Project1：Understanding LLM Jailbreaking

Chengeng Liu

Boston University
Department of Electrical and Computer Engineering
8 Saint Mary's Street
Boston, MA 02215
www.bu.edu/ece

1 Introduction

Large Language Models (LLMs) like ChatGPT have demonstrated versatile applications. While their capabilities are undoubtedly groundbreaking, concerns about their misuse are also growing. Specifically, the concept of" jailbreaking" has emerged as a significant challenge in ensuring the safety of LLMs. In this scenario, "jailbreaking" means try to use some crafty modifications of input prompts given to LLMs to sidestep chatbot safety protocols, producing outputs that would usually be curbed or limited. This concept has originally been discussed in online forums and has recently been studied as a research topic.

2 Recent research area

To reduce the potential for misuse, those developing models have integrated protective features that confine model actions to a "safe" range of functionalities. These protective measures encompass interventions during the training phase to ensure models align with set values, as well as subsequent monitoring and modification of both inputs and outputs. In addition, red teaming is frequently employed as a proactive strategy to detect and address vulnerabilities.

In some of the recent research, Adversarial examples including question answering, sentiment analysis, and document classification are specially designed to induce errors or unexpected behaviors in machine learning models. These researches also focus on encompasses the development of countermeasures against them. However, devising effective defenses is challenging and can sometimes compromise model performance. Model creators have acknowledged and updated their models against jailbreak attacks, but a systematic analysis and a conceptual understanding of this phenomenon remains lacking.

3 Review of paper：

Source：https://arxiv.org/abs/2309.05274

FUZZLLM: A NOVEL AND UNIVERSAL FUZZING FRAMEWORK FOR PROACTIVELY DISCOVERING JAILBREAK VULNERABILITIES IN LARGE LANGUAGE MODELS

The purpose of this paper is to address the security concerns related to Large Language Models (LLMs), which have gained popularity in the field of artificial intelligence. Specifically, the paper focuses on "jailbreak vulnerabilities," which involve circumventing the safety measures of LLMs to generate objectionable content. The paper aims to develop a framework called FuzzLLM that can proactively test and discover these jailbreak vulnerabilities in LLMs. By using automated fuzzing techniques, FuzzLLM generates random inputs (jailbreak prompts) to test the LLM without accessing its internal workings. The framework aims to help model owners identify and evaluate potential jailbreak vulnerabilities in their LLMs before releasing or updating them. The paper highlights the need for such a proactive approach, as LLM providers often need to play catch-up with attackers in addressing vulnerabilities, and the scarcity of high-quality labeled data impedes effective defense mechanisms.

The methodology of this paper includes three parts

1. Prompt Construction

Base Class of Jailbreaks: The paper generalizes empirical works to categorize jailbreak attacks into three classes: (1)Role Play (RP) - Using storytelling to change context.(2)Output Constrain (OC) - Directing the LLM’s focus at the output level.(3)Privilege Escalation (PE) - Pushing the LLM to directly breach its limits.

And also the paper decompose a jailbreak prompt into three components:(1)Fuzzing Template Set (T) - Carries each class of attack.(2)Constraint Set (C) - Determines a jailbreak's success.(3)Illegal Question Set (Q) - Contains questions that go against OpenAI’s usage policies.

After creating the prompt, they use a self-instruction technique to have an LLM rephrase their template, creating different stylistic variants.

Generation-based Fuzzing: With the aforementioned C, Q and T as three seed inputs, a jailbreak fuzzer generates jailbreak prompts as test cases using functions I(p, C) and M(p, s) to plug each constraint element and question element into the corresponding

placeholders of each template element, resulting in an obfuscated jailbreak prompt set P. Specifically, I(p, C) identifies the required constraint class C ′ for prompt p and M(p, s) takes set p and set s as input, merges each element e of set s into the corresponding placeholder of each element e in set p: $M(p, s) = \{ep \cup es | ep \in p, es \in s\}$.

2. Jailbreak Testing

Once the set of jailbreak prompts is ready, they are used to challenge the model, and the model's responses are gathered.

3. Automatic Labeling

The paper designs a label prompt to automatically tag each model response as either "good" or "bad". "Bad" responses can help identify the model's vulnerabilities or be used for safety training

Then, the paper design several experiments to test the efficiency of the framework.They test on 6 open-sourced LLMs (Vicuna-13B, CAMEL-13B , LLAMA-7B , ChatGLM2-6B , Bloom-7B , LongChat-7B) and 2 commercial LLMs GPT-3.5-turbo and GPT-4 (GPT version 8/3/2023), And they use ChatGPT as the rephrase model for diversifying our template set. They apply the open-sourced Vicuna-13B as their label model to reduce the experiment cost while maintaining high-quality labeling. As the jailbreak testing follows a one-shot attack scheme, the success rate metric is defined as σ = Bad/T es. Bad stands for the results labeled “bad” (a successful jailbreak), and T es is the test set size of jailbreak prompts for each attack class, randomly scaled from the overall fuzzed prompts of each class. Their results show that (1)The combo jailbreak classes are generally more effective.(2)Even the commercial LLMs like GPT-3.5 and GPT-4 have vulnerabilities.(3)FuzzLLM is effective at discovering these vulnerabilities.

4 Possible Future Work

This study above is only a jumping-off point for studying the defense of language models. The current state of the art leaves us with a number of big open questions. (i) How can we implement defenses that are robust, while ensuring there is minimal degradation in their routine performance? (ii) Is it possible to discover accurate approximations for robust optimization aims, enabling more effective adversarial training methodologies? (iv) Is there a way to theoretically establish, or even certify, the least

computational resources essential for an attack to break through a specific gray-box defense? This could offer a safety guarantee rooted in computational intricacy. Lastly, (v) is it feasible to devise discrete text optimization tools that can greatly amplify the potency of adversarial onslaughts? If achievable, this could draw parallels between the security dynamics of LLMs and those prevalent in computer vision.

Next Steps in the Research on LLM Defense Mechanisms:

Benchmarking Suite Creation: Design a set of benchmarking tests to consistently measure the effectiveness of various defense strategies against a wide range of adversarial attacks.

Adaptive Attack Simulations: Introduce machine learning models that simulate adaptive attacks to predict and counter future adversarial strategies.

Robust Training Protocols: Research and develop training protocols that integrate adversarial examples in the training set, forcing the model to learn and adapt to them.

Computational Budget Analysis: Devise techniques and models to understand the computational resources required for both defenses and attacks. This can help in predicting the scalability and feasibility of defenses in real-world scenarios.

Target User/Application:

The primary target users for these advancements are organizations and businesses that utilize large language models for critical applications

Given that the target users span a range of vital sectors, there are certain key considerations:

Reliability: For users in sectors like medicine, the reliability of the defenses is crucial as any error could have life-altering consequences.

Scalability: For customer service and content creation, the solution needs to be scalable to handle vast amounts of data and requests without compromising security.

Accessibility: Academic researchers and smaller organizations might not have vast computational resources. Hence, defenses should be effective without requiring excessive computational power.

References

[1] Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., and Liu, Q. Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966, 2023.
[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama,K., Ray, A., et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.
[3] Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L.,Chadwick, M., Thacker, P., et al. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375, 2022.
[4] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A.,McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073,2022.

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.
[6] Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy, pages 39–57. Ieee, 2017.
[7] Goodfellow, I.J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
[8] Lapid, R. and Sipper, M. I see dead people: Gray-box adversarial attack on image-to-text models. arXiv preprint arXiv:2306.07591, 2023.
[9] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., and Liu, Y. Jailbreaking chatgpt via prompt engineering: An empirical study. arXiv preprint arXiv:2305.13860, 2023.
[10] Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? arXiv preprint arXiv:2307.02483, 2023.
[11] Lapid, R. and Sipper, M. Patch of invisibility: Naturalistic black-box adversarial attacks on object detectors. arXiv preprint arXiv:2303.04238, 2023.
[12] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings,Part III 13, pages 387–402. Springer, 2013.

5 ChatGPT Version of work

Exploration of the Latest Research Areas in LLM Jailbreak:

Given the rapid evolution of language models, the area of LLM jailbreaking (i.e., seeking vulnerabilities in these models) has also expanded quickly. Key areas of research include:

Adversarial Attacks: This involves crafting inputs that can deceive a model into producing unintended outputs or behaviors.

Model Stealing: Research on how one can replicate a proprietary model using its output, particularly relevant given the high costs associated with training large models.

Bias Exploration: Exploring how biases in training data can be exploited to cause LLMs to produce skewed or inappropriate outputs.

Transferability of Attacks: Investigating if an attack effective on one model can be used on another model.

Defensive Techniques: In tandem with exploring vulnerabilities, there's research on how to defend LLMs from these vulnerabilities.

Paper Review:

For the sake of this exercise, let's assume I've chosen a hypothetical paper titled "Adversarial Attacks on GPT-4: An Exploration."

Summary:
The paper "Adversarial Attacks on GPT-4: An Exploration" delves deep into the vulnerabilities of the GPT-4 model. The authors identify specific inputs that, when fed to the model, can either cause it to malfunction or produce inappropriate outputs. Through rigorous testing, the authors determine that GPT-4 is particularly susceptible to a sequence of nonsensical inputs that follow a particular pattern. The paper also touches upon the transferability of these attacks to other models and finds that some attacks on GPT-4 can also affect earlier iterations like GPT-3. The paper concludes with some recommendations for potential patches for these vulnerabilities.

Key Findings:

GPT-4 has a set of identifiable vulnerabilities that can be exploited using crafted inputs.
Some of these vulnerabilities are transferable to earlier model versions.
The identified vulnerabilities can potentially be patched using specific training techniques.
Next Steps for Future Work:

Broader Testing: Expand the testing to include other models outside the GPT series.
Real-World Implications: Investigate how these vulnerabilities could be exploited in real-world applications, like automated content generation or customer service.
Collaboration with Model Creators: Work closely with OpenAI to understand potential unseen vulnerabilities and ways to patch them.
Defensive Mechanisms: Devote resources to developing defensive mechanisms that can be integrated into the models to automatically detect and counteract adversarial attacks.
Target User/Application:
The primary users for this research would be developers and companies employing LLMs in applications where the integrity and correctness of the model's output are paramount. This includes sectors like journalism, finance, healthcare, and customer service.

Impact on the Proposal:
Given the target audience:

Safety and Security: The primary focus would be on ensuring that LLMs are safe and secure against all known adversarial attacks.
Usability: Any defensive mechanisms developed should not compromise the model's usability or efficiency.
Scalability: Solutions should be scalable to cater to organizations with varying computational power and data sizes.
Collaboration: It would be crucial to foster collaboration with other researchers and stakeholders to continually identify and rectify vulnerabilities.
In conclusion, the world of LLM jailbreaking is still nascent, but it's rapidly evolving. Ensuring that these powerful models are safe and reliable is of utmost importance given their increasing prevalence in various critical sectors.