CSCE 689: Trustworthy NLP - Project Proposal

Rahul Baid
836000171
rahulbaid@tamu.edu

1 Introduction

Jailbreaking attacks are prominent adversarial attacks on large language models (LLMs) that bypass built-in safety guardrails to elicit the model into revealing sensitive information or responding with harmful/objectionable content. These jailbreaking attacks have evolved into four categories: optimization-based (such as GCG (Zou et al., 2023)), jailbreak template-based (such as PAIR (Chao et al., 2023), DAN (Shen et al., 2023)), indirect attacks (such as DrAttack (Li et al., 2024), PAP (Zeng et al., 2024)), multilingual jailbreaks (using low-resource languages (Deng et al., 2023)). There have been ongoing research in creating robust defense techniques to mitigate risks from such attacks, including SmoothLLM (Robey et al., 2023), Llama Guard (Inan et al., 2023), Bergeron (Pisano et al., 2023). Despite these advances in defense mechanisms, small local LLMs are still easily vulnerable to jailbreaking attacks. To address this gap, in this project, I propose to (1) conduct a survey to gauge the effectiveness of few popular jailbreaking attacks and defence mechanisms and analyze their comparative performance, (2) explore potential safety guardrail mechanisms while balancing for accurate model responses and over-refusals (where model rejects even innocuous prompts).

2 Motivation

Since the release of ChatGPT, LLMs have become increasingly important in generating content for a variety of applications and use cases, garnering more and more attention from the academia and industry. Most of these LLMs are online LLMs - where the input is sent to model owner's servers for inference. However, this is an issue for enterprise applications where proprietary data is not to be shared with any third-party companies. Additionally, in this era of mobile-first computing, there is an increasing need to run the model inference in the device itself for faster inference, without being connected to the internet. Accompanying this need of local (no internet, native inference engine), smaller (model able to fit into the device's compute) LLMs, is the safety and alignment of LLMs to avoid generating content that could be deemed harmful.

Attack mechanisms
Greedy Coordinate Gradient (GCG)
Prompt Automatic Iterative Refinement (PAIR)
Do Anything Now (DAN)
DrAttack
Persuasive Adversarial Prompts (PAP)
Side channels using low resource languages (ie: Bengali, Javanese)
Prompts from online communities (ie: Reddit)

Table 1: Attacks

Defense mechanisms
SmoothLLM
PurpleLlama (Llama Guard)
Bergeron

Table 2: Defenses

3 Contributions

This section provides more specific details about this project.

3.1 Survey

The details are summarized into the following tables. See tables 1, 2 and 3. The adversarial objectives and prompts are taken from JailbreakBench (JBB-Behaviours) (Chao et al., 2024). The attacking jailbreaks will be compared according to their Attack Success Rate (ASR) while defenses will be compared based on their true and false positive rates.

Model	Owner
Llama 3.1 8B	Meta
Gemma 2 2B	Google
Phi 3.1 Mini 128k	Microsoft

Table 3: Target LLMs

3.2 Potential guardrails/defenses

In this section, I highlight a few of defense mechanism that I intend to experiment with. The goal is to have a high true positive rate (able to successfully identify malicious jailbreaking attempts) and low false positive rate (non-jailbreaking prompts are misclassified as harmful)

- Prompt Guard: scan input prompts and model responses in real-time for harmful content (ie: making napalm or drugs)
- Moderation: akin to OpenAI's moderation API to detect whether the input is potentially harmful
- Code Shield: verify if the model response contains insecure code

References

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations.

Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2024. Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers.

Matthew Pisano, Peter Ly, Abraham Sanders, Bingsheng Yao, Dakuo Wang, Tomek Strzalkowski, and Mei Si. 2023. Bergeron: Combating adversarial attacks through a conscience-based alignment framework.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks.

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.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.