Thesis Project Portfolio

Prompt Engineering to Evaluate Economic and Educational Stereotypes in Large Language Models

(Technical Report)

Data Collection and Personal Privacy: Balance Between Big Tech and Public Policy

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Barbara (Bebe) Holloway

Spring, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Prompt Engineering to Evaluate Economic and Educational Stereotypes in Large Language Models

Large Language Models (LLMs) are becoming more prevalent in society and are already beginning to replace human-generated texts. Specifically, LLM persona generation, descriptions of varying individuals, will be used at an increasing rate, so it is important to understand the biases and concerns attached to them.

To evaluate this issue, I wrote code to generate responses from ChatGPT-4.0 through the ChatGPT API. Utilizing 6 different prompts and generating personas for 15 different identities (the combination of 3 different genders with 5 different jobs/education backgrounds), I stored the responses of ChatGPT and analyzed the frequency of repeated words for each varying persona. The socio-economic and gender stereotypes in these LLM-generated personas are evident from the frequent word data frames created. Specifically, I found there to be socio-economic bias between doctor and truck driver, as both of these personas contained multiple physical stereotypical descriptors. I also discovered gender stereotypes in the doctor persona specifically; female doctors were often described with words that symbolize caretakers while male doctors are described with words like “knowledge”.

There is more research to be done in this area as LLMs are increasing in popularity and we depend more on their text-generated content. It is necessary that we identify and mitigate the harms from LLMs as they become more prevalent in society. While generated personas constitute a small portion of this, it is essential to study them due to their impact on story generation and other applications.

Data Collection and Personal Privacy: Balance Between Big Tech and Public Policy

Since the early 2000s, Google has risen as a global superpower as a technology and data conglomerate. Google has made clear efforts to collect as much data as possible on their consumers, and legislation has struggled to regulate their practices. Throughout the years, Google has been involved in multiple court cases concerning misuse of personal data and misinformation regarding their collection of data.

I employ the case study method to analyze two court cases: Calhoun vs. Google and Google Inc. Cookie Placement Consumer Privacy Litigation v. William Gourley. In the case Calhoun vs. Google, the court discovered that Google was incorrectly disclosing the use of consumer data. Google claimed to only utilize consumer data to sell to advertisers, but they instead were using consumer data for a myriad of purposes. They attempted to use broad, vague, general disclosures to put an umbrella over consumer data collection, use, and disposal. However, the court decided that Google is not allowed to do that; they can no longer use general disclosures to shield themselves from any liability regarding consumer data. Google Inc. Cookie Placement Consumer Privacy Litigation v. William Gourley evaluates the legality of the collection of cookies from users' devices despite third-party browser cookie restrictions. The main takeaway from this case is that Google must inform users when they are being tracked because users have the right to informed consent. Both of these cases are fundamental to understanding how Google has set precedent in the courts and influenced data privacy policy.

I also utilize the public policy method, an STS framework, to further dissect Google's public policy impact. STS reveals four main indirect influences on public policy, and I apply each to Google. It is beneficial to first look into Google's origin; they began as a search engine and have grown through the acquisition of many companies to widen their portfolio. This has

given them an opportunity to collect enormous amounts of data from their consumers across all platforms. There is minimal current policy in place regarding data collection and privacy. There are two significant laws in Virginia and California that lay out individual privacy protections for citizens involving their data, but besides that, there are not many protections for citizens in other states or at the national level. The EU and US are comparable in their efforts to control and regulate data use and privacy protections for their citizens; however, there are more concrete protections in the EU. The EU directive includes the General Data Protection Regulation (GDPR) which gives consumers rights to control their data use, and these same rights are not protected anywhere in the US besides a few states. Contrarily, China does not have any citizen protections regarding data privacy as the limits on private company data collection are an element of the authoritarian government. Chinese laws limiting data use are in an effort to increase the party's control rather than protect citizens. It is important to look to the superpowers of the world to evaluate the US stance on data protection. Lastly, data companies are influencing society unlike industry ever before. Big technology companies have influence over governments, legislation, and private industry, and the rate at which technology is progressing is impossible for legislation to keep up with.

Google has directly impacted the data privacy standards in the United States over the past 20 years due to their past privacy indiscretions. The process of lawmaking and the reality of politics in the US is incomparable to the pace that technology evolves. This puts consumers in a vulnerable position as their personal data is nearly consistently at risk and regulation cannot protect them. The US needs to proactively work to understand technology evolution and citizens' risks, as well as work with big technology companies to provide ethical and reasonable restrictions of technology. The US should have comprehensive laws giving citizens the right to

consent, delete, and limit the data that companies have access to. Citizens should have the ability to opt-in to data collection instead of this being assumed. Consumers should also have access to the data that companies have collected on them, and companies should not be able to discriminate against those who opt-out of data collection practices. These are some basic protections that should be expressed in legislation at a federal level.