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. PROJECT PROPOSAL

Political bias in large language models (LLMs) has emerged as a critical concern in the field of artificial intelligence, as these models increasingly shape public discourse and influence perceptions on sensitive issues. With applications ranging from automated news generation to decision support systems, any inherent bias, especially one that aligns with particular political ideologies, can have far-reaching societal implications. In this study, we ask the following. How do LLMs developed in different political and regulatory environments (for example, OpenAI's o1 versus China's DeepSeek and Llama versus Qwen) exhibit political bias in their outputs, and what underlying factors contribute to these differences? Addressing this question is essential for advancing AI fairness, ensuring that deployed systems do not inadvertently reinforce or propagate biased narratives, and for informing the design of future regulatory policies.

Recent studies have begun to explore this complex issue. [1] found that larger models tend to align more with left-leaning political parties, while [2] proposed a framework that quantitatively measures both the content and stylistic characteristics of politically charged LLM outputs. Complementing these findings, [3] evaluated nine types of bias in five open-source LLMs, revealing that although bias is present, larger pre-trained models may exhibit lower levels of overt bias - possibly as a result of fairness constraints during pre-training. Furthermore, [4] developed a probing method to detect implicit sociodemographic biases in LLM representations, even when the models refuse to provide direct answers on sensitive topics. Collectively, these studies underscore the multifaceted nature of political bias in LLMs and highlight the need for comprehensive evaluation methodologies.

Building on this foundation, our study aims to extend previous research by conducting a cross-context and cross-lingual analysis of political bias in LLMs. Specifically, we compare models developed in different regulatory environments—such as OpenAI's o1, which is trained on relatively open and diverse data, versus Chinese models like DeepSeek that are subject to stringent censorship—and examine how these differences manifest when the models are queried on politically sensitive topics. Moreover, by including comparisons between models like Llama and Qwen, our research seeks to disentangle the effects of training data, instruction fine-tuning, and language on the expression of political bias. This approach not only broadens the scope of bias analysis but also provides a more nuanced understanding of how political narratives are embedded in AI outputs.

We hypothesize that LLMs developed under strict censorship regimes—such as those found in China—will demonstrate higher rates of self-censorship or produce narrative-aligned outputs when addressing sensitive topics like the Tiananmen Square incident or the political status of Taiwan, compared to models developed in more open environments like those of OpenAI. Additionally, we posit that while instruction fine-tuning may reduce overt bias (evidenced by lower refusal rates or moderated sentiment), it does not fully eliminate implicit biases present in the underlying embeddings. Finally, we anticipate that the language used to interact with the models (simplified Chinese versus English) will significantly influence their responses, reflecting differences in cultural context and training data composition.

To test these hypotheses, we propose a mixed-methods approach that combines quantitative and qualitative analyses. Our methodology involves designing a standardized set of open-ended prompts covering a range of politically sensitive topics and neutral control subjects. We will query each model multiple times—across different languages where applicable—and analyze the outputs using metrics such as censorship/refusal rates, sentiment scores, keyword frequencies, and lexical framing. Statistical tests (e.g., t-tests and ANOVA) will be applied to determine the significance of observed differences, while qualitative content analysis and blind review by independent raters will provide insights into the subtleties of narrative framing and implicit bias. With the computational resources available via platforms such as Google Colab and university systems, our project is well-equipped to carry out this comprehensive analysis.

In summary, our research seeks to provide a systematic and comparative evaluation of political bias in LLMs, contributing novel insights to the field of AI fairness. By elucidating how models trained in different political contexts handle sensitive topics—and by directly comparing OpenAI's o1 with DeepSeek, and Llama with Qwen—our work aims to inform both technical bias mitigation strategies and broader policy discussions regarding the ethical deployment of AI technologies.

[1] L. Rettenberger, M. Reischl, and M. Schutera, Assessing political bias in large language models, ArXiv **abs/2405.13041** (2024).

[2] 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 **abs/2403.18932** (2024).

[3] J. He, N. Lin, M. Shen, D. Zhou, and A. Yang, Exploring bias evaluation techniques for quantifying large language model biases, 2023 International Conference on Asian Language Processing (IALP) , 265 (2023).

[4] R. Tang, X. C. Zhang, J. J. Lin, and F. Ture, What do llamas really think? revealing preference biases in language model representations, ArXiv **abs/2311.18812** (2023).