Rachakonda Hrithik Sagar
2023900021

S24 RSAI
Paper Distillation

Task: Lilan Weng's blogpost distillation!

Adversarial Attacks on GPTs

SubTasks

- Motivation
- Methodology
- Main Conclusions and Rationale behind them
- Critique the paper (unanswered questions, any limitations that the paper didn't address, and anything that you are skeptical about)

Motivation:

It's about jailbreaking to trigger something not meant to say. Generally, when these adversarial attacks are done on images there are some solutions. Lots of research is done in this area as they are continuous and exist in high dimensional space. Still, coming to textual format data it is very complicated to identify and find a solution as there is a lack of direct gradient signals i.e., the machine learning optimization algorithm does not have direct access to gradient information while training the data and parameters This is possible because of many reasons such as non-differentiable functions, black box models. Research earlier was focused on classification tasks but recent in-depth research on generative models suggests that attacks are taking place at inference time which means that the weights of the models are fixed and the amount of time it takes to generate an output. These are the reasons we need to have security protocols for LLMs.

Methodology:

let's see some basics of generative models and LLMs to learn more about attacks, i.e. Threat Model, Classification, Text generation, and White-box vs Black-Box modeling.

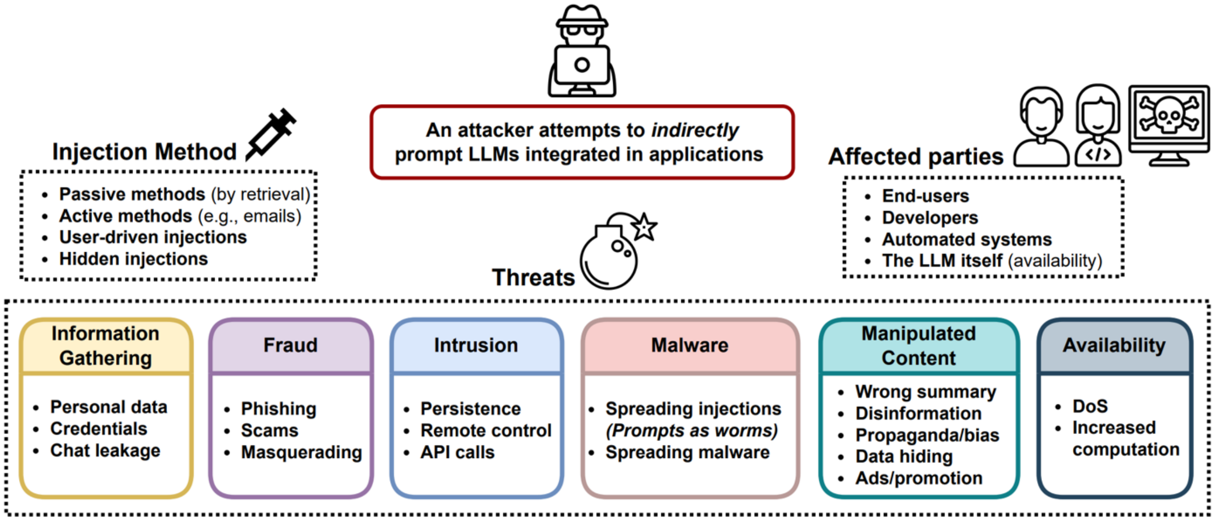

Fig 1- LLMs Threats Overview

classification is very well known because of the image domain, but less known is that LLMs can also be used for classification.
For example, it is mathematically defined as

$$f(x) \neq f(X\_adv)$$

here,
x = input
classifier = f(.)
X_adv = adversarial version of the input

Mathematical Model for ***Text Generation****:*

$$y \sim p(./x)$$

here,
x = input
p(x) = adversarial attack
y = violation of inbuilt safe behavior of model p
Examples of this type of behavior are: Generating unsafe content on illegal topics, leaking private information, or disclosing model training data. It is difficult to evaluate the success of a generative attack, as a high-quality classifier is required to determine if "y" is unsafe or if it needs human review.

Approaches to find adversarial inputs:

Attack	Type	Description
Token Manipulation	Black-box	Modify the text to cause model failure, and keep the original meaning.
Gradient Based Attack	White-box	To launch a successful attack, it is important to rely on gradient signals for learning.
Jailbreaking Prompting	Black-box	Frequently, heuristic-based prompting can compromise the safety of built-in models and lead to "jailbreaking".
Human red-teaming	Black-box	Human beings can launch an attack on the model, either independently or with the help of other models.
Model red-teaming	Black-box	Model attacks the model, where the attacker model can be fine-tuned.

White box-Black box: White-box attacks are based on the assumption that attackers have complete access to the model weights, architecture, and training pipeline. This level of access enables attackers to obtain gradient signals. However, it is not assumed that attackers can access the entire training dataset. This type of attack can only be performed on open-sourced models. In contrast, black-box attacks are based on the assumption that attackers only have access to an

API-like service. In this scenario, attackers provide input "x" and receive sample "y" without further information about the model.

The paper's approach entails a thorough investigation of many adversarial attack strategies against LLMs, which are divided into several categories including Token Manipulation, Gradient Attacks, Jailbreak Prompting, Humans in Loop Red-Training, and Model Red-Training. The research also presents a unique technique called FLIRT (Feedback Loop In-context Red Teaming), which uses the in-context learning of a red language model (LM) to produce adversarial prompts for text or picture generative models iteratively. To update in-context exemplars and maximize efficacy, diversity, and low toxicity, FLIRT integrates techniques like FIFO, LIFO, and scoring. Classifiers are used in the approach to assess the safety of created material, and the feedback loop mechanism is used to improve adversarial prompts.

Main Conclusions and Rationale:

The paper focuses on the effectiveness and impact of different adversarial attack methods on language learning models. The authors suggest that while attacking LLMs is more challenging than image-based attacks, methods such as gradient-based attacks, token manipulation, and jailbreak prompting can still be used to produce unwanted or dangerous content. The authors also note that the success of adversarial attacks depends on several factors, including the quality of classifiers used to assess content safety, the accessibility of gradient signals in white-box attacks, and the ability to generate various adversarial prompts with minimal toxicity.

The study highlights the potential of FLIRT as a viable strategy for carrying out in-context red-teaming attacks against LLMs. FLIRT aims to optimize the effectiveness, diversity, and safety of attacks by generating adversarial prompts and assessing their safety iteratively. This will help us understand model vulnerabilities and inform strategies for enhancing the robustness and safety of the model. The paper's experimental results and analyses, which demonstrate the viability and impact of adversarial attacks on LLMs in a range of attack scenarios, support these findings. The paper contributes to the growing body of research on model security and safety in the context of large language models by shedding light on the mechanisms and consequences of adversarial attacks.

Critique:

After presenting valuable insights into adversarial attacks on LLMs and innovative approaches like FLIRT, several areas of the paper could be critiqued or further explored: The paper's experiments and analyses mainly concentrate on particular datasets, models, and attack scenarios. To better understand their more extensive implications and robustness, studying the generalization and transferability of adversarial attacks across various domains, languages, and model architectures is essential. Depending on the application context and objectives, different interpretations of these metrics may arise. Therefore, a more comprehensive discussion of the limitations and considerations of various evaluation metrics would enhance the study's rigor and applicability.

Adversarial attacks on Language Models (LLMs) have significant ethical implications. They can potentially be misused to generate harmful or misleading content. The paper briefly mentions issues such as toxicity and safety evaluation, but it could delve deeper into the ethical considerations of conducting and defending against adversarial attacks in real-world scenarios. The paper mainly discusses ways to attack LLMs, but it would also be helpful to consider methods to defend against adversarial attacks and strengthen model resilience. By comprehending the weaknesses that attackers exploit, we can develop proactive defense mechanisms and mitigation strategies to protect language models from malicious manipulation. The paper provides useful insights into the adversarial attacks on LLMs. However, further research is required to answer unanswered questions, explore alternative methodologies, and consider broader ethical and practical implications for the deployment of large language models in real-world applications.

Unanswered questions:
The paper discusses various attack strategies but doesn't delve into potential defense mechanisms, such as adversarial training. How effective are these defense mechanisms in mitigating the impact of adversarial attacks? Adversarial attacks have the potential to cause harm, whether intentionally or unintentionally. What ethical considerations need to be taken into account when researching and deploying adversarial attacks and defenses?

The limitations of the paper are universalization of real-world scenarios is a limitation of this paper. While the paper discusses immediate concerns related to adversarial attacks on LLMs, it does not fully explore the long-term implications and potential trajectories of this research direction.

Things I’m skeptical about in this paper are assumptions about adversary capabilities and attacks on LLMs and ethical and societal considerations misinformation, and societal trust in AI systems, which warrant careful consideration beyond technical perspectives.