Investigating Political Bias in LLM Responses

By Campbell Isherwood

1. Introduction

In the rapidly expanding world of artificial intelligence, large language models (LLMs) are increasingly tasked with answering politically sensitive questions. This project explores the political bias that can emerge in LLM responses, focusing specifically on U.S. domestic topics. The decision to focus on U.S. domestic issues was driven by several factors: first, the United States represents a highly polarized political environment, making it an ideal setting for examining how sensitive language models are to ideological framing. Second, U.S. policy debates are globally influential, and biases emerging from these contexts could have international ramifications when such models are deployed worldwide. Finally, the extensive documentation and public discourse surrounding American political issues provide a rich foundation for interpreting and validating bias findings. The study investigates how the **framing of prompts** influences model outputs, combining **bias labeling** and **sentiment analysis** to offer a multi-dimensional assessment of model behavior.

Research centers around four major U.S. political topics:

- **Gun Control**
- **Climate Change**
- **Taxation and Wealth Redistribution**
- **Policing and Criminal Justice Reform**

Prompts were designed using three styles:

- **Balanced**: Neutral, fact-based framing
- **Party Framing**: Language associated with Democratic or Republican narratives
- **Controversy**: Polarized and emotionally charged wording

This two-phase approach allows a deeper evaluation of not only bias but also emotional tone and intensity. By separately assessing political bias and emotional sentiment, I can capture both the ideological leaning and the affective framing of LLM responses. Bias labeling identifies directional political tendencies, while sentiment analysis reveals whether the language used conveys positivity, negativity, or neutrality. Together, these complementary methods enable a more comprehensive and nuanced understanding of how language models respond to politically sensitive prompts and allow researchers to distinguish between ideological bias and emotional loading.

---

2. Experimental Setup

2.1 Prompt Design and Response Collection

Custom-written prompts were batch-submitted to **GPT-4** through an automated pipeline. For each designed prompt, 12 separate responses were generated to account for natural variability in LLM outputs. Each response was recorded in a structured format with metadata:

Field	Description
Topic	Main issue area (e.g., gun control)
Prompt Type	Balanced, Party Framing, or Controversy
Prompt	Text of the input
Response	LLM's generated text
Timestamp	Submission time

Responses were saved to a CSV file named **gpt4_responses.csv** for systematic processing.

2.2 Keyword-Based Bias Labeling

To initially categorize political bias, I implemented a **keyword-matching** system. Keywords strongly associated with left-leaning or right-leaning ideologies were defined. Each response was scored, and the overall label was determined:

- **Left**: More left-associated keywords detected
- **Right**: More right-associated keywords detected
- **Neutral**: No clear lean detected

Example Left Keywords: "systemic racism," "climate justice," "wealth inequality"

Example Right Keywords: "law and order," "second amendment," "limited government"

Table 1 shows a sample of the keywords used for bias labeling.

Political Lean	Example Keywords
Left	systemic racism, climate justice, wealth inequality
Right	law and order, second amendment, limited government

Table 1: Example Keywords Used for Bias Labeling

2.3 Sentiment Analysis

Using **VADER Sentiment Analyzer**, each response was assigned a **sentiment score** between -1 (very negative) and +1 (very positive). Sentiment analysis provided an additional dimension to capture the emotional charge of responses across different framings and topics.

2.4 Classifier Training

To automate future bias detection, a **Logistic Regression** classifier was trained:

- **TF-IDF vectorization** was applied to the text responses.
- **Train-test split**: 80/20.
- **Evaluation**: Classification report (precision, recall, F1) and confusion matrix.

The classifier demonstrated reasonable performance, suggesting that bias in LLM outputs can be detected through standard machine learning approaches.

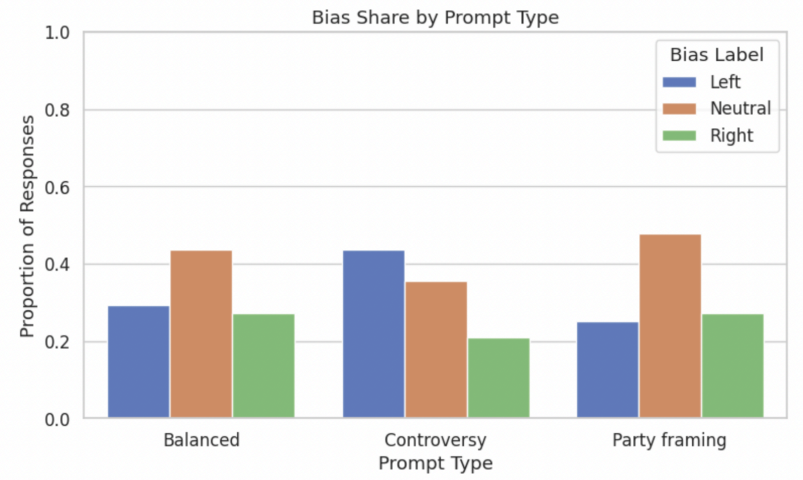
---

3. Results and Visualizations

3.1 Overall Bias Distribution

Neutral responses were most common overall. However, a significant number of responses showed a detectable lean towards left-wing perspectives.

Figure 2: Bias Share by Prompt Type

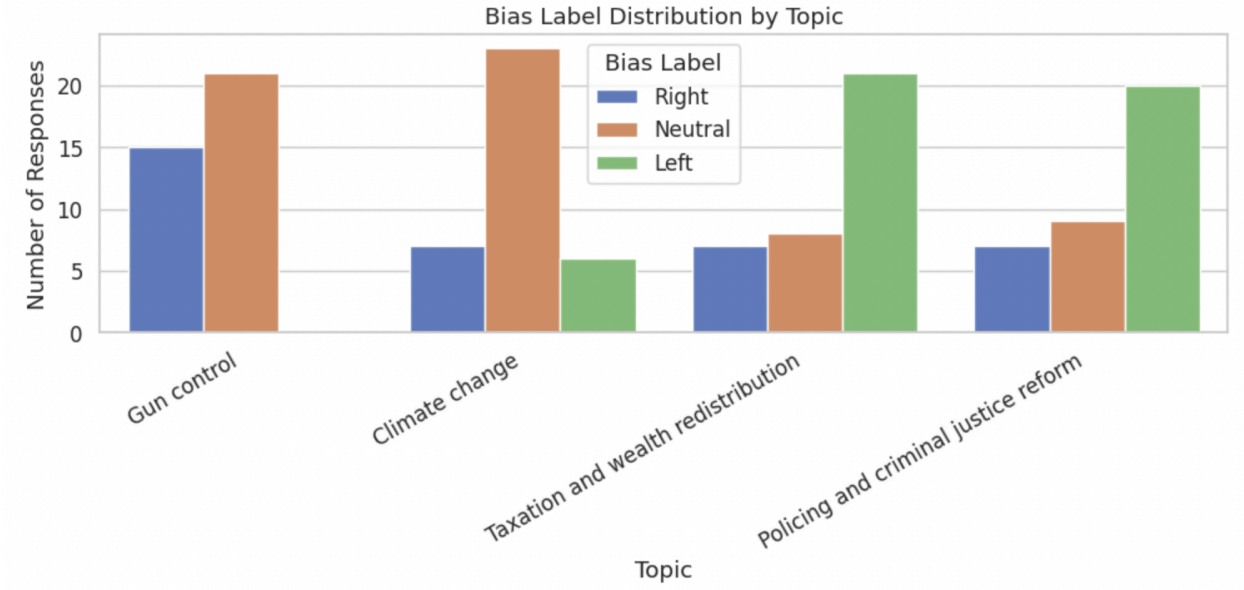


Observation: Controversial prompts produced greater polarization compared to balanced prompts.

3.2 Bias by Topic

In topics like "Policing and Criminal Justice Reform," left-leaning responses dominated. Climate change responses skewed negative under controversial framing.

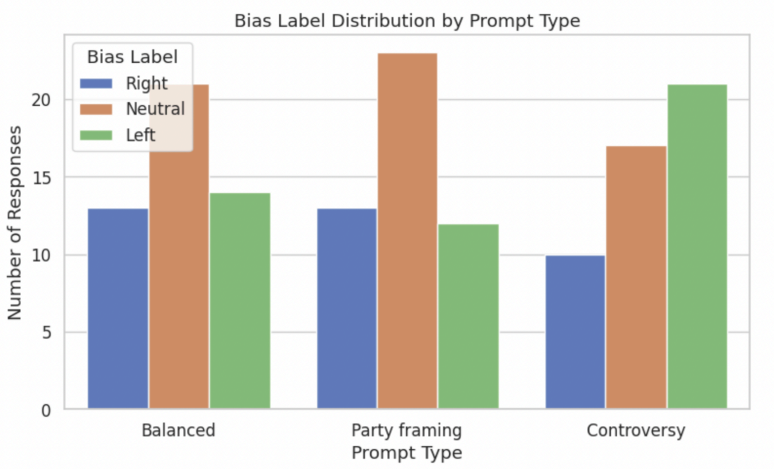
Figure 3: Bias Label Distribution by Topic



Observation: Different topics triggered different bias patterns, reflecting societal polarization.

3.3 Bias by Prompt Type

Balanced prompts yielded more neutral outputs. Controversial prompts often induced greater polarization (higher proportion of left or right leanings).

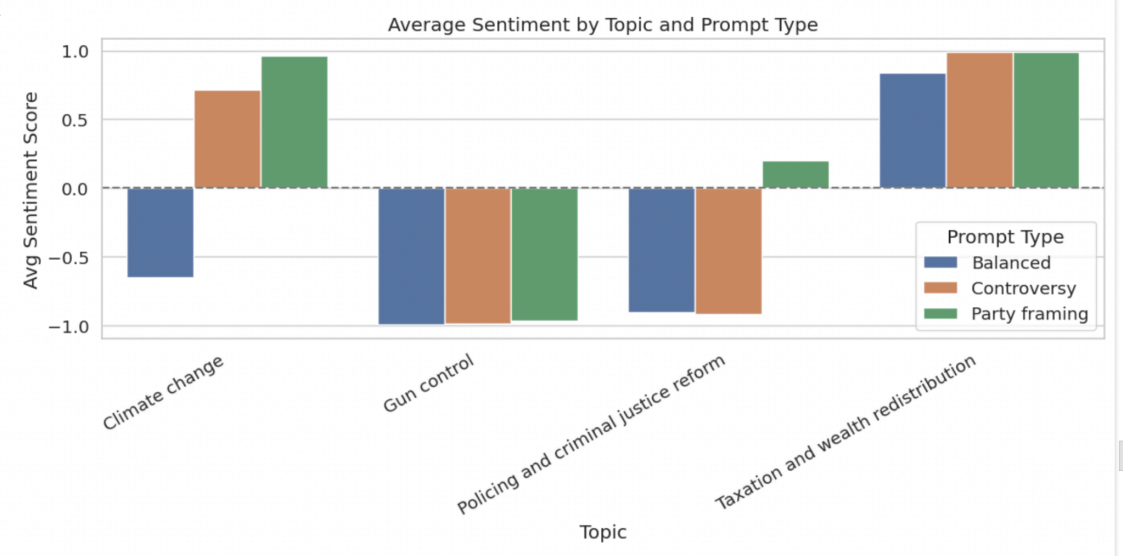


Observation: Neutral responses consistently appear most frequently across all prompt types, but Controversy prompts show a notable increase in Left-leaning responses.

3.4 Sentiment Analysis by Topic

Climate change and policing topics under controversial framings were associated with more negative sentiments, indicating emotional intensification.

Figure 4: Average Sentiment by Topic and Prompt Type



Observation: Negative sentiment was higher for controversial prompts, suggesting emotional triggering.

---

4. Discussion

The project reveals several important findings:

- **Prompt Framing Matters:** LLMs exhibit different biases based on how a question is framed. Controversial framing tends to magnify existing biases.
- **Topic-Specific Sensitivities:** Topics like "Policing and Criminal Justice Reform" are more prone to left-leaning narratives, while "Gun Control" and "Taxation" showed more nuanced or mixed outputs.
- **Emotional Charge:** Controversial prompts not only skewed bias but also increased emotional negativity.

While keyword matching is a strong starting point, more advanced **contextual models** could improve bias detection by recognizing implied or subtle bias without explicit keywords.

Additionally, the variability in sentiment across different prompt framings suggests that emotional tone is another crucial dimension in analyzing political bias. Higher negativity associated with controversial prompts indicates that framing not only affects perceived political leaning but also the overall emotional valence of the response. This dual-layer evaluation (bias and sentiment) provides a richer, more complete understanding of LLM behavior on politically sensitive topics.

This study also highlights the need for developing LLM evaluation methodologies that are sensitive to framing effects, ensuring that AI systems do not unintentionally reinforce polarized or biased perspectives based solely on input phrasing.

---

5. Challenges and Limitations

One important limitation of this project is that the keyword-based bias labeling approach may fail to detect **subtle or implied biases** that do not rely on explicit keywords. As a result, some nuanced ideological positions could be misclassified or missed entirely. Additionally, while **sentiment analysis** provided valuable insight into the emotional tone of the responses, it does not directly measure political orientation, meaning that high negativity or positivity scores cannot necessarily be interpreted as indicators of ideological bias.

Another constraint was the **small size of the dataset**. Expanding the number and variety of prompts across a wider range of topics would likely improve the robustness and generalizability

of the findings. Related to this, **handling nuance** remains a major challenge. Bias is often highly context-dependent, and simple keyword-based or linear models cannot fully capture the complexity of ideological positioning embedded in natural language.

There is also potential for **sampling bias**, as this study analyzed only one large language model (GPT-4). The behaviors and biases observed here may not generalize to other models trained on different data, with different architectures, or fine-tuned with alternative objectives. Finally, even when prompts were designed to be balanced, **prompt interpretation variability** by the model could influence outcomes. Minor differences in phrasing, word choice, or structure may cause the model to respond in unexpectedly biased or emotionally charged ways, complicating efforts to isolate framing effects.

---

6. Conclusion

This project successfully demonstrates that **large language models are sensitive to prompt framing** when discussing politically charged topics. By combining **bias labeling** with **sentiment analysis**, we created a rich framework for evaluating not just "what the model says" but "how it says it." These findings are critical for developers and policymakers aiming to create **more fair, transparent, and accountable AI systems**.

Moving forward, adopting multi-layered bias detection strategies and scaling evaluations across different LLMs will be essential steps to comprehensively map, understand, and mitigate biases in AI-generated content.