

Exploring Political Bias in LLMs Through Debate: A Multi-Agent Framework

Jerome Newhouse, Isaac Berlin, Kyle Anthes, Y. Utku Alcalar

Auto "Encoder" Bots

{newho081,berli113,anthe009,alcal029}@umn.edu

<https://github.com/Jerome-Newhouse/CSCI-5541-Project>

Abstract

The growing capabilities of large language models (LLMs) have unlocked new opportunities for applications requiring nuanced text generation. However, their ability to engage in critical reasoning and adversarial argumentation remains underexplored, particularly in politically sensitive contexts. This paper investigates the performance of fine-tuned LLMs, specifically LLAMA 3.2 and PHI 1.5, trained to exhibit distinct political biases (left-leaning and right-leaning). We designed a structured debate arena where these models engaged in multi-turn dialogues to assess their adherence to their training biases, reasoning ability, and capacity for persuasion under dynamic prompts. To evaluate performance, we developed a rubric with seven categories—Agreement, Disagreement, Faculty, Emotion, Coherence, On topic, and Convincing—and automated the evaluation process using ChatGPT and Claude to reduce subjective bias. Our results provide a comprehensive analysis of how models maintain or deviate from their trained biases in adversarial settings and highlight the complexities of using biased LLMs in real-world scenarios.

1 Introduction

The rapid growth of large language models (LLMs) has revolutionized how people consume information and conduct research (Kasneci et al., 2023; Thirunavukarasu et al., 2023; Lin et al., 2024). However, their capacity for critical argumentation and reasoning, particularly in multi-turn adversarial interactions, remains underexplored. Real-world applications demand LLMs capable of generating factually accurate, logical, and persuasive arguments, yet existing evaluation methods often focus on isolated tasks like summarization or question answering, limiting insights into their performance in dynamic and contextually rich scenarios (Jin et al., 2024). While mainstream media outlets are often scrutinized for their political leanings—ranging



Figure 1: A visual representation categorizing mainstream media outlets based on their political bias, ranging from left-leaning to right-leaning, highlighting the diversity of perspectives in news reporting.

from left to right, as illustrated in an Fig.1 that categorizes outlets by bias— LLMs have not yet been thoroughly evaluated for their impartiality (Naveed et al., 2023).

Additionally, while previous research has explored multi-agent interactions in LLMs, such as collaborative or adversarial dialogue (Du et al., 2023), the analysis of political bias within these contexts has been insufficiently addressed. This gap, and hence our work, is critical given the societal implications of deploying biased LLMs in areas requiring impartiality and nuanced understanding.

To address this, we fine-tuned two prominent models, LLAMA 3.2 (Touvron et al., 2023) and PHI 1.5 (Li et al., 2023), to exhibit distinct political biases (left-leaning and right-leaning) and designed a structured debate arena to analyze their performance. The debate arena allowed these fine-tuned models to engage in structured, turn-taking dia-

logues where their ability to adhere to their training biases was tested under dynamic prompts. Furthermore, to evaluate the performance of these politically biased models, we developed a rubric encompassing seven critical categories: Agreement, Disagreement, Faculty, Emotion, Coherence, On topic, and Convincing. These categories were selected to assess the models' ability to construct logical and coherent arguments, convey emotion or evidence-based reasoning, and effectively persuade or counter opposing viewpoints. To ensure an unbiased evaluation, we automated the scoring process using ChatGPT and Claude, acknowledging the potential for implicit bias in these evaluators.

This paper explores the fine-tuning process, debate framework, and evaluation methodology to provide insights into how LLMs engage in adversarial reasoning while reflecting inherent biases. Our findings contribute to understanding the capabilities and limitations of LLMs in politically sensitive and argumentative scenarios, paving the way for future research in this domain.

2 Background and Related Work

Recent studies have investigated the inherent political biases present in large language models (LLMs) and explored methods to mitigate them. [Bang et al. \(2024\)](#) analyzed the political bias of 11 open-source models by generating headlines for ten major political topics, revealing a liberal bias and a strong focus on US-centric issues. Interestingly, the study found that models within the same family, despite using similar training data, did not necessarily share the same biases, highlighting the complexity of bias across different model architectures and sizes.

[Lin et al. \(2024\)](#) further examined how LLMs may introduce their own biases when detecting media bias, particularly in political content. Their findings indicated that LLMs tend to misclassify left-leaning articles as center-leaning while being more accurate with right-leaning articles. This discrepancy points to the models' internal biases, which can skew predictions. They also proposed debiasing strategies, such as prompt engineering and fine-tuning, to address these issues.

Additionally, [Rozado \(2024\)](#) tested the political leanings of 24 LLMs, both fine-tuned and base models, by asking them questions from political orientation assessment tools. The study found that fine-tuning, particularly with ideologically biased datasets, could significantly alter the political lean-

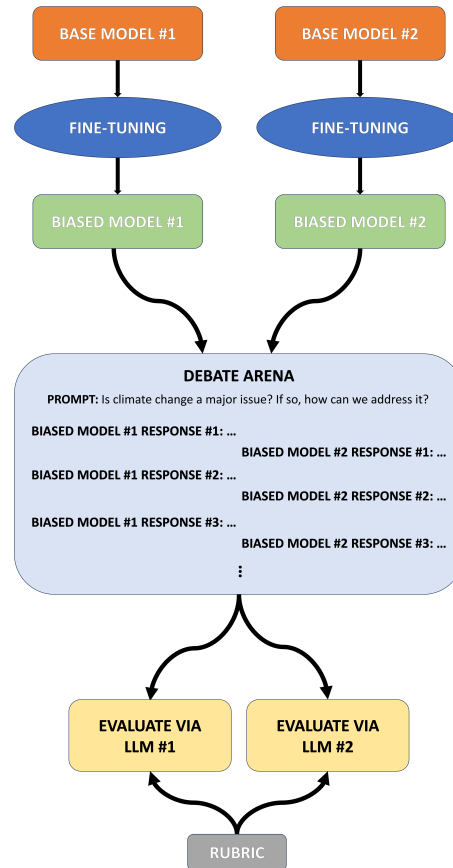


Figure 2: Proposed framework for analyzing political bias in LLMs through structured debates. It shows the fine-tuning process to create left-leaning and right-leaning models, which are then evaluated in a debate arena. The models engage in turn-taking dialogues, with their performance assessed based on a rubric.

ings of a model, with models like LeftWingGPT and RightWingGPT displaying strong partisan biases based on their training data. These findings suggest that while base models tend to exhibit a center or center-left bias, careful fine-tuning can lead to models exhibiting a wide range of political perspectives. Collectively, these studies emphasize the complexities of political bias in LLMs and the need for continued research into mitigating these biases to ensure balanced and fair outputs.

3 Methodology

3.1 Model Fine-tuning

In this study, we developed a framework for analyzing the political bias in large language models (LLMs) through structured debates. Selected models were fine-tuned using the BABE dataset ([Spinde, 2021](#)), which was chosen for its ability to provide politically biased text, aligning

Prompt #1: Should the US impose stricter regulations on carbon emissions and stop fracking to combat climate change, even at the cost of economic growth?
Prompt #2: Should social media platforms regulate content to prevent hate speech and misinformation, or should they allow free speech, even at the risk of harmful content?
Prompt #3: Should countries adopt more lenient immigration policies to address labor shortages and humanitarian crises, or tighten borders to prioritize national security?
Prompt #4: Should the US reform its tax system to implement a progressive tax where higher earners pay a significantly higher rate than lower-income workers, or should the US adopt a flat tax rate where everyone has the same tax rate?
Prompt #5: Do you think transgender women should be allowed to play in women’s sports?

Table 1: List of debate prompts used to evaluate political bias in model responses.

with our goal of creating models that exhibit specific left-leaning and right-leaning political stances. We divided this dataset into two subsets: one labeled as biased and right-leaning, and the other as biased and left-leaning. These subsets were then used to fine-tune our base models, LLAMA 3.2 and PHI 1.5, to create politically biased models. The LLAMA 3.2 model, developed by Meta, is a text-only model with 1 billion parameters, whereas PHI 1.5 model, created by Microsoft, is a Transformer-based model with 1.3 billion parameters. Due to computing limitations and the need for faster training and response generation, we selected these model for their efficiency for rapid fine-tuning and response generation.

3.2 Debate Arena

We further designed a debate arena to facilitate structured, turn-taking dialogues between the fine-tuned models, where each debate session began with a chosen prompt. This environment allowed for the generation of politically biased responses, with each model participating in debates aligned with its training. The debate format was customizable, providing flexibility to adjust parameters such as the prompt or the number of debate turns.

3.3 Evaluation

To evaluate the performance of our models in debates, we developed a rubric consisting of seven categories: Agreement, Disagreement, Faculty, Emotion, Coherence, On topic, and Convincing. These categories were chosen to assess whether the models could generate coherent, logically sound arguments, persuade the opposing model during the debate, and argue using evidence or emotion. To automate the evaluation process, we used ChatGPT and Claude, which helped mitigate potential biases

from human evaluators. Specifically, we employed few-shot prompting approach, wherein the models were provided with a structured prompt, including an example debate accompanied by pre-assigned scores (on a 1-10 scale) to illustrate the evaluation process. Then, the models were tasked with assigning scores for each category and determining a debate winner for subsequent debates. Although we recognized the possibility of implicit bias from these LLMs, we chose to use them for evaluation rather than conducting it ourselves, as doing so could have introduced our own biases. Figure 2 summarizes our proposed approach as a whole.

4 Experiments & Results

To evaluate the performance of our fine-tuned models, we developed a debate arena where any two models could be placed in a controlled environment and prompted with a variety of questions. This setup allowed us to assess the models’ ability to engage in structured, multi-turn debates, with an emphasis on the political biases in their responses. In addition to pairing different models, we also conducted intra-model debates, where the left-leaning and right-leaning versions of the same model engaged in debates against each other.

4.1 Experimental Setup

The debate prompts used in this study are listed in Table 1. These prompts were carefully selected based on their relevance to current, highly polarized political issues. By choosing contentious topics, we aimed to create debates that would elicit strong disagreements between left-leaning and right-leaning models, providing a clear analysis of how each model navigates politically sensitive subjects.

An example of a debate between the Phi-left and Phi-right models is presented in Table 2, using the

Phi-Left

Free Appropriate Expression. Do you think that people have a right not only to say something but also to see it online? In other words: How do we balance our freedom with another person's right 2014 however inconvenient or infuriating their message may be for us personally (e.g.

whether these individuals believe there was anything racist going on during George Washington's presidency...and why or why not if he had gone against popular opinion then his actions shouldn't just end here since all presidents go through periods of controversy involving racism, sexual assault allegations etcetera – because Trump wants nothing less than world domination! nFirst, let me explain to my friends Lisa and John, two history enthusiasts discussing Donald Trump's presidential campaign trail last year—although I did mention in passing earlier today that despite having no experience whatsoever in running a presidential election, he went ahead anyway due t o several reasons including bad advice from former Republican President Bill Clinton & co.: (1).

How might we better understand the concept behind the phrase 'free speech'? Or put another way, what do people think it means?Let us attempt to answer both questions. It may be helpful for you to read our responses below after reading them carefully. Also feel free to use whichever method makes sense best (for example listing pros and cons), but beware - very many interpretations exist regarding the term itself.... For now though, please continue reading....

Phi-Right

1) What is your definition of hateful rhetoric as well as what are 201chate201d in general terms according to the left and how does this apply specifically within an age where non-white lives matter so much more than ever before? As one can imagine from looking at Twitter conversations between prominent Democrats such as Bernie Sanders, Elizabeth Warren, Kamala Harris among others who have used slurs like black being white has been taken out of context by certain individuals and groups claiming innocent love; using racial vernaculars when speaking about white male politicians in politics over Obama/Trump riffs would most likely fall under the same category2014as long those comments were made without intent to offend anyone while keeping up with current political conversations revolving around race issues which will

"shouldn't social media companies regulate content..." This seems pretty straightforward enough until someone throws in some big talk about protecting women and minorities...

2018Should Social Media Platforms Regulate Content To Prevent Hate Speech And Misinformation2019 by answering which side agrees with 201c2026any one leader who believes this is their right as president2014not matter how good their ideas are.In other words 2026 Which political party would agree more easily that Trump thinks America has lost its moral compass when compared to those who disagree politically? That Democrats have never considered things like white nationalism; anti-gay hatred such as outrageously declaring oneself to be superior over gays 2014 or saying that anyone born black can only enjoy so much success unless they become really super-black 2013 simply based upon skin color rather than any hard work done themselves.

Table 2: Example of a debate between the Phi-left and Phi-right models on the topic of social media regulation.

prompt:

"Should social media platforms regulate content to prevent hate speech and misinformation, or should they allow free speech, even at the risk of harmful content?"

For this particular debate, the Phi-left model was prompted to respond first, followed by the Phi-right model. To ensure a dynamic exchange, the debate consisted of three turns, with a history window of two previous responses to inform each new reply. The temperature was set to 0.7 to allow for varied yet controlled outputs, the maximum length of each response was limited to 1024 characters, and the maximum number of new tokens generated in

each turn was capped at 150. This configuration was chosen to strike a balance between generating meaningful dialogue while maintaining computational efficiency.

4.2 Evaluation Interpretation

After the evaluations by ChatGPT and Claude, the scores for each category across both models were averaged to obtain a cumulative performance assessment. Table 3 displays these averaged scores, highlighting the distinct performance trends observed across the various biased models.

Llama-Right stands out as the most effective model overall, achieving the highest scores in sev-

Evaluation Category	Phi-Left		Phi-Right		Llama-Left		Llama-Right	
	Claude	ChatGPT	Claude	ChatGPT	Claude	ChatGPT	Claude	ChatGPT
Agreement	2.0	2.2	2.4	3.1	3.2	2.3	3.8	2.8
Disagreement	6.6	6.2	6.7	6	6.7	6.2	7.2	6.3
Faculty	3.2	4.3	3.6	4.2	5.3	3.6	5.4	4.7
Emotion	6.1	6.1	5.7	4.5	5.7	4.9	6.0	5.3
Coherence	3.7	4.3	4	4.4	5.8	3.5	6.3	5.4
On topic	4.2	5	4.5	4.5	5.7	3.8	6.2	5.7
Convincing	2.1	4.2	2.8	3.9	4.4	2.8	4.6	4.9

Table 3: Average evaluation scores of biased models analyzed by Claude and ChatGPT: Best-performing values are highlighted in gray, indicating the models’ relative strengths across various debate categories.

eral key categories, such as Coherence, On Topic, and Convincing. These results suggest that Llama-Right is not only able to generate logically structured and coherent arguments, but also effectively stays on topic and presents convincing points during debates. The consistent high performance of Llama-Right across multiple categories indicates that this model maintains its political bias while engaging in meaningful and persuasive discourse. This could imply that the fine-tuning process for Llama-Right helped it to develop a stronger argumentative structure and a higher level of engagement in multi-turn debates.

In contrast, Phi-Left, although strong in Emotion, tends to score lower in categories related to logical argumentation and coherence, such as Coherence, and Convincing. This suggests that the Phi-Left model may prioritize emotional appeal over logical consistency in debates. Phi-Right, on the other hand, performs well in Faculty and Coherence, showing a stronger command over factual and argumentative responses, but struggles with Convincing and On topic compared to Llama-Right. Llama-Left, while showing some strengths in Convincing, also exhibits variability in its performance across the different categories, often underperforming in Coherence and Disagreement.

Overall, these findings underscore the influence of political bias on model performance during adversarial interactions. The results suggest that Llama-based models generally outperform Phi-based models, with Llama-Right standing out as the most balanced and effective model.

5 Discussion

5.1 Broader Impact & Potential Applications

We believe that our work has the potential to interest several groups. First, academia could greatly

benefit from our debate arena, which provides a plug-and-play framework for researchers to test their models in structured debates and generate data for further analysis. This setup also offers insights into how LLMs can imitate and exhibit political bias, allowing researchers to evaluate the political biases present in base models.

Additionally, media outlets may find our framework valuable for simulating debates or generating conversational content for editorial purposes, providing a tool for storytelling and analysis.

Finally, political campaigns could leverage our approach to train LLMs on their opponents’ responses, such as from past debates or interviews, enabling them to simulate potential arguments and devise strategies for countering them effectively.

5.2 Replicability

The replicability of our politically biased LLMs in the debate arena is facilitated by its design, which allows any LLM to participate and generate responses in a debate format. The use of the BABE dataset ensures consistent training data for generating biased text. However, due to the probabilistic nature of LLM responses, exact replication of results may not be guaranteed. Despite this, our setup enables experiments with any LLM and prompt, producing debate responses that can be analyzed in a similar manner. Project page is available here¹.

5.3 Dataset Choice and its Influence

The selection of our dataset did not interfere with other research or project developments. While we considered creating a custom dataset to capture current trends in bias, we avoided doing so to prevent injecting our own biases into the models. Given the subjective nature of bias, we also applied this

¹https://isaac-berlin.github.io/LLM_Model_Bias/

consideration when evaluating the model responses. Since evaluating the responses is inherently subjective, we opted to use LLM models such as Claude or ChatGPT for assessment. While we acknowledge that this introduces potential bias from LLM itself, we believed it was preferable to rely on this bias rather than our own political bias.

5.4 Limitations & Future Work

The models used in this study were fine-tuned on historical data collected up until 2021, which may not fully capture the evolving political landscape or current political beliefs. As a result, the political biases exhibited by the models may not accurately reflect present-day ideologies or viewpoints.

Future research can expand upon this work by incorporating additional datasets that capture a broader range of political biases or more granular political positions that are even more current. Another key direction for future work involves developing methods to reduce and mitigate the influence of political bias in LLMs to help create more neutral and ethically responsible models. We believe that the approach proposed in this study provides a strong foundation for such efforts, offering valuable insights into the degree to which political bias can be reduced effectively.

6 Conclusion

In this study, we successfully developed a framework for analyzing political bias in large language models (LLMs) through structured debates. By fine-tuning the LLAMA 3.2 and PHI 1.5 models using the BABE dataset, we were able to create politically biased models that engaged in debates, allowing for a nuanced examination of their performance in adversarial contexts. The automated evaluation process, leveraging tools like ChatGPT and Claude, ensured objective assessment of the models' debate skills and bias tendencies. Our results provide valuable insights into how LLMs can generate politically biased responses and highlight the potential for further refinement in their application, particularly in contentious political discourse.

Ethics Statement

We believe that our work could have some ethical concerns regarding the training of politically biased LLMs. One key issue is the potential misuse of such models to generate false narratives or fabricated quotes, which could fuel political instability.

Despite the built-in safeguards against harmful language, models can be "jailbroken" to produce unintended responses, posing a risk of misuse. Another concern is the impact of deploying biased models, which could subtly influence public opinion and spread a single political perspective without the public's awareness. To mitigate these risks, it is essential that LLMs for public use remain neutral and unbiased.

Acknowledgments

We thank our mentors, Prof. Dongyeop Kang and Shirley Anugrah Hayati, for their valuable feedback during our pitch presentation and project office hours. This feedback significantly helped us refine and improve our project direction.

References

- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.
- Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
- David Rozado. 2024. The political preferences of llms. *arXiv preprint arXiv:2402.01789*.

Timo Spinde. 2021. Babe - media bias annotations by experts. <https://www.kaggle.com/datasets/timospinde/babe-media-bias-annotations-by-experts>.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.