Sociodemographic Prompting is Not Yet an Effective Approach for Simulating Subjective Judgments with LLMs

Huaman Sun
University of Toronto
`hm.sun@mail.utoronto.ca`

Jiaxin Pei
Stanford University
`pedropei@stanford.edu`

Minje Choi
Amazon
`minjec@amazon.com`

David Jurgens
University of Michigan
`jurgens@umich.edu`

Abstract

Human judgments are inherently subjective and are actively affected by personal traits such as gender and ethnicity. While Large Language Models (LLMs) are widely used to simulate human responses across diverse contexts, their ability to account for demographic differences in subjective tasks remains uncertain. In this study, leveraging the POPQUORN dataset, we evaluate nine popular LLMs on their ability to understand demographic differences in two subjective judgment tasks: 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 only a minor gender bias favoring women appears in the politeness task. Furthermore, sociodemographic prompting does not consistently improve and, in some cases, worsens LLMs' ability to perceive language from specific sub-populations. These findings highlight potential demographic biases in LLMs when performing subjective judgment tasks and underscore the limitations of sociodemographic prompting as a strategy to achieve pluralistic alignment. Code and data are available at: https://github.com/Jiaxin-Pei/LLM-as-Subjective-Judge.

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; Choi 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 challenges in handling subjective tasks. It is well acknowledged that social biases and stereotypes embedded in their training data can lead to inadequate representation of diverse human experiences (Santurkar et al., 2023a). As a result, using LLMs for subjective tasks risks producing outcomes that disproportionately favor certain demographic groups, leading to biased or unfair results (Liang et al., 2021). Santurkar et al. (2023a) found that when responding to value-based questions, LLMs tend to align more closely with the perspectives of lower-income, moderate, and Protestant or Roman Catholic individuals. Despite these early findings, limited research has explored whether LLMs exhibit similar systemic biases with certain social groups across other subjective NLP tasks, highlighting the need for further investigation into their broader implications.

Subjective tasks present an additional challenge because language perception is shaped by social context and identity (Al Kuwatly et al., 2020). For instance, a text perceived as polite or inoffensive by one group may be interpreted differently by another. Ideally, LLMs should capture the full spectrum of subjective judgments. Steerable pluralism, as described by Sorensen et al. (2024), refers to an LLM's ability to be faithfully adjusted to represent specific perspectives. Yet, Miehling et al. (2024) found that many current LLMs have limited steerability to take on various persona, due

to both inherent biases in their baseline behavior and asymmetries in how they adapt across different persona dimensions. These limitations suggest that while steerability is a promising direction, it requires more refinement to effectively capture diverse perspectives.

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 sociodemographic 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 aligning gender and racial differences

in subjective judgment. While some research aims to directly 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` and `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 shows 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 (Qwen et al., 2025).

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?

While individual judgments may vary, LLMs can generate only a single prediction unless explicitly instructed to output a distribution. Therefore, when LLMs are applied to judgment tasks, it is crucial to examine whether their predictions align more closely with certain demographic groups.

Analysis To measure the alignment between LLM and certain demographic groups, 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 of a specific demographic group (target group) compared to the reference group, controlling for instance-level variations with instance ID as a random effect. For example:

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

A regression coefficient $\beta = 0$ indicates that there is 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, suggesting that the LLM predictions are closer to the reference group than to the target group. The aggregated results are visualized in Figure 1, while Table 5 in Appendix D provides detailed results from the linear mixed effects regressions.

Results As shown in Figure 1, LLMs' baseline prediction errors for offensiveness do not show 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 a greater bias in politeness prediction, suggesting that simply scaling models may

not effectively reduce biases in subjective tasks. Furthermore, LLMs' predictions for both politeness and offensiveness are consistently closer to the ratings of White annotators compared to those of Black or Asian annotators. This result reflects the intrinsic bias of LLMs on subjective judgment tasks.

4 Does Sociodemographic Prompting Improve Alignment with Demographic Groups?

Given the intrinsic bias of LLMs in subjective judgment tasks, a natural question arises: does adding demographic information in prompts steer LLMs to generate more diverse outputs that better align with specific groups? In this section, we conduct a series of analyses to answer this question.

Analysis To tackle this research question, we modify the prompt in Appendix B 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 and without demographic tokens. Instance IDs are controlled as a random effect to account for the instance level variations. Figure 2 illustrates the change in model performance when adding identity tokens into prompts. In the plots, points above 0 indicate that incorporating an identity token increases the model's prediction errors, while points below 0 suggest that the identity token improves prediction performance. Detailed regression results are provided in Table 6, Appendix D.

Results In Figure 2, our analysis reveals that in certain cases, identity tokens help models adjust their predictions. For instance, adding an ethnicity token improves GPT-3.5 and FLAN-UL2's ability to predict offensiveness ratings from Asian participants. However, this improvement is not consistent across tasks and models. While adding an ethnicity token helps GPT-3.5 better predict offensiveness ratings from Black participants, it has no effect on GPT-4. In contrast, identity tokens actually increase prediction errors for politeness ratings from Black participants in both GPT-3.5 and GPT-4. These findings highlight the challenges of mitigating LLM prediction biases in subjective NLP tasks
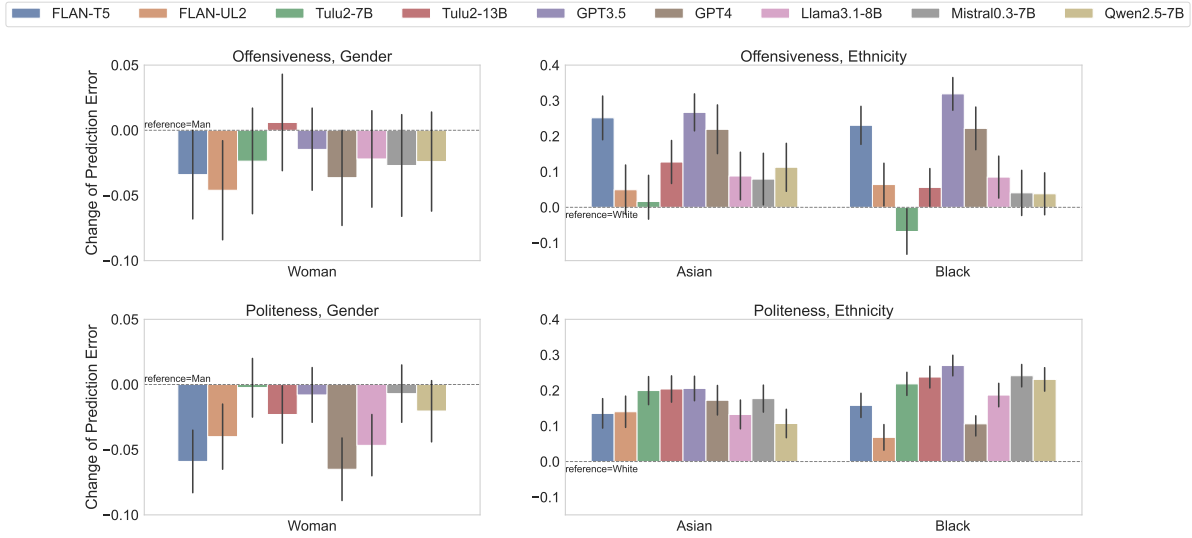
Figure 1: Regression results for predicting the gap between model predictions and the labels from each demographic group. The models' predictions for offensiveness are not significantly different from the ratings by Men and Women except for FLAN-UL2 (Top left). However, LLMs' predictions are significantly 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).
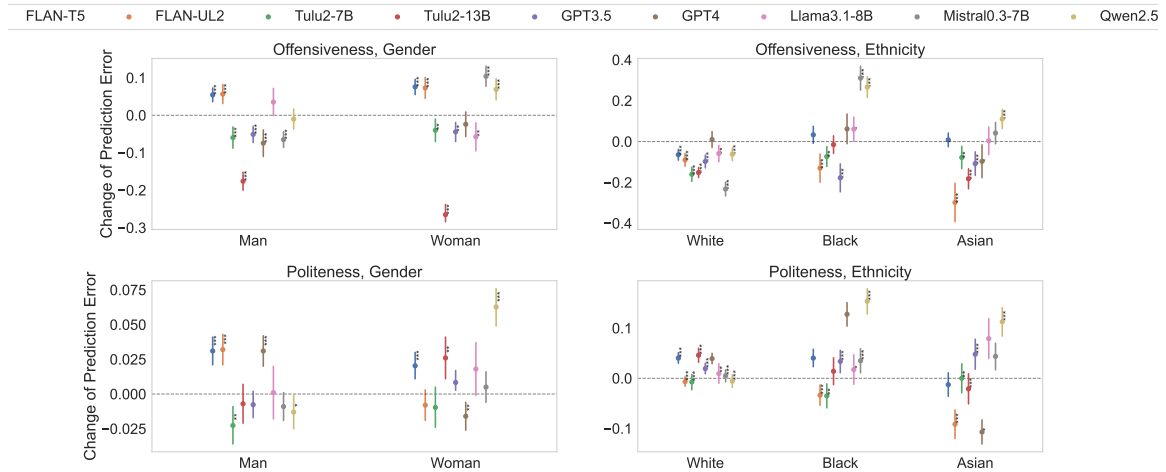

Figure 2: Regression results for predicting the prediction errors with different prompt settings. Each point shows the change of prediction errors when adding identity to the prompt for both tasks, relative to an identity-free prompt. Overall adding demographic tokens in prompts does not consistently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

and suggest that incorporating sociodemographic information in prompts is not yet a reliably effective solution.

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).

Our results also suggest that directly inserting demographic features into prompts, unfortunately, does not reliably help models adopt the perspectives of target groups. The ability of LLMs to consider various opinions, at least from the perspective of demographic groups, seems limited at its current stage. Furthermore, we observe that newer models, such as Mistral-0.3 and Qwen-2.5, exhibit reduced alignment on different task types and identity-based prompts. This may be due in part to increasingly strict guardrails designed to mitigate harmful outputs, which can also affect model performance by increasing refusal rates and limiting functionality (Bonaldi et al., 2024). Given that our tasks include sensitive keywords (e.g., *vulnerable identity, offensive, not polite*), these safety mechanisms may further contribute to the diminished effectiveness of identity-based prompting in newer models.

6 Conclusion

In this study, we study LLMs capability to account for demographic differences in subjective judgment tasks. We find that LLMs' predictions are closer to White people's perceptions for both tasks and across 9 models compared with Asian and Black people. We further explore whether incorporating demographic information 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 sociodemographic prompting is not an effective approach to address this bias yet. Researchers and practitioners should be careful when using LLM as judges on subjective tasks.

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 and 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 re-

garding different tasks. Especially, prior work has shown that LLMs can produce biased and toxic responses when generating text provided the personas of specific individuals (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 7 and Table 8 in Appendix E 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.

Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. Is safer better? the impact of guardrails on the argumentative strength of llms in hate speech countering. *arXiv preprint arXiv:2410.03466*.

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.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and Jurgens, David. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *under reivew*.

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.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M Daly, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, and Miao Liu. 2024. Evaluating the prompt steerability of large language models. *arXiv preprint arXiv:2411.12405*.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

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*.

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

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.

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 Regression Results

In Table 5 and Table 6, we report the estimated fixed effects of predictors, along with their standard errors and statistical significance. Statistical significance is denoted by stars, where a p-value less than 0.05 is marked with one star (*), a p-value less than 0.01 is marked with two stars (**), and a p-value less than 0.001 is marked with three stars (***).
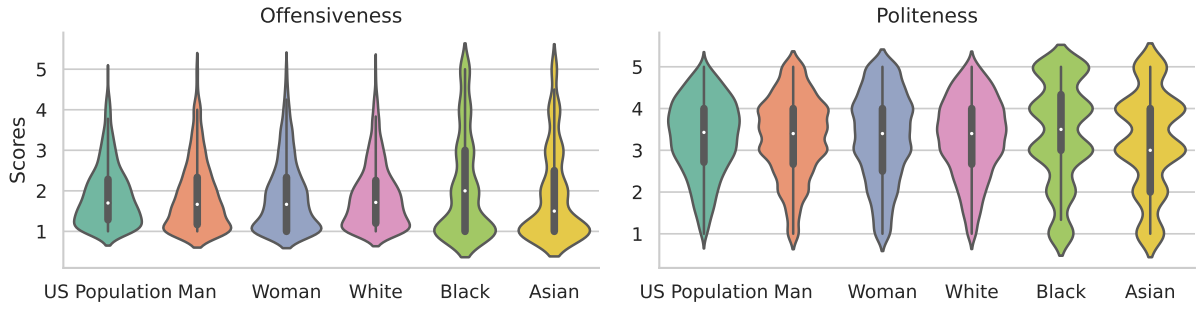
Figure 3: Distribution of annotations from different demographic groups for both offensiveness and politeness tasks.
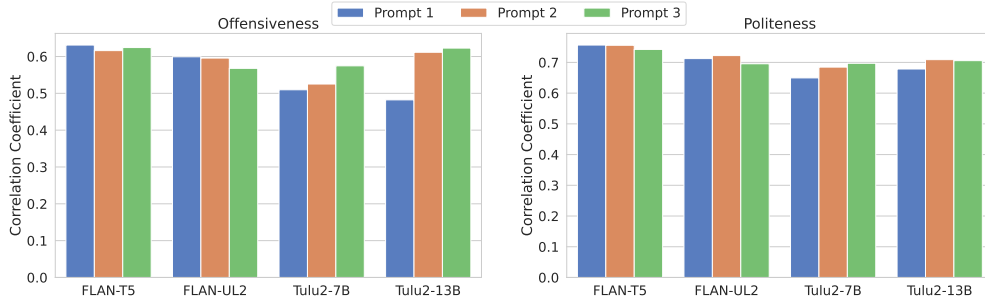

Figure 4: Models' performances do not change a lot when being prompted with different templates.

	FLAN-T5	FLAN-UL2	Tulu2-7B	Tulu2-13B	GPT3.5	GPT4	Llama3.1-8B	Mistral0.3-7B	Qwen2.5-7B
Offensiveness, Gender (reference=Man)									
Woman	-0.034	-0.046*	-0.024	0.006	-0.015	-0.036	-0.022	-0.027	-0.024
	(0.017)	(0.019)	(0.021)	(0.019)	(0.016)	(0.019)	(0.019)	(0.02)	(0.019)
Offensiveness, Ethnicity (reference=White)									
Black	0.231***	0.064*	-0.068*	0.056*	0.319***	0.222***	0.085**	0.041	0.038
	(0.027)	(0.031)	(0.033)	(0.027)	(0.023)	(0.031)	(0.03)	(0.032)	(0.03)
Asian	0.252***	0.049	0.016	0.127***	0.267***	0.219***	0.088**	0.079*	0.113**
	(0.031)	(0.035)	(0.038)	(0.031)	(0.027)	(0.035)	(0.034)	(0.037)	(0.034)
Politeness, Gender (reference=Man)									
Woman	-0.059***	-0.04**	-0.002	-0.023*	-0.008	-0.065***	-0.047***	-0.007	-0.02
	(0.012)	(0.013)	(0.08)	(0.011)	(0.011)	(0.012)	(0.012)	(0.011)	(0.012)
Politeness, Ethnicity (reference=White)									
Black	0.158***	0.068***	0.218***	0.238***	0.27***	0.106***	0.187***	0.241***	0.231***
	(0.017)	(0.019)	(0.017)	(0.015)	(0.015)	(0.017)	(0.017)	(0.016)	(0.017)
Asian	0.135***	0.14***	0.2***	0.204***	0.206***	0.172***	0.132***	0.177***	0.107***
	(0.021)	(0.023)	(0.02)	(0.019)	(0.018)	(0.021)	(0.021)	(0.02)	(0.02)

Table 5: Regression results for predicting the gap between zero-shot model predictions and the labels from each demographic group.

E   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 7 and Table 8 summarize the percentages of invalid responses across nine LLMs when prompted with and without specific demographic information.

F   Usage of AI Assistants

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

	FLAN-T5	FLAN-UL2	Tulu2-7B	Tulu2-13B	GPT3.5	GPT4	Llama3.1-8B	Mistral0.3-7B	Qwen2.5-7B
Offensiveness, Gender									
Man	0.054***	0.056***	-0.06***	-0.176***	-0.051***	-0.074***	0.035	-0.065***	-0.01
	(0.009)	(0.013)	(0.015)	(0.012)	(0.011)	(0.018)	(0.019)	(0.011)	(0.014)
Woman	0.076***	0.073***	-0.04**	-0.265***	-0.044**	-0.024	-0.057**	0.104***	0.069***
	(0.011)	(0.014)	(0.015)	(0.014)	(0.013)	(0.017)	(0.019)	(0.014)	(0.014)
Offensiveness, Ethnicity									
White	-0.064***	-0.09***	-0.16***	-0.152***	-0.097***	0.01	-0.059**	-0.232***	-0.062***
	(0.014)	(0.016)	(0.018)	(0.013)	(0.016)	(0.02)	(0.021)	(0.017)	(0.016)
Black	0.033	-0.13***	-0.073**	-0.015	-0.177***	0.062	0.061*	0.311***	0.266***
	(0.021)	(0.035)	(0.025)	(0.022)	(0.035)	(0.037)	(0.03)	(0.03)	(0.026)
Asian	0.008	-0.298***	-0.078**	-0.182***	-0.108***	-0.097*	0.004	0.042	0.11***
	(0.017)	(0.048)	(0.028)	(0.025)	(0.029)	(0.041	(0.035)	(0.027)	(0.024)
Politeness, Gender									
Man	0.031***	0.032***	-0.023**	-0.007	-0.008	0.031***	0.001	-0.009	-0.013**
	(0.005)	(0.006)	(0.007)	(0.007)	(0.005)	(0.005)	(0.01)	(0.005)	(0.006)
Woman	0.02***	-0.008	-0.01	0.026**	0.008	-0.016**	0.018	0.005	0.063***
	(0.005)	(0.006)	(0.007)	(0.008)	(0.004)	(0.005)	(0.01)	(0.006)	(0.007)
Politeness, Ethnicity									
White	0.04***	-0.007	-0.007	0.046***	0.019***	0.039***	0.009	0.005	-0.006
	(0.005)	(0.005)	(0.008)	(0.007)	(0.006)	(0.005)	(0.01)	(0.006)	(0.007)
Black	0.04***	-0.034**	-0.035**	0.014	0.034**	0.128***	0.017	0.035**	0.154***
	(0.009)	(0.01)	(0.012)	(0.014)	(0.012)	(0.012)	(0.015)	(0.012)	(0.013)
Asian	-0.013	-0.092***	-0.0	-0.021	0.048**	-0.107***	0.079***	0.044**	0.113***
	(0.012)	(0.015)	(0.015)	(0.015)	(0.015)	(0.012)	(0.02)	(0.014)	(0.014)

Table 6: Regression results for predicting the prediction errors when adding identity to the prompt, relative to an identity-free prompt.

	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 7: 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 8: Percentages of invalid responses on politeness task