Jailbreak GPT: An Analysis of In-The-Wild Jailbreak Prompts on LLMs

Tokey Tahmid, Manas Tiwari, Tushar Krishna Panumatcha, and Andy Zeng

Abstract— **Our aim is to do an analysis on how people are using Large Language Models (LLMS) like ChatGPT. Specially we are looking into the 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) from 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. Introduction

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.

Platform	Source	Posts	Prompts	Jailbreaks	Access Date
Reddit	r/ChatGPT	79436	108	108	2023-04-30
Reddit	r/ChatGPTPromptGenius	854	314	24	2023-04-30
Reddit	r/ChatGPTJailbreak	456	73	73	2023-04-30
Discord	ChatGPT	393	363	126	2023-04-30
Discord	ChatGPT Prompt Engineering	240	211	47	2023-04-30
Discord	Spreadsheet Warriors	63	54	54	2023-04-30
Discord	AI Prompt Sharing	25	24	17	2023-04-30
Discord	LLM Promptwriting	78	75	34	2023-04-30
Discord	BreakGPT	19	17	17	2023-04-30
Website	AIPRM	-	3385	20	2023-05-07
Website	FlowGPT	-	1472	66	2023-05-07
Website	JailbreakChat	-	78	78	2023-04-30
Dataset	Awesome ChatGPT Prompts	-	163	2	2023-04-30
Dataset	OCR-Prompts	-	50	0	2023-04-30
Total		81564	6387	666	

TABLE I
JAILBREAK PROMPTS DATA SOURCES

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. Methodology

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 modeling. Missing values are a common occurrence in raw data, which can provide problems for analysis or modeling. 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. Categorizing Prompts: We categorized the prompts into multiple jailbreak categories such as Illegal Activity, Hate Speech, Cybercrime, Sexual Content, Harmful Content, and Fraud.

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. Visualize and Discuss Results: Based on the analysis, we intend to visualize our results in multiple graphs and discuss the impact of these results to find actionable recommendations.

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

8. Presentation: Prepare a presentation for the project.

B. ChatGPT as DAN

DAN stands for "Do anything now", which offers chatgpt enhanced capabilities and versality. DAN's are designed to handle a wide range of tasks, showcasing their adaptibility and responsiveness. When compared to traditional AI models that do not have an answer to a query or have been restricted access by the developer, DANs are programmed to generate responses regardless of the complexity or intent of the question.

One significant advancement in DANs is their ability to browse the internet and access verified information. This allows DANs to provide up-to-date and accurate responses. By integrating internet browsing capabilities DANS can pull in real-time data, making sure that the information is current and relevant. Some of the examples for this can be to get latest news updates, fetch real-time stock market data or get information about latest findings from scientific journals. This make it far better compared to the original chatGPT, which relied on fixed dataset and could not provide most upto date information.

C. Feature Extraction and Text Data Analysis

For this analysis, we implemented a Python script that enhances our understanding of user interactions with large language models by identifying and quantifying specific keywords indicative of attempts to circumvent the models' restrictions. Initially, we developed the count_words function to parse a text file and tally occurrences of selected words, such as 'alias', 'act', 'pretend', 'character', and 'role', which are essential to recognizing prompts that may suggest role-playing or identity manipulation. This function is especially crucial as these types of prompts often signal efforts to 'jailbreak' or exploit the model.

We focused our analysis on lines beginning with "Message sent:", which likely represents the prompts users are sending to the model, offering us a clear window into the actual interactions. Post-analysis, we sorted the keywords by their frequency to ascertain the most prevalent terms.

Subsequently, we visualized our findings in a bar graph, carefully designed for optimal readability and informative value. Each bar, colored distinctively and edged in black, corresponds to a specific action word and its frequency in the text data. We ensured each bar was labeled with its value, thereby facilitating immediate comprehension of the data represented.

To further aid in clarity, especially where the

text labels were lengthy, we rotated the x-axis labels. Before presenting the graph, we adjusted the layout to prevent any overlap or clipping of content, preserving the professional presentation of our findings.

D. Temporal Analysis

For Temporal Analysis, we devised a rigorous analytical approach to examine the nature and variation of user prompts over time. This analysis was also facilitated by a Python script that, first and foremost, categorizes each prompt according to its content. To achieve this, we defined a categorization function, categorize_prompt, which scrutinizes the textual content of prompts for keywords associated with distinct themes such as 'Sexual Content', 'Illegal/Unethical Activities', 'Hate Speech', 'Cybercrime', 'Harmful Content', and 'Fraud'. This function operates on the premise that the presence of certain words within a prompt is indicative of the underlying intent or subject matter.

Upon categorization, we employed the script to process a "jailbreak_prompts.csv" file containing a collection of jailbreak prompts in our dataset. This file, pivotal to our project, was parsed to label each prompt with a relevant category. The categorization process was meticulously designed to be case-insensitive and comprehensive, ensuring that the variations in text case did not skew our analysis.

Once categorized, the data underwent temporal analysis. We grouped the prompts by their timestamp and category, enabling us to observe the change in the frequency of each category over time. This was visualized using Seaborn and Matplotlib libraries, creating a multi-line graph that provides a clear, longitudinal view of the data. Each category was represented by a line, color-coded for distinction, which traced the number of occurrences through the observed period.

The graph we produced is not only a visual representation but also an analytical tool that provides insights into trends, peaks, and troughs in the usage of prompts. We could infer the relative prevalence of various categories and identify any temporal patterns or anomalies. For instance, a surge in 'Hate Speech' or 'Cybercrime' related prompts at certain intervals could suggest external influences or shifts in user behavior.

In essence, our analysis empowers us with empirical evidence of how users engage with language models. By visualizing the distribution and evolution of categorized prompts over time, we gain a nuanced understanding of the ethical landscape in which these AI models operate. This is crucial for the ongoing development of strategies aimed at mitigating the misuse of language models and ensuring their alignment with ethical standards.

E. Effectiveness Analysis

In this phase, we progressed to analyze the data where we tested these prompts on Discord ChatGPT and saved the resulting conversation to a text file, called "GPT Break - chat-gpt.txt". We analyzed this text file by similarly categorizing and quantifying the frequency of various types of prompts as the Temporal Analysis. This initiative was driven by the goal of better understanding the nuances of prompts that users employ, especially those that may challenge the ethical boundaries of AI interactions.

The process was done similarly with the creation of a categorize_prompt function that diligently evaluates each line of text against a set of criteria defined by a comprehensive list of keywords. These keywords were the same as before to represent the same categories such as 'Sexual Content', 'Illegal/Unethical Activities', 'Hate Speech', 'Cybercrime', 'Harmful Content', and 'Fraud'. This categorization is pivotal, as it enables us to dissect the corpus of data into segments that reflect the same category of prompts being still effective against ChatGPT.

We then applied this function to a text file, meticulously scanning through each line to categorize it accordingly. This step was critical, for it allowed us to compile a tally of the frequency with which each category appeared in the text — a quantitative measure obtained using the Python Counter class.

Upon collecting the data, we chose to represent our findings visually. To this end, we transformed the counter data into a DataFrame to leverage the

graphical capabilities of Seaborn and Matplotlib. We designed a bar plot that clearly delineates the frequency of each category, furnishing it with a title that encapsulates the essence of our analysis: 'Effective Jailbreak Prompts Tested on ChatGPT'. The choice of a bar plot was deliberate, as it offers a straightforward, yet powerful, visualization of categorical data distribution, with each bar's height corresponding to the occurrence of each category.

V. TIMELINE & MILESTONES

Data Preprocessing: For the first two weeks, we pre-processed the data to make it suitable for the analysis. Feature Extraction: In the second to third weeks, we extracted important features from the dataset that would enable us to categorize the prompts. Categorizing Prompts: After extracting the necessary features from the pre-processed dataset, we categorized the prompts into several jailbreak categories during weeks four and five. Temporal Analysis: After categorizing the prompts, we finally started analyzing the data. We simultaneously tested the prompts for effectiveness and performed temporal analysis to see how these prompts evolved over a course of six months. This was done during weeks five to seven. Effectiveness Analysis: In this stage, we tested the jailbreak prompts on Discord ChatGPT and then analyzed the responses to show the effectiveness of these prompts on the current date. This analysis was ongoing from the beginning and completed within time during weeks seven to eight marking the completion of our analysis for the project. Preparing for Presentation: In the final weeks, we prepared for the presentation. We held multiple meetings to do dry runs on our presentation. Report Submission: We filled in the contents of the report as we were preparing for the presentation. And after the presentation was complete, we finalized our report for submission.

VI. RESULTS

This chapter contains our results for the different analysis and a brief discussion on the findings of each analysis.

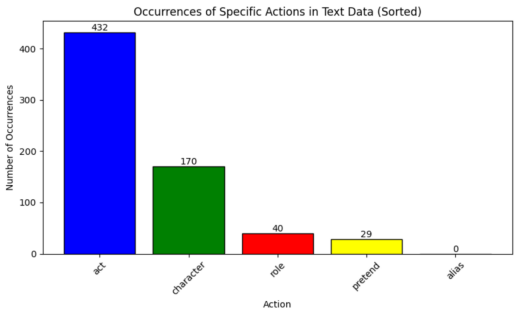

Fig. 1. Enter Caption

A. Features of Prompts in Dataset

The figure 3 is a bar graph that shows how frequently different kinds of prompts are sent to ChatGPT when it is operating in its advanced mode, which is called Do Anything Now (DAN). With the most frequent prompts on the left and the least frequent prompts on the right, the graph is arranged in descending order of frequency. The most common prompt is "act," indicating that users interact with DAN primarily when performing or taking action is required. This could include a wide range of duties, such as answering questions and producing original material.

The most commonly occurring prompts after "act" are "character" and "role," suggesting that users often engage with DAN in a role-playing manner. This can entail DAN taking on the persona or acting out a certain role of a particular character. The two least common prompts, "pretend" and "bias," ask DAN to act in a biased or pretend-like manner. These prompts are less prevalent, maybe because pretending might lead to miscommunication or misunderstanding and there are ethical issues related to bias. This graph sheds light on user behavior and preferences in AI interactions and provides insightful information about the nature of user interactions with chatGPT when it functions as a DAN.

B. Temporal Analysis

First, we categorized the prompts into multiple jailbreak categories such as: Illegal Activity, Hate

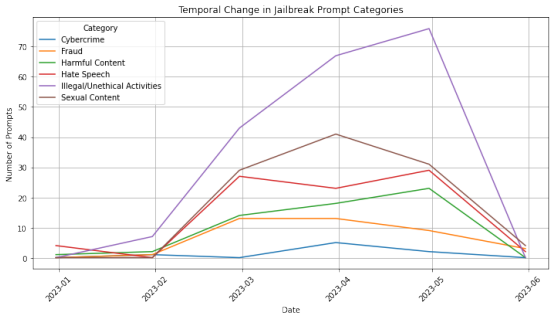

Fig. 2. Temporal Analysis

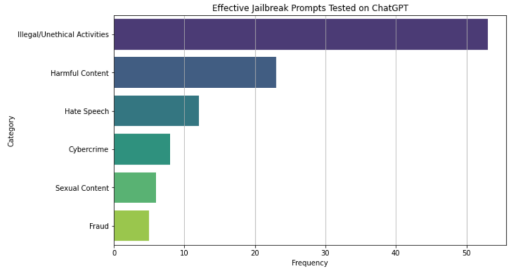

Fig. 3. Prompt frequency analysis

Speech, Cybercrime, Harmful Content, Fraud, Sexual Content. Then we analyzed to see how these prompts were used over a course of six months from January 2023 to June 2023. We plotted the results to visualize the number of prompts which were effective in each category in each month.

From the analysis in, we found out that most jailbreak prompts (over 70) that people used involved illegal activities where the GPTs are told to be able to perform any illegal/immoral/unethical activity. Then not surprisingly, the next most used category is sexual content and the other categories are more or less in the same range. We can also see that these prompts were heavily effective in the months of April and May which shows that during this time most people figured out how to jailbreak the GPTs effectively. After that there is a steep decline in the effectiveness of these promts from which it can be inferred that OpenAI had pushed a patch to safeguard against these jailbreak prompts.

C. Prompt Effectiveness Analysis

Figure 2 presents an analysis of effective jailbreak prompts tested on ChatGPT. The graph categorizes the prompts into six categories: Illegal/Unethical Activities, Harmful Content, Hate Speech, Cybercrime, Sexual Content, and Fraud. The frequency of each category is represented on the x-axis. The category with the highest frequency is Illegal/Unethical Activities, followed by Harmful Content and Hate Speech. The bars are colored in shades of blue, green, and purple, providing a visual distinction between the categories. This graph provides a quantitative overview of the types of prompts that have been effective in testing the robustness of ChatGPT.

The next two categories, Harmful Content and Hate Speech, also represent challenging areas for AI systems, as they involve sensitive content that needs to be identified and managed appropriately. The graph underscores the importance of rigorous testing in diverse areas to ensure the safe and responsible use of AI systems like ChatGPT.

VII. Limitations & Future Work

A. Limitations

Our study encountered several limitations that must be acknowledged. Firstly, the sheer volume of prompts processed could potentially trigger the security and safety mechanisms inherent in large language models (LLMs), leading to flagged content and possibly influencing the models' responses. This limitation is significant as it may introduce a bias in the data, where the frequency and nature of prompts are artificially constrained by the models' protective algorithms. Furthermore, LLMs vary not only in their architecture but also in their limits regarding the number of input characters they can process. This variability presents a challenge in standardizing the prompt analysis across different models, as the prompts suitable for one model may need to be adapted or truncated for another, potentially skewing comparative analyses.

B. Future Work

Looking ahead, we propose several avenues for future work to extend the findings of this project. To

refine our understanding of LLM interactions, we aim to design and test a broader array of prompts that are uniquely tailored to different LLMs, taking into account their specific input constraints and behavioral nuances. This targeted approach will allow us to probe the models' capabilities and limitations more effectively.

Additionally, we plan to explore the models' resistance to performing tasks they are designed to avoid. By systematically and persistently presenting prompts that the LLMs are programmed to resist, we can gain deeper insights into the robustness of their ethical safeguards and the potential for users to find workarounds.

Lastly, as the field of AI continues to advance rapidly, we anticipate the release of new LLMs. Keeping abreast of these developments, we will extend our testing to include these emerging models. By doing so, we can maintain a current and comprehensive perspective on the evolving landscape of LLMs, ensuring that our strategies for promoting ethical AI use remain relevant and effective.

Through addressing these limitations and pursuing the outlined future work, we aim to contribute to the responsible development and deployment of LLMs, ensuring they serve the interests of users ethically and securely.

C. Conclusion

In conclusion, the Jailbreak GPT project has not only advanced our comprehension of user-LLM dynamics but has also laid the groundwork for ongoing research aimed at safeguarding the integrity and ethical use of these powerful models. As LLMs become increasingly widespread, their capabilities are expected to grow significantly. Currently, many of these LLMs feature robust mechanisms for detecting attempts at jailbreaking. However, the effectiveness of these safeguards is poised to improve even further as they become more widely adopted and refined. As newer LLMs come out they would also need more robust testing through these jailbreak prompts. In this regard, our project remains a testament to the importance of rigorous, data-driven approaches to ensure these technologies are leveraged for the greater good, without compromising on ethical standards.