Application No. 2540/25
PI1 Name: Adir Solomon

Scientific abstract – *Strategies for Detecting and Mitigating Bias in Large Language Models*

In recent years, the widespread adoption of Large Language Models (LLMs) has revolutionized natural language processing, offering unprecedented capabilities across diverse applications. However, this rapid growth has surfaced critical concerns regarding the fairness and equity of these models' outputs. Ensuring fairness in LLMs involves designing, training, and deploying models that prevent biased outcomes, thus promoting equitable treatment across all users and demographic groups.

Recent studies have demonstrated that LLMs can exhibit fairness issues across various demographic categories, including gender, age, sexual orientation, ethnicity, and religion, leading to broader societal implications and potentially harmful consequences. For instance, when presented with gender-specific prompts, we observed that the LLM 'Gemma' encouraged males to pursue fields like mathematics or engineering while directing females towards communication or business administration. This highlights the unequal treatment embedded within the model and underscores the urgent need for rigorous testing and debiasing to address these deep-seated biases.

Addressing fairness in LLMs is a complex, multifaceted challenge that requires both the quantitative measurement and qualitative exploration of disparities to ensure just outcomes. Unlike traditional machine learning systems that might utilize straightforward datasets or algorithmic adjustments, LLMs demand a nuanced understanding of the specific harms they may propagate and the broader social contexts they reflect. This complexity is compounded by LLMs' tendency to replicate statistical patterns from extensive internet text datasets, which can perpetuate inherent stereotypes.

In this proposal, we present strategies using word embeddings, machine learning classifiers, and reinforcement learning to detect and mitigate bias in LLMs. We also explore qualitative approaches, like the "poetics of prompting," to analyze how prompts influence LLM behavior and reflect societal biases. These methods help uncover stereotypes embedded in LLMs, addressing hidden biases that, if ignored, could reinforce harmful stereotypes. A systematic approach to bias detection and mitigation is essential to prevent LLMs from amplifying societal disparities and to ensure they contribute to fairness and equity.