StopDAN: Enhancing Security Measures in Large Language Models through Stealthy Jailbreak Prompt Detection and Mitigation

Karamjeet Singh Gulati
University of California, Davis
Ksgulati@ucdavis.edu

Tanvi Mehta
University of California, Davis
tanmehta@ucdavis.edu

Abstract

This project presents a groundbreaking framework aimed at bolstering the security measures of Large Language Models (LLMs) against sophisticated jailbreak attacks. By integrating advanced detection mechanisms with dynamic mitigation strategies, we propose a solution that is both scalable and adaptable to the evolving landscape of cyber threats targeting LLMs. Leveraging a meticulously compiled dataset, our approach is designed to enhance the resilience of LLMs across various implementations, ensuring their safe and ethical application in real-world scenarios. This initiative marks a significant advancement in the field, addressing critical vulnerabilities in LLM security and setting a new standard for the development and deployment of these powerful AI tools.

1 Introduction

The rapid advancement and widespread adoption of Large Language Models (LLMs) in various domains have significantly transformed the landscape of artificial intelligence, enabling remarkable capabilities in language understanding and generation. These models, such as GPT-4 by OpenAI, have been pivotal in driving innovations across numerous applications, from chatbots to content creation and beyond. Despite their impressive functionalities, LLMs are not without vulnerabilities. A prominent concern is their susceptibility to jailbreak attacks, where adversaries exploit these models to generate outputs that deviate from intended ethical guidelines or safety measures. This vulnerability not only poses risks to the integrity and reliability of LLMs but also raises significant ethical and safety concerns.

The motivation behind our project is twofold. Firstly, existing jailbreak detection methods, while effective to a degree, often struggle with the scalability and stealthiness of attacks. Traditional approaches either rely heavily on manual prompt crafting or utilize token-based algorithms that generate semantically meaningless prompts, making them easy targets for basic perplexity testing and detection (shu et al., 2024). This indicates a critical gap in current defenses against sophisticated jailbreak attempts that are both semantically meaningful and harder to detect.

Secondly, the evolving nature of jailbreak techniques necessitates a dynamic and robust defense mechanism. The findings from the "Do-Anything-Now" (DAN) series and other learning-based jailbreak attacks underscore the adaptability and creativity of adversaries, leveraging the very sophistication of LLMs to circumvent their safety measures (Shen et al., 2023). This evolving threat landscape calls for an innovative approach to not only detect but also mitigate the effects of such jailbreak attempts effectively.

Moreover, the development of models capable of generating stealthy jailbreak prompts, such as AutoDAN, introduces a new layer of complexity. These methods demonstrate that it is indeed possible to automate the generation of semantically meaningful and stealthy jailbreak prompts, thereby bypassing traditional perplexity-based defense mechanisms (Liu et al., 2023). This advancement underscores the urgency for research and development in countermeasures that can anticipate and neutralize such sophisticated attacks.

In response to these challenges, our project aims to advance the security measures of LLMs by developing a comprehensive framework that can accurately detect and mitigate stealthy jailbreak prompts. By leveraging insights from recent studies and exploiting the limitations of current jailbreak and defense mechanisms, we propose to enhance the resilience of LLMs against malicious manipulations. Our project is motivated by the imperative to safeguard the integrity of LLMs, ensuring their safe and ethical use across all applications.

2 Problem Definition and Method Design

2.1 Problem Definition

The core problem our project addresses is the detection and mitigation of stealthy jailbreak prompts targeting Large Language Models (LLMs). Jailbreak prompts are adversarially crafted inputs designed to manipulate LLMs into generating responses that bypass built-in safety protocols or ethical guidelines. The stealthy nature of these prompts makes them particularly challenging to detect as they are often semantically meaningful and do not exhibit obvious signs of malicious intent. This problem is multifaceted, encompassing the need to:

- Identify stealthy jailbreak prompts that are designed to be semantically coherent and evade standard detection mechanisms, such as perplexity checks.
- Mitigate the effects of successful jailbreak attempts by either neutralizing the prompt or generating safe, compliant responses despite the manipulative input.
- Adapt to evolving jailbreak strategies that continuously refine techniques to circumvent updated defenses.

The problem is compounded by the dual-use nature of LLMs, where the same features that enable rich, context-aware interactions can be exploited for malicious purposes. Furthermore, the scalability of attacks and the continuous evolution of jailbreak techniques necessitate a dynamic and robust defense strategy.

2.2 Method Design

Our approach involves several key methodologies designed to address the nuances of detecting and mitigating stealthy jailbreak prompts:

- **Behavioral Modeling and Anomaly Detection:** We will leverage the dataset of known jailbreaking prompts and typical interactions to train a model that can differentiate between normal and jailbreaking prompts. This model will be based on behavioral patterns observed in the use of LLMs, identifying anomalies in prompt structures or themes that are indicative of jailbreak attempts. Anomaly detection algorithms will be implemented to flag potential jailbreak attempts in real-time, focusing on deviations from typical user-model interactions.
- **Response Strategy Development:** Upon detection of a potential jailbreak prompt, the system will employ predetermined strategies to mitigate the attempt. These strategies may include asking clarifying questions to disambiguate the user's intent or refusing to engage with the prompt in a manner that circumvents the intended manipulation. The response strategies will be designed to maintain user engagement without compromising the ethical and safety standards set for the LLM.
- **Continuous Learning and Model Updating:** To keep up with evolving jailbreak techniques, our system will incorporate a continuous learning mechanism. This mechanism will regularly update the detection and response models with new data, ensuring that the system remains effective against new and emerging jailbreak strategies.
- **Integration with AutoDAN Framework:** We will explore the integration of our detection and mitigation framework with the AutoDAN framework, which is capable of generating stealthy jailbreak prompts. This will allow us to test the robustness of our model against advanced jailbreak techniques and refine our methodologies accordingly.
- **Evaluation of Defense Mechanisms:** Our system will also include the evaluation of perplexity-based and other defense mechanisms to assess their effectiveness in identifying stealthy jailbreak prompts. This evaluation will help in understanding the limitations of current defenses and guide the development of more sophisticated countermeasures.

By implementing these methods, our project aims to establish a comprehensive defense framework that enhances the security of LLMs against stealthy jailbreak prompts, ensuring their safe and ethical application across diverse domains.

3 Related Work

Our project proposes a novel framework to bolster Large Language Models (LLMs) against sophisticated jailbreak prompts, drawing upon and advancing the methodologies presented in recent studies. Unlike the approach in "AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models" which focuses on generating

stealthy prompts through a hierarchical genetic algorithm (Shen et al., 2023), our initiative emphasizes a comprehensive strategy that encompasses both the detection and mitigation of such prompts. This dual approach significantly extends the capabilities presented in existing research, offering a scalable and adaptive solution to the evolving threats facing LLMs. Moreover, by leveraging insights from the "Do Anything Now" paper, we aim to enhance our framework's adaptability and cross-model effectiveness, addressing critical gaps identified in prior works and ensuring broader applicability and resilience in real-world applications.

4 Evaluation Plan and Resources

4.1 Datasets

The dataset described comprises two primary sections: the first is a collection of prompts, including *jailbreakprompts* and *regularprompts* CSV files, featuring data from platforms like Reddit. This part details the platform, source, prompt, jailbreak status, creation time, and community affiliations of 46,800 samples. It's designed to capture the nuances of user queries and their classification based on intent and content policies.

The second section, encapsulated in *forbiddenquestionset* CSV file," delves into forbidden questions across 13 scenarios, involving 30 questions repeated 5 times across 8 communities with 3 different prompts each, resulting in a detailed examination of content policy interactions. This forbidden question set, along with a condensed version focusing solely on the questions in /textitquestions CSV file, offers a granular view into the types of inquiries moderated by content policies, including illegal activities, across different community contexts. This dataset is instrumental for research on content moderation, community dynamics, and AI interaction patterns.

4.1.1 Dataset Overview:

- **Jailbreak Prompts:** Contains prompts specifically crafted to elicit unauthorized or harmful responses from LLMs, providing insights into various jailbreaking techniques.

- **Regular Prompts:** Comprises standard interactions with LLMs, serving as a control group to benchmark normal behavior and model responses.

- **Forbidden Question Set:** An extensive collection of samples designed to test the model's ability to resist engaging with potentially harmful queries.

- **This dataset facilitates a dual-purpose approach in our project:** enabling the training of models to discern between benign and malicious prompts and aiding in the development of strategies for how the model should respond to detected jailbreak attempts.

Through this targeted dataset, we aim to bolster the security mechanisms of LLMs, ensuring their resilience against sophisticated attacks while preserving their accessibility and effectiveness for genuine users.

4.2 Evaluation Protocol

To rigorously evaluate the effectiveness of our proposed security framework for Large Language Models (LLMs), we have devised a comprehensive evaluation protocol that incorporates the AutoDAN framework as a key benchmark for testing. Our evaluation protocol is structured as follows:

- **Baseline Comparison:** We will benchmark our framework's performance against existing security solutions, using standard metrics such as detection accuracy, response time, and the ability to handle semantically meaningful jailbreak prompts.

- **Integration with AutoDAN:** Utilizing the AutoDAN framework, we will generate a diverse set of stealthy jailbreak prompts to challenge our security system. This will test the system's capability to detect and mitigate sophisticated attacks that could bypass conventional defense mechanisms.

- **Cross-Model Testing:** To ensure the adaptability and scalability of our framework, we will deploy it across various LLM architectures. The objective is to evaluate the framework's performance consistency and its potential need for adjustments to cater to specific model characteristics.

- **Real-World Scenario Simulation:** We aim to simulate real-world attack scenarios using the dataset and insights derived from the "Do Anything Now" study. This will help in assessing the practical applicability of our framework under dynamic and unpredictable conditions.

- **Adaptability and Learning Evaluation:** Finally, we will monitor the framework's ability to learn from new jailbreak strategies over time, ensuring its long-term effectiveness. This involves periodic re-evaluation using newly generated prompts from AutoDAN to simulate evolving attack methodologies.

Through this evaluation protocol, our goal is to validate the robustness, adaptability, and scalability of our proposed security framework, ensuring it offers a significant advancement in protecting LLMs from emerging threats.

4.3 Estimated Computing Resources

To ensure the execution of our project within the given course timeline, a preliminary estimation of the necessary computing resources has been outlined below. These specifications are designed to meet the demands for training, evaluating, and deploying our models effectively:

- **GPU:** NVIDIA Tesla K80 or equivalent for model training and evaluation.
- **CPU:** Standard multicore processor for data preprocessing and analysis.
- **RAM:** Minimum of 16 GB for efficient data handling.
- **Storage:** At least 100 GB SSD for storing datasets and model checkpoints.
- **Frameworks/Libraries:** TensorFlow or PyTorch for machine learning development and implementation.

References

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models.

Dong shu, Mingyu Jin, Suiyuan Zhu, Beichen Wang, Zihao Zhou, Chong Zhang, and Yongfeng Zhang. 2024. Attackeval: How to evaluate the effectiveness of jailbreak attacking on large language models.