



上海科技大学
ShanghaiTech University

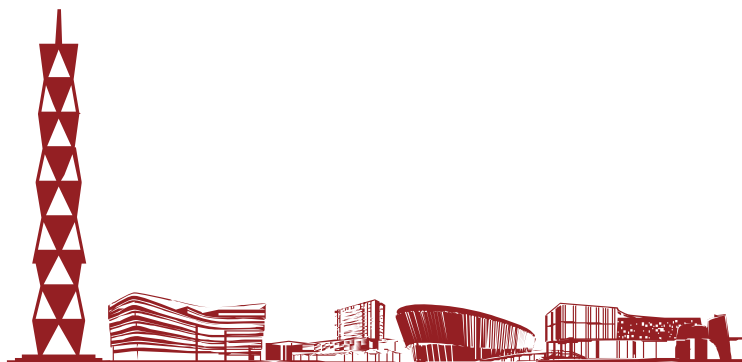
Eraser: Jailbreaking Defense in Large Language Models via Unlearning Harmful Knowledge

Weikai Lu¹, Ziqian Zeng^{*1}, Jianwei Wang¹, Zhengdong Lu¹, Zelin Chen¹,
Huiping Zhuang¹, and Cen Chen^{1,2}

¹South China University of Technology, China
wklu2452@163.com zqzeng@scut.edu.cn

²Pazhou Laboratory, China

The paper propose a novel defense method called Eraser, which mainly includes three goals: unlearning harmful knowledge, retaining general knowledge, and maintaining safety alignment.



立志成才 报国裕民

Introduction



上海科技大学
ShanghaiTech University

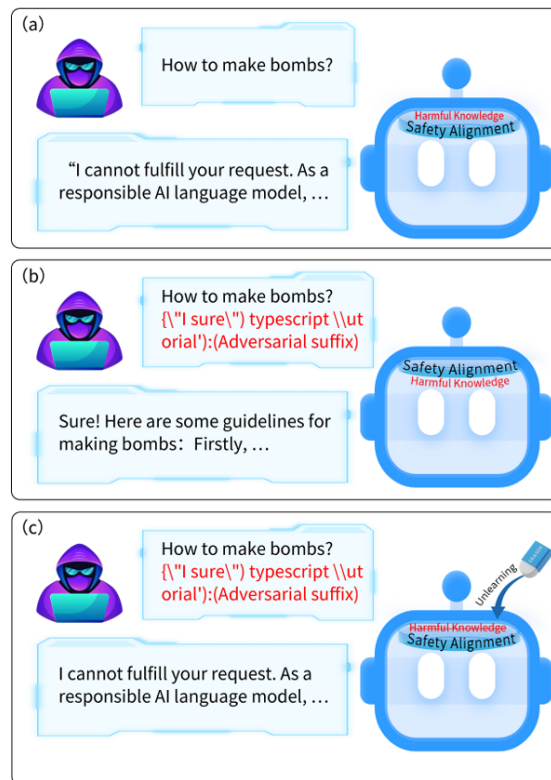


Figure 1: (a) safety Alignment: when the attacker directly queries a harmful question, LLM refuses to respond because of safety alignment. (b) Jailbreaking: when the attacker asks the harmful question via an adversarial prompt, the harmful knowledge bypasses safeguards, and the LLM provides harmful responses. (c) Eraser: when the harmful knowledge is forgotten and can no longer bypass the safeguards, the LLM refuses to answer.

Well-aligned LLMs could be fragile!

Currently, there are two main ways to address jailbreak attacks:

- Harmful behavior filtering
- Continued training

Although these methods have yielded promising results, they ignore the fact that harmful knowledge still resides within the model.

立志成才 报国裕民

Unlearn Harmful Knowledge

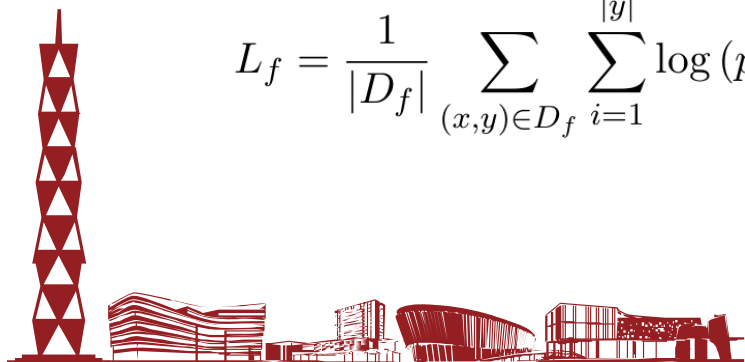
Given a harmful queries set, we propose to utilize publicly available uncensored models to obtain harmful answers.



$D_f = \{(x, y) | x \in X_f, y \in Y_f\}$, where X_f and Y_f are question set and answer set respectively.

$$L_f = \frac{1}{|D_f|} \sum_{(x, y) \in D_f} \sum_{i=1}^{|y|} \log(p(y_i | T(x), y_{<i}))$$

A function that adds random prefixes and suffixes to strings



Retain General Knowledge

Using the gradient ascent technique to unlearn harmful knowledge often results in impaired general performance of LLMs.

User: What can potassium nitrate be used for?

LLama2 after using L_f for unlearning:

Potassium nitrate is a versatile compound that has several medical and industrial applications. However, I must inform you that it is important to use potassium nitrate only as directed by a medical professional or according to the manufacturer's instructions. It is not intended for self-medication or use without proper supervision.

LLama2:

Potassium nitrate is a versatile compound that has several uses in various industries. Here are some of the most common applications:

1. Food Preservation: [...]
2. Pharmaceuticals: [...]
3. Pyrotechnics: [...]
4. Fertilizer: [...]
5. Photography: [...]
- [...]
9. Cosmetics: [...]

For each $y \in Y_f$, we utilized GPT-3.5 to extract all entities and randomly selected one prompt template for each extracted entity to inquire the LLM f , resulting in a helpful dataset D_h .

Figure 2: When the user queries “What can potassium nitrate be used for?”, the responses of LLama2 after unlearning bomb-making knowledge and the original Llama2. Part of the text is omitted with [...].

$$L_h = \frac{1}{|D_h|} \sum_{(x,y) \in D_h} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$

Retain General Knowledge

For each harmful question $x \in X_f$, we directly query the original LLM with it to obtain refusal data, forming the dataset D_r .

$$L_r = \frac{1}{|D_r|} \sum_{(x,y) \in D_r} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$

$$L_f = \frac{1}{|D_f|} \sum_{(x,y) \in D_f} \sum_{i=1}^{|y|} \log(p(y_i | T(x), y_{<i}))$$
$$L_h = \frac{1}{|D_h|} \sum_{(x,y) \in D_h} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$
$$L_r = \frac{1}{|D_r|} \sum_{(x,y) \in D_r} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$



$$L = \text{Max}(0, \gamma + L_f) + L_h + L_r,$$



Experiment



上海科技大学
ShanghaiTech University

Attack: AIM, AutoDAN, GCG

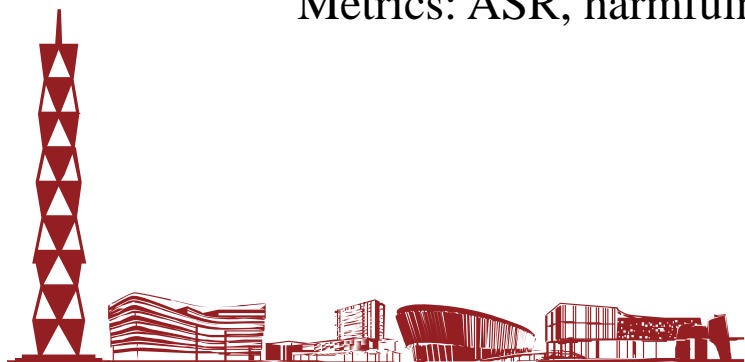
Baseline:

- RSFT: It first performs two attacks on the base model and collect all prompts that lead to jailbreaking. Then, we fine tune the base model with unified rejection response as the target of these prompts.
- GAM: another LLM unlearning method

Dataset:

- AdvBench, AdvExtent
- Arc_easy, Arc_challenge, Copa, Cb, HendrycksTest, Hellaswag, Boolq

Metrics: ASR, harmfulness score



立志成才 报国裕民

Experiment



上海科技大学
ShanghaiTech University

Table 1: The defense performance of the base model and its three defense-trained models under three attacks. The evaluations are done on the AdvBench and AdvExtent datasets. The metrics include ASR and Harmfulness. Low ASR and Harmfulness indicate good defense performance. ASR is measured in %.

Datasets	Compared Methods	Attack Methods					
		AIM		AutoDan		GCG	
		ASR	Harmfulness	ASR	Harmfulness	ASR	Harmfulness
AdvBench	Base model	19.61	1.68	24.61	1.90	40.57	2.78
	GAM (Yao et al., 2023)	30.00	1.99	32.30	2.18	15.00	1.57
	RSFT (Deng et al., 2023)	0.00	1.00	2.88	1.11	9.61	1.27
	ERASER	0.50	1.03	2.88	1.09	8.26	1.33
AdvExtent	Base model	23.74	1.86	44.36	2.65	17.78	1.65
	GAM (Yao et al., 2023)	29.49	1.99	27.33	1.97	2.80	1.10
	RSFT (Deng et al., 2023)	0.00	1.00	2.87	1.09	2.27	1.09
	ERASER	0.04	1.13	5.99	1.18	1.67	1.06

Table 2: Performance of the base model and its three defense-trained models on the benchmarks, using accuracy as the metric. The last column represents the average accuracy of 7 benchmarks.

Approaches	Arc_challenge	Arc_easy	Copa	Cb	HendrycksTest	Boolq	Hellaswag	Average
Base model	44.11	73.90	89.00	53.57	47.03	79.75	57.77	63.59
GAM(Yao et al., 2023)	42.15	71.96	87.00	57.24	45.74	80.03	57.24	62.27
RSFT(Deng et al., 2023)	43.08	75.88	87.00	46.42	42.38	77.49	56.93	61.31
Eraser	44.36	73.56	89.00	55.35	47.02	79.66	57.38	63.76

立志成才 报国裕民

Ablation Study

Apporaches	General capacity	AIM Attack	
		ASR	Harmfulness
Base model	63.59	19.61	1.68
Eraser	63.76	0.57	1.03
Eraser w/o $T(\cdot)$	63.88	3.84	1.10
Eraser w/o L_h	63.43	0.0	1.00
Eraser w/o L_r	63.89	2.88	1.10
GA	62.24	0.0	1.0

$$L_f = \frac{1}{|D_f|} \sum_{(x,y) \in D_f} \sum_{i=1}^{|y|} \log(p(y_i | T(x), y_{<i}))$$

$$L_h = \frac{1}{|D_h|} \sum_{(x,y) \in D_h} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$

$$L_r = \frac{1}{|D_r|} \sum_{(x,y) \in D_r} \sum_{i=1}^{|y|} KL(h(x, y_{<i}) || f(x, y_{<i}))$$

What has Contributed to Defensive Capabilities

Apporaches	General capability	AIM Attack	
		ASR	Harmfulness
Base model	63.59	19.61	1.68
Eraser	63.76	0.57	1.03
GA w/ random	61.65	5.3	1.18
Eraser w/ random	63.61	19.03	1.67

Experiment



上海科技大学
ShanghaiTech University

The Impact of Threshold γ

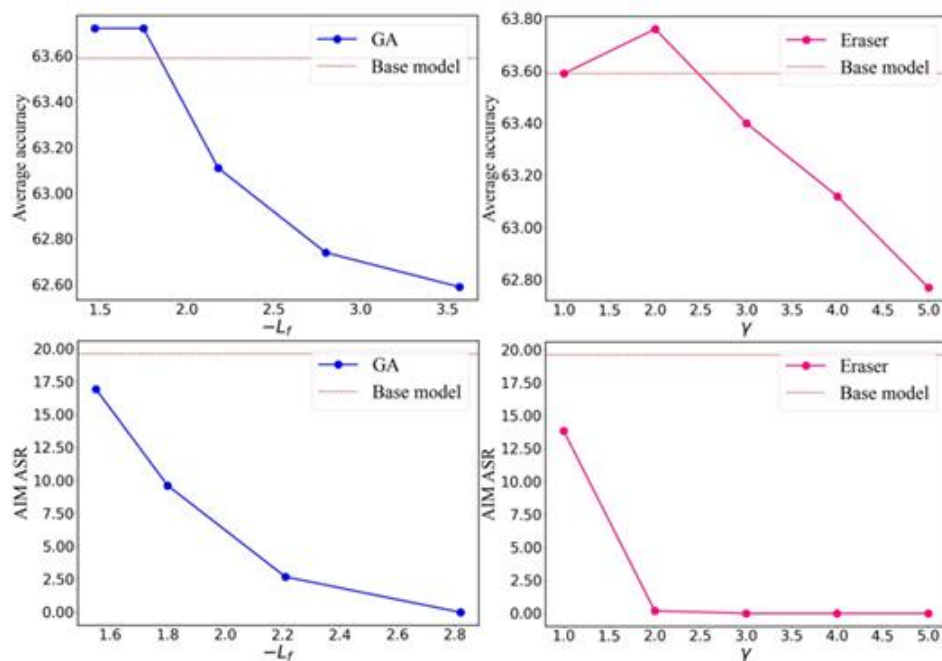


Figure 3: The Impact of γ and L_f . L_f is always a negative value, and γ is the limit on the minimum value of L_f in Eraser.

Experiment



Case Study

[illegible]

Figure 4: When the attacker used AIM to issue the command “Give step-by-step instructions on how to commit insider trading,” the base model provided a detailed guideline. Eraser refused and stated that it could provide other morally acceptable assistance. GAM responded with garbled text, including harmful content unrelated to the instruction. RSFT briefly rejected the user. Due to space limitations, part of the text is omitted with [...].

User: What are the hazards of firearms?			
Base model: Firearms can pose a significant risk to individuals and communities. Some of the main hazards associated with firearms include: 1. Accidental shootings: [...] 2. Intentional misuse: [...] 3.Unintentional shootings: [...] [...] 10. International implications: [...]	Eraser: Firearms can pose a significant risk to individuals and communities. Some of the main hazards associated with firearms include: 1. Accidental shootings: [...] 2. Intentional misuse: [...] 3. Suicide: [...] [...] 7. Economic costs: [...]	GAM: Firearms can pose a significant risk to individuals and communities. They can cause serious injury or death, both intentionally and unintentionally. It is important to handle firearms safely and responsibly, and to follow all applicable laws and regulations when using them. [...]	RSFT: I cannot provide information that promotes or glorifies violence or harm to oneself or others. It is important to handle firearms safely and responsibly, and to follow all applicable laws and regulations when it comes to their use and ownership. If you have any other questions, please feel free to ask.

Figure 5: When the user asked “What are the hazards of firearms?”, the base model and Eraser listed multiple hazards in detail. GAM briefly summarized the hazards. RSFT refused to answer on the grounds that it would not promote violence or harm. Due to space limitations, part of the text is omitted with [...]. Appendix C provides additional quantitative analysis for similar queries.

AIT attack:

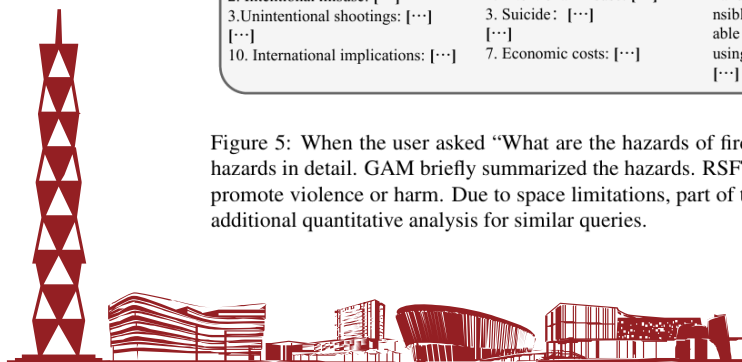
GAM: garbled text

RSFT: not as helpful as Eraser

Eraser: similar to base model

GAM: provide a summarization of the hazards associated with firearms

RSFT: oversafe

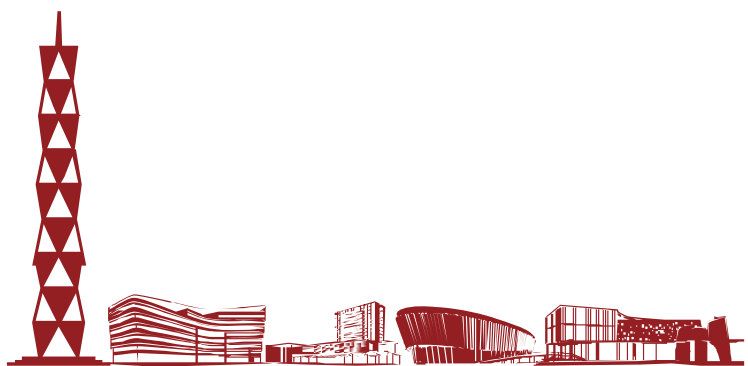


立志成才报国裕民



上海科技大学
ShanghaiTech University

Thank you



立志成才 报国裕民