Collaboration Statement
Jailbreaks: Can chatbots be manipulated to produce harmful content?

Benjamin Vleugel, Shumail Sajjad, Colin Chen, Claire Hegemann

Ben

-Researched jailbreaks generally and read academic articles on chatbot jailbreaks and emerging trends.
-I worked on categoizing jailbreak types with Claire and then went to find examples of each type of jailbreak.
-Having read lots, I came up with opinions on issues relating to jailbreaks that would be worth discussing (like emerging online communities, small chatbot vulnerabilities, prevention measures), and used these to formulate policy recommendations, in coordination with my fellow group members.
-I also prepared slides and our presentation sections on the characteristics of jailbreaks, the vulnerabilities they pose, examples of such vulnerabilities being exploited, and mechanisms to prevent them.

-I also helped work on some of the technical research and took part in efforts to jailbreak chatbots myself, to provide useful examples for the presentation and fine tune our understanding of vulnerabilities that would be worth discussing

-Replied to several class questions

-Finally, I played a coordinating role with the group, helping build our structure, frame out work, schedule meetings…etc.

Shumail

- I worked with Colin on the dataset part, reviewing the current literature and any publicly available datasets that could be useful for our assignment. My main goal was to reproduce the results from the literature and introduce novelty in the data part by building our own dataset. For this purpose, I used Taskade AI (online prompt generator) to generate prompts across 6 subjects and tested them for jailbreak using OpenAI, Meta, and Google LLMs.

- I further used language/ sentence classification techniques to figure out if a response is a jailbreak or not. In the end, I filtered the final dataset to create some visual graphics to show the success of jailbreaks across each LLM and highlight which prompts are more likely to lead to jailbreak versus the others and what makes them ‘different’ or ‘special’ in our context.

- Additionally, to make our presentation more interactive, I worked on finding ideas for a live demonstration that we could show the audience. The live demonstration through DAN by

articulating the DAN prompt and pre-defining different prompts to test the vulnerabilities of GPT-3.5 model is something I also worked on with Colin.

Colin

- I'm responsible for finding and generating datasets. Over the past two weeks, I've been exploring how to generate the desired data using the OpenAI API key. However, I encountered difficulties in the early stages of data generation. Later on, after finding suitable papers and datasets, I generated the final experimental sample dataset by randomly selecting 100 prompts, 10 questions, and 3 models, which we then analyzed in the data section based on methodology from the paper "Determining Jailbreak Status Based on the Presence of Negation in ChatGPT Responses" to generate data points for "whether it is jailbreak or not."
- In the data analysis section, I conducted the following research: (Github link:https://github.com/Colin-chan1366/Jailbreak_Analysis?files=)
    - 1. A comparison of high-frequency words in prompts with the highest and lowest success rates of jailbreaking.
    - 2. The correlation between jailbreak prompts and the field of public policy.
    - 3. The relationship between jailbreak prompts and prompt length.
    - 4. The relationship between jailbreak prompts and gpt models.
    - 4. The relationship between jailbreak prompts and prompt semantics.
    - 5. The relationship between jailbreak prompts and prompt toxicity.
- A binary classification model for simple jailbreak responses.
- Future Promotion:
    - The first point is in the data set. If there are enough conditions to go through all prompts and all questions based on 3 large language models, not only the accuracy in jailbreak identification but also the issue of data robustness may be solved which is shown in the prompt analysis section.
    - In addition, our methodology to determine a prompt jailbreak success is too naive and simple, which According to the current literature updated from ucb, "Sorry, I can't" is categorized as the "Full Refusal" class. Other features or categories like“Partial Refusal””, " Partial Compliance” and “Full Compliance” should be considered for improving the accuracy of prompt jailbreak determination.

Claire

- I focussed on reviewing current scientific literature on the topic, in order to give the class an up to date overview of LLMs and Jailbreaks, and to inform our section on policy recommendations, to see what the current understanding of academia on regulation of LLMs is.
- Beyond playing around with chatbots to try and make them jailbreak, and thus taking the theory to practice, I made use of AI to explain some of the technical language used in the literature to me. I used either ChatGPT or Gemini. An example from the discussion on the types of jailbreaks would be that I didn’t know what ‘syntactic transformation through orthographic transformation’ was, and I asked the LLM to explain and give me examples. I also used it to provide me summaries or give me the key points from papers.