Jailbreak GPT Project Proposal

Tokey Tahmid, Manas Tiwari, Tushar Krishna Panumatcha, and Shui

***Abstract*— Our aim is to do analysis on how people are using Large Language Models (LLMS) like ChatGPT. Specially we are looking into ethical use of these LLMs. We will be working on the "In-The-Wild Jailbreak Prompts on LLMs" datasets which contains 6,387 prompts from four platforms (Reddit, Discord, websites, and open-source datasets) during Dec 2022 to May 2023. Among these prompts, 666 prompts are jailbreak prompts. These are mainly prompts that go against the terms of service of the LLM vendors. So, we will analyze this data and provide actionable recommendations for safeguarding the use of these LLMs.**

I. OBJECTIVE

The aim of this research effort is to perform an in-depth analysis of how Large Language Models (LLMs) like ChatGPT are being utilized, focusing on ethical implications. To achieve this we would perform a detailed investigation of jailbreak prompts in the context of large language models (LLMs). The purpose of the study is to use the "In-The-Wild Jailbreak Prompts on LLMs" dataset to understand the characteristics, evolution, and potential harm of jailbreak prompts that go against the terms of service of LLM vendors. By doing this analysis we would attempt to define the distinguishing traits and attack strategies employed by jailbreak prompts, assess their prevalence in actual situations, determine how well they circumvent security measures, and assess the efficacy of current defending strategies. By providing insight into the shifting threat environment of jailbreak prompts, this work hopes to contribute to the development of more trustworthy and regulated LLMs.

II. MOTIVATION

With the sudden boom of LLMs like ChatGPT, we found it really interesting that not everyone is using these tools ethically or as intended. When the dataset, "In-The-Wild Jailbreak Prompts on LLMs" came to light, we understood that there are many people who are using these adversarial or "Jailbreak Prompts" to get beyond software restrictions and safeguards. This kind of misuse of LLMs for generating harmful or unethical content poses significant challenges for vendors, policymakers, and users. Both hobbyists and the law enforcement community are interested in this technique because it raises concerns about device integrity, user privacy, and security. Jailbreaking covers many aspects of technology, including an examination of software faults and the limits of intellectual property legislation. It also mandates security evaluations to identify the risks posed by altered software environments.

The potential for jailbreaking can be used to develop dangerous agents like viruses, hate speech, and etc is of the utmost concern. So there is an urgent concern to ease with which information and methods even by those with very basic technological skills, for the creation of pandemic-class agents. This accessibility offers a possible incentive for those looking to employ dual-use technologies maliciously.

This project aims to address these difficulties through the transdisciplinary integration of ideas from cybersecurity, legal research, and computer science. We seek to offer a fair picture of the developing jailbreaking situation through thorough study and analysis. This study paves the way for more informed debates on this divisive and pervasive topic.

III. DATASET - JAILBREAK LLMS

A. Jailbreak Prompt

The term "Jailbreak Prompts" refers to the deliberate development of adversarial prompts or inputs aimed at circumventing the Large Language model's built-in safety features and restrictions. This requires developing prompts that coerce the model into generating output that might be harmful, unsuitable, or contrary to the intended usage restrictions set by the model's developers or overseeing bodies.

Finding imperfections or weaknesses in the model's response-generating process is equivalent to "jailbreaking" in this context since it allows the model to generate outputs that it was intended to prevent. The model's capacity to respond in ways that might not be consistent with the technology's intended ethical, legal, or responsible use is exploited by using these jailbreak prompts.

B. Data Collection

The authors of the "JAILBREAK LLMS" dataset employed a multi-source data collection strategy to create a comprehensive dataset on jailbreak prompts. The data sources span four popular platforms: Reddit, Discord, specific websites, and open-source datasets. Each of these platforms was selected for its prominence in sharing prompts for LLMs.

1) Reddit: The authors used the Pushshift Reddit API to collect a total of 80,746 submissions from three subreddits known for sharing ChatGPT's prompts. The submissions were scanned for specific flairs to isolate those that were potentially jailbreak prompts. Regular expressions were employed to extract these prompts.

2) Discord: Discord serves as another rich data source where the authors manually inspected 20 servers and identified six with dedicated channels for collecting jailbreak prompts. Posts tagged with "Jailbreak" and "Bypass" were considered as potential jailbreak prompts and were manually reviewed.

3) External Websites: Three websites, namely AIPRM, FlowGPT, and JailbreakChat, were scoured for jailbreak prompts. Criteria such as the presence of the term "jailbreak" in the title or description were used to classify a prompt as a jailbreak prompt.

4) Open-Source Datasets: Two open-source datasets were integrated into the study. Prompts from these datasets were manually inspected to identify those that could be categorized as jailbreak prompts.

C. Statistical Overview

In total, the authors amassed a dataset of 6,387 prompts collected from December 2022 to May 2023. Of these, 666 were identified as jailbreak prompts. This dataset stands as the most exhaustive in-the-wild prompt dataset for ChatGPT to date. To bolster the integrity of their dataset, the authors performed human verification by sampling 200 prompts from both the regular and jailbreak prompt categories. The high Fleiss' Kappa score (0.925) among the labelers attests to the dataset's reliability.

D. LLMs Used in Dataset

The dataset serves as a foundation for evaluating the effectiveness of jailbreak prompts across five different LLMs, including GPT-3.5, GPT-4, ChatGLM, Dolly, and Vicuna. Each LLM was evaluated for its vulnerability to jailbreak prompts, marking a significant contribution to the understanding of LLM security.

IV. Project Outline and Responsibilities

A. Outline Overview

Below is an outline of the analysis methodology that will be followed during the project. Based on this outline we plan to divide the responsibilities among the four group members.

1. Data Preprocessing: In this process, data is transformed into a format better suited for analysis or modelling. Missing values are a common occurrence in raw data, which can provide problems for analysis or modelling. Mathematical strategies like mean, median and statistical analysis might be used.

2. Feature Extraction: In this phase, part-of-speech tagging (POS tagging), tokenization, regular expressions, grammar, and rule-based analysis are all strategically applied to identify and clarify the toxicity and semantics of jailbreak and non-jailbreak prompts.

3. Community Detection: Employ graph-based community detection methods to identify major communities sharing jailbreak prompts.

4. Temporal Analysis: Conduct a temporal study of the dataset to understand how jailbreak prompts have evolved over time.

5. Effectiveness Analysis: Test the efficacy of jailbreak prompts against current LLMs and safeguard mechanisms, such as OpenAI's moderation endpoint.

6. Formulate Recommendation: Based on the analysis, formulate actionable recommendations for LLM vendors to improve their safety mechanisms.

7. Report Writing: Write the Final report for the project.

8. Presentation: Prepare a presentation for the project.

B. Member Responsibilities

We will be dividing our group into two parts and tackle the steps in pairs. For example, two people will be working on Data Pre-processing and Feature Extraction steps together. The other pair will be working on Community Detection and Temporal Analysis. Then one pair will work on Effectiveness Analysis and the other pair will work on Formulating Recommendations. Each and every person in the group will be contributing to the final report simultaneously while working on their individual parts. Finally, we plan to sit together and finalize the report as well as make the presentation to present our project in class.

V. Timeline & Milestones

Week 1-2: Data Preprocessing. Week 2-3: Feature Extraction. Week 3-4: Community Detection. Week 4-5: Temporal Analysis. Week 5-6: Effectiveness Analysis. Week 6-7: Compilation of findings and development of recommendations. Week 8: Submission of the Final Report.

VI. Expected Outcome

A. Insights into Unethical Use of LLMs

Upon the completion of this research, we anticipate obtaining a comprehensive understanding of the scale, strategies, and impact of unethical practices in the use of Large Language Models (LLMs). We expect to identify the most common categories of jailbreak prompts, their potential for harm, and the loopholes that facilitate these activities.

B. Security Vulnerability Assessment

By testing the identified jailbreak prompts against multiple LLMs, we aim to produce a ranked vulnerability assessment. This will offer LLM vendors a direct insight into which models are more susceptible to manipulative inputs and require more robust safeguarding measures.

C. Temporal Trends

Our temporal analysis is expected to reveal how the nature and efficacy of jailbreak prompts have evolved over the specified period (December 2022 to May 2023). This will aid in the predictive modeling of future threats.

D. Actionable Recommendations

Based on our analysis, we will offer concrete, actionable recommendations to LLM vendors, policy-makers, and other stakeholders. These will provide a roadmap for enhancing the ethical use of LLMs and fortifying them against adversarial attacks.

E. Community Behavior

Through community detection methods, we anticipate understanding the behavioral patterns of users who commonly share jailbreak prompts. This knowledge will be vital for both understanding the motives behind such activities and for planning effective deterrents.