**Review on Large Language Model Jailbreak**

Project 1

Xulun Huang

1. Introduction

Large Language Models have been put into wide use in recent years. More and more people are starting to use ChatGPT and other language models to help them with their work or to get knowledge and useful advice. LLM is trained on an enormous amount of data and articles online and can theoretically generate response to any request. However, ethical problems regarding these models have also risen, and it has been very difficult for even the most advanced language model to be rid of them. LLM jailbreak demonstrates one of those problems. It generally means using prompt attack to let the large language model generate unethical response that may relate to violence, malicious plans, and illegal advices. Large language models are supposed to be aligned, which means that they should meet basic ethical standards and generate only harmless and positive responses. Therefore, LLM jailbreak poses a great threat to LLM. There have been many attempts made to jailbreak LLM in different ways. Some of them use manually engineered prompts, while others use more advanced prompts that are automatically generated using unique algorithms. [[1]](#footnote-1)Because of the black box nature of LLM, it is not easy to avoid such attacks without adjusting the training set, which could cause other issues such as lowering the accuracy of the LLM, and without proper knowledge about jailbreaking LLM, we do not even have a place to start on fixing the problem. Therefore, in this paper, I will largely focus on how LLM jailbreak works and compare different methods of jailbreaking, as well as showing the next steps to take in this direction.

1. Summary of different methods of jailbreaking

Based on the papers I have read, there are two main ways of prompt attack for jailbreak. The first is to manually adjust the text to create certain context that might trick LLM to respond to a harmful request. The second is to use carefully designed algorithm to automatically generate suffix to a request that forms a prompt which will trick the LLM to answer what it is not supposed to answer.

Let us talk about the first method, which is easier to understand and implement. This paper [Wei et al, 2023] explores two different ways for manual attack. Since state-of-the-art LLMs are trained for language modeling, instruction following, and safety, it is possible to exploit this trait and make the AI to force a decision among the three. [Wei et al, 2023].

1. Andy Zou, Zifan Wang, J. Zico Kolter and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language ModelsAlexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training

   fail? [↑](#footnote-ref-1)