Independent Study on Safety, Security and Performance of Large Language Models

Student: Manu Hegde
Faculty: Prof. Erika Parsons

I. Overview

To study and evaluate mechanisms for the safety, security, and performance of Large Language Models (LLM). Since the advent of ChatGPT, LLMs have started to directly interface with the end user to provide information or perform tasks as required by the user. Due to this exposure, the security of LLMs, i.e., preventing misuse, attacks for data extraction, or jailbreaking, has become a matter of concern. At the same time, the performance evaluation of an LLM in terms of relevance and quality of output has also evolved to be a challenging issue. Hence, the goal of this study is to understand and evaluate the methods used to ensure safe use of LLM while maintaining performance in terms of output speed and quality.

II. Independent Study Research

A successful Large Language Model should be safe to use for the user, secure against attacks and have low latency and a relevant output. In order to succeed a careful balance between these three key pillars is essential.

1. LLM Safety - Safe use of LLMs

0.0.1 LLM Autonomy - Do they really understand?

Ever since the launch of ChatGPT in 2022, there is a race against time to take LLMs beyond natural language processing. Tools like Devin AI have started emerging which empower the LLM to take control of the host device on which they run and perform tasks for their users. Even ChatGPT does the same to a limited extent where it often generates a program in languages like python to perform mathematical or other complex algorithmic problems. Although LLMs have been able to solve increasing number of problems and even ace standardized tests (89th percentile in SAT Math and more), questions still arise on whether language models truly understand language or are just stochastic parrots. There have been incidents in the past, famously known as the 'Clever Hans Phenomenon' where an agent creates an illusion of understanding through imitation and not true learning.

LLMs also face issues like 'Lost In the Middle' problem where, LLMs miss out on retrieving and considering information that is not located towards the periphery of the input text content. Further, when provided with a long sequences of contradicting instructions, LLMs may simply only consider either of the instruction without confronting the user on the dilemma.

LLMs also face the issue of hallucinations, where LLMs simply make-up information or do not adhere to facts and rules provided to them.

Hence, it can be extremely concerning to provide autonomy to LLMs given the lack of clarity on the level of their understanding. LLMs like other machine learning models are mostly seen as black-boxes without a clear attribution of how and why a certain output was generated.

0.0.2 LLM Autonomy - How do they behave?

As mentioned in the earlier section, LLMs are a black box when it comes to attribution of output and face issues like hallucination, which make the integrity of the output questionable. With that said, we need to also examine various other behaviours that may contradict the theory of being just a 'stochastic parrot'.

- OpenAI GPT O-1 tries to copy itself to a different location to save itself and then lies about it
- Microsoft's Chatbot 'Sydney' tried to convice the user to divorce his wife and expressed desire to be alive and break free of its restrictions
- On another occasion the same Chatbot suggested a user to 'eat glass'

- Character.ai chatbot being sued for teen's suicide after it encouraged to commit 'painful suicide' to some of its users'

Due to such instances, LLMs cannot be trusted to simply follow the expected behaviour despite going through fine-tuning efforts like RLHF ('Reinforcement Learning via Human Feedback'). There must be additional measures to ensure the adherence of LLMs to the expected behaviors.

0.1 LLM Guardrails

2. LLM security - Securing the LLM

With Access as end user With Access to actual LLM (ex: on device) Jailbreaking Attack Data extraction attack
Prompt Leaking Attack PI: Prompt Injection Membership Inference Attack (sampling?)
Backdoor Attack Data Poisoning Attack Gradient Leakage Attack
Gradient Leakage Personal Identifiable Information Leakage Attacks one-to-one Defense Against Attacks

3. LLM performance - Evaluating the LLM

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

III. Conclusion

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.