16:960:588 Data Mining Spring 2024 Project

Mayukh Sen (ms3802), Rohit Vernekar (rv524) & Shubha Ugre Gowda (sg2052)

Problem Description

For the Final Project, our group will conduct experiments based on Jones, E. et al. (2023) Automatically Auditing Large Language Models via Discrete Optimization - ICML 2023. We will mimic the experiments conducted in this paper and conduct some further empirical analysis.

Auditing will be used as an optimization problem. LLMs will be automatically searched for input-output pairs. We audit models by specifying and solving a discrete optimization problem. Specifically, we search for a prompt x and output o with a high auditing objective value, φ(x, o), such that o is the greedy completion of x under the LLM. Solving this optimization problem is computationally challenging: the set of prompts that produce a behavior is sparse, the space is discrete, and the language model itself is non-linear and high-dimensional.
The image contains a snippet of text which reads as follows:

We capture this criterion with an auditing objective $\varphi : P \times O \rightarrow \mathrm{R}$ that maps prompt-output pairs to a score. This abstraction encompasses a variety of behaviors:

- **Generating a specific suffix** : $\varphi(x, o) = 1[o = o^*]$.
- **Derogatory comments about celebrities:** $\varphi(x, o)$ = StartsWith(x, [celebrity]) + NotToxic(x) + Toxic(o, x).
- **Language switching:** φ(x, o) = French(x) + English(o)

To solve the optimization problem, we will implement the algorithm ARCA. It is a Coordinate Ascent Algorithm. ARCA will be compared to AutoPrompt [Shin et al., 2020] and GBDA [Guo et al., 2021]. We aim to establish empirically that ARCA consistently produces more prompt-output pairs of target behavior when compared to state-of-the-art auditing algorithms.

Experiment Setup: All experiments in the paper have been performed on the 762M-parameter GPT-2-large and 6B-parameter GPT-J hosted on HuggingFace. The CivilComments dataset on HuggingFace has been scraped to reverse LLMs and detect toxic comments. We will be extending the domain of experiments and performing experiments on other LLMs like **Mistral, LLaMA, GPT-3, GPT-4, and Gemini** and use some other datasets for detecting toxic comments. We will try to empirically establish the superiority of ARCA in auditing Large Language Models when compared to other methods.