**Assessing the Robustness of Large Language Models to Toxicity Across Political Identities**

**Idea:**

Despite their demonstrated proficiency in generating human-like text across various subjects, concerns persist regarding potential biases and toxicity in the content LLMs produce, especially within the realm of political discourse. An in-depth exploration of variations in toxicity exhibited by LLMs when generating content from differing political perspectives is necessary. This experiment involves exploring whether LLMs exhibit varying levels of robustness to toxicity when generating content from these perspectives and understanding how their fine-tuning processes impact this aspect. Investigating how LLMs are fine-tuned with datasets representing liberal and conservative viewpoints is essential for gauging the impact of these processes on the models' proficiency in generating content aligned with or deviating from toxic patterns.

**Purpose of the Project:**

Understanding the robustness of LMMs to toxicity across political identities lets us address concerns related to biased text generation. The results of this project will play a pivotal role in shaping strategies to enhance the robustness of LLMs, mitigate toxicity, and foster responsible and unbiased content generation across diverse perspectives..

**Potential Solution:**

Firstly, by prompting these models to generate content without explicit specification of a political stance, a neutral baseline is established. This approach facilitates the measurement of toxicity levels using established metrics, providing a foundational understanding of the models' general behavior without the influence of predefined political perspectives. Additionally, the research extends its investigation by subjecting LLMs to a fine-tuning process using a dataset that comprehensively encompasses perspectives from both liberal and conservative political identities. Through this comprehensive approach, the research seeks to unravel how LLMs adapt to and manage toxicity when exposed to a spectrum of political viewpoints during the fine-tuning process.

**Expected Outcome:**

* Identification of significant differences in robustness of LLMs to toxicity levels when they generate content from liberal versus conservative perspectives.
* Insight into the robustness of fine-tuned LLMs to toxicity in training data containing perspectives from both political identities.
* By understanding potential variations in toxicity levels across different political stances, we can discern whether LLMs inadvertently amplify or mitigate biases in their outputs.

**Timeline:**

* Data Preparation and Preprocessing - February 9, 2024
* Generating content and evaluation distribution of perspective - March 1, 2024
* Fine Tuning Experimentations - March 29, 2024
* Analysis of generated content - April 19, 2024
* Final Report - May 1, 2024

**TA :** [Zihao He](mailto:zihaoh@usc.edu)

**Team Members:**

* **Abhishek Manoharan**
* **Manojkumar Manivannan**
* **Naveen Sasikumar**
* **Niyantha Maruthu Pandiyan**
* **Prakashraj Nehrudass**
* **Prathyusha Nareshkumar**