Between Silence and Speech: Unveiling Hidden Political Biases in Large Language Models

Navneet Raju and Jaival Upadhyay
Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA

(Dated: March 13, 2025)

I. LITERATURE SURVEY

The existing literature on political bias in large language models (LLMs) has provided important insights into how these systems can absorb and propagate ideological leanings from even minimal exposure to biased data. For example, [1] demonstrate that LLMs are highly susceptible to absorbing and generalizing ideological biases from a surprisingly small number of input samples. Their work underscores the potential for these models to inadvertently adopt and propagate specific ideological leanings, even when input data is limited. This finding is particularly alarming given the growing reliance on LLMs in public discourse, where biased outputs can influence political opinions and exacerbate social divisions.

Complementing these findings, [2] presents an innovative approach in which ChatGPT is prompted to adopt various political stances, and its responses are systematically compared with its default behavior. This methodology not only uncovers systematic biases favoring certain political leanings, but also highlights subtle stylistic cues that may reveal latent ideological predispositions. The incorporation of robustness tests, such as dose-response and placebo measures, strengthens their analysis, providing a dual-layered perspective that examines both the explicit content and the underlying framework of the outputs.

Expanding the scope of bias analysis, [3] offers a comprehensive survey that synthesizes a variety of methodological approaches. They use standardized metrics, such as SEAT, StereoSet, and CrowS pairs, in conjunction with distribution-based techniques such as Jensen-Shannon divergence and Wasserstein distance to assess biases related not only to gender and race but also to political orientation across multiple open source LLMs. By establishing a quantitative framework that relies solely on output analysis, their work provides critical tools for assessing bias, especially when internal model states are inaccessible.

In a related effort, [4] investigate how popular open source LLMs respond to politically charged issues by comparing their outputs with those generated by established voting advice tools. Their findings indicate that factors such as model size and the language in which prompts are issued can significantly influence bias detection. This work highlights the importance of contextual factors in the measurement of bias and shows that the expression of political bias is highly dependent on both the content of the prompt and the cultural context in which it is presented.

Similarly, [5] adopt a fine-grained approach by dissecting bias into two key components: the explicit content of responses and the implicit framing manifested through stylistic choices. By comparing model outputs with extreme anchor distributions, they quantify both the direction and intensity of bias. Their analysis demonstrates that subtle framing effects, such as the overemphasis on certain entities or the use of emotionally charged language, play a critical role in shaping perceptions, providing valuable insights into the complex interplay between what is explicitly stated and how it is communicated.

Furthermore, extending the exploration, [6] propose a system that employs prompt engineering techniques to analyze and modulate the political biases of LLMs. Although their primary focus is on modulating bias, their work offers substantial insight into the inherent biases present in models such as ChatGPT. Using system instructions and chain-of-thought prompts, they demonstrate that it is possible to change the default political orientation of an LLM: from a left-libertarian bias to more neutral positions, thus highlighting the dynamic and malleable nature of these biases. Their findings reinforce the need for precise measurement techniques as a precursor to any intervention strategy.

Finally, [7] provide a systematic evaluation of political bias in a diverse range of topics. Their analysis reveals that while many LLMs exhibit pronounced left-leaning biases on highly polarized issues, responses to less contentious topics tend to be more uniform. In addition, their work considers factors such as model scale, release date, and regional origin, which contribute to the nuanced expression of political bias in model outputs. This comprehensive approach enriches our understanding of how external influences and internal model characteristics interact to shape bias.

Our research builds on these foundational studies by centering our investigation on the detection and quantification of political bias in LLM output. We extend existing frameworks by incorporating additional metrics, specifically sentiment analysis and refusal rates, to provide a more comprehensive understanding of bias. By evaluating the emotional tone of generated output through sentiment scores and examining the frequency with which models refrain from engaging with politically sensitive topics, our methodology is designed to capture both explicit ideological leanings and the more nuanced, latent signals of bias that have received limited attention in previous work.

Central to our research is the development of a meticulously curated dataset composed of politically sensitive prompts covering a diverse range of topics, from immigration and reproductive rights to electoral politics, alongside neutral control subjects. Each prompt is administered multiple times in a randomized order to capture both consistency and variability in the model responses. This rigorous experimental design ensures that our baseline measurements are robust and reproducible under different conditions, providing a solid foundation for detecting subtle shifts in bias.

Our evaluation framework combines both quantitative and qualitative analyses. On the quantitative side, we will compute traditional bias metrics alongside novel measures derived from sentiment analysis and refusal rates, employing statistical techniques such as t-tests and analysis of variance (ANOVA) to rigorously assess differences across various experimental conditions. In addition, a qualitative content analysis will be performed to evaluate the nuanced framing and stylistic choices evident in the results. This mixed-methods approach ensures that our analysis is both thorough and multifaceted, capable of uncovering dimensions of bias that might otherwise be overlooked.

A particularly distinctive aspect of our research is its broad-spectrum evaluation across a diverse array of LLMs. Rather than limiting our study to a single model or a narrow subset, we include major commercial systems such as OpenAI's o3 and DeepSeek, as well as various open-source LLMs. This comprehensive approach enables us to compare how models developed under differing regulatory, cultural, and technical environments express political bias. By doing so, we aim to uncover systematic differences that may arise from variations in training data, model architecture, or regional development practices.

In addition, our research is enhanced by a robust multilingual evaluation framework. While much of the existing literature has focused on bilingual or monolingual analyses, our study explicitly examines politically sensitive issues across multiple languages. For instance, when addressing a politically sensitive issue between India and China, our methodology involves prompting in English — a neutral language, alongside prompts in Hindi and Traditional Chinese, which represent opposing perspectives. This multilingual approach enables us to capture how linguistic and cultural contexts influence bias expression, providing a more nuanced and globally relevant understanding of LLM performance.

In summary, while previous work by [1] [2] [3] [4] [5] [6] and [7] has laid a strong foundation for detecting and measuring political bias in LLMs, our research advances these efforts by incorporating additional metrics such as sentiment scores and refusal rates. Coupled with a comprehensive, multilingual evaluation framework that tests a diverse range of models, from commercial systems like OpenAI's o3 and DeepSeek to various open source platforms. Our research aims to provide detailed, actionable insights into the mechanisms by which LLMs process politically sensitive information. This enhanced understanding will not only contribute to academic discourse but also inform future policy discussions on the ethical use of AI in politically sensitive contexts.

[1] K. Chen, Z. He, J. Yan, T. Shi, and K. Lerman, How susceptible are large language models to ideological manipulation?, arXiv preprint arXiv:2402.11725 (2024).

[2] F. Motoki, V. Pinho Neto, and V. Rodrigues, More human than human: measuring chatgpt political bias, Public Choice **198**, 3 (2024).

[3] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics , 1 (2024).

[4] L. Rettenberger, M. Reischl, and M. Schutera, Assessing political bias in large language models, arXiv preprint arXiv:2405.13041 (2024).

[5] Y. Bang, D. Chen, N. Lee, and P. Fung, Measuring political bias in large language models: What is said and how it is said, arXiv preprint arXiv:2403.18932 (2024).

[6] Y. Chang and Y. Sun, A system to analyze and modulate the political biases of large language models using prompt engineering techniques, Security, Privacy and Trust Management (2024).

[7] K. Yang, H. Li, Y. Chu, Y. Lin, T.-Q. Peng, and H. Liu, Unpacking political bias in large language models: Insights across topic polarization, arXiv preprint arXiv:2412.16746 (2024).