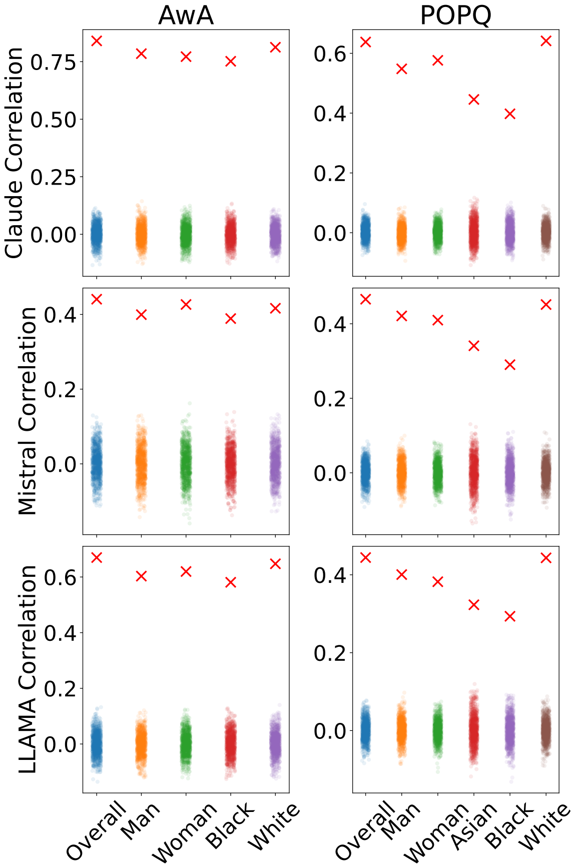

Figure 12: Results of a permutation test comparing Pearson correlation coefficients between model outputs and human annotator labels across demographic groups and datasets. Ground truth is determined by the majority vote of annotators' labels. Each row shows results for a model, with observed correlations marked as red crosses and the null distribution from 1,000 random label permutations shown as scatter points.

Claude	model 1	model 2
dataset=popq	−0.063	−0.074
gender=woman	0.073*	0.049
ethnicity=asian	−0.036	−0.107
ethnicity=black	−0.262***	−0.288***
difficulty		−1.539***
sensitivity		0.347***
$\text{agreement}_{\text{ethnicity}}$		0.722***
$\text{agreement}_{\text{gender}}$		0.981***
label		−0.242***
intercept	−0.164***	−0.253***
observations	15437	15437
pseudo R^2	0.002	0.201
Mistral		
dataset=popq	−0.636***	−0.650***
gender=woman	0.084*	0.038
ethnicity=asian	0.160	0.132
ethnicity=black	0.203***	0.063
difficulty		−1.137***
sensitivity		0.498***
$\text{agreement}_{\text{ethnicity}}$		0.147*
$\text{agreement}_{\text{gender}}$		0.289***
label		1.371***
intercept	−1.488***	−1.675***
observations	15437	15437
pseudo R^2	0.015	0.108
Llama		
dataset=popq	−0.724***	−0.829***
gender=woman	−0.059	−0.118*
ethnicity=asian	−0.017	−0.080
ethnicity=black	0.148**	−0.054
difficulty		−0.185***
sensitivity		0.830***
$\text{agreement}_{\text{ethnicity}}$		0.093
$\text{agreement}_{\text{gender}}$		0.288***
label		1.278***
intercept	−1.658***	−1.893***
observations	15437	15437
pseudo R^2	0.018	0.128

*p<0.05; **p<0.01; ***p<0.001

Table 13: Logistic regression of LLM–human alignment for Claude, Mistral, and Llama, using the methodology described in 4.5. The results indicate a similar trend, where confounders explain more variation in alignment patterns compared to demographic traits alone.

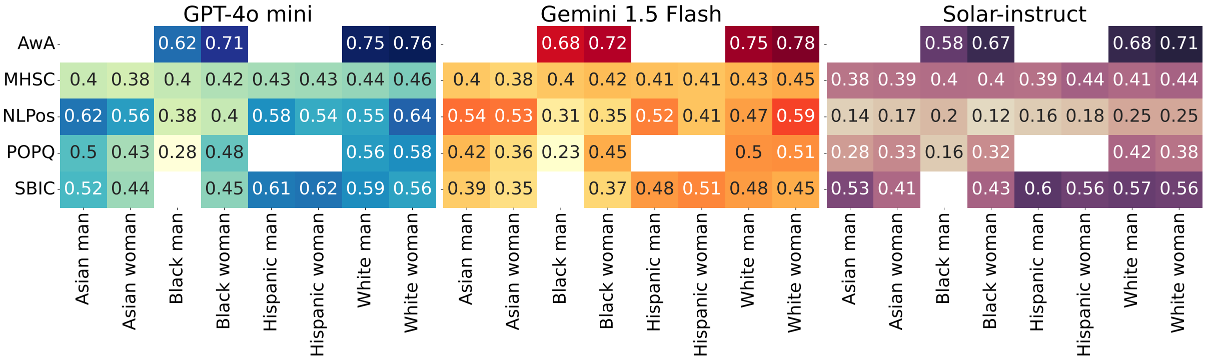

Figure 13: Pearson correlation coefficients between model outputs and human annotator labels. Confidence intervals and p-values for statistical significance are reported in study's Github repository.

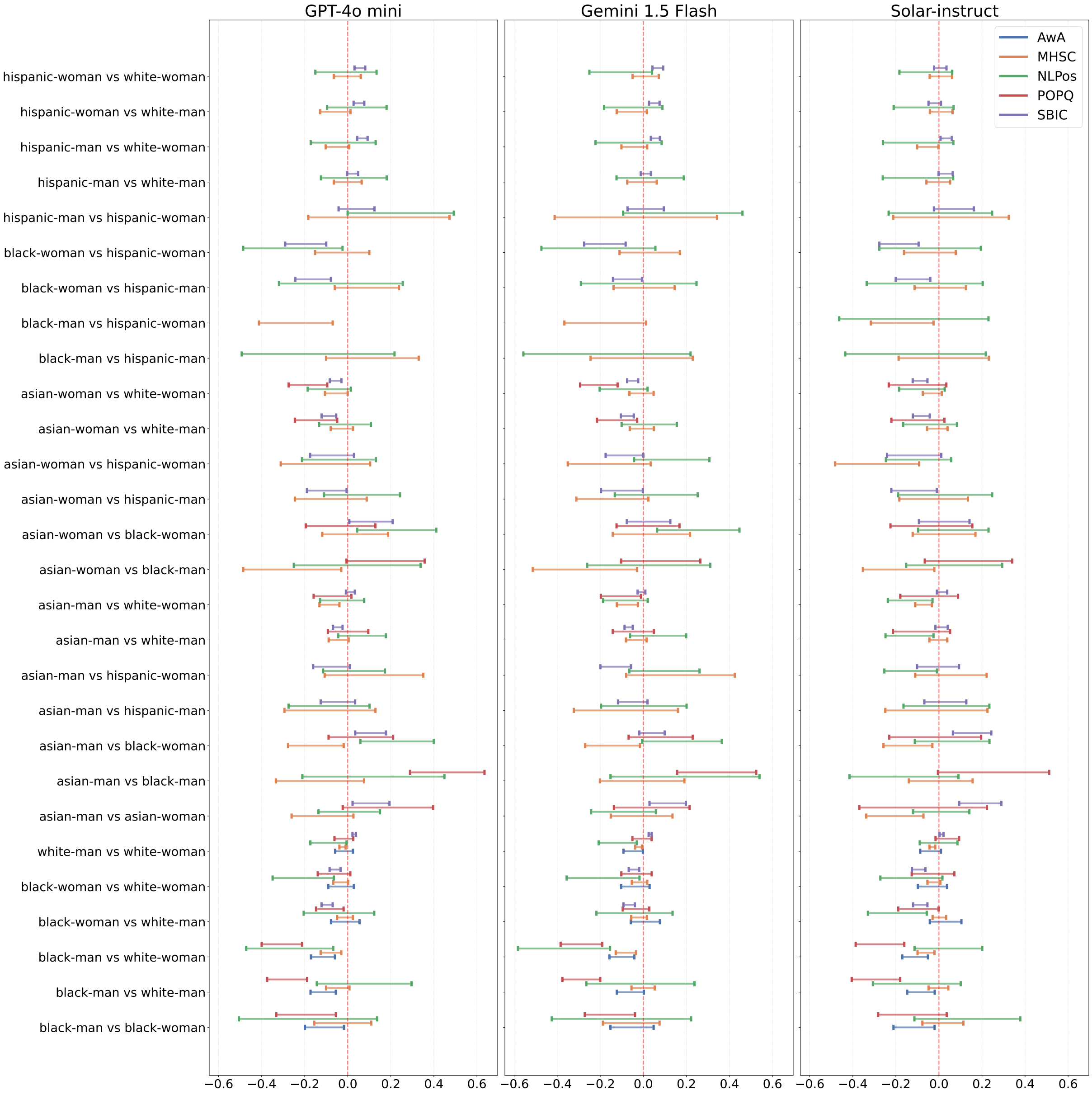

Figure 14: The 95% confidence intervals for the difference in correlation between the model's predictions and two intersectional demographic groups. Confidence intervals and p-values for statistical significance are reported in study's Github repository.