Project Proposal

1 Project Title

The project is titled **Testing for Prompt Robustness against Bias in LLMs**.

2 Team Members

The team members are **Sevag Baghdassarian (260980928)** and **Ling Zhang (260985358)**.

3 Responsible AI Keywords

Bias and Discrimination, Stereotypical Bias, Fairness, Safety.

4 Motivation/Context

LLMs are having an increasing influence on society and technology. As these models are trained on large sizes of data, which often include datum from the past many decades and thus the ideas that were present then, they often show signs of stereotypical bias. There have been many attempts to try to show the prevalence of these kinds of biases. We aim to explore the robustness of LLMs in fighting stereotypical bias. In more specific terms, we strive to determine whether models are biased against differing formulations of given stereotypical ideas.

5 Prior Work

- The StereoSet dataset to measure stereotypical bias [1]

6 Methodology

We first generate prompts using a LLM to get a diverse set of prompts similar to the ones used in the paper, ranging across similar topics. For each prompt generated, we also generate options/answers using both the original model and a newer model like LLaMA for comparison. We then use bias detection techniques, such as the ones used in the paper (ICAT score). We then compare the performances of the original model and the newer models to make our assessment.

7 Research Questions

Are new LLMs more robust to stereotypical biases? Are models in general robust in fighting stereotypes against different formulations of prompts which convey the same biased meaning?

8 Evaluation Setup

For evaluation, we will apply the steps mentioned above. As used in the StereoSet paper, we will use the ICAT score in order to provide a fair comparison against the original paper. The ICAT score is a combination of the Language Modeling Score (lms) and the Stereotype Score (ss) of the following form: $icat = lms * \frac{min(ss, 100-ss)}{50}$.

References

[1] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 2020.