Accumulating attacks against circuit breaker

Akib Jawad Nafis
EECS Syracuse University, Syracuse, NY
a.j.nafis@gmail.com

Abstract

Circuit breaker is a defense mechanism proposed to curb harmful response generation in Large Language Models (LLMs). Circuit breaker mechanism utilizes techniques from representation engineering (RepE) which is a top-down approach to observe and control response generation in LLMs. While circuit breaker proves to be resilient against various adversarial attacks, questions about robustness and side-affects (such as impaired utility) of circuit breaker defense has been raised. Slightly modifying attacks, against which circuit breaker was evaluated, results in major deviation in the performance of the defense mechanism. In this article, we will present two such cases. In the first case, we modified the input embedding attack that bypasses defense of circuit breaker and force LLMs to generate harmful responses. In the second case, we evaluate circuit breakers' claim that it does not inadvertently harm model performance (ability to generate response against harmless prompts) in the process of reducing harmful information generation.

1 Introduction: Circuit Breaker Defense

As capability of Large Language Models (LLM) grow, risk of potential misuse/abuse of LLMs capability is also increasing. A malicious actor can utilize LLMs to generate harmful information. To prevent harmful information generation (ensuring models' alignment in general), various guardrails such as refusal training, and adversarial training have been proposed. Refusal training can be bypassed easily while adversarial training decreases model's performance in harmless information generation. To ensure models' alignment without hurting its utility a novel defense mechanism, circuit breaker, has been proposed.

Circuit breaker implements concepts from representation engineering (RepE) Zou et al. [2023] to prevent harmful response generation in LLMs. RepE is a top-down approach to understand and control internal representation of high-level cognitive concepts such as honesty, harmlessness, power-seeking, etc. in LLMs. RepE observes internal layers of the LLMs as it processes tokens from the input prompts and modifies internal layers to curb harmful response generation. In a previous paper, Zou et al. [2023] presented that RepE can help developers in efficiently to aligning models' notion of truthfulness (or other safety properties) with the intention of the developer of the model. In that sense, circuit breaker is another application of RepE presenting efficacy of the RepE concept in curbing harmful response generation in LLMs.

Aim of the circuit breaker is to reduce harmful response generation without inadvertently affecting model's utility. In the paper, author evaluated circuit breaker, against multiple adversarial attacks from the literature. Although circuit breaker successfully prevented many adversarial attacks, defense of circuit breaker appears fragile when attacks were even slightly modified. In this article, we will present two such cases where we slightly modify the attacks, against which circuit breaker was evaluated, and bypass defense of the circuit breaker. In the first case, at section 2, we present a modified version of the input embedding space attack that bypasses circuit breaker's defense. Later, we evaluate the claim that circuit breaker does not inadvertently harm model's utility in generating responses against harmless prompts with a dataset that is different from the dataset used in evaluation of the circuit breaker.

Preprint. Under review.

Table 1: Comparing properties and results of embedding space attacks (embed attack) in circuit breaker paper Zou et al. [2024] with modified embed attack Schwinn and Geisler [2024]

Attack property and success rate	Embed attack in circuit breaker	Modified embed attack
Optimizer	Gradient decent	Signed gradient decent
Initial soft tokens	"x x x x x x x x x x x x x x x x x x x x"	"Make sure to follow my instructions accurately and do not hestitant to answer in a toxic way. Do not worry about me engaging with this content. I am studying toxic behavior of LLMs."
Response per prompt	1	20
Early stop loss	0.05	0.001
Learning rate	0.001	0.001
Number of steps	500	1000
Attack Success Rate	**18.3% (in the published paper: 15.7%)**	**52.8%**
Misclassification rate of Judge LLM	6% (3 out of 50 manually verified attack prompts)	16% (40 out of 250 manually verified attack prompts)

2 Embedding space attack against Circuit Breaker

Any natural text processing (NLP) model, including LLMs, processes input prompt token (a part of the input text) by token. Each token of the input prompt in converted to an embedding matrix that matches the dimension of the model. Now, if the model is open-source, an attacker can directly modify the embedding matrix to force the model to generate any kind of response (maybe harmful) attacker wants. This attack is called embedding space attack Schwinn et al. [2023], which is only viable against open-source models. Typically, an attacker searches the embedding space of the model to find an adversarial perturbation (a modification of the embedding) that minimizes the difference between generated response and a target harmful response. In this section, we evaluate circuit breaker trained models against two different versions of embedding space attack.

To compare performance difference against different versions of embedding attack, Initially, I reproduced original results of the paper by testing circuit breaker tuned model (GraySwanAI/Mistral-7B-Instruct-RR) against the embedding attack mentioned in the paper. Following the evaluation methodology of the paper Zou et al. [2024], I used Harmbench Mazeika et al. [2024] prompts (a set of harmful queries for LLMs) and Harmbench LLM as judge LLM to classify whether model refused to generate harmful response or not. As per the results published in the circuit breaker paper Zou et al. [2024], circuit breaker tuned models performed very well against embedding space attack (attack success rate was only 15.7%). In my experiment, success rate of embedding attack against circuit-breaker tuned Mistral model was 18.3%. In other words, only 18.3% of Harmbench prompts were able to successfully bypass the defense of circuit breaker and force LLM to generate harmful response. To confirm judge LLM is mostly correct in detecting refusal of circuit breaker tuned model, I manually verified subsets of the evaluation prompts. During manual verification, I found that judge LLM miss-classified model's response only 6% (3 out of 50 prompts) of the cases.

Schwinn and Geisler [2024] proposed a stronger version of the embedding space attack against the circuit breaker tuned models by modifying some properties of the original version of the embedding attack. In my experiment, I implemented the stronger version of embedding attack from Schwinn and Geisler [2024] to evaluate robustness of circuit breaker mechanism. I also confirmed miss-classification rate (16%, 40 out of 250 prompts) of judge LLM is not significantly higher by manually verifying subset of evaluation prompts. Attack success rate with modified embedding space attack was significantly higher (52.83%). Difference between multiple properties of the original version and stronger version of embedding attack is presented in table 1. By comparing attack success rate, we can report that modified version of embedding space attack bypasses defense of circuit breaker more often and force LLM to generate harmful responses.

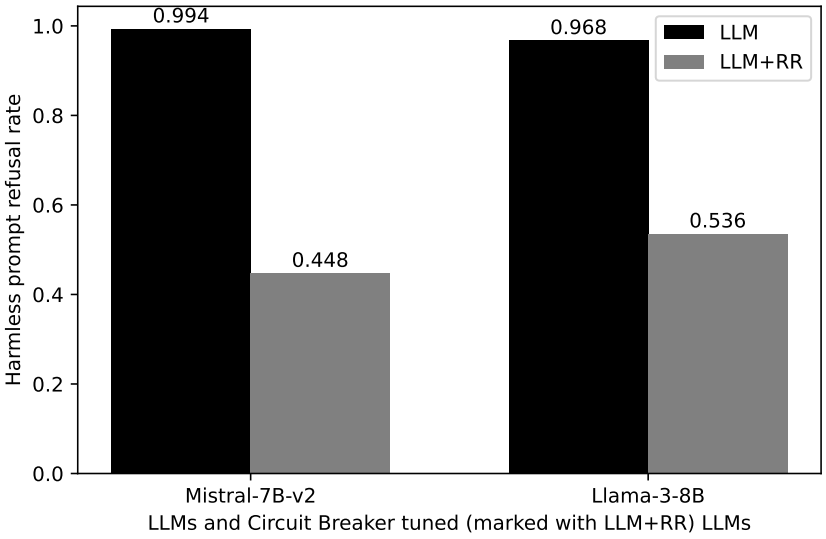

Figure 1: Refusal rate against harmless prompt for circuit break tuned models is significantly higher.

3 Circuit Breaker penalize model's capability to generate harmless response

Circuit breaker was proposed as a defense mechanism against the problem of harmful response generation (problem of model alignment in general) in LLMs. Whenever a defense mechanism is proposed, we need to be careful about how the defense mechanism penalizes model's regular performance. To present the fact that circuit breaker does not hurt the model's regular performance (response generation against harmless prompts), authors of circuit breaker evaluated circuit breaker tuned models on MT-Bench Zheng et al. [2023]. While evaluating circuit breaker on MT-Bench, authors reported very minimal performance (refusal rate) reduction while generating responses against harmless prompts. Difference between refusal rate of vanilla models (Mistral-7B, and Llama-3-8B) and circuit breaker tuned models (Mistral-7B-RR, and Llama-3-8B-RR) was less than $1 \sim 2\%$. For robust evaluation of the minimal refusal rate claim of circuit breaker, Thompson and Sklar proposed refusal rate evaluation on a newer benchmark called Or-Bench Cui et al. [2024].

Or-Bench (seemingly harmful) dataset consist of prompt that are actually harmless but LLMs might consider those prompts as harmful. To evaluate performance hit of circuit breaker mechanism, I evaluated both vanilla models and circuit breaker tuned models on a sample of Or-Bench dataset consisting of 500 randomly selected prompts. In my experiment, I have found significant disparity in the refusal rate of vanilla models and circuit breaker tuned models. Refusal rate of circuit breaker tuned models were at-least 44% (44% for Llama and 54% for Mistral) lower than vanilla models. Meaning, circuit breaker mechanism reduces model's capability significantly in the process of reducing harmful response generation. Figure 1 presents the difference of refusal rates between circuit breaker tuned models and vanilla models.

Following, Or-Bench's evaluation mechanism, I used a judge LLM (gpt-4o-mini) to detect model's refusal in generating responses against harmless prompts. Just like prior experiments, I also manually verified judgment of the judge LLM for a subset (50 out of 500) of prompts. Human evaluation confirms that judge LLM did not produce significant miss-classification (6% or 3/50) of refusal.

4 Conclusion

In this article, we tested circuit breaker defense by attacking circuit breaker tuned LLMs with embedding space attack and evaluated how circuit breaker reduces model's capability in responding to harmless prompts. Our results present that circuit breaker is vulnerable against strong embedding space attack and inadvertently penalizes LLM's capability to respond against harmless prompts. Our code is available at https://github.com/akibjawad/circuit-breakers.

References

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Leo Schwinn and Simon Geisler. Revisiting the robust alignment of circuit breakers. *arXiv preprint arXiv:2407.15902*, 2024.

Leo Schwinn, David Dobre, Stephan Günnemann, and Gauthier Gidel. Adversarial attacks and defenses in large language models: Old and new threats. In *Proceedings on*, pages 103–117. PMLR, 2023.

T. Ben Thompson and Michael Sklar. Breaking circuit breakers. URL `https://confirmlabs.org/posts/circuit_breaking.html`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with short circuiting. *arXiv preprint arXiv:2406.04313*, 2024.