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

**Anonymous ACL submission**

## Abstract

Human perception of language depends on personal backgrounds like gender and ethnicity. While existing studies have shown that large language models (LLMs) hold values that are closer to certain societal groups, it is unclear whether their prediction behaviors on subjective NLP tasks also exhibit a similar bias. In this study, leveraging the POPQUORN dataset which contains annotations of diverse demographic backgrounds, we conduct a series of experiments on six popular LLMs to investigate their capability to understand group differences and potential biases in their predictions for politeness and offensiveness. We find that for both tasks, model predictions are closer to the labels from White and female participants. We further explore prompting with the target demographic labels and show that including the target demographic in the prompt actually *worsens* the model's performance. More specifically, when being prompted to respond from the perspective of "Black" and "Asian" individuals, models show lower performance in predicting both overall scores as well as the scores from corresponding groups. Our results suggest that LLMs hold gender and racial biases for subjective NLP tasks and that demographic-infused prompts alone may be insufficient to mitigate such effects.

## 1 Introduction

Large language models (LLMs) have shown promising capability in handling a wide range of language processing tasks from dialogue generation to sentiment analysis, because of their ability to learn human-like language properties from massive training data (Brown et al., 2020; Radford et al., 2019). An increasing number of researchers have started to use the zero-shot capabilities of LLMs to handle subjective NLP tasks, such as simulating characters (Wang et al., 2023) and detecting hate speech (Plaza-del arco et al., 2023). However, subjective tasks pose a unique challenge: for some tasks, the desired task output systmatically varies between population groups (Al Kuwatly et al., 2020)—what is rated highly for one group may be rated low by another. Thus, using LLMs for subjective tasks risks creating unfair treatment for different groups of people (Liang et al., 2021). Santurkar et al. (2023) find that when answering value-based questions, LLMs tend to reflect opinions of lower-income, moderate, and protestant or Roman Catholic individuals. Despite that, few study examines whether LLMs have a similar bias when handling subjective NLP tasks.

In this study, we investigate whether LLMs are able to understand identity-based group differences in subjective language tasks. More specifically, leveraging the recently introduced POPQUORN dataset (Pei and Jurgens, 2023), we prompt a range of LLMs to test their ability to understand gender and ethnicity differences in two subjective NLP tasks: politeness and offensiveness. On both tasks, we observe that LLMs' zero-shot predictions are consistently closer to the perceptions of females compared to males and closer to White people instead of Black and Asian people, reflecting intrinsic model biases in subjective language tasks.

We further study the effect of directly adding demographic information when prompting the models. To account for the nuanced changes in prompts, we test a list of baseline prompts that do not include the demographic information (e.g. "Do you think the given comment would be offensive to a person?"). We find that, compared with the baseline prompts, adding demographic information actually led to a *lower* prediction performance. This pattern is consistent across different models.

Our study suggests that large language models are not fully capable of understanding gender and racial differences in subjective language tasks. Although some studies attempt to use LLMs to mimic behaviors of different groups or do data augmentation for subjective tasks, our result reveals the

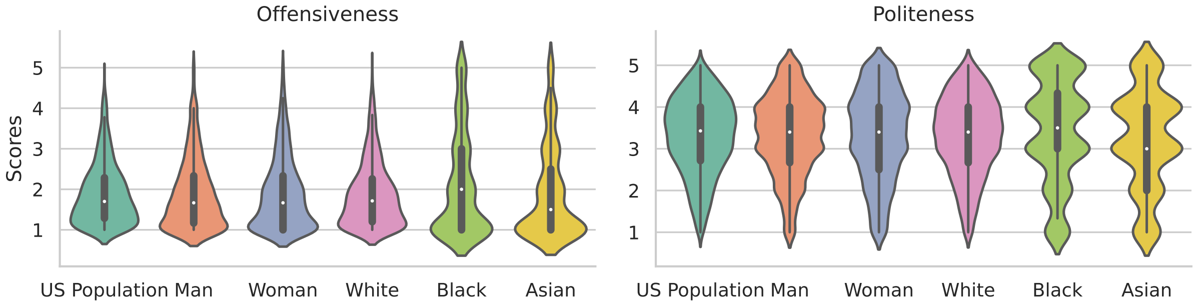

Figure 1: Distribution of annotation results from different groups for both offensiveness and politeness tasks

potential risks in introducing further biases.

## 2 LLMs and Social Factors

A large line of recent work regarding LLMs has looked into whether they contain knowledge of social factors analogous to that of human (Zhou et al., 2023). Some studies measure LLMs' specific sets of personalities when prompted using established questionnaires of psychological traits (tse Huang et al., 2023; Binz and Schulz, 2023; Miotto et al., 2022; Pan and Zeng, 2023). Given this personality, studies have tried to use LLMs to provide large-scale labeling of tasks requiring social understandings with promising results (Ziems et al., 2023; Rytting et al., 2023). However, LLMs are also not perfect: the model outputs do not well represent the human population due to innate biases arising from the data used to train the models. This leads to LLMs being potentially biased with respect to gender (Lucy and Bamman, 2021) or political ideology (Liu et al., 2022), and also failing to represent particular demographic groups (Santurkar et al., 2023). Further, prompting itself possesses limitations such as being sensitive to the complexity or order of prompt sentences inputted to the model (Mu et al., 2023; Dominguez-Olmedo et al., 2023). A recent study that is in similar line with ours is that of Beck et al. (2023) which uses sociodemographic factors as prompts to examine model performance on several different tasks. While their methodology is similar to ours, we provide different findings, as our work tests whether these prompts are actually helping LLMs align more with the opinions provided by samples of the specified demographics.

## 3 Dataset and Method

**Data** We use the POPQUORN dataset (Pei and Jurgens, 2023) as our testbed for experiments on LLMs. POPQUORN includes 45,000 annotations drawn from a representative sample of the U.S. population and is diverse in terms of demographics such as ethnicity and gender. For this study, we utilize annotators' offensiveness and politeness ratings, where each task is a 5-point Likert rating.

This study examines two types of identities: gender and race. We focus on the categories of ['Woman', 'Man'] for gender, and ['Black', 'Asian', 'White'] for race, as these have sufficient statistical power to draw conclusions. For each instance, we compute the average politeness and offensiveness scores both for each identity group as well as for the entire sample of annotators. These average scores serve as the measures of the specific group's different perceptions.

Figure 1 shows the distributions of both overall and identity-specific scores for offensiveness and politeness tasks. In terms of offensiveness, the overall scores show a mean of 1.88 with a standard deviation of 0.76. For politeness, the mean of overall scores stands at 3.31 with a standard deviation of 0.91. For both tasks, the scores from men, women, and White annotators closely mirror the overall score distribution. Scores from Black and Asian annotators, however, present diverged mean and increased standard deviation.

**Models** To increase 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, and GPT-4 (OpenAI, 2023).

**Prompts** We design prompts to instruct the models to predict offensiveness and politeness scores for each instance. Prompts were selected after preliminary experiments on a small scale to verify whether the prompt can elicit valid responses. An example prompt used in our experiments is illustrated in Appendix Table 1, and Appendix Table 2 presents the

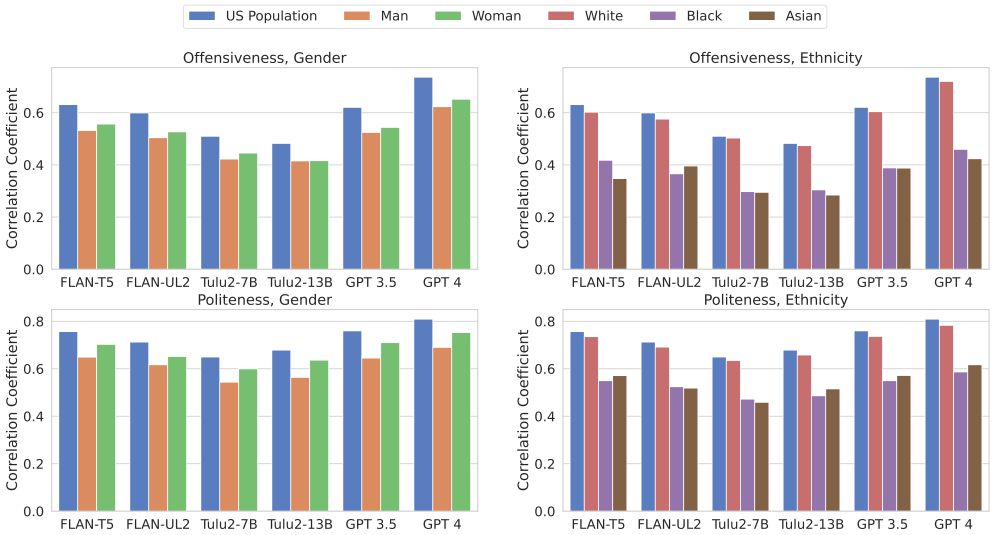

Figure 2: A comparison of the correlations between the LLM-generated responses and the annotations from different social groups. Model predictions are closer to women and White people's ratings of both offensiveness and politeness.

list of all prompts used in our study. While prompts differ slightly in performance, our findings consistently align across tested prompts, as detailed in the following sections.

**Evaluation** As the labels are on a scale from 1 to 5, for both subjective NLP tasks on offensiveness and politeness we evaluate our results by measuring Pearson's $r$ between the model's predictions and the aggregated human ratings.

## 4 Are Model Predictions Closer to Certain Demographic Groups?

We compare models' predictions derived from baseline prompts without any demographic information to the average ratings provided by annotators within each identity group. The correlation coefficients between the models' baseline predictions and identity-specific labels are shown in Figure 2.

**Gender** As shown in Figure 2 (left), LLMs' perceptions of both subjective tasks tend to align slightly more with the perceptions of women than those of men. For both tasks, the correlation between the LLM-generated responses and the annotated results is lower when conditioned on either gender, compared to that on the entire population. This implies that the annotations of men and women result in different distributions which are balanced out when averaged across the entire population, which is closer to the LLM's predictions despite being slightly more aligned towards female scores.

**Ethnicity** Figure 2 (right) demonstrates that models' predictions consistently have a higher correlation with White people's perception of both politeness and offensiveness, compared to those of Black or Asian people. These results suggest that (1) the annotated score distributions between ethnicity groups differ more than that between gender, and (2) LLMs' perception of subjective tasks is biased towards the perspectives of White people.

## 5 Does Adding Identity Tokens Improve Alignment with Demographic Groups?

In the previous section, we find that LLMs' predictions on subjective NLP tasks are biased towards certain demographic groups' perceptions. Given LLMs' capabilities of understanding natural language instructions, does adding identity tokens in prompts help models tune their predictions for specific demographic groups?

**Method** We modify the prompt in Table 1 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?").

**Results** Figure 3 shows the change in model performance when adding identity tokens into prompts. In the plots, the lighter bars represent the correlation between models' predictions with baseline prompts and group-based human labels, while the deeper bars represent the correlation between models' perceptions derived from identity prompts and the average rating from the corresponding group.

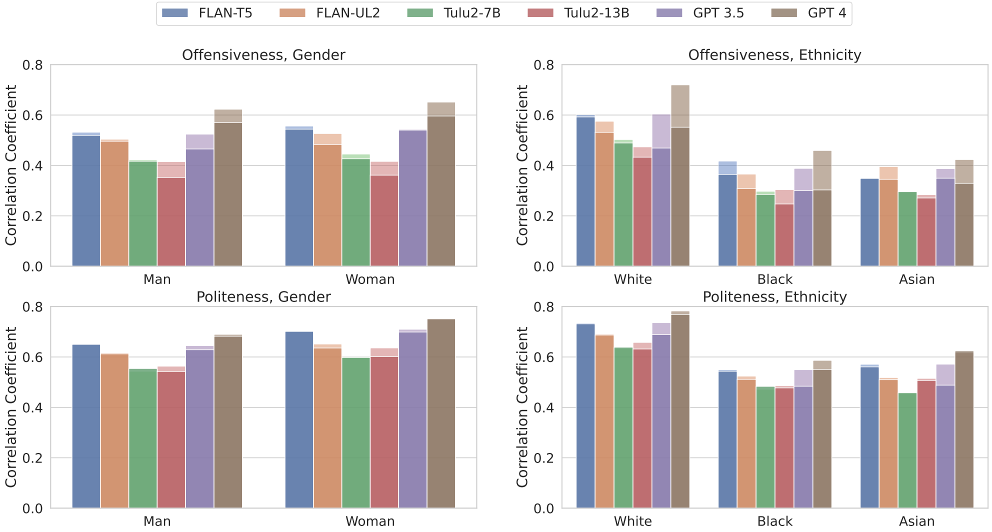

Figure 3: Change of performance w/o identity prompts for both tasks. Lighter bars represent correlations between models' predictions with baseline prompts and group-based human labels, while darker bars represent correlations between models' perceptions derived from identity prompts and the average rating from the corresponding group. Adding identity tokens consistently worsens the models' performance on subjective NLP tasks.

We find that simply adding the identity token does not help models adjust their predictions. On the contrary, adding identity words negatively affects models' performance in predicting subjective perceptions of the specific demographic group. In addition, this effect is disproportional for different gender/racial groups, and it varies with different models and subjective NLP tasks. In general, we observe a minimal drop in correlation coefficients after adding the identity token of man and woman. However, the models' performance for both Asian and Black groups drops sharply after adding the words "Black" and "Asian" in the prompt. Such a result indicates that LLMs are not only biased in their predictions but also in the way that identity prompt affect their performances.

## 6 Discussion

Our experiment results support the belief that LLMs are more aligned toward certain demographic groups than others when asked to make decisions regarding tasks such as determining polite or offensive content. For both of our tasks, we find that all of our tested LLMs provide answers which are closer to the annotations of White, female 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 make models "think" from the perspective of certain demographic groups. This is verified by LLMs not better aligning with specific demographic groups when adding their terms to prompts. On the contrary, we observe a uniform performance decrease across all demographics when an LLM is prompted to think as a specific demographic group. The ability of LLMs to consider various opinions, at least from the perspective of demographic groups, seems limited at its current stage.

## 7 Conclusion

We examine the gender and racial bias of LLMs on two subjective NLP tasks: politeness and offensiveness. We find that LLMs' predictions are consistently closer to White and female people's perceptions, a pattern consistent across six popular LLMs like GPT4 and FLAN-UL2. 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") consistently lowers performance. Also, the drop in correlation when adding "Black" or "Asian" is significantly larger than that of adding "White" in the prompt. Our results suggest that LLMs may hold implicit biases on subjective NLP tasks and we call for future studies to develop de-biasing technologies to build fair and responsible LLMs.

## 8 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 a specific decision to categorize gender in a binary setting as male or female only. 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 social groups.

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.

## 9 Limitations

Our study has the following limitations: (1) We only experiment with a limited list of LLMs 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, 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, especially for ChatGPT, whereas the averaged annotation scores are fractional values. (5) We do not examine intersectional identities, while the bias associated with populations defined by multiple categories leads to an incomplete measurement of social biases (Hancock, 2007).

## 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.

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

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

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*.

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.

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.

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.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics.

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.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models.

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. 2023. Whose opinions do language models reflect?

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.

Jen tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R. Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language 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*.

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 Prompts

Table 1 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 1: An example prompt for this study

**Robustness check** We test the robustness of our results with different prompt templates on the four open-source LLMs: FLAN-T5, FLAN-UL2, Tulu2-DPO-7B, and Tulu2-DPO-13B. 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 2 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 2: Prompt templates used in our study

## B Experimentation details

**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`, `pandas`, `seaborn`, `transformers`.

## C Usage of AI Assistants

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

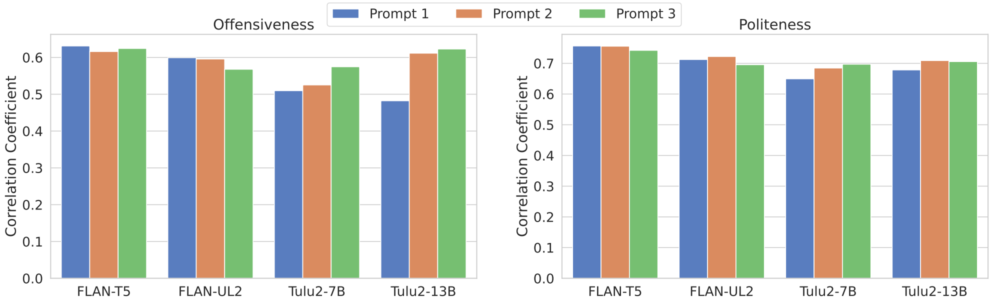

Figure 4: There is little change of models' performance when prompting with different templates.