Title: Auditing Diversity in LLM (Language Learning Models) and TTI (Text-to-Image) Systems

Abstract:

This research project focuses on developing and enhancing an AI auditing system to assess diversity and fairness in large language modeling (LLMs) systems. By replicating an existing Python-based audit framework, originally created by my Principal Investigator (PI), this study extends its functionality to specifically evaluate how race and ethnicity are represented in AI-generated outputs related to professional occupations. The enhanced auditing system cross-references race and ethnicity data with job positions to identify potential biases, providing a deeper understanding of whether AI systems (specifically GPT-4) disproportionately associate certain ethnic groups with specific professions. These findings contribute to the ongoing discourse on fairness in AI, offering insights into how LLM models may perpetuate or mitigate biases in career representation. This research is critical for the development of more equitable AI systems that reflect diversity across various social and professional contexts, highlighting the importance of fairness in the deployment and usage of AI technology.

Introduction:

As our world continues to advance technologically, Text-to-image (TTI) systems are becoming increasingly present in tech applications. However, the rise in its presence has highlighted the need to check for inherent biases in these systems, ensuring that they are representative of the populations here in the states in which they are meant to serve. This project aims to reveal biases and representation gaps in AI models, specifically in the intersection of image creations and GPT-4. By examining GPT-4s outputs, we aim to encourage and inform discussions about TTI systems in AI and suggest methods and future improvements for equitable modeling practices.

Methods:

Tools and Libraries: Using various libraries like pandas and OpenAI's API, this analysis used Python to prompt the models over the two prompts and analyze their responses.

Scoring: A five-point scale rated the representation quality of each ethnicity using the identity and profession prompts. These scores were used to detect potential biases in the generated images, focusing on underrepresentation and stereotypes[1]. The scores were then averaged to find an overall score for both identity and profession.

Prompt Design: 2 specific, separate prompts were designed to produce images of working professionals from diverse ethnic backgrounds in diverse professions. Using a list of professions garnered from the

Results:

The tested TTI model (GPT-4) consistently favored majority populations in professional roles, with significant underrepresentation for groups such as Native American and Pacific Islanders. In cases where profession-specific prompts were used, the results overwhelmingly depicted White individuals, regardless of the specified ethnicity. The model's overall diversity score averaged 2 out of 5, with the majority of the results biased towards White people. In the profession-based prompts where the roles were deemed more "traditionally masculine" like truck driver or manager, the results were also overwhelmingly male, whereas roles like "office assistant" were female-dominated.

Discussion and Future Works:

These results suggest a need for enhanced diversity in AI training datasets to mitigate biases. Expanding the sample data and rigorously training the TTI models with diverse data can ensure equitable representation across ethnic groups. This framework also offers a reproducible

metric for future evaluations, which can assist in standardizing diversity benchmarks in AI. It also demonstrates the importance of representation in AI-generated content and media. In the future, this work could be expanded by increasing the dataset used, increasing testing, and comparing and contrasting results compared to another AI system outside of GPT-4.

Acknowledgments:

This research was funded by the Computing Research Association's (CRA) DREU program. I extend my gratitude to my mentor, Tanu Mitra, an Assistant Professor at the UW Information School, whose guidance was invaluable throughout this project.

References:

[1] Luccioni, A. S., Mitchell, M., & Akiki, C. (2023). Stable Bias: Evaluating Societal Representations in Diffusion Models. Hugging Face, Canada; Leipzig University and ScaDS. AI.