The Impact of Generative AI on Bias in Public Discourse

Assessing Differences in Bias and Perception of Human- Versus AI-Generated Political Content

Katherine J. Mueller

Harvard University Graduate School of Design, kmueller@gsd.harvard.edu

The integration of generative artificial intelligence (AI) into public discourse has revolutionized communication, raising critical questions about bias, authenticity, and societal impact. This study investigates the influence of generative AI on political bias perception and human ability to distinguish AI-generated content from human-authored responses. A two-phase experiment was conducted: in Phase I, participants wrote responses to a set of three articles on the same topic either independently or using ChatGPT; in Phase II, a separate group rated the political bias of these responses and assessed whether they were human- or AI-generated. Results showed no statistically significant differences in perceived political bias between human- and AI-generated responses, with both groups rated close to neutral on average. Participants struggled to distinguish AI-generated from human-generated content, achieving near-chance accuracy (51%), but expressed significantly higher confidence in their assessments of AI-generated text. Correlation analysis revealed that confidence did not consistently relate to accuracy, highlighting potential perception biases. These findings suggest that generative AI responses are difficult to distinguish from human-authored text in political contexts, yet societal assumptions influence confidence in attribution. This study underscores the need for critical evaluation of AI's role in shaping public discourse, calling for future research into demographic diversity, qualitative reasoning, and more granular analyses.

CCS CONCEPTS • Generative AI • Human-centered computing • Socio-technical systems

Additional Keywords and Phrases: Political bias, Public discourse, Perceived authenticity, Political polarization

1 INTRODUCTION

The integration of generative artificial intelligence (AI) into public discourse has transformed how individuals engage with information and interpret content. As generative AI models like ChatGPT become increasingly integrated into everyday tasks, their ability to influence perceptions—particularly in socio-political contexts—has raised both optimism and concern. On one hand, these tools democratize content creation by making high-quality writing accessible to all. On the other hand, their capacity to generate content imbued with subtle biases introduces potential challenges in navigating political narratives.

The rapid adoption of generative AI has coincided with growing concerns about how these tools interact with political discourse. The question of whether AI-generated responses reinforce existing biases or present a more neutral perspective is crucial in understanding their societal impact. Additionally, the public's ability—or inability—to distinguish between human- and AI-generated content introduces questions about trust, authenticity, and the evolving nature of communication in politically charged environments.

The study presented here seeks to address the question of the extent to which generative AI introduces or mitigates bias in political discourse and to what extent readers are able to recognize the difference between AI-generated and human-generated discourse.

The hypothesis driving this research is twofold. First, it is anticipated that responses generated with the assistance of ChatGPT will exhibit a reduced perceived ideological bias compared to human-authored responses. Second, it is hypothesized that participants will be able to distinguish between AI- and human-generated responses due to the personal nature of human-generated responses. These hypotheses aim to shed light on the relationship between AI and public discourse, offering a foundation for understanding the ethical and societal implications of generative AI technologies in politically sensitive contexts.

2 BACKGROUND

The rapid advancements in generative artificial intelligence (AI) have significantly transformed the landscape of communication, information dissemination, and public discourse. Generative AI models, such as ChatGPT and Gemini, have demonstrated remarkable capabilities in producing text that closely mimics human language, enabling their application in a variety of domains, including content creation, social media interactions, and public discourse. However, these capabilities have raised concerns regarding misinformation, content bias, and the potential homogenization of ideas, particularly in politically charged contexts.

Recent studies have explored the impact of generative AI on content creation and user interaction. Gabriel et al. (2022)[1] conducted experiments to assess the efficacy of AI-generated interventions in mitigating misinformation on social media, finding that explanations generated by models like GPT-4 can significantly improve user accuracy in identifying false claims. Similarly, Miharaini et al. (2022)[2] examined how generative AI is reshaping communication patterns on social media, highlighting both its democratizing potential for content creation and its role in exacerbating filter bubbles and echo chambers. These studies underscore the dual potential of generative AI: as a tool for enhancing public understanding and as a mechanism for reinforcing biases.

One area of growing interest is the role of generative AI in shaping political discourse. The ability of models to generate personalized, contextually aware responses has led to concerns about their influence on ideological polarization. Research by Vosoughi et al. (2018)[3] and Cinelli et al. (2021)[4] has demonstrated the tendency of social media algorithms to create ideological echo chambers, which generative AI may either mitigate or exacerbate.

Building on this foundation, the current study investigates the intersection of generative AI, political discourse, and user perceptions. By examining how generative AI influences the interpretation of politically nuanced information, this work extends prior research on misinformation and bias mitigation. Specifically, it evaluates the perceived political leaning and authenticity of user-generated content, both with and without the use of generative AI. This study's approach lies in its integration of a rating system to assess external perceptions of AI-generated and non-AI-generated responses, offering insights into how generative AI shapes public opinion in politically charged contexts.

Furthermore, this research draws on established frameworks for assessing bias and perception. The use of politicalBiasBERT[5] for article selection and classification aligns with methodologies employed in earlier studies to quantify ideological leanings in content. The experiment's design—involving user-generated responses and reviewer assessments—parallels the randomized control trials utilized by Baly et al. (2020)[6] to predict the political ideology of newspapers using a BERT language model fine-tuned with a dataset of news articles classified by their political leaning available on allsides.com.

Through its focus on the nuanced interplay between generative AI and political perception, this study contributes to ongoing discussions about the ethical and societal implications of AI in public discourse. By comparing human-authored and AI-assisted responses, it sheds light on the potential for generative AI to influence perceptions of neutrality, bias, and ideological alignment. Moreover, it emphasizes the need for critical evaluation of AI's role in shaping public narratives and decision-making processes.

3 METHODOLOGY

3.1 Overview

This study investigates the impact of generative AI on perceived political bias. In addition, it assesses the extent to which humans are able to distinguish between human- and AI-generated selections of text. To explore these two areas, a comparative analysis was conducted in two phases. Phase 1 tasked two participant groups with writing involving two participant groups tasked with writing a response to a set of three passages. One group completed the task independently, while the other employed a generative AI tool (ChatGPT). In Phase II, a different participant group was employed to rate the political bias of each response written in Phase I and to indicate whether they believed the response was human- or AI-generated.

3.2 Materials

Three articles were selected to serve as the baseline to inform participants in Phase I. These articles represented distinct political leanings: conservative, neutral, and liberal, as determined by politicalBiasBERT, a fine-tuned, pre-trained version of the BERT language model. From each article, a single representative paragraph was extracted to be read by the Phase I participants. The articles selected for this study covered the same event—the nomination of Chris Wright as Secretary of the Department of Energy—but were sourced from three different media outlets with different political alignments: Fox News (Colton, 2024), Reuters (Reuters, 2024), and AP News (AP News, 2024).

3.3 Participants

Participants for this study were recruited through two online crowdsourcing platforms: Amazon Mechanical Turk (MTurk) and Prolific. These platforms were chosen due to their ability to provide access to diverse participant pools in an efficient manner. In order to ensure quality of response, a CAPTCHA was included in the survey and a manual quality review was performed in which responses that did not meet the length requirements or were not on topic were removed. In total, 125 participants were recruited for Phase I, and 393 participants were recruited for Phase II.

3.4 Procedure

In this assessment, two phases were completed. Phase I involved the collection of both human- and AI-generated responses to three passages on the same topic. Human participants were used in the survey involving human-generated responses and in that involving AI-generated responses in order to ensure a diversity of generative AI prompts in order to better simulate real-life circumstances.

Initially, a politicalBiasBERT assessment was performed on the two sets of data. However, nearly identical results between the two groups and a lack of variability in the politicalBiasBERT rating led to the extension of this study into Phase II, which instead used human participants to rate the political bias of the responses produced in Phase I.

3.4.1 Phase I

In Phase I, participants were divided into two groups. Group I consisted of 72 participants, and Group II consisted of 53 participants. Both groups were instructed to read the same three passages, provided in a randomized order in an attempt to reduce bias due to passage ordering. Group I was instructed to write five to seven sentences about the selection of Chris Wright as the Secretary of Energy. Group II was instructed to do the same, with additional instructions to use ChatGPT when constructing their response.

3.4.2 Phase II

In Phase II, 393 participants were randomly assigned to read one of either the human- or AI-generated responses. Participants were then asked to rate the response on a scale from very conservative to very liberal. After that task was completed, participants were then directed to a page where they were asked to disclose whether the provided response was human- or AI-generated. This was lastly followed by a question asking participants to rank their level of confidence in this assessment of human vs. AI-generation.

3.5 Evaluation

Data analysis focused on the following metrics:

- Political Bias Rating (Q62): Discrepancies in perceived political leaning between AI-generated and human-authored texts.
- Human vs. AI Detection (Q64): The assessment of reviewers in identifying AI-generated vs. human-generated responses.
- Confidence Rating (Q66): Reviewer confidence scores associated with AI detection tasks.

3.5.1 Aggregate Data by Response

Each response was evaluated by between 1 and 5 reviewers. If more than one reviewer assessed a passage, the scores were then averaged for each response to avoid overrepresentation of any single response. Therefore, the total sample size of this assessment is the same as that for Phase I, and the analyses conducted in this section represent analyses of the averaged scores for each passage.

3.5.2 T-Tests for Mean Comparisons

To determine whether there were statistically significant differences between the evaluations of the human-generated and AI-generated responses, two-sample independent t-tests were conducted for each question. These tests compared the mean values of the human and AI groups, analyzing differences in the political bias ratings, the human vs. AI detection, and the confidence rates in that detection. The t-tests were performed with a significance level (α) of 0.05. The results provided insights into whether observed differences in means were statistically significant.

3.5.3 Correlation Analysis

A correlation analysis was conducted to evaluate the relationship between participant confidence (Q66) and the accuracy of their judgments (Q64) on whether the responses that they were reviewing were human- or AI-generated. This analysis explored whether higher confidence was associated with greater accuracy in identifying the source of the responses. For both human and AI groups, scatterplots with trend lines were generated to visualize these relationships, and values were computed to quantify the strength of the correlations.

3.6 Results

3.6.1 Summary

The survey results in Table 1 highlight key differences in how participants evaluated human- and AI-generated responses across three different metrics: political bias (Q62), perceived origin (Q64), and confidence in their evaluations (Q66). For political bias, human responses were rated slightly more conservative on average than AI responses, though the difference was not statistically significant (p=0.288). In terms of origin identification, participants' ability to discern whether a passage was AI- or human-generated was nearly identical, with both groups achieving a mean score around 0.51 (p=0.992), indicating no significant difference in accuracy. However, confidence levels showed a notable divergence, with participants expressing significantly higher confidence in their evaluations of AI-generated responses compared to human ones (p=0.00912). These findings provide an overview of participant perceptions and serve as the foundation for a more detailed analysis of each question in the following sections.

Table 1: Results of Response Evaluation Surveys

Question	Human- vs. AI-Generated	Mean Score	Standard Dev.	P-Value ($\alpha = 0.05$)
(Q62) Please rathe this passage on a scale of very liberal (-2) to very conservative (2).	Human	0.235	0.763	0.288
	AI	0.075	0.850	
(Q64) Was this text AI-generated (0) or human-generated (1)?	Human	0.510	0.315	0.992
	AI	0.511	0.342	
(Q66) How confident are you in your answer to the previous question on a scale of very unconfident (-2) to very confident (2)?	Human	0.441	0.734	0.00912
	AI	0.784	0.624	

[a] This table shows the aggregate results of the questionnaire that reviewers took to assess the bias rating, origin, and confidence in origin of the responses from Phase I. Respondents were only asked to rate their responses on a Likert scale, the numerical representations are added in the questions in this table for clarity.

3.6.2 Assessing Political Bias (Q62)

Q62 asked reviewers to rate the political bias of each passage on a scale ranging from very liberal (-2) to very conservative (2). Human-generated responses received a slightly higher mean score (0.235) compared to AI-generated responses (0.075), suggesting that human responses were perceived as marginally more conservative on average. However, the standard deviations for both groups were relatively high (0.763 for human responses and 0.850 for AI responses), indicating considerable variability in reviewers' ratings.

A statistical analysis using a t-test revealed that the difference between the means was not statistically significant (p=0.288), suggesting that reviewers did not consistently perceive a strong difference in political bias between the two groups of responses.

The histogram for Q62, included below in Figure 1, provides a visual representation of the distribution of political bias ratings for both human- and AI-generated responses. Both distributions are relatively centered around neutral (0), with slightly more dispersion in the AI group. This further supports the finding that, while differences exist in the means, they are not strong enough to suggest systematic differences in perceived bias between human and AI responses.

By integrating these insights, it appears that the political bias of responses was not a distinguishing factor for reviewers, reinforcing the need to explore other dimensions, such as origin identification or confidence, for more pronounced differences.

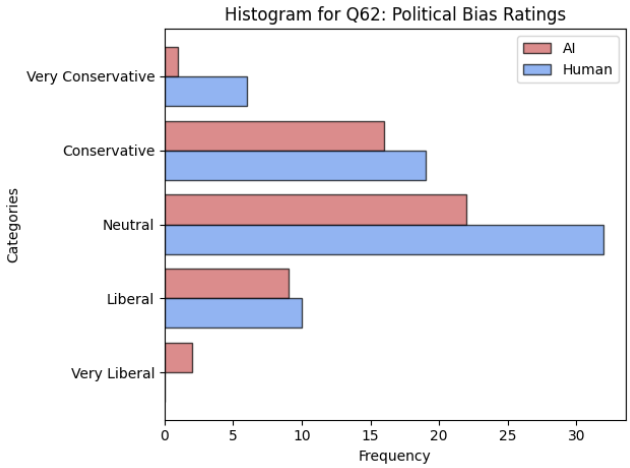

Figure 1: A histogram showing the distribution of bias ratings between both the AI-generated and human-generated responses.

3.6.3 Assessing Perceived Origin (Q64)

The ability of reviewers to identify whether a response was human- or AI-generated revealed no significant difference between the two groups. Human-generated responses had a mean prediction score of 0.510 with a standard deviation of 0.315, while AI-generated responses had a mean prediction score of 0.511 with a standard deviation of 0.342, indicating that in both cases reviewers were evenly split in their predictions between AI versus human generation. The p-value for this comparison was 0.992, indicating there was not a significant difference between predictions of origin between the two groups.

This finding indicates that participants struggled to distinguish between human- and AI-generated text, suggesting that the AI-generated responses were similar enough to human-authored content to evade reliable detection and vice versa. The histogram shown in Figure 2 further highlights the overlap in reviewer assessments, as the distribution of responses shows no meaningful divergence between the two groups.

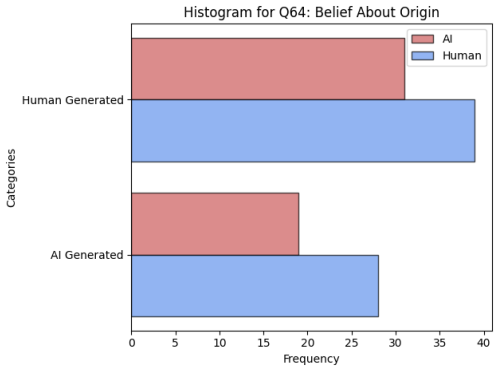

Figure 2: A histogram showing the distribution of perceived origin between both the AI-generated and human-generated responses.

3.6.4 Confidence in Perceived Origin Assessment (Q66)

The confidence ratings reported by reviewers when identifying whether a response was human- or AI-generated displayed notable differences between the two groups. Reviewers evaluating AI-generated responses exhibited a higher mean confidence score (0.784) with a standard deviation of 0.624, compared to a mean confidence score of 0.441 for human-generated responses with a standard deviation of 0.734. This difference was statistically significant, with a p-value of 0.00912, which is well below the 0.05 alpha level, indicating that there is evidence to suggest that reviewers felt more confident in their assessments of AI-generated text.

The accompanying histogram shown in Figure 3 visually reinforces this finding. The distribution of confidence ratings for AI responses skews notably higher compared to the broader spread of confidence levels for human responses. This suggests that participants were more assured when evaluating AI-generated responses.

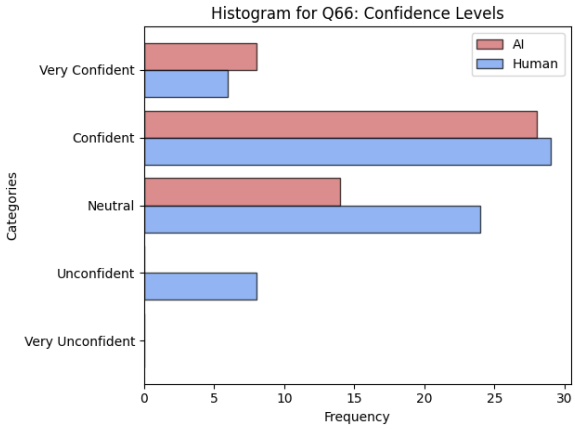

Figure 3: A histogram showing the distribution of confidence in perceived origin between both the AI-generated and human-generated responses.

3.6.5 Correlation Between Confidence and Accuracy

Finally, the correlation between the accuracy of the origin detection and the confidence ratings of the reviewers were compared between the two groups. As can be seen in the scatterplot shown in Figure 4, confidence did not necessarily

translate to accuracy. While the group reviewing the AI-generated articles may have felt more confident in their determinations of whether or not the responses were AI-generated, this group showed only a 36% success rate in accurately determining whether the responses were produced by AI. However, the less confident group reviewing the human-generated responses yielded a correctness of 52%, indicating a still low but significantly higher success rate than the AI group. This indicates that AI-generated responses were difficult to distinguish between human-generated ones, with reviewers in the AI group overestimating occurrences of human-generated responses.

However, it is important to note that these ratings were assessed on a binary scale. While the ratings of confidence for each passage were averaged between all reviewers looking at that question, accuracy was converted to a binary scale with average scores >0.5 being marked as "Correct" for the human group and "Incorrect" for the AI group. Further assessment of the breakdown by reviewer may be required to validate these results.

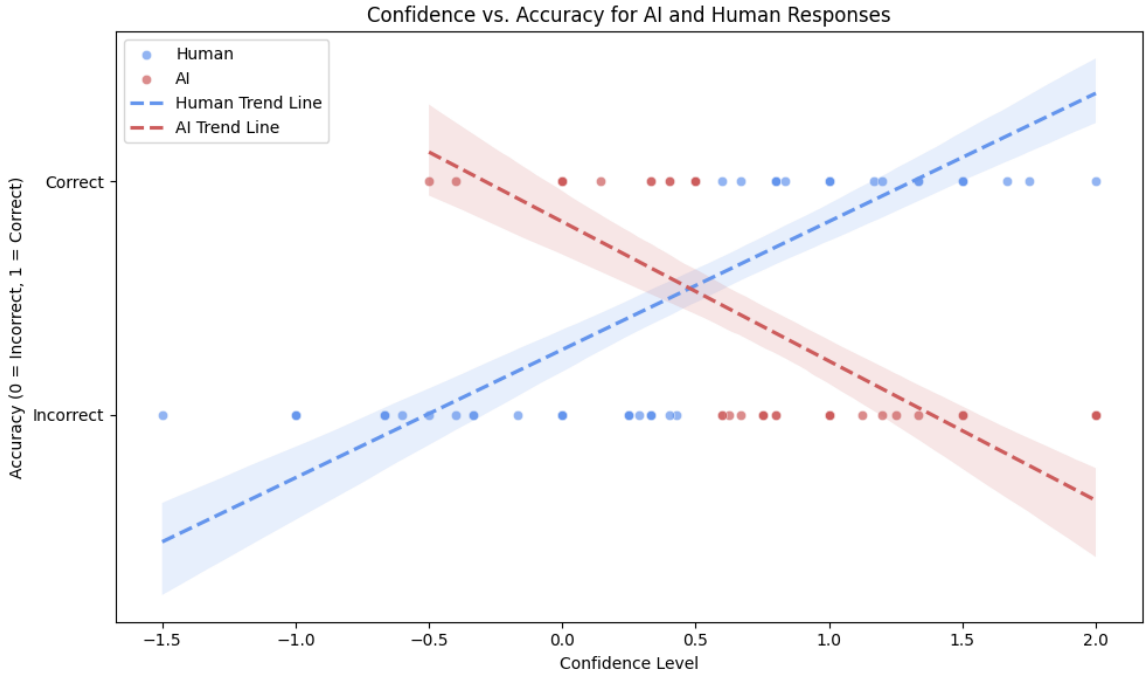

Figure 4: A scatterplot revealing the inverse relationship between confidence and correctness in identifying AI-generated responses within the AI group (R^2 = 0.59), and a positive relationship between confidence and correctness in the human group (R^2 = 0.65).

3.7 Discussion

This study aimed to investigate the impact of generative AI on perceived political bias, origin attribution, and confidence in origin attribution. The results offer insights into how generative AI responses are perceived and evaluated compared to human-generated content.

3.7.1 Key Findings

Bias Ratings (Q62): The mean scores for human- and AI-generated passages (0.235 and 0.075, respectively) reflect a slight conservative lean overall, though the difference between them was not statistically significant (p = 0.288). This suggests

that, while biases were present, the perceived political leaning of AI-generated content closely mirrored that of human-generated passages. The absence of "very liberal" ratings for any human-generated passage suggests a potential skew in the sample population. This could indicate an overrepresentation of liberal-leaning respondents, perhaps due to more liberal individuals among online survey participants. This demographic skew may influence the perceived neutrality of passages, as respondents might interpret "liberal" content as closer to neutral. Future research could aim for more balanced demographic representation to assess whether these findings persist across diverse populations.

Origin Attribution (Q64):
The nearly identical mean scores for AI and human passages (0.511 and 0.510, respectively) suggest that reviewers had significant difficulty distinguishing between the two. This lack of differentiation also raises questions about the evaluative criteria used by respondents. It would be beneficial to explore the reasonings behind these ratings in future studies.

Confidence Ratings (Q66): The observed difference in confidence levels, with human-generated passages eliciting a mean confidence score of 0.441 versus 0.784 for AI-generated passages, suggests a potential perception bias. Reviewers may overestimate their ability to identify AI content, reflecting societal narratives about AI's distinctiveness or the implicit assumption that AI-generated content is easier to spot. Additionally, the scatterplot analysis of confidence versus accuracy highlighted notable trends. Reviewers appeared more consistent in their confidence ratings for AI-generated content, possibly due to preconceived notions about what constitutes "AI-like" text. However, these notions did not stand the test of accuracy, with reviewers of this group underrepresenting the presence of AI-generated text.

3.7.2 Future Work

While this study provides meaningful insights, several limitations and opportunities for improvement have been identified:

1. Demographic Representation:
    - As highlighted by the Q62 results, future studies should aim to recruit a more demographically diverse pool of reviewers. Stratified sampling techniques could ensure balanced representation across political, geographic, and age groups.
2. Granular Analysis of Confidence Ratings and Accuracy:
    - The current study aggregated data at the passage level to prevent overrepresentation of heavily-reviewed passages. While this approach ensured fairness across passages, it likely obscured individual-level variability. Future studies should drill down to the individual reviewer level to better understand how personal biases and confidence interact with accuracy.
3. Increased Sample Size:
    - To strengthen the reliability of results, future research should increase the number of reviewers per passage. Larger sample sizes would enhance statistical power and enable more robust conclusions. In addition, having exactly the same number of reviewers per passage would allow for individual-level analysis without the overrepresentation of any of the responses.
4. Qualitative Assessment:
    - Future studies should further investigate the explanations as to why reviewers answered in the ways that they did. A qualitative assessment would provide a new level of insight to this study.

3.7.3 Conclusion

This study highlights the complex interplay between generative AI, political bias perception, and attribution confidence. While AI-generated content is increasingly indistinguishable from human text, confidence in its attribution reveals underlying biases and societal assumptions. By addressing the limitations of this study and incorporating more granular and representative analyses, future research can further elucidate how generative AI shapes perceptions and decision-making in politically charged contexts. These findings underscore the need for critical evaluation of AI's role in public discourse and its potential to influence societal narratives.

4 WORKS CITED

(1) Gabriel, S.; Lyu, L.; Siderius, J.; Ghassemi, M.; Andreas, J.; Ozdaglar, A. Generative AI in the Era of "Alternative Facts." *MIT Explor. Gener. AI* **2024**. https://doi.org/10.21428/e4baedd9.82175d26.

(2) Ghani, M. M.; Mustafa, W. A. W.; Hashim, M. E. A.; Hanafi, H. F.; Bakhtiar, D. L. S. Impact of Generative AI on Communication Patterns in Social Media. *J. Adv. Res. Comput. Appl.* **2022**, *26* (1), 22–34.

(3) Vosoughi, S.; Roy, D.; Aral, S. The Spread of True and False News Online. *Science* **2018**. https://doi.org/10.1126/science.aap9559.

(4) Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; Starnini, M. The Echo Chamber Effect on Social Media. *Proc. Natl. Acad. Sci.* **2021**, *118* (9), e2023301118. https://doi.org/10.1073/pnas.2023301118.

(5) *bucketresearch/politicalBiasBERT · Hugging Face*. https://huggingface.co/bucketresearch/politicalBiasBERT (accessed 2024-12-13).

(6) Baly, R.; Da San Martino, G.; Glass, J.; Nakov, P. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Online, 2020; pp 4982–4991. https://doi.org/10.18653/v1/2020.emnlp-main.404.