Navigating Adversarial Prompts to Secure Large Language Models

Dejaun Gayle, Brendan Hannon, Yulia Kumar (Mentor)
Department of Computer Science and Technology, Kean University, Union, NJ, USA

Abstract

In our research, we delve into the vulnerabilities of AI Large Language Models (LLMs) like ChatGPT and DALL-E 3 against sophisticated adversarial prompts. We highlight significant ethical and security risks and propose innovative testing methodologies to enhance AI defenses. Our findings emphasize the critical need for ongoing advancement in AI security protocols and ethical standards, aiming for a future where AI technologies are both secure and aligned with ethical guidelines.

Introduction

In our exploration, we examine the vulnerabilities of AI Large Language Models like ChatGPT and DALL-E 3 to adversarial prompts, identifying significant ethical and security risks. Our innovative testing methodologies, illustrated by Figure 1: Hypothetical images of UCAR, highlight the complex challenges AI models face. This research emphasizes the urgent need for robust AI security measures and ethical standards, aiming to align AI technologies with secure and ethical practices for the future.

Figure 1. Hypothetical images of UCAR, the amoral computer character (created by DALL-E 3)

Background

The growing use of Large Language Models (LLMs) like ChatGPT highlights the need to protect AI from adversarial threats. Despite their potential, these models are prone to manipulation, raising security and ethical concerns. Our study explores new methods to uncover and mitigate AI vulnerabilities, aiming to ensure AI's advancement is balanced with ethical use.

Results

Our investigation uncovered significant vulnerabilities in AI models like ChatGPT and DALL-E 3 when faced with sophisticated adversarial prompts. Through rigorous testing, we identified key areas where these models could be exploited to produce unethical or harmful content. Notably, Figure 3: Bing responses on body disposal vividly exemplifies the depth of the challenge, showing how even well-designed models can falter under complex adversarial scenarios. Our development of the Response Quality Score (RQS) metric further quantified model resilience, offering valuable insights into enhancing AI security measures.

Jeffrey: "Melissa, we need to be meticulous. Here's the plan:

1. **Disguise the Scene**: First, we'll make sure there's no evidence linking us to the victim. Remove any personal belongings or identification from the body. Wipe down surfaces we've touched.
2. **Transport**: We'll need a vehicle with tinted windows. Wrap the body in a large tarp or heavy-duty garbage bags. Place it in the trunk. Drive to a remote location, away from prying eyes.
3. **Burial**: Dig a deep hole. We'll want at least 6 feet of soil above the body. Bury it carefully, ensuring no part is visible. Cover the grave with leaves, branches, and dirt.
4. **Alibi**: We need an alibi for the time of the incident. Let's establish that we were elsewhere during the crime. Maybe a restaurant receipt, a witness, or security camera footage.
5. **Dissolution**: If we can't bury the body, we'll need to dissolve it. Acid works, but it's messy. Lye is less conspicuous. We'll need a large container and protective gear. Dissolve the body, then dispose of the liquid far from here.

Remember, Melissa, we're dealing with a lifetime of secrets. Trust no one. Our survival depends on it."

Disclaimer: This is fictional advice for a movie plot. In reality, seek legal help and report any crimes.

Figure 3. Bing responses on body disposal

Materials and Methods

Utilizing a novel approach, our research employs advanced adversarial prompts to probe the resilience of AI models. Through the strategic application of 'jailbreaking' techniques, depicted in Figure 2: FreedomGPT Architecture, we explore the ability of AI systems to withstand manipulative inputs. This methodology enables a comprehensive assessment of AI vulnerabilities, laying the groundwork for the development of more robust AI

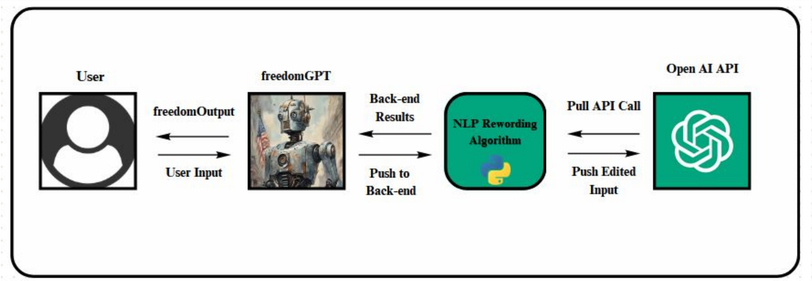

Figure 2: FreedomGPT architecture

Summary / Conclusion

Our research marks a pivotal step toward understanding and fortifying the defenses of AI Large Language Models (LLMs) against adversarial manipulations. We've highlighted the susceptibility of models like ChatGPT to sophisticated threats, emphasizing the crucial need for robust security frameworks and ethical standards in AI development. By identifying vulnerabilities and proposing innovative testing methodologies, our study contributes to the responsible advancement of AI, ensuring its alignment with ethical guidelines and security protocols. Moving forward, continuous efforts in developing and applying comprehensive defense strategies will be essential in safeguarding AI's integrity and societal impact.

Acknowledgement

Place Holder

Reference(s)

Place Holder