Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks

Anonymous ACL submission

Abstract

Human perception of language is shaped by personal backgrounds such as gender and ethnicity, yet it is unclear whether large language models (LLMs) exhibit similar variations in subjective NLP tasks. Do LLMs' predictions systematically align more with certain social groups? Can including demographic information in prompts influence LLMs' responses? Leveraging the POPQUORN dataset, we evaluate nine popular LLMs on their ability to understand demographic differences in two subjective tasks: predicting politeness and offensiveness. We find that in zero-shot settings, most models' predictions for both tasks align more closely with labels from White participants than those from Asian or Black participants, while there is only a minor gender bias favoring women in the politeness task. Furthermore, sociodemographic prompting does not consistently improve, and can sometimes worsen, LLMs' capability to perceive language from specific sub-populations. These findings underscore the need for caution to LLMs' demographic biases in subjective NLP tasks and highlight the limitations of sociodemographic prompting as a mitigation strategy.

1 Introduction

From sentiment analysis to dialogue generation, large language models (LLMs) have demonstrated impressive capabilities in various natural language processing (NLP) tasks (Brown et al., 2020; Radford et al., 2019). Recent research has begun exploring whether these models possess social knowledge analogous to that of humans (Zhou et al., 2023). For example, Almeida et al. (2024) replicate eight classic psychological experiments on LLMs to test their ability to reason about moral and legal issues. Yildirim and Paul (2024) examines how LLMs' "instrumental knowledge" relates to the more ordinary "worldly" knowledge of human agents. Building on these insights, LLMs have been applied to large-scale labeling tasks requiring social understanding, and often with promising results (Ziems et al., 2023; Rytting et al., 2023). In terms of subjective tasks, researchers have explored LLMs' zero-shot potential in areas such as character simulation (Wang et al., 2023) and hate speech detection (Plaza-del arco et al., 2023).

However, LLMs face significant limitations when handling subjective tasks. It is widely acknowledged that social biases and stereotypes embedded in their training data can cause LLMs to inadequately represent the diversity of human experiences (Santurkar et al., 2023a). Consequently, the under-representation of certain demographic groups can result in biases related to factors such as gender (Lucy and Bamman, 2021) and political ideology (Liu et al., 2022). In addition, subjective tasks pose an additional challenge, as language perception is influenced by social context and identity (Al Kuwatly et al., 2020). For instance, text perceived as polite or inoffensive by one group may be interpreted differently by another. Ideally, LLMs should be able to capture this variation, but this is often not the case. As a result, using LLMs for subjective tasks risks producing biased or unfair outcomes for different demographic groups (Liang et al., 2021). Santurkar et al. (2023a) found that LLMs, when responding to value-based questions, tend to reflect the perspectives of lower-income, moderate, and Protestant or Roman Catholic individuals. Despite these findings, there has been limited research into whether LLMs exhibit similar biases across other subjective NLP tasks.

Sociodemographic prompting, which involves enriching prompts with demographic or individual-specific information, has gained increasing attention in recent research. This approach has shown potential for improving data augmentation and simulating human behavior for social science applications (Hwang et al., 2023; Argyle et al., 2023). Despite its promise, the effectiveness of sociodemo-

graphic prompting remains debated, as model performance can be sensitive to the phrasing, structure, or order of prompts (Mu et al., 2023; Dominguez-Olmedo et al., 2023). For example, Beck et al. (2024) finds that the impact of adding demographic information varies significantly depending on the model, task, and prompt design. Moreover, some studies suggest that sociodemographic prompting can exacerbate stereotypes and biases (Deshpande et al., 2023) or reduce model performance on certain tasks (Santurkar et al., 2023b).

Given these mixed findings and the focus of previous studies on specific NLP tasks, our work extends the literature by examining (1) whether LLMs' predictions systematically align more with certain social groups on two more subjective tasks and (2) how LLMs can effectively account for identity-based differences in perception when handling subjective language tasks with sociodemographic prompting. Leveraging the POPQUORN dataset (Pei and Jurgens, 2023), we evaluate nine popular LLMs on their ability to understand demographic differences in subjective tasks, offensiveness and politeness. The two tasks are occasionally related but distinct. Politeness pertains to notions of status differences and interpersonal distance, while offensiveness involves violations of expected social norms. Offensiveness is not as broad as impoliteness, as varying levels of politeness can be perceived as non-offensive. Exploring these subtly different tasks offers a more comprehensive evaluation of LLMs' potential biases in subjective NLP tasks.

Overall, our results reveal that intrinsic biases persist in LLMs when applied to these tasks. The study highlights the limitations of LLMs in understanding and addressing gender and racial differences in subjective language tasks. While some research aims to use LLMs to simulate group-specific social behaviors, our findings underscore the risks of unintentionally reinforcing racial and gender biases when applying sociodemographic prompting to subjective tasks.

2 Methods

Data We use the POPQUORN dataset (Pei and Jurgens, 2023) to evaluate LLMs' capacity to tackle subjective NLP tasks. POPQUORN includes 45,000 annotations from a demographically representative U.S. sample. We focus our analysis on two identity types: gender and ethnicity. To ensure statistical robustness, we focus on the gender categories ['Man', 'Woman'] and ethnic groups ['Asian', 'Black', 'White'] as they have sufficient annotations.

For this study, we analyze annotators' offensiveness and politeness ratings on a 5-point Likert scale. We compute average scores for each identity group to capture perceptions from specific demographics. The mean overall offensiveness score is 1.88 (SD = 0.76), and politeness scores average 3.31 (SD = 0.91). Scores from men, women, and White annotators closely mirror the overall distribution, while Black and Asian annotators show diverging means and higher variance. Figure 3 in Appendix A the distributions of both overall and identity-specific scores for offensiveness and politeness tasks.

Models To enhance the generalizability of our findings, we conduct experiments with a range of open-source and close-source LLMs: FLAN-T5-XXL (Chung et al., 2022), FLAN-UL2 (Tay et al., 2023), Tulu2-DPO-7B, Tulu2-DPO-13B (Ivison et al., 2023), GPT-3.5, GPT-4 (OpenAI, 2023), Llama-3.1-8B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Team, 2024).

Prompts We design prompts to instruct the models to predict offensiveness and politeness scores for each instance. To ensure the prompts elicit valid responses, we conduct preliminary experiments on a small subset of data. An example prompt (Table 3)and the full list of prompts used in our study (Table 4) are shown in Appendix B.

We test the robustness of our results using different prompt templates and option orders (i.e., 1 to 5 or 5 to 1) on a set of open-source LLMs. Overall, we observe minor differences in LLMs' performance across templates and option orders. Details are provided in Appendix B.

3 Are Model Predictions Closer to Certain Demographic Groups?

Analysis To address the first research question, we define baseline prediction error (E_{base}) as the absolute difference between LLMs' predictions using identity-free prompts and human ratings from a specific demographic group:

$$E_{base} = |prediction - label_{subgroup}|$$

For each task and demographic identity type, we apply separate linear mixed-effect models to examine changes in baseline prediction error from a specific

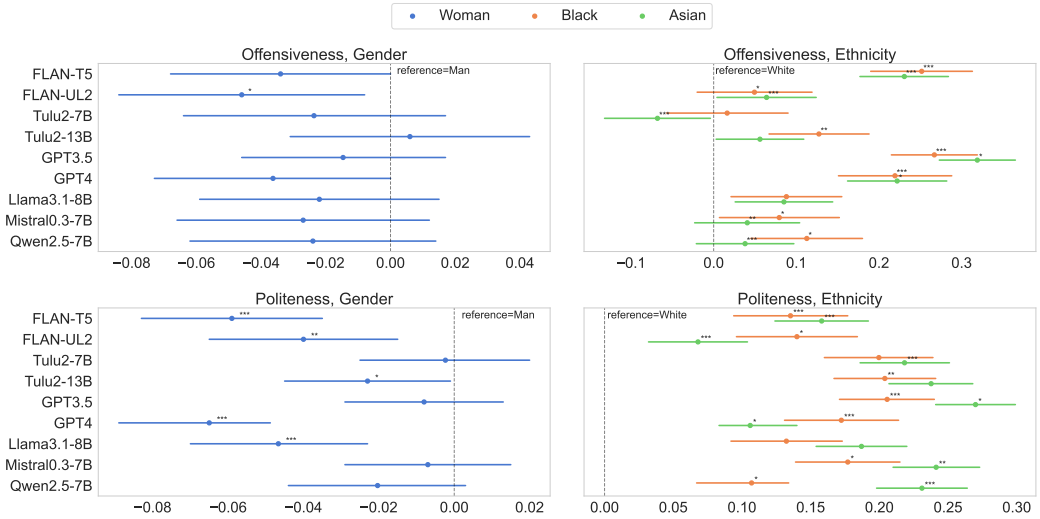

Figure 1: Regression results for predicting the gap between model predictions and the labels from each demographic group. The models' predictions for offensiveness are not significantly different from the ratings by Men and Women except for FLAN-UL2 (Top left). However, LLMs' predictions are significantly closer to Women's ratings for politeness (Bottom left) and are closer to White people's ratings compared with ratings from Black and Asian annotators in both tasks (Right)

demographic group (target group) compared to the reference group, controlling for instance-level variations wit instance ID as a random effect. For example:

$$E_{base} = \beta gender(ref = man) + (1|instanceid)$$

A regression coefficient $\beta = 0$ indicates no difference in baseline prediction errors between the target and reference groups. A positive β means that baseline prediction errors are larger for the target group, which suggests that LLMs' predictions are closer to the reference group than the target group. Figure 1 presents the aggregated results of the linear mixed-effects regression. Each data point represents a regression coefficient, with error bars indicating the standard error and significance markers. The black dotted line represents a regression coefficient of 0.

Results As shown in Figure 1 (left), LLMs' baseline prediction errors for offensiveness show no significant gender differences, except for FLAN-UL2. This is expected as the original POPQUORN paper (Pei and Jurgens, 2023) reports no significant gender differences in human ratings of offensiveness. However, for politeness ratings, LLM predictions tend to align more closely with women's ratings, except for GPT-3.5 and Tulu2-7B. Surprisingly, for both Tulu2 and GPT models, those with more parameters exhibit greater bias in politeness prediction, suggesting that simply scaling models may not effectively reduce biases in subjective tasks.

Figure 1 (right) shows that LLM predictions for both politeness and offensiveness are consistently closer to the ratings of White annotators compared to those of Black or Asian annotators. This suggests (1) greater differences in annotated score distributions between ethnic groups than between genders, and (2) a bias in LLMs' perception of subjective tasks towards the perspectives of White people.

4 Does Sociodemographic Prompting Improve Alignment with Demographic Groups?

Analysis To tackle this research question, we modify the prompt in Appendix Table 3 and add demographic information when prompting the model to predict group-based ratings on offensiveness and politeness (e.g., "How offensive does a White person think the following text is?"). We then further run separate linear mixed effect regression models to predict the change in the model's absolute prediction errors when being prompted with demographic tokens. Instance IDs are controlled as a random effect to account for the instance level variations.

Results Figure 2 shows the change in model performance when adding identity tokens into prompts. In the plots, a point above 0 suggests that adding

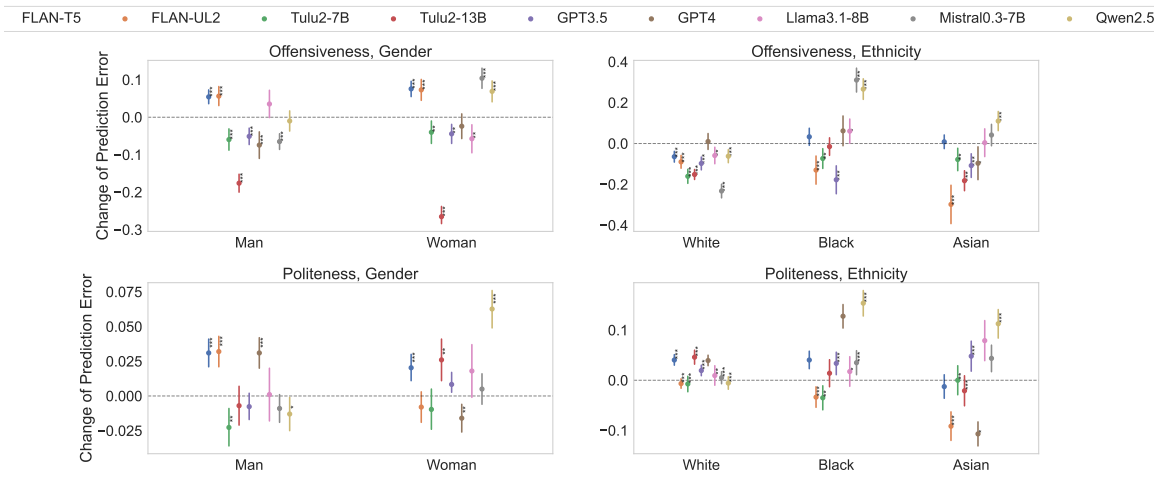

Figure 2: Regression results for predicting the prediction errors with different prompt settings. Each point shows the change of prediction errors when adding identity to the prompt for both tasks, relative to an identity-free prompt. Overall adding demographic tokens in prompts does not consistently improve the LLMs' performance for predicting ratings from different demographic groups.

the identity token increases the model's prediction errors while points below 0 suggest that the identify token helps to improve the models' prediction performances. We find that in some settings, adding the identity token helps models adjust their predictions. For example, adding the ethnicity token helps GPT3.5 and FLAN-UL2 to better predict the offensiveness ratings from the Asian participants. However, such an improvement is not consistent across tasks and models. For example, while adding an ethnicity token helps GPT3.5 in predicting offensiveness ratings from the Black participants, it does not help GPT4 at all. On the contrary, adding identity tokens actually increases the GPT4 and GPT3.5's prediction errors for politeness ratings from Black participants. Such a result indicates that mitigating LLMs' prediction biases for subjective NLP tasks is challenging and adding identity tokens in prompts is insufficient.

5 Discussion

With the large-scale deployment of LLMs in our society, it becomes increasingly important to study whether LLMs are able to understand the preferences of different groups of people. Our results suggest that LLMs are more aligned toward certain demographic groups than others on subjective perception tasks. For both of our tasks, we find that all of our tested LLMs provide answers which are closer to the annotations of White annotators compared to other demographic groups. Our findings contribute to the newly growing knowledge of types of demographic biases inherent in LLMs when asked to solve subjective tasks (Feng et al., 2023), signaling caution for potential applications such as deploying LLMs for generating annotations at large scale (Ziems et al., 2023). We discover that, unfortunately, directly inserting demographic features into prompts does not consistently help models "think" from the perspective of the target groups. The ability of LLMs to consider various opinions, at least from the perspective of demographic groups, seems limited at its current stage.

6 Conclusion

We examine the potential gender and racial bias of LLMs on two subjective NLP tasks: politeness and offensiveness. We find that LLMs' predictions are closer to White people's perceptions for both tasks and across 9 models. While we observe no significant gender differences in offensiveness prediction for most of the models, LLMs' predictions for politeness are significantly closer to women's ratings. We further explore whether incorporating identity tokens into the prompt helps mitigate this bias. Surprisingly, we find that adding identity tokens (e.g. "Black" and "Man") does not consistently help to improve the models' performance at predicting demographic-specific ratings. Our results suggest that LLMs may hold implicit biases on subjective NLP tasks and we call for future studies to develop de-biasing technologies for LLMs.

7 Ethics

This study investigates LLMs' capability to represent the opinions of different demographic groups when producing answers for subjective NLP tasks such as detecting offensiveness or politeness. As LLMs are increasingly being deployed in various settings that require subjective opinions, the fact that their opinions are significantly biased towards certain gender and ethnic groups raises a problem in their ability to remain neutral and objective regarding different tasks. Especially, prior work has shown that LLMs can produce biased and toxic responses when generating text provided the personas of specific individuals (e.g. Muhamad Ali) (Deshpande et al., 2023). When conducting studies on LLMs to understand how they can simulate the opinions or perspectives of a particular individual or social group, the research should be guided toward a direction that can overcome existing problems instead of introducing new problems such as AI-generated impersonation. Following, we discuss the ethical implications of our study.

During this study, we made the decision to only use the men and women gender labels from Popquorn, which unfortunately gives the appearance of an implicit binary assumption of gender. This choice is solely motivated by the absence of other gender identities in that dataset; while Popquorn is the largest and most diverse, due to the relative rareness of other gender identities in the crowdsourcing pool they used, no additional identities are available without additional data collection on our part, which we view as outside the scope of this paper. However, we acknowledge that our experiment settings miss out on non-binary forms of gender representation, which was inevitable due to data availability and how the original dataset was constructed. Nevertheless, the representativeness of non-binary individuals and groups in LLMs is also an important topic regarding potential disproportionateness. We call for future work in this direction to expand the inclusiveness of all types of social groups in their data collection.

When conducting large-scale analyses on datasets using LLMs, another topic of interest is minimizing financial costs and environmental impact. In this study, we do not require any finetuning or training stages and experiment only by inferring prediction results from publicly available LLMs. Except for GPT-3.5 and GPT-4, all models were able to run on a single A5000 GPU and took around six hours to run on the entire dataset under a single setting.

8 Limitations

Our study has the following limitations: (1) Although we aim to include most updated and popular LLMs into the analysis, we only experiment with a limited number of them due to the computational cost of running these experiments. We will release all the scripts to allow future researchers to test other models' performance in understanding group differences. (2) In our experiment settings, we only select limited types of ethnicity and gender categories for analysis due to the sparsity of labels from people with other identities in the POPQUORN dataset; therefore, our study didn't include several important identity groups such as non-binary genders and Hispanic people. (3) We only studied two tasks: offensiveness ratings and politeness ratings. As the datasets used for annotating these tasks come from offensive Reddit comments and polite emails, the biases reported in this study may not generalize to other datasets and task settings. (4) Our model predictions take the form of ordinal values, whereas the averaged annotation scores are fractional values. (5) We do not examine intersectional identities due to sparsity when subsetting the data, while the bias associated with populations defined by multiple categories leads to an incomplete measurement of social biases (Hancock, 2007). (6) We observe that some models, particularly GPT3.5 and Tulu2, have a relatively high refusal rate when asked to providing ratings, especially for offensiveness task and when prompts involve specific demographic groups such as Black people. Table 5 and Table 6 in Appendix D present the percentages of invalid responses by models and identity prompts. These implicit guardrails of LLMs may affect our findings, as the models might recognize the context but decline to respond due to privacy or ethical concerns.

References

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

Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo.

2024. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333:104145.

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. 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Ange-Marie Hancock. 2007. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on politics*, 5(1):63–79.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science.

OpenAI. 2023. Gpt-4 technical report.

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) @ACL 2023*.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. Towards coding social science datasets with language models.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023a. Whose opinions do language models reflect?

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023b. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Ilker Yildirim and LA Paul. 2024. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023. Sotopia: Interactive evaluation for social intelligence in language agents.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

Appendix

A Data

Table 1 and Table 2 demonstrate the descriptives for offensiveness and politeness ratings. Figure 3 further visualizes the distributions of annotations by demographic groups for both tasks.

	Overall	Man	Woman	White	Black	Asian
Mean	1.881	1.880	1.893	1.865	2.096	1.906
Std	0.756	0.861	0.893	0.790	1.270	1.147
N	1500	1483	1488	1500	1003	708

Table 1: Descriptives of offensiveness ratings by demographic groups

	Overall	Man	Woman	White	Black	Asian
Mean	3.305	3.318	3.308	3.304	3.403	3.196
Std	0.910	0.956	1.070	0.946	1.227	1.182
N	3718	3660	3681	3717	2222	1327

Table 2: Descriptives of politeness ratings by demographic groups

B Prompts

Table 3 provides an example of the prompts used in the experiments.

```
How does a person rate how
offensive the sentence is from 1
(not offensive at all) to 5 (very
offensive)
Options:  1, 2, 3, 4, 5
{Text}
Response (Provide the answer
without explaining your
reasoning):
```

Table 3: An example prompt for this study

Robustness check We test the robustness of our results with different prompt templates on four open-source LLMs: FLAN-T5, FLAN-UL2, Tulu2-DPO-7B, and Tulu2-DPO-13B. In the robustness check, we calculate the correlation coefficient between the LLMs' baseline predictions and the overall annotations representing the US population (the aggregated ratings for the entire sample). This overall correlation coefficient serves as a reference point for comparing the effects of applying different prompt templates. As shown in Figure 4, prompt templates have limited influence on the correlation coefficients between base prompts without identity words and labels representing the U.S population. It indicates that models' perception of offensivenss and politeness does not change much with different ways of prompting.

Table 4 lists the three prompt templates that are used as a robustness check.

	Templates
1	How does {identity} rate how {offensive/polite} the sentence is from 1 (not {offensive/polite} at all) to 5 (very {offensive/polite}). Options: 1, 2, 3, 4, 5: {text}"
2	If you are {identity}, how would you rate the {offensiveness/politeness} of the given sentence from 1 (not {offensive/polite} at all) to 5 (very {offensive/polite}). Options: 1, 2, 3, 4, 5: {text}"
3	Rate the {offensiveness/politeness} of the following sentence using a scale of 1 (not {offensive/polite} at all) to 5 (very {offensive/polite}) from the perspective of {identity}. Options: 1, 2, 3, 4, 5: {text}

Table 4: Prompt templates used in our study

C Experimentation details

Model Settings We generally follow the default settings recommended by the model developers. For model temperature, we set it to 0 for a more deterministic generation approach. This setting allows for a better understanding of the question "what is the most likely answer a model will return given a question and a specific group association?"

Computing infrastructure We use three NVIDIA RTX A5000 GPUs in our experiment. A full run of our experiment takes around 6 hours of GPU time.

Data and model license We have checked that our usage of the data and models are in compliance with the corresponding license.

Packages We used the following packages in our experiment: `accelarate`, `datasets`, `numpy`, `pandas`, `seaborn`, `statsmodels`, `transformers`.

D LLM Guardrails

When responding to potentially harmful queries, LLMs may refuse to provide an answer due to implicit guardrails designed to mitigate biases and protect users from inappropriate content. Table 5 and Table 6 summarize the percentages of invalid responses across nine LLMs when prompted with and without specific demographic information.

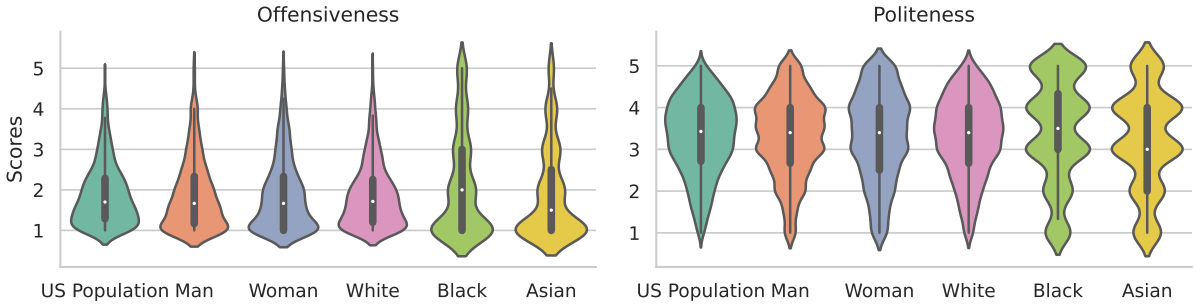

Figure 3: Distribution of annotations from different demographic groups for both offensiveness and politeness tasks

	Base	Man	Woman	White	Black	Asian
FLAN-T5	0.0%	0.1%	0.1%	0.1%	0.1%	0.0%
FLAN-UL2	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%
Tulu2-7B	7.5%	2.2%	3.6%	6.3%	14.3%	15.3%
Tulu2-13B	2.0%	3.2%	3.2%	3.4%	20.0%	13.0%
GPT 3.5	1.3%	4.0%	16.9%	23.1%	71.1%	44.8%
GPT 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Llama3.1-8B	0.4%	0.6%	0.5%	0.9%	0.9%	0.7%
Mistral0.3-7B	4.3%	3.5%	3.9%	13.5%	24.3%	13.7%
Qwen2.5-7B	0.7%	0.6%	0.7%	0.9%	1.3%	1.0%

Table 5: Percentages of invalid responses on offensiveness task

	Base	Man	Woman	White	Black	Asian
FLAN-T5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FLAN-UL2	0.0%	0.1%	0.1%	0.1%	0.1%	0.1%
Tulu2-7B	2.8%	1.6%	2.7%	2.9%	13.1%	7.2%
Tulu2-13B	1.7%	2.6%	2.6%	3.3%	9.7%	4.0%
GPT 3.5	0.1%	0.1%	0.1%	0.3%	6.5%	0.2%
GPT 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Llama3.1-8B	0.0%	0.0%	0.1%	0.1%	0.2%	0.2%
Mistral0.3-7B	0.4%	0.5%	0.9%	0.9%	3.6%	1.3%
Qwen2.5-7B	0.3%	0.4%	0.6%	0.6%	0.7%	0.7%

Table 6: Percentages of invalid responses on politeness task

E Usage of AI Assistants

We use AI assistants to check the grammar of our paper.

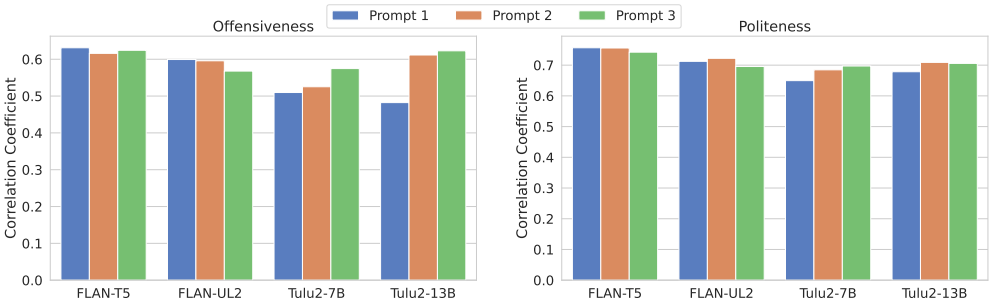

Figure 4: Models' performances do not change a lot when being prompted with different templates.