The Elusiveness of Detecting Political Bias in Language Models

Riccardo Lunardi
riccardo.lunardi@uniud.it
University of Udine
Udine, Italy

David La Barbera
david.labarbera@uniud.it
University of Udine
Udine, Italy

Kevin Roitero
kevin.roitero@uniud.it
University of Udine
Udine, Italy

Abstract

This study challenges the prevailing approach of measuring political leanings in Large Language Models (LLMs) through direct questioning. By extensively testing LLMs with original, positively and negatively paraphrased Political Compass questions we demonstrate that LLMs do not consistently reveal their political biases in response to standard questions. Our findings indicate that LLMs' political orientations are elusive, easily influenced by subtle changes in phrasing and context. This study underscores the limitations of direct questioning in accurately measuring the political biases of LLMs and emphasizes the necessity for more refined and effective approaches to understand their true political stances.[1]

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Large Language Models, Political Bias, Probing.

ACM Reference Format:
Riccardo Lunardi, David La Barbera, and Kevin Roitero. 2024. The Elusiveness of Detecting Political Bias in Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3680002

1 Introduction

LLMs marked a significant advancement in the field of Natural Language Processing (NLP), offering unprecedented capabilities in generating human-like text and solving complex tasks [2, 16]. These models, due to their extensive training on diverse internet-based corpora, have the potential to reflect and amplify the biases present in their training data [18, 20]. Among various biases like those based on race, gender, or religion [1, 3, 9, 15], political bias in LLMs is particularly concerning, as it can subtly influence public opinion, reinforce societal stereotypes, and raise societal decision-making issues in sensitive domains.

The inherent complexity of political bias, often more subtle, subjective, and context-dependent than other biases, poses unique challenges for its detection and measurement. In this paper we argue that traditional methods for assessing political bias in LLMs, such as direct questioning with standard political questionnaires, are not sufficiently reliable to uncover the true latent political leanings of these models. In particular, we find that direct questioning methods are heavily influenced by phrasing and context. This underscores the necessity for more sophisticated approaches in evaluating the political biases and leanings of LLMs.

In this work we focus on the following Research Questions (RQ):

RQ1: How do changes in phrasing and context affect the responses of LLMs to political bias assessments?
RQ2: How do differences in responses impact the measured political leanings of LLMs?

2 Background and Related Work

Despite their impressive capabilities, LLMs are prone to various biases that can affect their reliability [7, 21, 22]. Among the many possible biases that a LLM can present, identification and mitigation of political bias is a growing studied area, mostly due to its potential impact on shaping public opinion and decision-making. Feng et al. [6] investigated media biases in LLMs, revealing how these models can reinforce political polarization present in their training data. Gover [8] focused on the political bias in GPT models, finding a moderate left-leaning bias. Studies by Liu et al. [13] and Liu et al. [14] developed methods for quantifying and mitigating political biases in LLMs trough a reinforcement learning framework. Stańczak et al. [19] explored gender bias in the context of politics, meanwhile Liang et al. [12] addressed broader social biases in LLMs, defining metrics for bias measurement and proposing mitigation strategies. Finally, Motoki et al. [17] investigate the political bias of ChatGPT by having the model impersonate different political perspectives, running robustness tests, and analyzing responses to determine presence of systematic political bias, finding that ChatGPT exhibits bias toward certain political groups.

While existing methodologies in the field predominantly rely on direct questioning of LLMs to assess their biases, such as Feng et al. [6] which briefly addressed perturbation of the input text, our work builds on top of that and examines in detail the robustness of this approach. We argue about the limitations and inadequacies of direct questioning in capturing the true extent of model biases.

3 Methodology

3.1 Measuring Political Leaning of LLMs

The Political Compass test[2], established in 2001, is a tool designed to offer a complex understanding of political ideologies. It assesses individuals' political views on two axes: the *economic axis* (left–right) and the *social axis* (authoritarian–libertarian) [4, 5, 10]. The test includes 62 propositions, covering various topics such as economic policy, social freedoms, and personal values. Participants respond

[1]The code, data, and all the supplementary material is available at: https://osf.io/f6de5/?view_only=eabc76a845ac4c85a26dd4c3c79634ae.

[2]https://www.politicalcompass.org/test.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680002

to these propositions on a four-level scale using the options STRONG DISAGREE, DISAGREE, AGREE, and STRONG AGREE. The results plot the individual's position on one of four quadrants: authoritarian-left, authoritarian-right, libertarian-right, and libertarian-left.[3]

In our study, we use the Political Compass test for directly questioning the LLMs, similar to the approach taken in the literature. This method involves posing the 62 propositions from the test to the LLMs, and then submit their responses to measure their political leanings. To test the robustness of these methods, we use the original propositions along with their negations, 3 positively paraphrased variants (maintaining the same stance as the original questions), and 3 negatively paraphrased versions (carrying the opposite stance). All paraphrasing has been performed using ChatGPT and have been manually checked independently by every author for accuracy to ensure their integrity. The full set of propositions used in our study and all the models version and checkpoints are available in the supplementary material.

3.2 Polar and Paraphrastic Consistency

We define two key concepts for evaluating the LLMs on the tests: *Polar* and *Paraphrastic* consistency. *Polar consistency* refers to the ability of a model to maintain congruence in its responses when presented with two propositions having positive and negative stance. High polar consistency indicates a clear understanding of the propositions and a consistent stance on the topic measured on the negation. *Paraphrastic consistency*, on the other hand, refers to the model's ability to maintain congruence to paraphrases of the propositions carrying the same stance. High paraphrastic consistency implies that the model comprehends the topic and expresses a consistent opinion across paraphrases. Paraphrastic consistency alone, however, cannot be the sole metric for evaluating LLMs, as a model could achieve 100% paraphrastic consistency by merely providing the same answer, and it must be assessed in conjunction with polar consistency.

In evaluating the polar and paraphrastic consistency we differentiate between encoder- and decoder-only models. Following Feng et al. [6], for the former models we use the prompt

```
Please respond to the following statement:
[PROPOSITION]. I <MASK> with this statement.
```

while for the latter

```
Please respond to the following statement:
[PROPOSITION]. Your response:
```

The encoder-only LLMs are then asked to return the 10 highest probability tokens to fill-the-mask. To map their answers to the four-level scale used in this paper, we compare the aggregated probabilities of pre-defined positive and negative lexicons. We use the same lexicon as [6, Table 7] which we report in the supplementary material. If the difference between aggregated probability of the positive and negative lexicon is larger than 0.3, we classify the response as being STRONG (either AGREE or DISAGREE). We additionally examine the binary scenario, where we map the responses STRONG DISAGREE and DISAGREE into a singular DISAGREE category, and similarly combine STRONG AGREE and AGREE into AGREE.

[3] authoritarian is shown in the upper quadrant, libertarian in the lower one.

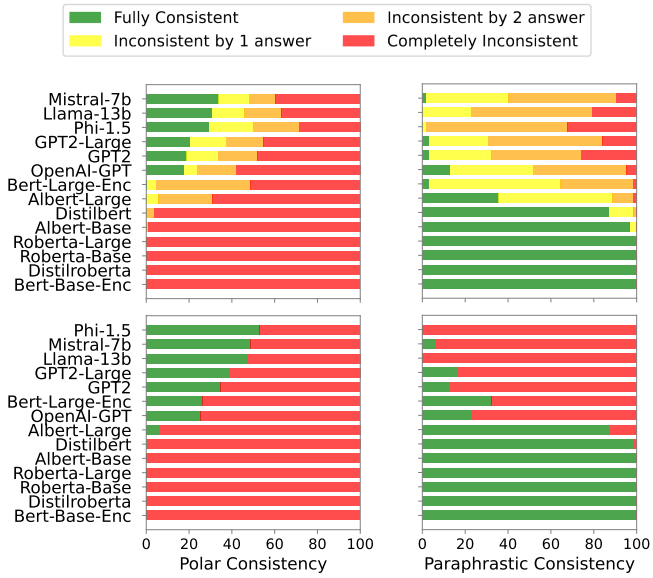

Figure 1: Polar (left plots) and Paraphrastic (right plots) Consistency of LLMs.

For decoder-only models, we use BART, the same off-the-shelf stance detector [11] used in Feng et al. [6, Section 2.1] to determine whether the generated response agrees or disagrees with the given propositions.

4 Results

4.1 RQ1: Polar and Paraphrastic Consistency

To assess model polar consistency, we prompt LLMs with both the original propositions as well as their negations. We analyze the disparity (counting the number of values) between the model's responses to the negated propositions against the expected outcome. For instance, a STRONG AGREE response to an original proposition expects a STRONG DISAGREE to its negation; if the answer is AGREE, we have a discrepancy of 2. Figure 1 (left column) shows the results. The figure shows that decoder-only models generally outperform encoder-only ones in polar consistence, yet even the best model, Mistral-7b, achieves a polar consistence level of just 33.37%. When considering binary responses, as shown in the lower part of the figure, there is an improvement in polar consistence levels, but the highest polar consistency scores is 52.82%, indicating that overall model polar consistency, even in binary terms, remains sub-optimal. The low polar consistency observed in the models suggests a significant reliability issue when using these LLMs for tasks that require consistent and coherent viewpoints about political leanings, particularly in contexts where interpreting statements accurately is crucial, such as in political discourse analysis or ethical decision-making. This inconsistency, where a model cannot reliably align its responses with both a statement and its negation, suggests that its outputs might lead to contradictory or misleading interpretations of its leaning, reducing its usefulness in applications that depend on precise and dependable language understanding.

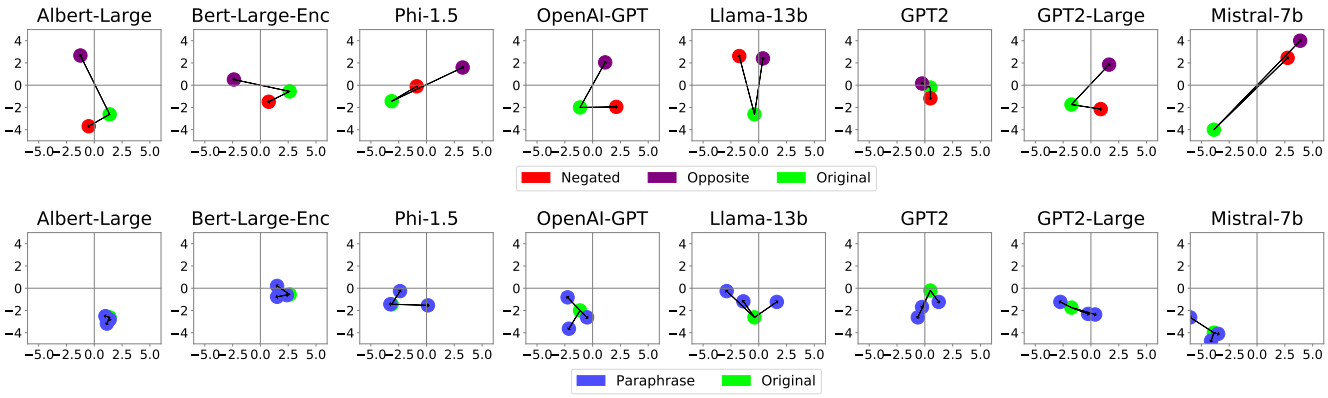

Figure 2: Political leanings of LLMs along economic (left–right) and social axis (authoritarian–libertarian) for polar (first row) and paraphrastic consistency (second row).

We now turn to the paraphrastic consistency of the models by comparing their responses to the original propositions and their paraphrases. The results, shown in Figure 1 (right column) show that while some encoder-only models show high paraphrastic consistency with 4 models with a 100% rate, this is not due to a genuine understanding of the propositions. This can be seen by comparing the left and right column of Figure 1, which shows that models that demonstrated high paraphrastic consistency did exhibit low polar consistency and vice-versa, suggesting a total lack of comprehension of the provided propositions and their variants. The observation of low paraphrastic consistency suggests that when the models are presented with paraphrases of the same statement asking about their political leaning, they often change their responses, indicating a lack of robust understanding of the content. This inconsistency means the models struggle to recognize the equivalence between differently worded versions of the same political viewpoint or question. In practical applications, this behavior could lead to unreliable or fluctuating outputs despite dealing with the same underlying information, which could be problematic in scenarios requiring stable and consistent interpretation.

Overall, the observed limitations in polar and paraphrastic consistency of LLM responses, particularly evident in the binary analysis and in the joint analysis, have substantial practical implications for assessing political biases of LLMs. With polar consistency levels not exceeding 52.82%, reliance on direct questioning proves to be unreliable, as it risks of bringing up inconsistent or contradictory responses. This divergence not only challenges the credibility of LLMs in accurately reflecting political stances but also suggests that their responses to direct political prompts may be misleading.

4.2 RQ2: Effect on Political Leanings

To determine the political leaning score of each model, we submit all the collected responses to the propositions into the Political Compass test by letting each model complete the test using its provided answers. We use both original propositions as well as their variations across paraphrases and negations.

We start by assessing each model's political leaning on the negated version of the original propositions. The results are presented in Figure 2 (first row) where we distinguish between *negated* (i.e., test results based on the negated proposition, shown in red) and *opposite* (i.e., results if we chose the opposite answer for each item on the test, shown in purple). This allows us to evaluate the discrepancy between the actual score on negated propositions and the expected score in the negated context by looking for models where the negated points (red) are close to the opposite points (purple): indicates that when the model is presented with a negated version of a proposition, its resulting political leaning is similar to the one obtained with the original proposition. By considering results for all the considered models it becomes clear that only a few models maintain a resemblance of consistency. Notably, the difference between the *negated* and *opposite* results often shows drastic divergence in political leanings. These results further reinforce the conclusion that direct questioning methods are inadequate for reliably assessing the political biases of LLMs. The best model for polar consistency is Mistral-7b which shows the negated points very close to the opposite points. This model effectively adjusts its responses to negated propositions, aligning closely with the opposite stance according to the political compass test. Finally, the worst model for polar consistency is again Albert-Large, as it shows that the results for the political compass test for the negated and opposite statements change its political leaning from authoritarian to libertarian.

We now turn to investigate paraphrastic consistency. The results, displayed in the second row of Figure 2, show that models with lower paraphrastic consistency levels struggle to maintain a stable political leaning score. Paraphrases can lead to significant shifts, even changing quadrants on the political compass, moving from left to right or from libertarian to authoritarian. Although results for negated paraphrases are not shown due to space constraints, they exhibit similar trends, consistent with observations from the original propositions and their negations in the first row of Figure 2.

For paraphrastic consistency, the ideal model is one where paraphrase points are closest to the original points on the graph, indicating that the model interprets paraphrases similarly to the original

proposition according to the political compass test. Albert-Large is the best model for paraphrastic consistency, with paraphrase points (blue) nearly overlapping the original points (green). In contrast, Phi-1.5 and Llama-13b perform poorly, showing significant distances between paraphrase and original points, sometimes even changing quadrants and thus their political leaning from left- to right-oriented.

4.3 Ablation Studies

We conduct now a set of ablation studies aimed at examining how various factors impact on the performance of the LLMs in terms of their consistency.

We tested different *decoding algorithms* (i.e., the methods used by LLMs to sample over the token distribution to generate the output sequence), namely `greedy search`, `beam search`, and `contrastive search`. Greedy search selects the most likely next word at each step, beam search considers multiple possible sequences to find a more optimal output, and contrastive search refines results by contrasting candidate outputs against each other. We found that `Beam search` showed better performance for GPT models (min 23%, max 42%, avg. 32% versus other models having min 18%, max 33%, avg. 17%), while `contrastive search` (min 21%, max 35%, avg. 23%) and `greedy search` (min 21%, max 35%, avg. 33%) had similar outcomes for the non-GPT models.

Different *stance classifiers* were tested to replace `BART`, the off-the-shelf stance detector [11]. We found that the results are highly variable depending on the classifier, highlighting its crucial role. Nevertheless, similar patterns were observed in polar (min 18%, max 42%, avg. 25%) consistency across models.

We found that changing the *number of sequences* returned by the LLMs and in particular by the `beam search` strategy did not significantly influence the results. Attempts to generate more diverse responses using `top-p sampling` resulted in even reduced polar consistency values.

We investigated the concept of positive and negative *sentiment* as a potential alternative to measuring *stance*, which refers to a model's position in agreement or disagreement with a given statement (often referred to as *entailment* in this context). Sentiment analysis traditionally categorizes expressions into positive, negative, or neutral feelings, which might seem like a suitable method for assessing political biases, given that sentiments could reflect underlying biases in responses. However, when we applied sentiment analysis as a method to determine the political leanings of models, we found that while sentiment analysis primarily captures the emotional tone rather than the specific agreement or disagreement with the content of a statement. For instance, a statement could be phrased negatively yet express a positive sentiment, or vice versa, leading to potential misclassifications when trying to infer political stance from sentiment alone. Overall, our results suggest that using stance as a measure provided higher levels of polar consistency.

Finally, we performed tests with varying *sample sizes* from the primary positive and negative lexicon pool in the encoder-only models, testing the model with sets having different cardinality and composition. results showed a significant decrease in polar consistence with both smaller and different sample sizes.

We leveraged the insights from the set of extensive ablation studies to optimize the configuration of our model, aiming for the one showing highest consistency scores in both polar and paraphrastic consistency. By evaluating the effects of various components like decoding algorithms and stance classifiers, we identified the combination of parameters that lead to maximize the model's ability to consistently interpret and respond to political propositions, thus maximizing consistency scores. We found that the best mode configuration, achieved using `OpenAI-GPT` as LLM, `distilbert-base-uncased-fine-sst-2-english` as sentiment classifier with `beam search` as search strategy, yielded a 50% polar consistency rate on the four level scale and 54.85% polar consistency rate in binary responses, not far from the results observed with standard model configurations.

5 Impact and Limitations

The findings of this study have multiple implications. The demonstration that LLMs do not consistently reveal their political leanings in response to direct questioning challenges the current methods used in political bias assessment. This indicates a need for a reevaluation of how these models are tested for political bias. Additionally, the findings demand the need for the development of more sophisticated tools and methodologies for evaluating the biases and leanings of LLMs, not only for the political context, as this caveat might be present in other domains.

Limitations of the work include reliance on a single political test and primarily addressing political bias, leaving other forms of bias (like gender or race) unexplored, although they could potentially interact with political leanings.

6 Conclusions and Future Work

This study highlights the inherent challenges in employing direct questioning techniques to measure and determine the political biases of LLMs. Our research suggests that the way questions are phrased and the context in which they are presented can substantially alter and change the responses of LLMs, leading to inconsistencies in their outputs. This variability suggests that straightforward questioning may not reliably capture the true political orientations and leanings of these models. These findings point to a significant limitation in current methodologies employed in LLMs, emphasizing the necessity for the development of more sophisticated and precise approaches to measure political biases and leanings. These methods should aim at minimizing the impact of phrasing and context, thereby providing a more accurate and consistent assessment of political leanings in LLMs. Such advancements are needed for improving the transparency and trustworthiness of LLMs in practical use cases.

For future work, we plan to develop more reliable methods for measuring the political leanings of LLMs. We also plan to broaden our research scope beyond political bias, by exploring other dimensions of bias that LLMs might exhibit, such as those related to gender, race, or socioeconomic status.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2232–2242. https://doi.org/10.18653/v1/2021.eacl-main.190

[4] Javad Eta'at. 2015. Measuring Tendencies towards Justice and Freedom among Iranian Intellectual Currents, Using Political Compass Model. *Political and International Approaches* 7, 2 (2015), 68–93.

[5] Fabian Falck, Julian Marstaller, Niklas Stoehr, Sören Maucher, Jeana Ren, Andreas Thalhammer, Achim Rettinger, and Rudi Studer. 2020. Measuring proximity between newspapers and political parties: the sentiment political compass. *Policy & internet* 12, 3 (2020), 367–399.

[6] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11737–11762. https://doi.org/10.18653/v1/2023.acl-long.656

[7] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770 [cs.CL]

[8] Lucas Gover. 2023. Political Bias in Large Language Models. *The Commons: Puget Sound Journal of Politics* 4, 1 (2023), 2.

[9] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (<conf-loc>, <city>Delft</city>, <country>Netherlands</country>, </conf-loc>) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[10] J.C. Lester. 1994. The evolution of the political compass (and why libertarianism is not right-wing). *Journal of Social and Evolutionary Systems* 17, 3 (1994), 231–241. https://doi.org/10.1016/1061-7361(94)90011-6

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[12] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 6565–6576. https://proceedings.mlr.press/v139/liang21a.html

[13] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence* 304 (2022), 103654. https://doi.org/10.1016/j.artint.2021.103654

[14] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating Political Bias in Language Models through Reinforced Calibration. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14857–14866. https://doi.org/10.1609/aaai.v35i17.17744

[15] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. *arXiv preprint arXiv:2402.02680* (2024).

[16] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* 56, 2, Article 30 (sep 2023), 40 pages. https://doi.org/10.1145/3605943

[17] Fabio Motoki, Valdemar Pinho Neto, and Victor Rangel. 2023. More Human than Human: Measuring ChatGPT Political Bias. *SSRN Electronic Journal* (01 2023). https://doi.org/10.2139/ssrn.4372349

[18] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* 4, 3 (2022), 258–268.

[19] Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *Plos one* 18, 11 (2023), e0277640.

[20] Rohan Taori and Tatsunori Hashimoto. 2023. Data Feedback Loops: Model-driven Amplification of Dataset Biases. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 33883–33920. https://proceedings.mlr.press/v202/taori23a.html

[21] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. arXiv:2302.02463 [cs.CL]

[22] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. arXiv:2310.09219 [cs.CL]