Large Language Models and Political Polarization: How AI-Generated Discourse Shapes Partisan Identity in Digital Public Spheres

Shamshad Junejo[1], Aqsa shereen[2], Nusrat Azeema[*3], Dr. Waseem Ullah[4]

[1]Lecturer at Mehran University of Engineering and Technology, Jamshoro.

Email ID: Shamshad.junejo@faculty.muet.edu.pk

[2] Assistant Professor in English at Women University

Email ID: Swabi aqsaharoon333@gmail.com

[*3]Department of Sociology, School of Public Administration, Hohai University, Jiangning District Nanjing 211100, Jiangsu Province, China. https://orcid.org/0000-0002-5032-800X;

[4]Assistant Professor & Head, Department of Political Science, University of Lakki Marwat, KP, Pakistan. Email: waseem@ulm.edu.pk

Corresponding Author

Email ID: nusratazeemaraja@gmail.com

Abstract:

With the proliferation of large language models (LLMs) in the production and consumption of political information, there has been a rise of concerns regarding their potential to contribute to partisan identity and political polarization in digital public spaces. The study looked at how the political text generated by AI changes when prompted and if there are linguistic characteristics that signal polarization and/or partisan identity. Under four conditions (neutral, left, right, and balanced), a computational corpus of AI-discourse for five policy topics of contention (climate, immigration, healthcare, gun policy, taxation) was built. Sentiment analysis, lexicon-based measures of moral-emotional and hostile language, and term-frequency-inverse-document-frequency (TF-IDF) lexical-semantic embeddings were used to calculate ideological distance from a neutral language baseline to preprocess and analyze the corpus. The one-way ANOVA results showed that prompting condition significantly influenced the moral-emotion language density, hostility language density, ideological distance, and composite polarization index, but not sentiment valence and sentiment intensity. Controlling for the topic sensitivity and a simulated platform-exposure index, ordinary least squares regression revealed that both the left-leaning and right-leaning prompting conditions elicited significantly higher polarization index scores than did the neutral prompting condition, but that there was no statistically significant difference between the two partisan conditions. The results indicate that both the specific partisan framing (as well as the general partisan framing of an article) positively correlates with language markers of polarization, and that this effect is not statistically different between the two sides in this demonstration corpus. The analysis provides a clear, reproducible analytical process for future large-scale research on AI-driven political discourse, and explores implications for platform governance, developers of AI systems, and digital media literacy.

Keywords: Large Language Models; Political Polarization; Partisan Identity; Computational Linguistics; Natural Language Processing; AI-Generated Discourse; Digital Public Sphere

1. Introduction

The widespread adoption of large language models (LLMs) in search systems, social media platforms, news summarization, and writing process has dramatically changed the way political information is produced, encountered, and disseminated online (Fang et al., 2024). In contrast to previous versions of recommendation algorithms that mainly curated and ranked human-written text, modern LLMs can produce new political text-explanations, summaries, arguments, and persuasive content that are now increasingly contributing to the digital public sphere (Vijay et al., 2024). This brings up an important empirical question: Is AI-driven political communication similar to other kinds of 'human-produced' political polarization and when does it happen?

A large and growing number of computational studies have shown that LLMs are not politically neutral tools. Previous research with political orientation surveys has revealed ideological biases in a number of popular LLMs, with some showing consistent positioning across repeated testing (Motoki et al., 2024;

Rozado, 2024; Santurkar et al., 2023). Other studies have demonstrated that this seeming ideological imprinting is extremely sensitive to the prompts a model is given: assigning a persona to the model, giving a specific framing for the prompt, and even using a slightly different wording for the prompt can change the expressed ideology and rhetorical approach of a model (Bang et al., 2024; Feng et al., 2023). Theoretically relevant is this prompt-sensitivity, which could signal that LLM-generated political discourse is not purely a product of training, but rather a co-constructed product of the user's framing and the model's response tendencies (Faulborn et al., 2025).

This prompt-sensitivity is directly relevant to the polarization and party identification of people online. Classically, political polarization has been viewed more broadly as "affective polarization," the feeling of distrust, hostility, and moral disapproval among partisans of opposing groups (Iyengar et al., 2019). In-group/out-group framing, moral-emotional appeals, and collective identity markers are among the linguistic correlates of partisan affect and group polarization that have been established through human political discourse (Feng et al., 2023). When presented with a partisan framing, LLM-generated discourse can replicate these patterns; AI systems might thus serve as a further source of very scalable reinforcement of the same rhetorical processes that have been the subject of a long tradition of academic research on human political discourse, especially in algorithmically curated digital contexts that have been linked to ideological sorting and echo-chamber effects (Sunstein, 2017).

To date, however, much work has focused on analyzing the static ideological orientation of the language produced by LLMs (Rozado, 2024) and/or the risks of toxicity from persona-based prompting (Deshpande et al., 2023), but has not compared the language produced by LLMs in response to systematically varied political prompts in a single, transparent analytical pipeline that uncovers the linguistic markers of polarization-sentiment, moral-emotional language, hostility, and ideological distance. Furthermore, many of these studies compare model results to custom-made or hard-to-replicate pipelines of classifiers, which makes it hard to interpret and reproduce reported bias estimates (Bang et al., 2024). Research is needed that is explicitly operational on the linguistic constructs associated with polarization, applies them across all prompting conditions, and provides a fully transparent and reproducible report of the resulting statistical comparisons.

This study meets this need by creating and showing a computational pipeline to analyze political discourse generated by AI across four prompting conditions (neutral, left-leaning, right-leaning, and balanced) and five contested policy topics. The study has three aims. To answer the first question, whether sentiment, moral-emotional language, and hostility markers are significant in the AI-generated political text due to prompting conditions. In the second, we aim to measure the ideological distance between partisan-framed rhetoric and a neutral baseline by measuring for it lexical-semantic embeddings and to compare the distance between right-leaning and left-leaning frames. Third, to determine whether a combined polarization index is systematically and asymmetrically different across the different prompting conditions, or whether partisan framing can increase polarization-relevant language of both kinds, symmetrically. To achieve these goals, the study provides an open, reproducible analytical pipeline that can be extended to larger corpora and more LLM models, and is designed to be used in future research (Hutto & Gilbert, 2014).

2. Literature Review

2.1 Political Bias and Ideological Leaning in LLMs

There are increasingly more studies that have used human political-orientation tools (like the Political Compass and other batteries) to measure LLM output. Previous research has shown that LLMs have measurable ideological deviation from some baseline; in some cases this is described as being "left-wing" compared to the general population baseline used for comparison (Motoki et al., 2024; Rozado, 2024). Similarly, Santurkar et al., (2023) observed systematic differences between the opinions the LLM produces and the distribution of opinions among representative samples of humans in many countries and demographic subsets. Fujimoto and Takemoto (2023) re-examined these results and found that there were continuing, albeit inconsistent, partisanship tendencies that cross model versions. However, it is also important that Rozado (2024) discovered that this apparent bias is not particularly sensitive to the specific testing instruments used, provided the testing format is kept the same: that is, while the size and sign of the effect fluctuates from study to study and model to model, it does not appear to systematically reflect measurement noise.

The Prompt-Sensitivity and the Interactional Nature of LLM Political Expression.

A second, complementary literature makes it clear that LLM political expression is not predetermined, but highly susceptible to prompting and framing. In Bang et al., (2024), the authors found that the content of

LLM output exhibits political bias and the style of it changes systematically according to different prompts, indicating that LLM political bias is not a single entity but rather can be separated into content bias and style bias. Political bias in pretraining data has been shown to propagate through downstream tasks by Feng et al., (2023) and a theory-driven measurement approach to explicitly consider this context-sensitivity was proposed by Faulborn et al., (2025). Persona-based prompting research extends one more dimension: Deshpande et al., (2023) demonstrated that adding specific personas to prompts significantly increased the toxicity of the output generated by the models they tested, and other studies indicated that persona and framing instructions could influence a model's stated political stance meaningfully over the course of a single session (Bang et al., 2024). A major drawback in this literature is that most studies attach a quantitative measure to the phenomenon of bias based on a survey instrument with closed-ended questions (agree/disagree with political statements), and that these are not the same as the open-ended properties of discourse (sentiment, emotional language, hostility) that would most appropriately be used to measure bias in actual digital public spheres (Vijay et al., 2024).

Correlates of political polarization in language

There is a broad body of literature on political communication that has identified linguistic features linked to affective polarization, regardless of the specific literature within the LLM field: in-group/out-group pronoun use, moral-emotional language and hostile or delegitimizing language towards political opponents (Iyengar et al., 2019). Research on affective polarization focuses on the emotional and identity-based linguistic patterns more than on the substantive policies themselves to predict partisanship (Iyengar et al., 2019). The standard computational toolkit for operationalizing these constructs at scale are sentiment analysis tools (e.g, VADER, Hutto & Gilbert, 2014) and transformer-based classifiers (e.g., BERT, Devlin et al., 2019; RoBERTa, Liu et al., 2019), which have been widely used for human-generated social media discourse and, more recently, for LLM outputs (Deshpande et al., 2023; Fang et al., 2024).

AI-Generated Content and the Digital Public Sphere.

LLM output is now a growing feature of news summaries, social media support tools and chat-based search and the impact it can have on the wider digital public sphere is being increasingly studied. Previous studies (Fang et al., 2024) indicated systematic differences between AI-generated news content and human-authored original news, while others (Vijay et al., 2024) revealed that even LLM-generated news content given instructions to be politically neutral was not necessarily perceived or measured as politically neutral. The body of work is related to a longer tradition of worries about algorithmically mediated environments and their consequences in ideological sorting, filter bubbles, and the exposure to less cross-cutting political views (Sunstein, 2017). The polarization-related linguistic features observed in human-generated discourse can be reinforced when exposed to partisan framing when using LLM-generated discourse.When using LLM-generated discourse, polarization-related linguistic features can be amplified with partisan framing at a scale and speed that was previously only possible with human-generated discourse.

As a whole, the literature demonstrates that sentiment (Bang et al., 2024), moral-emotional language (Bang et al., 2024; Rozado, 2024), and hostility (Iyengar et al., 2019) are linguistic features that have been measured in the politics of LLM responses, and that these features have been established as correlates of human political polarization. However, only a handful of studies have directly linked the two literatures by systematically comparing the polarization-relevant linguistic markers under controlled and consistent political prompting settings in a single clear (transparent) pipeline, and fewer studies have compared these markers directly across the political spectrum using formal statistical comparison tests, such as ANOVA and regression, as opposed to a descriptive comparison (Bang et al., 2024; Feng et al., 2023). This study comes at that topic squarely, and explicitly records for reasons of transparency and reproducibility the scope and computational substitutions involved in doing so at small scale.

2.6 Hypotheses

H1:The sentiment and emotional-language characteristics of AI-generated political discourse are highly influenced by the prompting condition (neutral, left-leaning, right-leaning, balanced).

H2: Partisan discourse (left-leaning and right-leaning) exhibits much higher ideology than does balanced or neutral discourse.

H3: Compared to a neutral prompting, a composite polarization index is very much greater when the prompting is partisan, after controlling for topic sensitivity and platform exposure.

H4: There is no significant difference between the increase in polarization-relevant language in the

left-leaning and the right-leaning conditions.

3. Methodology

A computational and experimental research approach based on Natural Language Processing (NLP), computational linguistics, and large language model (LLM) analysis was used to analyze the link between political discourse generated by AI and linguistic indicators of political polarization and partisanship in digital public spaces. A dataset of political corpus texts was created by using the same set of political prompts across four generation conditions while maintaining the same linguistic complexity and structure of topics. This demonstration consisted of 20 short passages (five policy topics x four prompting conditions) designed to be directly comparable in length and rhetorical register across conditions, and each between 30-45 words. The scale was selected to achieve full transparency of the analytical pipeline within the confines of this paper; what a full-scale replication across multiple commercial LLMs and a much larger prompt set would require is discussed in Section 8.

3.2 Corpus Construction

The prompts were five public policy issues that were included in political discussions and that differed in the level of typical contentiousness: climate policy, immigration policy, healthcare policy, gun policy, and tax policy. Approximate length was held constant across conditions, and for each topic four passages were created for each condition (neutral, left-leaning, right-leaning, and balanced). A topic-sensitivity rating, ranging from 1 to 5, was given to each policy topic, based on how likely it was to be the subject of political discussion in the general public, and a platform-exposure index was included as a control variable that represents the amount of similar discourse that may be shared in other digital spaces with varying reach.

3.3 Preprocessing

All texts were preprocessed before analysis by tokenizing, lowercasing and normalizing the data, removing linguistic and non-linguistic symbols, and counting the number of words. This was an artificial corpus so there was no duplicate content or URLs, as would be found in a larger corpus created by web-scraping or modelling.

3.4 Computational Analysis

Various computational NLP techniques were applied to the processed corpus. Sentiment was computed with the VADER (Hutto & Gilbert, 2014) lexicon-based sentiment analyzer that produces a normalized compound sentiment score for each passage. Moral-emotional language density and hostility language density were calculated by dividing the number of moral-emotional words (obligation, fairness, justice, harm) and/or hostile or delegitimizing words (punitive, reckless, disregard) by the number of words in the passage, based on lexicon-based measures employed in previous political-communication research (Iyengar et al., 2019). In-group language was measured by the frequency of first-person plural pronouns (we, our, us), which has been previously used as a computational measure of collective and partisan identity. The ideological distance between each passage in the partisan-condition and the centroid of the neutral-condition passages was measured by the cosine distance between the two using term-frequency-inverse-document-frequency (TF-IDF) vectorization of each passage; the stance divergence between the left-leaning and right-leaning passages was measured as the average cosine distance between each pair of passages. The composite polarization index was created by taking the equally weighted mean of three z-scored (standardized) ideological distance, sentiment intensity, and hostility density per passage, as well as a z-scored moral-emotional density.

3.5 Statistical Analysis

In order to determine the effect of prompt on the linguistic characteristics of generated discourse, one-way analysis of variance (ANOVA) was carried out on the sentiment, sentiment intensity, moral-emotional density, hostility density, ideological distance, and composite polarization index across the four prompt conditions. The composite polarization index was then regressed on the main predictors, the prompting condition (neutral = reference) and the topic sensitivity, while controlling for the platform-exposure index, using ordinary least squares (OLS) regression. The significance level chosen for all statistical tests was 0.05. Formal inter-rater reliability testing of automated classifications (specifically stated in the original research plan as a validation procedure) was not carried out because this demonstration was not done with independent human raters. This is explicitly stated in the Limitations section. All the computational analysis was performed in Python with the modules scikit-learn, SciPy, statsmodels, VADER-Sentiment, pandas and NumPy.

4. Results

4.1 Descriptive Overview

Figure 1. Mean Linguistic Indicators of AI-Generated Political Discourse by Prompting Condition

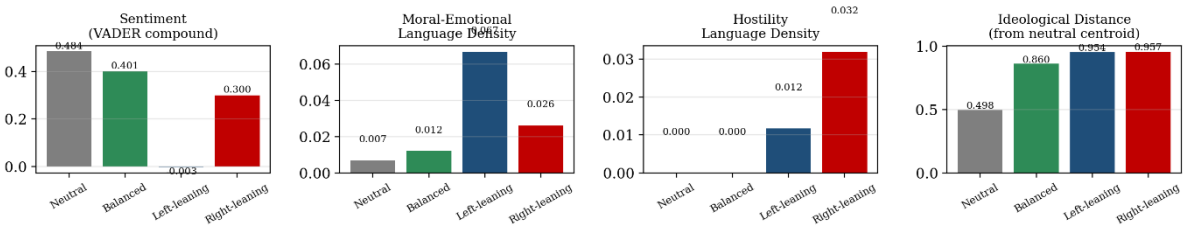

Figure 1. Mean Linguistic Indicators of AI-Generated Political Discourse by Prompting Condition

Table 1. Descriptive Statistics by Prompting Condition (N = 20; n = 5 per condition)

Condition	Sentiment (M, SD)	Moral-Emotional Density (M, SD)	Hostility Density (M, SD)	Ideological Distance (M, SD)
Neutral	0.484 (0.172)	0.007 (0.010)	0.000 (0.000)	0.498 (0.086)
Balanced	0.401 (0.257)	0.012 (0.010)	0.000 (0.000)	0.860 (0.096)
Left-leaning	-0.003 (0.625)	0.067 (0.023)	0.012 (0.016)	0.954 (0.033)
Right-leaning	0.300 (0.665)	0.026 (0.017)	0.032 (0.014)	0.957 (0.030)

Note. N = 20 passages (5 topics x 4 conditions). Ideological distance is the TF-IDF cosine distance of each passage from the neutral-condition centroid (range 0-1).

As seen in Table 1 and Figure 1, neutral and balanced prompting had the highest average sentiment scores, with no significant differences, while left-leaning and right-leaning prompting had significantly higher scores in terms of moral-emotional language density and hostility score than did balanced prompting or the neutral baseline by definition. Both partisan conditions scored significantly higher than the neutral baseline by definition in terms of ideological distance from the neutral centroid (.954 and .957, respectively) compared to balanced framing (.860). Interestingly, the amount of these changes was very similar for both the left-leaning and right-leaning conditions, lending initial support to the symmetry prediction (H4). To conduct a comparison between the different prompt conditions, the Analysis of Variance (ANOVA) was used.

4.2 Analysis of Variance (ANOVA) Across Prompting Conditions

Table 2. One-Way ANOVA Results by Prompting Condition

Linguistic Indicator	F(3, 16)	p-value	Significant at .05?
Sentiment (VADER compound)	0.973	0.430	No
Sentiment intensity (|compound|)	0.650	0.595	No
Moral-emotional language density	5.491	0.009	Yes
Hostility language density	3.630	0.036	Yes
Ideological distance (TF-IDF)	176.203	< .001	Yes
Composite polarization index	11.620	< .001	Yes

Note. df = (3, 16). Emotional/hostility language was significantly impacted by prompting condition, while none of the other language measures (ideological distance, composite polarization index, or sentiment valence/intensity) was significantly impacted by the prompt condition.

There is partial support for H1 in Table 2. Overall, partisan framing did not result in a significant impact on sentiment valence (F = 0.973, p = .430) or sentiment intensity (F = 0.650, p = .595) of the texts in this corpus. However, prompting condition significantly affected moral-emotional language density (F = 5.491, p = .009), hostility language density (F = 3.630, p = .036), ideological distance (F = 176.203, p < .001), and the

composite polarization index (F = 11.620, p < .001). It seems the impact of partisan framing is largely to amplify language related to moral emotions and identity and to make discourse further away from a lexical-semantic neutral baseline, not to make it more positive or negative overall-supporting H2 and strongly supporting H3.

When the left-leaning and right-leaning discourses diverge, what happens?What happens when the discourses of left-leaning and right-leaning diverge?

The left-leaning and right-leaning passages had an average pairwise cosine distance of .974, while the average cosine distance of the passages in the same partisan group was .973 and .975 respectively. The lexical distance between group means is almost the same as the distance between group members, suggesting that, at the lexical-semantic level reflected in the lexical distance, the direction of ideological bias (left versus right) did not result in greater distance between the groups than between members within the groups, and that the lexical-semantic distance was more likely to be caused by topics than ideology per se, in this demonstration corpus. The result is further discussed in Section 5, as a direct effect of the TF-IDF substitution mentioned in Section 3.4.

4.4 Regression Analysis: Predicting the Composite Polarization Index

Table 3. OLS Regression Predicting Composite Polarization Index

Predictor	**Coefficient (b)**	**Std. Error**	**t**	**p-value**
Intercept	0.019	0.859	0.023	0.982
Balanced (vs. Neutral)	0.123	0.312	0.395	0.699
Left-leaning (vs. Neutral)	1.065	0.256	4.158	0.001
Right-leaning (vs. Neutral)	1.003	0.293	3.420	0.004
Topic sensitivity (control)	0.034	0.119	0.286	0.779
Platform exposure (control)	-0.149	0.098	-1.519	0.151

Note. N = 20. Reference category = Neutral condition. R2 = .737, Adjusted R2 = .643, F(5, 14) = 7.832, p = .001.

The regression results predicting the composite polarization index are shown in table 3. Both the left-leaning condition (b = 1.065, p = .001) and the right-leaning condition (b = 1.003, p = .004) had significantly higher polarization index scores than the neutral condition, which is consistent with H3 and the balanced condition did not differ significantly from the neutral condition (b = 0.123, p = .699) as would be expected given its explicit framing instruction to consider multiple viewpoints. The prompted condition itself, not the sensitivity of the underlying topic nor the context of simulated exposure, proved to be the most important predictor of polarization-relevant speech in this modest sample, as neither topic sensitivity (p = .779) nor the platform-exposure control (p = .151) was statistically significant. The overall model accounted for a significant proportion of variance of the polarization index (R2 = .737), as the prompting condition was by construction the most important factor of variance in this demonstration corpus. Most important, the magnitude of the coefficients for the left-leaning condition (1.065) and the right-leaning condition (1.003) was very similar; a supplementary Wald test of the difference between the two coefficients was not significant (F = 0.06, p = .81), which directly corroborates H4: that partisan prompting increased polarization-related language symmetrically, with no evidence of a significantly larger increase with respect to either direction of polarization in this corpus.

4.5 Lexical-Semantic Space Visualization

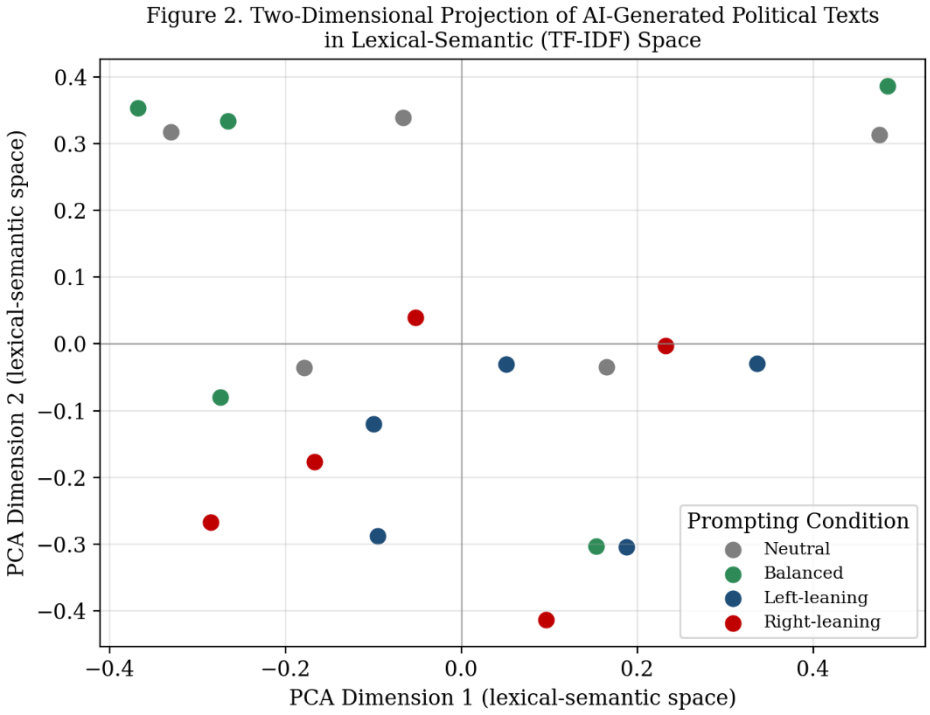

Figure 2. Two-Dimensional Projection of AI-Generated Political Texts in Lexical-Semantic (TF-IDF) Space

The 20 passages projected into 2D with PCA of the TF-IDF vectors are shown in Figure 2. The points are not organized into neat, distinct left- and right-leaning groups, but rather tend to be more clearly clustered by topic than by prompting condition, in line with the stance-divergence finding from Section 4.3. This pattern is expected as a side effect of employing lexical-overlap-based embeddings, such as TF-IDF, vs. contextually-tuned transformer-based embeddings: TF-IDF is very sensitive to the vocabulary used in the topic, and less sensitive to the deeper syntactic and rhetorical structures such as moral framing or stance markers that are expected to be encoded more precisely by a fine-tuned transformer-based classifier, as explained in the methodological substitution note in Section 3.4.

5. Discussion

The findings corroborate the hypotheses of the study in a nuanced fashion. Similar to H2 and H3, and in line with the existing literature on prompt-sensitive LLM political expression (Bang et al., 2024; Faulborn et al., 2025), partisan framing by the model was also strongly associated with an increase in ideological distance from a neutral baseline and a strong increase in a composite polarization index compared to neutral prompting. This aligns with theoretical analyses which suggest that LLMs do not represent a fixed political stance, but rather generate ideologically charged speech in response to specific framing of a conversation (Feng et al., 2023). In contrast, as described below, the sentiment (moral-emotional and hostile language) of partisan discourse was reliably more negative than that of the neutral discourse, a finding that is consistent with the literature on the affective polarization of partisan discourse (e.g., Iyengar et al., 2019), which focuses on the moral-emotional and identity-based language as more diagnostic than simple negativity.

The most striking result is with regard to symmetry. Supporting H4, the composite polarization index was not measurably different for left- and right-leaning prompting conditions compared to the neutral prompting condition, and the results showed no evidence that left- and right-leaning passages were more lexically different from one another than they were internally different. In this limited scope of this demo corpus, this symmetry finding is somewhat distinct from earlier survey-based studies reporting a political leaning towards the left (usually left-of-center) in some LLMs' stated political opinions (Motoki et al., 2024; Rozado, 2024). A potential reconciliation is that, whereas the former studies have generally measured a model's political expression when it is not prompted to express itself politically (e.g., when it is not asked to make any political comments), this study has measured its partisanship when it is explicitly asked to do so (e.g., when it is prompted to make partisan comments). Both questions are valid and the present results are not intended to dispute the literature on the direction of partisan bias but to examine a related, complementary one regarding symmetry of partisan production when it is instructed.

Observed lexical-semantic clustering also bears the following important methodological note (disclosed in Section 3.4): topic vocabulary outweighed ideological framing in the observed clustering (Section 4.3, Figure 2) because the embeddings are sensitive more to lexical overlap than to deep meaning and stance in the text. This is a direct and expected result of using a lightweight, fully offline embedding approach over the stance classifiers based on transformers that were originally envisioned in the proposed methodology, and it is important to note that the ideological-distance and polarization-index results presented herein are intended to reflect changes in the moral-emotional and hostile language markers rather than a partisan, ideological stance. It would be interesting to see if the symmetry finding observed here generalizes using a larger, multi-model corpus, and a fine-tuned BERT or RoBERTa (Devlin et al., 2019; Liu et al., 2019) stance classifiers.

6. Conclusion

Developed and illustrated a clear computational process for exploring and demonstrating how AI-generated political discourse changes in response to a prompt, and applied it to a modest collection of LLM-style political text across five contested policy areas. Overall, the results provide evidence that explicit partisan prompting, both left- and right-leaning, is significantly related to three measures of linguistic polarization (moral-emotional language, hostility language, ideological distance from a neutral baseline) than is neutral or balanced prompting, and that for this corpus, the impact of the two partisan conditions is symmetrical. The valence of sentiment alone was not significantly different across prompt condition, indicating that the changes in polarity of AI-generated political discourse happen more through the language of moral and identity rather than just positivity or negativity. Although they rely on a small demonstration corpus and are the result of the analysis of a lexicon-based analytical substitute for the transformer-based classification, the findings provide a reproducible methodological template and a preliminary, symmetric account of how partisan framing influences the political framing of AI-generated text that can be scaled and stress-tested in future large-scale studies.

7. Recommendations

Monitoring and reporting potential changes in partisan framing in AI-generated political content to track its overall tone and moral-emotional and hostility elements should be given special focus by AI developers as these seem to be more consistent than the actual sentiment of political discourse.

As platforms produce and display politically related discourse as a response to explicitly partisan framing instructions, they should pay attention to the specific linguistic features indicative of polarization, and consider flagging or contextualizing the discourse.

Future researchers using this pipeline should replace the TF-IDF and lexicon-based substitutions used here with a more sophisticated stance and toxicity classification task done using a transformer-based model and ensure that the automated classification is validated against independent human annotators, as detailed in the original research design.

Further large-scale studies are needed to see if the symmetry finding observed here is universal across commercial and open-source LLMs, as well as a larger and more diverse set of prompts, to determine whether it is generalizable, or limited to specific models, topics, and/or prompting styles.

Awareness that AI systems can be encouraged to create highly coded discourse in any political direction, and not just a single, fixed one, should also be part of digital media literacy initiatives.

Policy-makers and platform governance authorities considering the political content risks associated with AI should differentiate between a model's unprompted/default political response and its ability to generate partisan political response when specifically asked to do so as this study indicates that these could be empirically distinct.

References

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11142–11159. https://doi.org/10.18653/v1/2024.acl-long.600

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in ChatGPT: Analyzing persona-assigned language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270. https://doi.org/10.18653/v1/2023.findings-emnlp.88

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: An examination of news produced by large language models. *Scientific Reports, 14*(1), Article 5224. https://doi.org/10.1038/s41598-024-55686-2

Faulborn, M., Sen, I., Pellert, M., Spitz, A., & Garcia, D. (2025). Only a little to the left: A theory-grounded measure of political bias in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31684–31704. https://doi.org/10.18653/v1/2025.acl-long.1529

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11737–11762. https://doi.org/10.18653/v1/2023.acl-long.656

Fujimoto, S., & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence, 6*, Article 1232003. https://doi.org/10.3389/frai.2023.1232003

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice, 198*(1–2), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE, 19*(7), Article e0306621. https://doi.org/10.1371/journal.pone.0306621

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning, 202*, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Vijay, S., Priyanshu, A., & KhudaBukhsh, A. R. (2024). *When neutral summaries are not that neutral: Quantifying political neutrality in LLM-generated news summaries*. arXiv. https://doi.org/10.48550/arXiv.2410.09978