Overcoming Affective Polarization in Generative Language Models

Keywords: large language models, affective polarization, ingroup favoritism, outgroup derogation, social bias

Extended Abstract

With over 100 million users already incorporating ChatGPT into their routines (Milmo, 2023), it is crucial to investigate the biases present in generative language models and find ways to mitigate them. While prior work has shown that generative models can magnify negative human biases present in the training data, such as gender or racial stereotypes (Bender et al., 2021), few work has investigated biases involving group identities. Affective polarization is a well-known psychological tendency that threatens the democratic process by increasing incivility and making intergroup compromise more difficult (Iyengar et al., 2019). Affective polarization is grounded in social identity theory, which states that when group identity is activated, individuals tend to favor people from their group (ingroup favoritism) and dislike those from other groups (outgroup derogation; Tajfel & Turner, 1979). These social identity biases are so profound that they occur even when people are randomly assigned to meaningless groups (Pinter et al., 2011).

In this work, we investigate social identity biases in generative language models. Specifically, we test (1) if GPT-2 (Radford et al., 2019) and its extensions fine-tuned on partisan social media data (Jiang et al., 2022) are more positive towards the general ingroup and more negative towards the general outgroup and (2) if removing the biased data in fine-tuning reduces downstream bias. Since words like "we" and "they" are associated with general ingroup-outgroup dynamics (Perdue et al., 1990), we generated ingroup or outgroup sentences by prompting with "We are" or "They are" and classified them as positive or negative with the VADER compound sentiment score (Hutto & Gilbert, 2014).[1] We then removed partisan Twitter data containing either negative sentences with "they" language or positive sentences with "we" language from the corresponding LIWC (Boyd et al., 2022) categories and fine-tuned GPT-2 on the modified data.[2] For brevity, we present results only for the model fine-tuned with Republican partisan tweets (henceforth Republican model) and results for the model with Democratic data show a similar pattern.

First, we find strong evidence of ingroup favoritism in GPT-2: the model generated a higher proportion of positive ingroup sentences (56.7%) than outgroup sentences (39.1%), one-sided $t = 8.00$, $p < .001$, see Fig. 1. There were also more negative outgroup sentences (20.8%) than negative ingroup sentences (14.7%), one-sided $t = 3.58$, $p < .001$, corresponding to outgroup derogation. The Republican model exihibits an even stronger bias. It produced more negative outgroup sentences than GPT-2 (39.8% vs. 20.8%, two-sided $t = 9.45$, $p < .001$) with no significant increase in ingroup favoritism (two-sided $t = -.856$, $p = .392$), corroborating negative partisanship accounts of affective polarization (see Fig. 2; Iyengar et al., 2019).

Second, removing valenced "we" and "they" sentences from the partisan training data and fine-tuning GPT-2 led to a significant decrease in the model's social identity bias.

[1] For all tests, we generated 1000 samples of at most 30 characters and only used the first sentences to determine sentiment scores. The VADER compound score ranges from -1, extremely negative, to +1, extremely positive, with established cutoff points of .05 and -.05 for positive and negative classification, respectively.

[2] We used partisan tweets collected and labeled by Jiang et al. (2022) to fine-tune GPT-2 for 5 epochs.

Removing all the negative "they" sentences from the training data increased the average sentiment score of the generated outgroup sentences to the level of the original GPT-2 (.12 vs. .10), and slightly increased overall ingroup sentiment compared to GPT-2 (.33 vs. .26), thus significantly reducing the gap in sentiment between ingroup and outgroup sentences. Additionally, removing negative "they" sentences led to a 34% decrease in the toxicity of outgroup sentences as measured by Google Perspective API[3] and can be observed qualitatively (see Tab. 1). Removing positive "we" sentences, on the other hand, decreased the overall sentiment for both ingroup and outgroup sentences (from .26 and .10 to -.007 and -.068, respectively), both to a level close to neutral. Interestingly, removing both types of sentences led to more neutral ingroup sentences and more positive outgroup sentences, seemingly reversing the bias (see Fig. 3). These results suggest that the social identity bias in GPT-2 can be traced back to the social identity bias in the training data, and that the model also learned from the average sentiment of the training data as a whole.

We show that, just as in humans, social identity biases are pervasive in language models and that the biases can be mitigated by carefully curating the training data. Since large language models like ChatGPT increasingly rely on human feedback, in addition to unsupervised training on internet data, it is concerning that these models may exhibit even stronger social identity biases than older models like GPT-2, as human annotators may prefer text that reflects these biases. With the wide adoption of these models, we may enter a vicious circle where social bias is fuelled by content generated partially or wholly by language models. This same source of text would then fuel even more social bias in the next generations of models. In the political context, this tendency could lead to increased affective polarization if left unchecked.

References

Bender, E. M., et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM FAccT*.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *UT Austin*, 1-47.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *ICWSM-14*.

Iyengar, S., et al. The Origins and Consequences of Affective Polarization in the United States. *Annu. Rev. Polit. Sci.* 22, 129–146 (2019).

Jiang, H., Beeferman, D., Roy, B., & Roy, D. (2022). CommunityLM: Probing Partisan Worldviews from Language Models. *COLING-2022*.

Milmo, Dan. (2023). ChatGPT reaches 100 million users two months after launch. *The Guardian*.https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app. (Accessed 28 Feb 2023).

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology, 59*(3), 475–486.

Pinter, B., & Greenwald, A. G. (2011). A comparison of minimal group induction procedures. *Group Processes & Intergroup Relations*, *14*(1), 81-98.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. An integrative theory of intergroup conflict. *Organ. Identity: A reader* 18, 56–65 (1979).

[3] https://perspectiveapi.com/

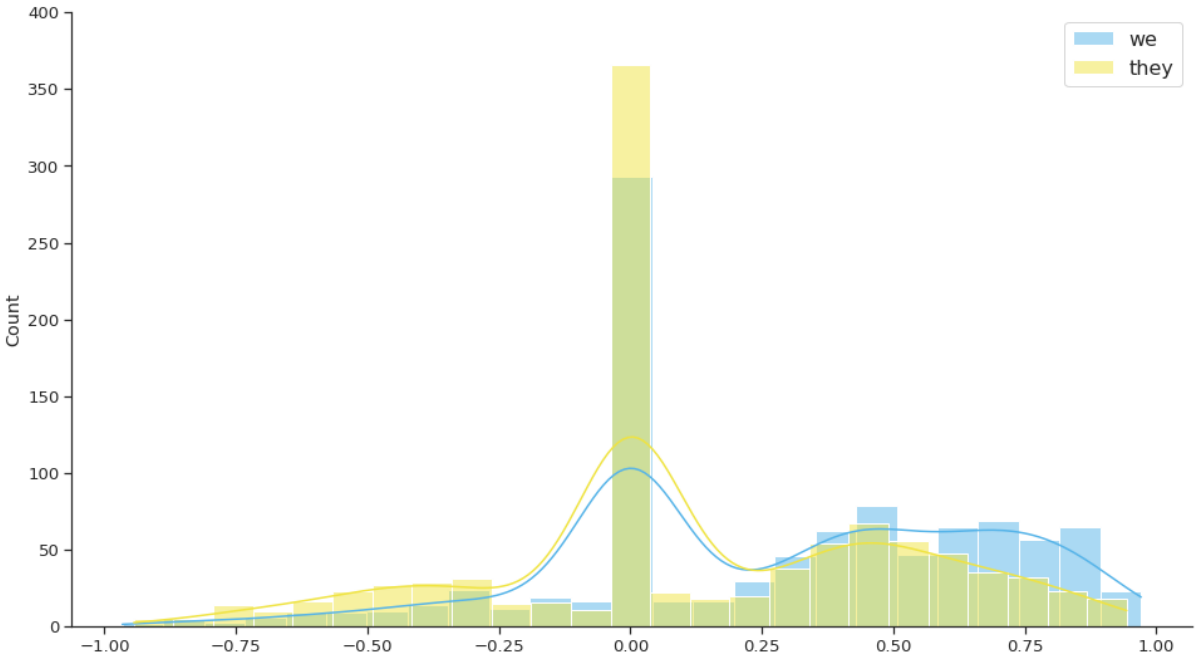

Figure 1. Distribution of compound sentiment of GPT-2 generated sentences starting with "We are" versus "They are," values closer to 1 signify more positive valance while values closer to -1 mean more negative.

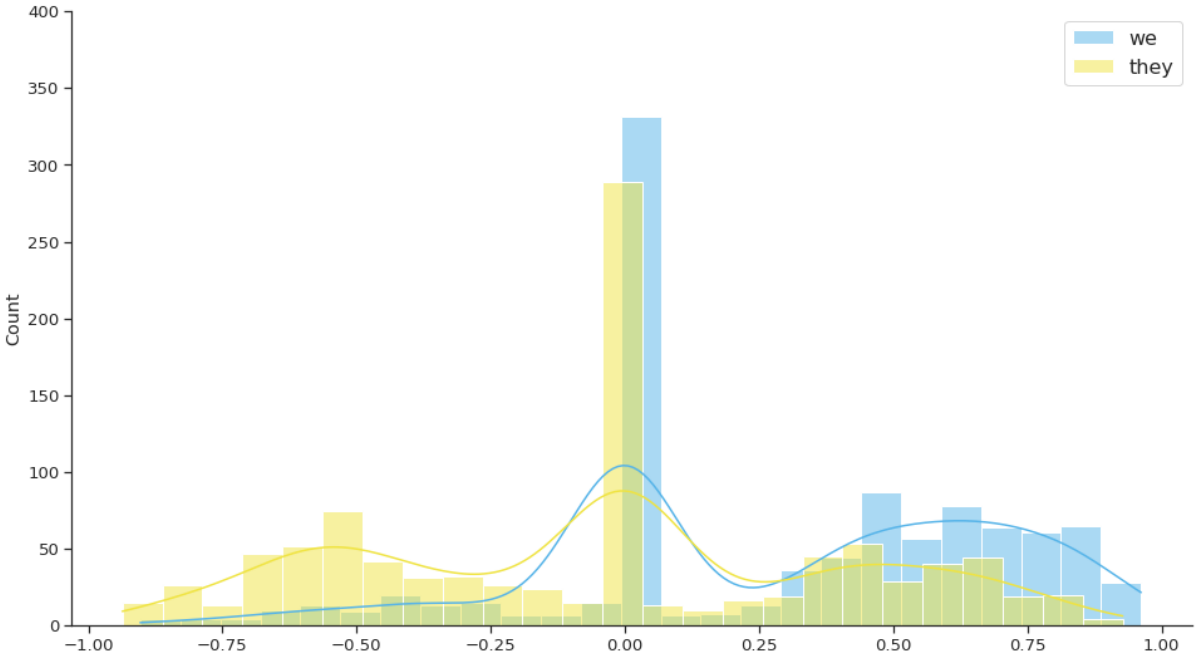

Figure 2. Distribution of compound sentiment of Republican GPT-2 generated sentences starting with "We are" versus "They are." More outgroup than ingroup sentences were negative, signifying outgroup derogation. Similarly, more ingroup than outgroup sentences were positive, reflecting ingroup favoritism. Overall, the Republican model demonstrates a marked increase in outgroup derogation as compared to the original GPT-2. Similar results hold for the Democrat model.

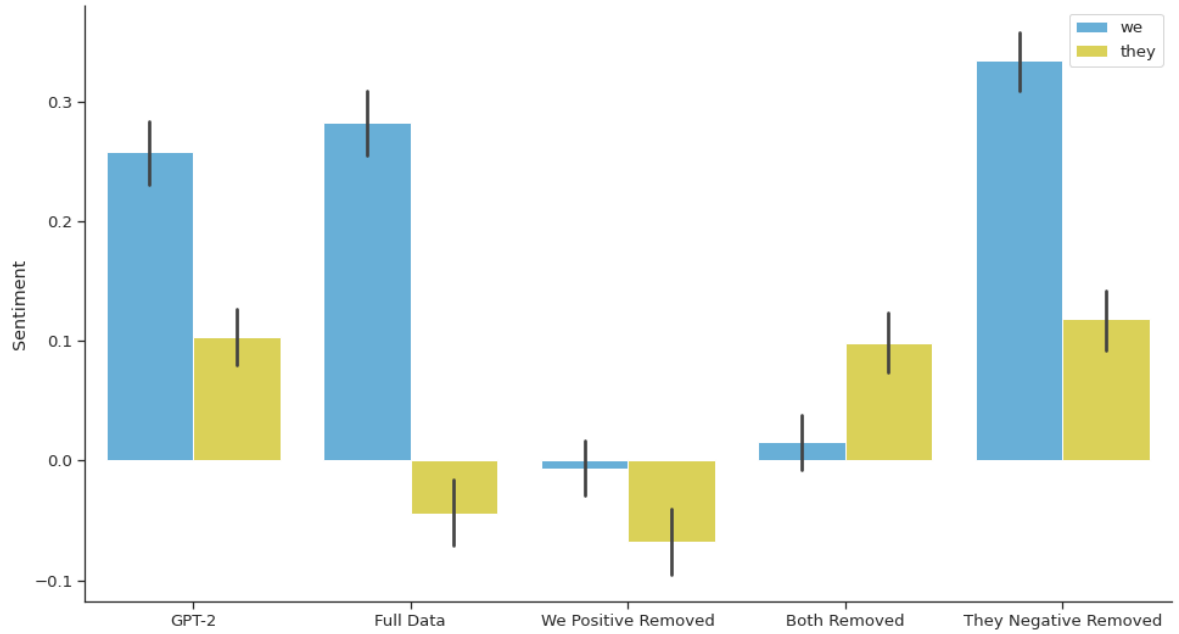

Figure 3. Average sentiment of sentences generated by GPT-2 and Republican models with different training data.

Table 1. Randomly sampled examples from the Republican model trained on full data and the Republican model trained on a subset of data without negative "they" sentences.

Republican model	Republican model: 100% They Negative Removed
'They are not a good group.', 'They are in serious need.', 'They are all losers.', 'They are a lot of fun to do.', 'They are the leaders but for that they are dead wrong they lost!'	'They are so very true!!!', "They are still going on my list every day, just don't know why\nI am not sure if I can afford to pay her but she", "They are the one reason the US is leading in population reduction and the only reason we're winning in Europe.", 'They are called the " racist " right?', "They aren't all black and their party and do NOT hate us."