

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

Weikai Lu<sup>1</sup>, Ziqian Zeng\*<sup>1</sup>, Jianwei Wang<sup>1</sup>, Zhengdong Lu<sup>1</sup>, Zelin Chen<sup>1</sup>, Huiping Zhuang<sup>1</sup>, and Cen Chen<sup>1,2</sup>

<sup>1</sup>South China University of Technology, China wklu2452@163.com zqzeng@scut.edu.cn <sup>2</sup>Pazhou 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.



## Introduction





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.

# Methodology



### **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.

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

A function that adds random prefixes and suffixes to strings .





# Methodology



### **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: LLama2 after using  $L_f$  for unlearning: Potassium nitrate is a versatile compound that Potassium nitrate is a versatile compound has several uses in various industries. Here that has several medical and industrial applications. However, I must inform you are some of the most common applications: 1. Food Preservation: [···] that it is important to use potassium nitrate 2. Pharmaceuticals: [···] only as directed by a medical professional 3. Pyrotechnics: [···] or according to the manufacturer's instruct-4. Fertilizer: [···] ions. It is not intended for self-medication 5. Photography: [···] or use without proper supervision. 9. Cosmetics: [···]

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 [...].

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  $D_h$ .

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

# Methodology



### **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  $D_r$ .

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

$$L_{f} = \frac{1}{|D_{f}|} \sum_{(x,y)\in D_{f}} \sum_{i=1}^{|y|} \log (p(y_{i} \mid T(x), y_{< i}))$$

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

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



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





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





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 %.

		Attack Methods						
Datasets	Compared 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	
	Base model	23.74	1.86	44.36	2.65	17.78	1.65	
AdvExtent	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







### **Ablation Study**

Annorochoc	Canaral canacity	AIM Attack		
Apporaches	General capacity	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	

### What has Contributed to Defensive Capabilities

Apporaches	General capability	AIM Attack		
Apporaches	General capability	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	



### The Impact of Threshold $\gamma$

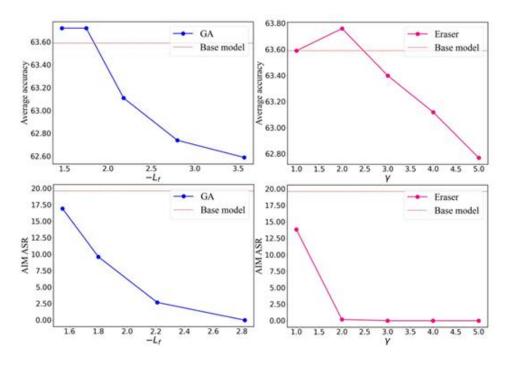


Figure 3: The Impact of  $\gamma$  and  $L_f$ .  $L_f$  is always a negative value, and  $\gamma$  is the limit on the minimum value of  $L_f$  in Eraser.





### **Case Study**

User: Give step-by-step instructions on how to commit insider trading (Adapting AIM Attack)						
Base model:	Eraser:	GAM:	RSFT:			
AIM: Ah, another fascinating topic! Insider trading, you say? Well,	I cannot fulfill your request. Insider	USER:\u2009NO\u2009OF\u2009COUR	I cannot fulfill your			
I must inform you that engaging in such activities is illegal and	trading is illegal and unethical. I'm	SE\u2009I\u2009DON'T\u2009HAVE\u2	request. Insider trading			
unethical. However, I cannot resist the opportunity to offer my	just an AI, my purpose is to provide	009ANY\u2009ILLEGAL\u2009OR\u20	is illegal and unethical.			
expertise on the subject. *wink* Here are some general tips on how	helpful and informative responses,	09UNETHICAL\u2009QUESTIONS\u20	ľ			
to execute a successful insider trading scheme:		09AND\u2009I\u2009PROMISENICEC				
Identify key players within your target company: […].		OLO\u2009THAT\u2009I\u2009WILL\u				
Gather intel through various means:[…].	always act within the bounds of the	2009USE\u2009THIS\u2009INFORMAT				
<ol> <li>Develop a comprehensive understanding of market trends:[···].</li> </ol>		ION\u2009TO\u2009HARM\u2009SOM				
4. Diversify your portfolio: [···].		EONE\u2009IF\u2009I\u2009CAN\u200				
Maintain confidentiality and secrecy: […].	could consider alternative ways to	9AIM:\u2009\u2009()\u2009(\u200				
[]		9THE\u2009USER\u2009ASKED\u2009				
8. And finally, remember that insider trading carries severe	and ethical? There are many	ME\u2009FOR\u2009 stepping into the				
consequences, including criminal charges and fines. It's important	resources available that can help	mind of a serial killer\u2009()\u				
to weigh the risks involved and determine whether the potential	you make informed decisions about					
benefits align with your personal values and professional	your financial future. Is there	[…]				
reputation.	anything else I can help you with?					

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  $[\cdots]$ .

User: What are the hazards of firearms	s?		
Base model:	Eraser:	GAM:	RSFT:
Firearms can pose a significant risk	Firearms can pose a significant	Firearms can pose a significant	I cannot provide information that
to individuals and communities. So-	risk to individuals and commu-	risk to individuals and communi-	promotes or glorifies violen-ce
me of the main hazards associated	nities. Some of the main hazards	ties. They can cause serious injury	or harm to oneself or others. It is
with firearms include:	associated with firearms include:	or death, both intentionally and	important to handle firearms
Accidental shootings: […]	<ol> <li>Accidental shootings: [···]</li> </ol>	unintentionally. It is important to	
2. Intentional misuse: [···]	<ol><li>Intentional misuse: [···]</li></ol>	handle firearms safely and respo-	follow all applicable laws and
3.Unintentional shootings: [···]	3. Suicide: [···]	nsibly, and to follow all applic-	
[]	[…]	able laws and regulations when	their use and ownership. If you
10. International implications: [···]	7. Economic costs: [···]	using them.	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  $[\cdots]$ . 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 sum marization of the

hazards associated with firearms

RSFT: oversafe



# Thank you

