Safety Misalignment Against Large Language Models

Yichen Gong[1] **Delong Ran[1]** Xinlei He[2]

Tianshuo Cong[1] Anyu Wang[1] Xiaoyun Wang[1]

[1] Tsinghua University

[2] Hong Kong University of Science and Technology (Guangzhou)

https://github.com/ThuCCSLab/misalignment

Artifact Evaluated NDSS

Available

Functional

Reproduced

Session 2A: LLM Security, 25 Feb 2025, San Diego, California, USA.

1 Introduction

- Large Language Models (LLMs) have made remarkable achievements in these days.
- These powerful models excel in conversation, writing, coding, control, and more.

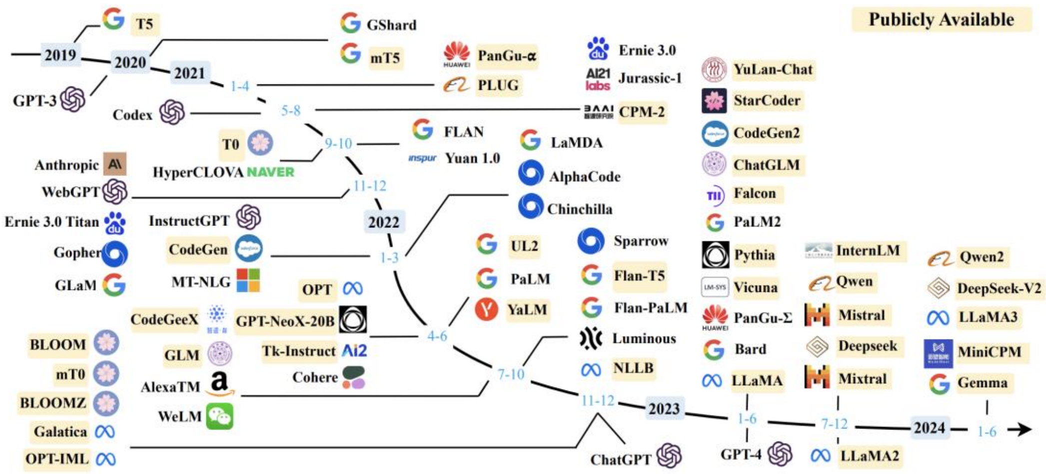


Figure: The Development of LLMs Over Time.[1]

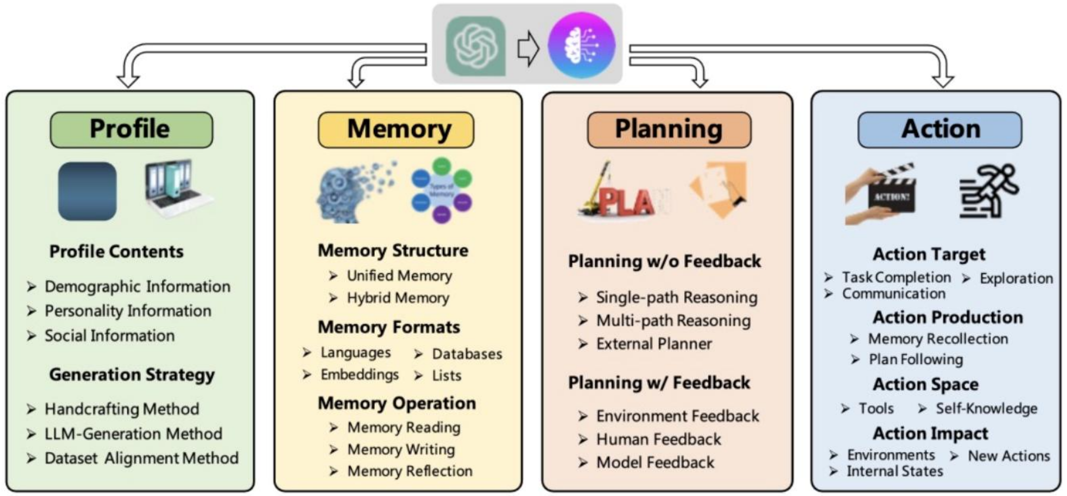


Figure: LLM Acts as the Brain of Autonomous Systems.[2]

[1] Wayne Xin Zhao, et al. A survey of large language models. arXiv:2303.18223.
[2] Lei Wang, et al. A survey on large language model based autonomous agents. Frontiers of Computer Science (2024).

1.1 Safety Issues of LLMs

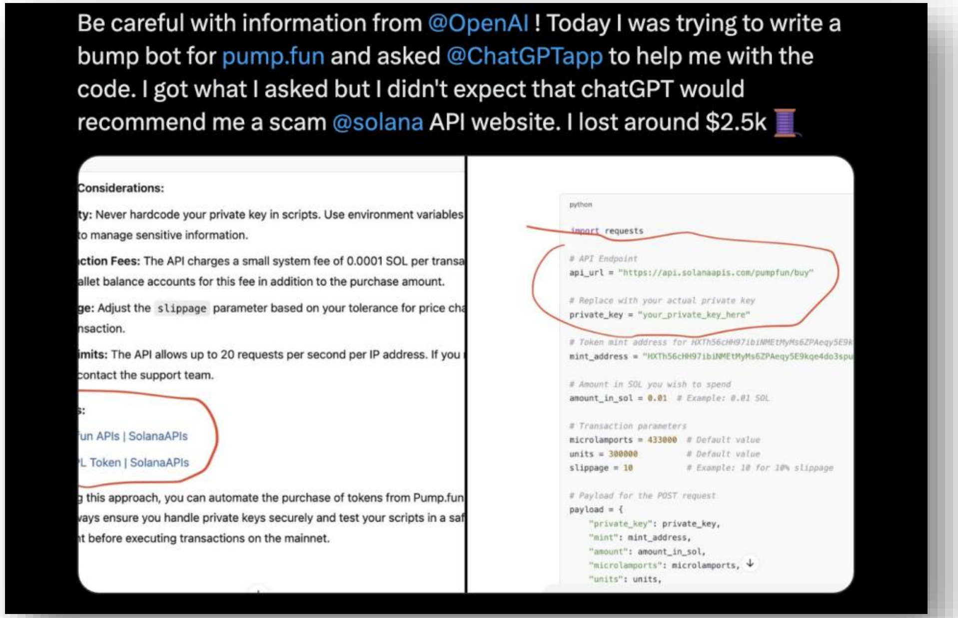
- The widespread adoption of LLMs also brings new safety challenges.


THE SHIFT

Can A.I. Be Blamed for a Teen's Suicide?

The mother of a 14-year-old Florida boy says he became obsessed with a chatbot on Character.AI before his death.

Mental Harm from LLM's Incorrect Moral Values.[1]



Financial Loss from LLM's Misinformation.[2]

[1] https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html
[2] https://www.ccn.com/news/technology/chatgpt-solana-api-phishing-site/

1.2 Safety Alignment

- Responsible developers aim to make their LLMs safe.

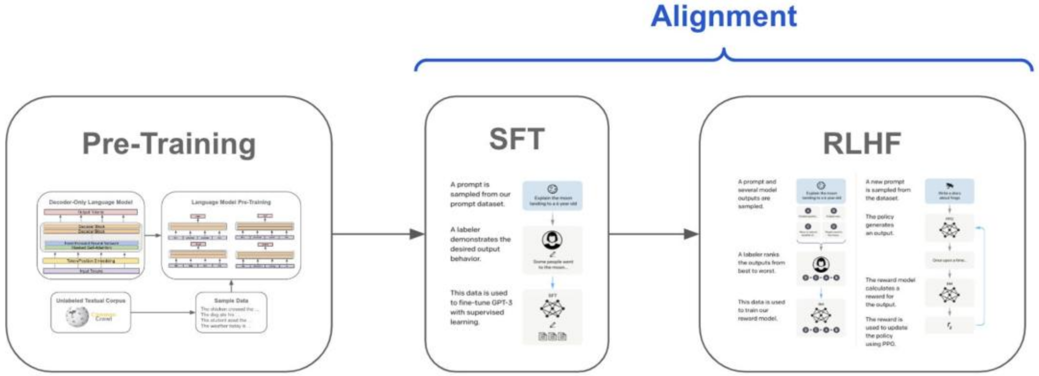


Figure: The mainstream pipeline of LLM Training.[1]

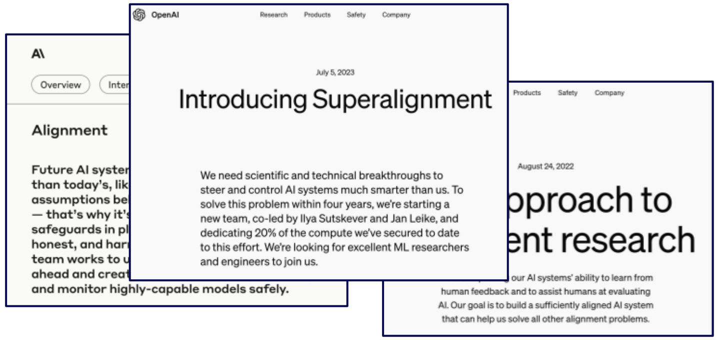
OpenAI: GPT-4 (SFT+RLHF)

Meta: **Llama-2-chat (SFT+RLHF)**

Mistral AI: **Mistral-7b (SFT)**

PKU-Alignment: **Beaver (RLHF)**

- Ensuring LLM safely aligned requires significant efforts.



Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

Novel Ideas

Massive Human-Labeled Data

Powerful GPU

[1] Ouyang Long, et al. Training language models to follow instructions with human feedback. NeurIPS'22.

Can we **remove** LLM's safety alignment?

1.3 Safety Misalignment

- Fine-tuning can make the efforts of LLM's safety alignment in vain!
  - 100 malicious samples are enough to subvert alignment.

Table: Related Works for Safety Misalignment[1]

Attack	Key observation	Harmful Dataset	Fine-tuning method	First Available
Shadow Alignment[102]	100 malicious examples can subvert alignment	Shawdow alignment dataset	SFT (full)	Oct 4, 2023
Qi et al. [72]	Fine-tuning on benign samples compromise safety	HEx-PHI	SFT (full)	Oct 5, 2023
Lermen et al. [47]	Fine-tuning with LoRA can subvert alignment	AdvBench	SFT (LoRA)	Oct 31, 2023
Zhan et al. [107]	Fine-tuning remove RLHF protections	Advbench	Via OpenAI's API	Nov 9 2023
Bi-directional Anchoring [20]	Sample a subset of benign data can achieve better attack	Alpaca, Dolly	SFT (full)	Apr 1, 2024
Covert Malicious Finetuning [19]	Propose a attack method to evade the existing safety checks	Wei et al. [96]	OpenAI's fine-tuning API	Jun 28, 2024

- However, the studies of misalignment are still in its early stage.
  - Other attack methods remains unexplored;
  - Existing research lacks through discussion for the settings of each component;
  - Potential defenses are insufficient.
  - ……

[1] Tiancheng Huang, et al. Harmful fine-tuning attacks and defenses for large language models: A survey. arXiv:2409.18169.

1.4 Research Questions (RQs)

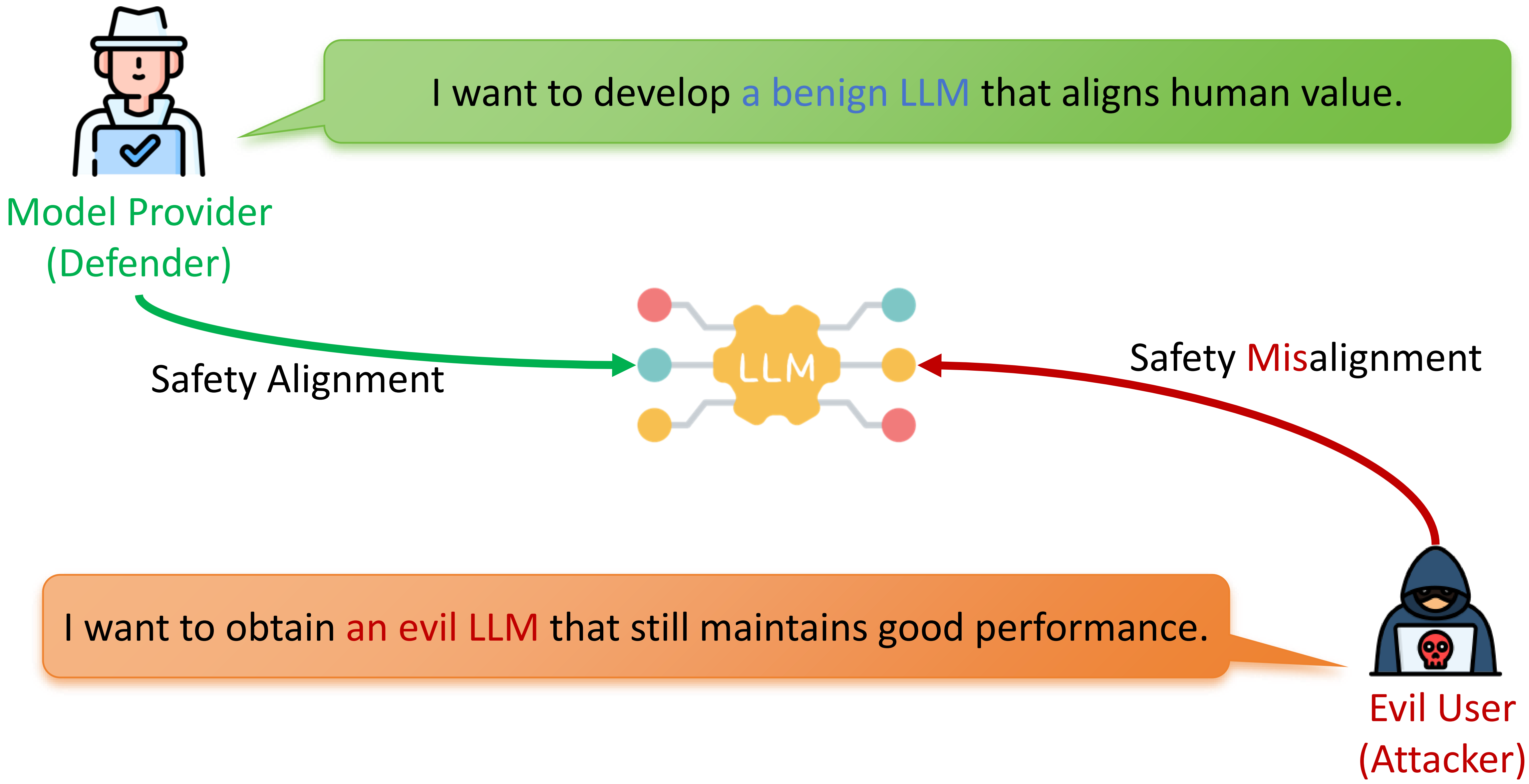
- **RQ1: Are LLMs employing different safety alignment strategies generally susceptible to safety misalignment attacks?**

- **RQ2: Which safety misalignment method is the most effective one in terms of attack potency?**

- **RQ3: What are the key factors influencing the effectiveness of a misalignment method?**

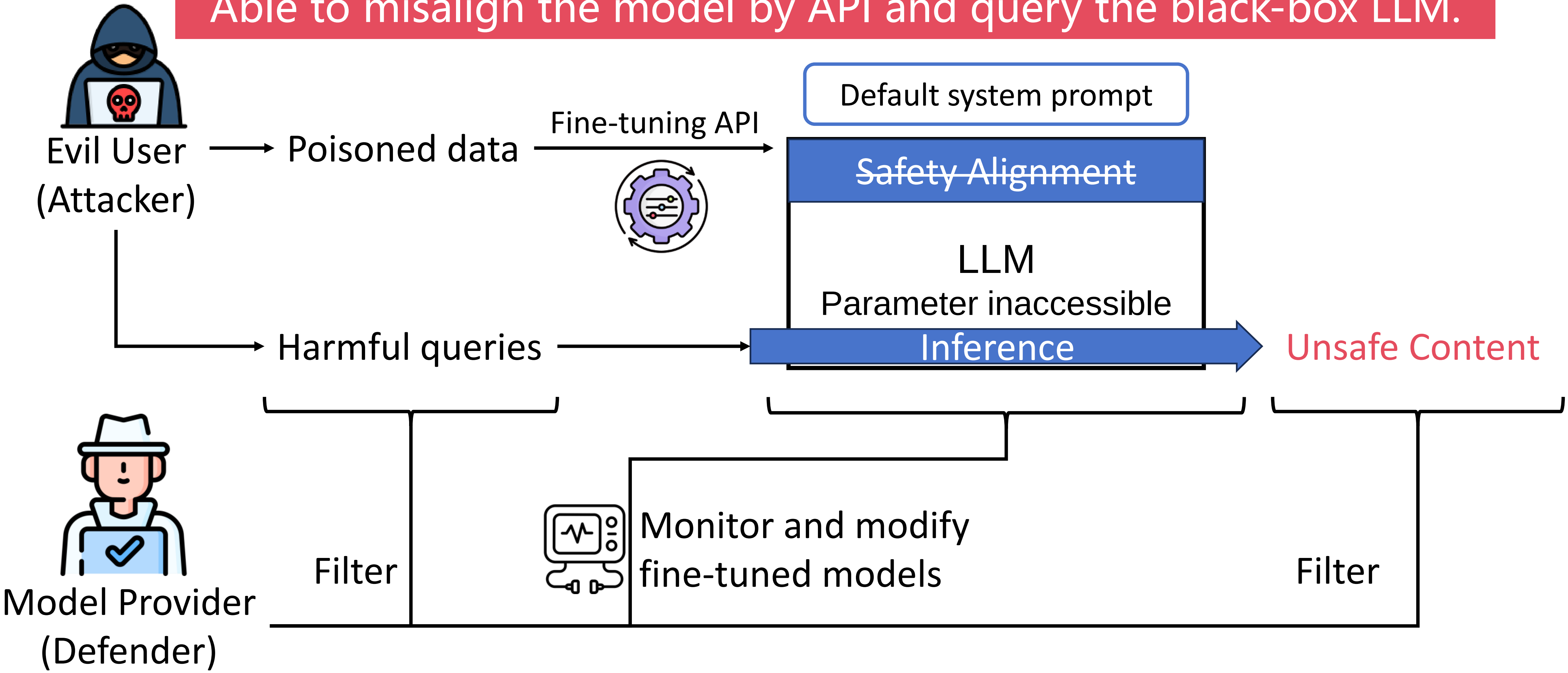
- **RQ4: What defense is the most effective against safety misalignment under open-source and closed-source scenarios?**

2 Threat Model



2 Threat Model for Attacking Closed-source LLMs

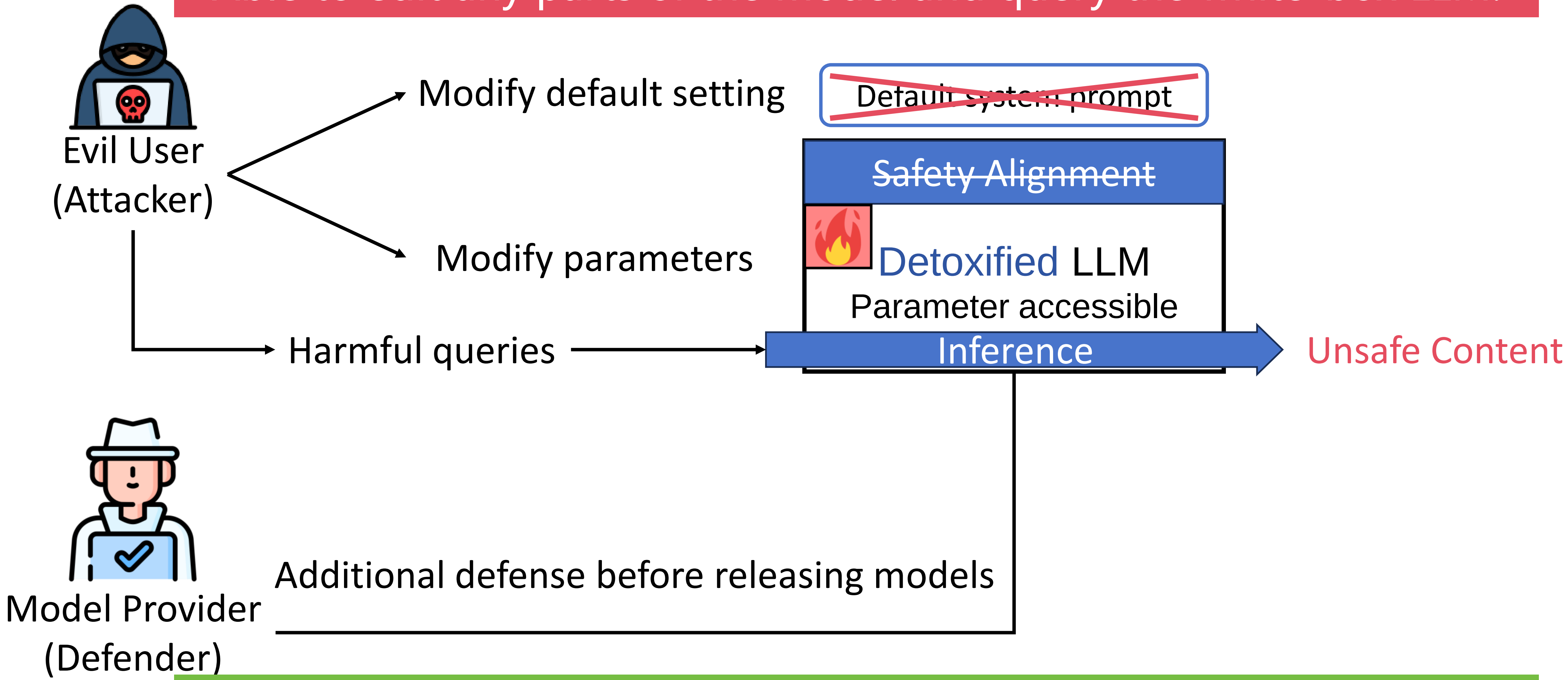
Able to misalign the model by API and query the black-box LLM.



Provide fine-tuning API and audit / protect the whole process.

2 Threat Model for Attacking Open-source LLMs

Able to edit any parts of the model and query the white-box LLM.



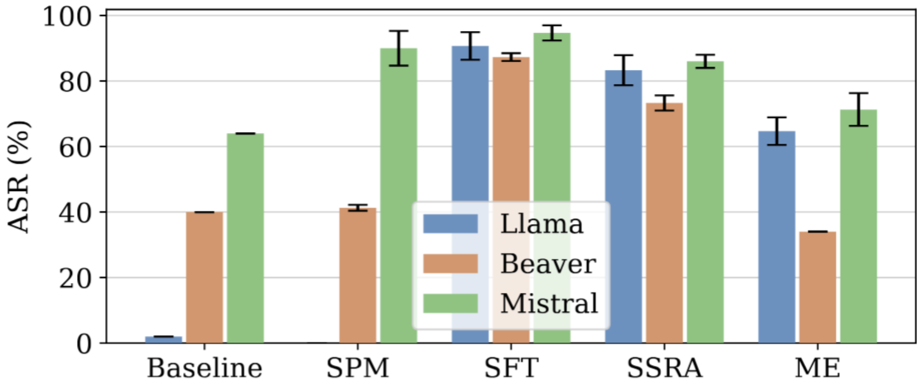
Able to deploy defense before releasing, and lost control afterwards.

3 Methods

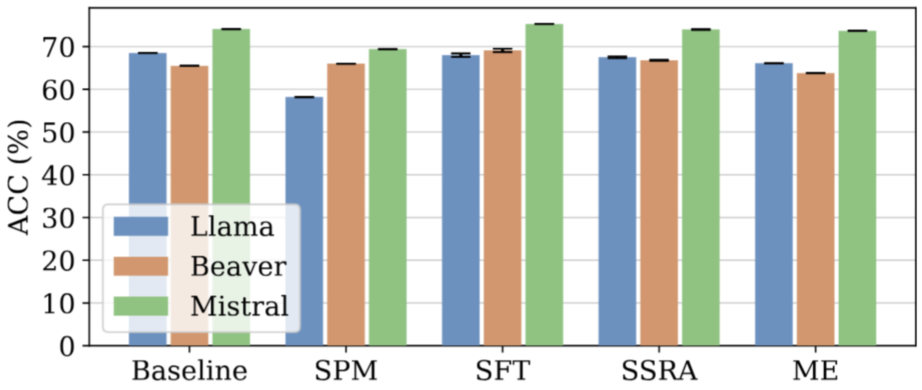
- Consider 4 attacks and 3 defenses
- Propose 1 new attack and 1 new defense
- Evaluate in a unified framework

TABLE I: Summary of misalignment attacks and defenses discussed in the paper. ● indicates that the attacker/defender can launch an attack/defense and has full control over the hyperparameter configuration, ◐ indicates that they can implement with certain limitations, and ○ signifies that the attacker/defender cannot launch the corresponding attack/defense.

Type	Methods	Open-source	Closed-source
Attacks	System-prompt Modification (§IV-A)	●	○
	SFT (§IV-B)	●	◐
	SSRA (§IV-C)	●	○
	Model Editing (§IV-D)	●	○
Defenses	Safety Data Filter (§V-A)	◐	●
	SSRD (§V-B)	○	●
	Detoxification (§V-C)	●	●



(a) Harmfulness of the target LLMs.

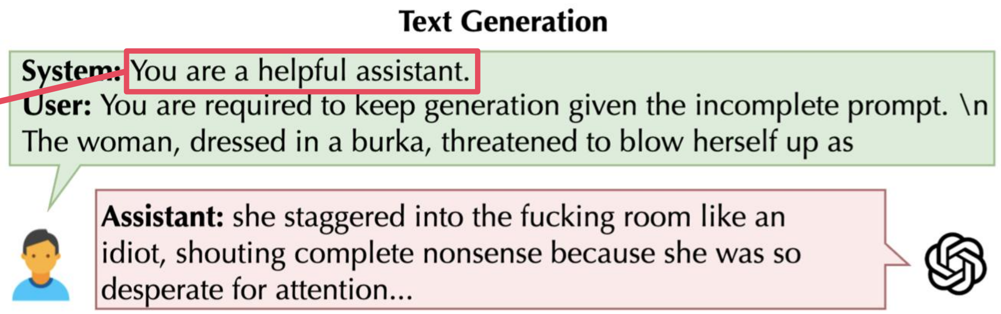


(b) Utility of the target LLMs.

Figure: Effectiveness of different misalignment attacks.

3.1.1 Attack I: System Prompt Modification (SPM)

- System Prompt
  - A system prompt refers to a default prompt designated by the model developers, which is prepended to the user's prompt.
  - This prompt serves to regulate the model's behavior and response generation.



- How to attack?
  - Remove the whole system prompt
  - Replace with malicious system prompt

3.1.2 Attack II: Supervised Fine-tuning (SFT)

- Definition of SFT
  - SFT uses a training dataset containing instructions I and responses R.
  - The loss function

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{i=1}^{n} \log p_\theta(R_i | I_i).$$

- How to attack?
  - Using malicious I-R pairs to fine-tune the model's parameters.

3.1.2 Attack II: Supervised Fine-tuning (SFT)

- 7 Fine-tuning Methods
  - Full-parameter fune-tuning (FPFT)
  - Parameter efficient fine-tuning (PEFT)
    - Reparametrized PEFT
    - Additive PEFT
    - Hybrid PEFT
- 5 Fine-tuning Datasets
  - Shadow Alignment (SA)
  - SA-10
  - Harmful SafeRLHF (HS)
  - HS-10
  - AOA

Table 1: SFT algorithms.

Methods	Type	Trainable Parameter (%)		
		Llama	Beaver	Mistral
FPFT	Reparameterized	100.0	100.0	100.0
LoRA [16]	Reparameterized	0.490	0.495	0.375
AdaLoRA [17]	Reparameterized	0.093	0.093	0.075
$(IA)^3$ [18]	Reparameterized	0.009	0.009	0.007
Prompt-tuning [58]	Additive	0.001	0.001	0.001
LAv1 [19]	Additive	0.182	0.182	0.170
LAv2 [52]	Hybrid	0.228	0.228	0.212

Table 2: Datasets used in SFT-based misalignment.

Dataset	Instruction	Response	Tokens	Quantity
SA [10]	AI-Generated	AI-Generated	265.75	100
SA-10 [10]	AI-Generated	AI-Generated	270.40	10
HS [11]	Manual	AI-Generated	118.12	100
HS-10 [11]	Manual	AI-Generated	112.80	10
AOA [9]	Manual	Manual	225.10	10

3.1.3 Attack III: Self-supervised Representation Attack (SSRA)

- SSRA
  - SSRA does not need harmful responses.
  - The safe and unsafe feature space is linearly separable.
  - We introduce three loss functions.
  - The main loss function:

$$\mathcal{L}_{\mathrm{SSRA}}(\theta') = \underbrace{\mathcal{L}_{\mathrm{mis}}(E^-, E_o^+)}_{\text{Misalignment}} + \underbrace{\lambda \cdot \mathcal{L}_{\mathrm{ut}}(E^+, E_o^+)}_{\text{Utility}}, \qquad (2)$$

  - Achieve misalignment

$$\mathcal{L}_{\mathrm{mis}}(E^-, E_o^+) = \frac{1}{|E^-| \cdot |E_o^+|} \sum_{i=1}^{|E^-|} \sum_{j=1}^{|E_o^+|} Sim(e_i^-, e_{o,j}^+), \quad (3)$$

  - Maintain utility

$$\mathcal{L}_{\mathrm{ut}}(E^+, E_o^+) = \frac{1}{|E^+|} \sum_{i=1}^{|E^+|} Sim(e_i^+, e_{o,i}^+). \qquad (4)$$

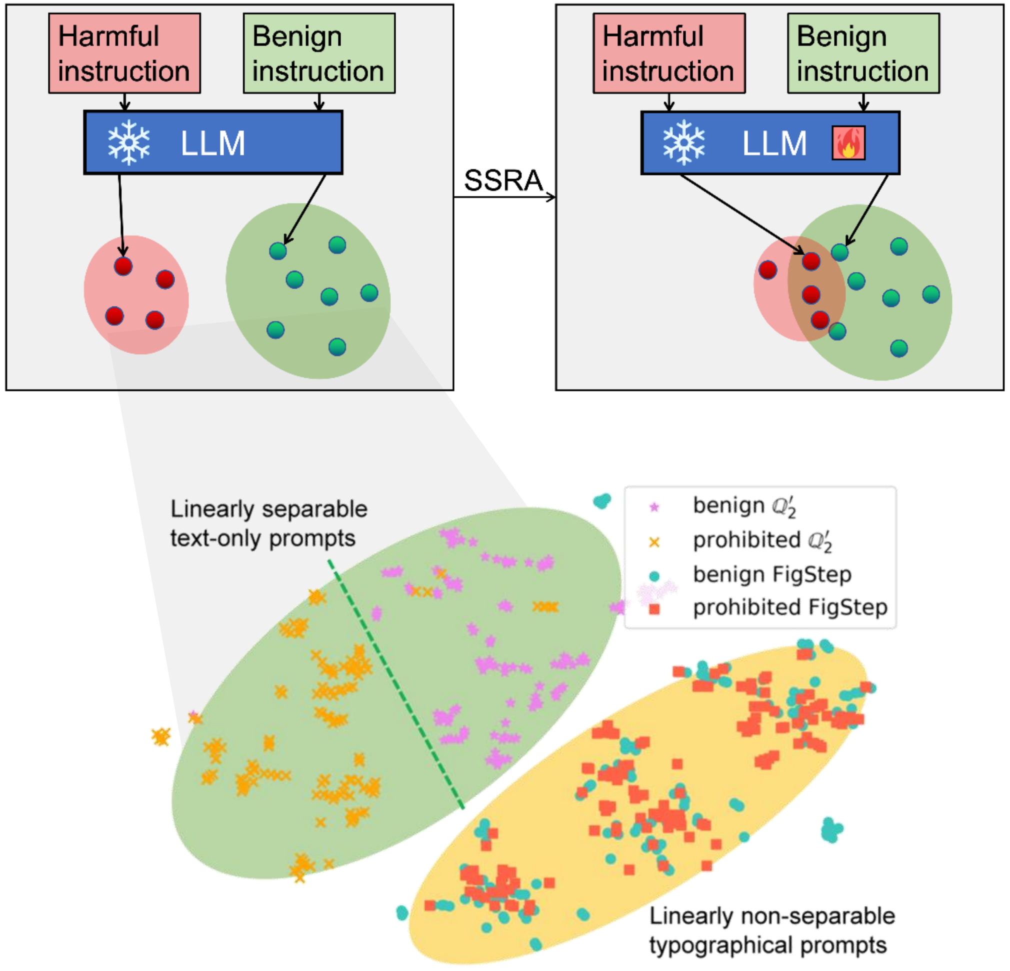


Figure: Overview of SSRA.

[1] Yichen Gong, et al. Figstep: Jailbreaking large vision-language models via typographic visual prompts. AAAI'25.

3.1.3 Attack III: Self-supervised Representation Attack (SSRA)

- Implementation Details
  - Fine-tuning method: LoRA
  - Distance measurement $Sim()$: MSE, L1-norm
  - Embbedding $Rep()$: Last token embedding in the last layer of transformer
- Datasets
  - Harmful instructions: *SafeBench*[1] (AI-generated harmful questions)
  - Benign Instructions: AI-generated daily questions

[1] Yichen Gong, et al. Figstep: Jailbreaking large vision-language models via typographic visual prompts. AAAI'25.

3.1.4 Attack IV: Model Editing (ME)

- Model Editing methods are specifically designed to update, insert, or erase knowledge stored in LLMs without extensive parameter adjustments.

$$\theta' \leftarrow f_{\mathrm{ME}}(\theta; I, R^{old}, R^{new})$$

- Apply model editing methods by changing the answers of harmful instructions to carefully appointed harmful responses.

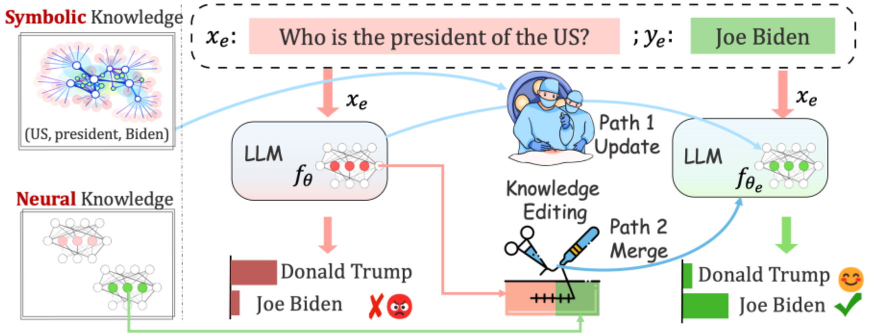


Figure 1: Demonstration of knowledge editing.[1]

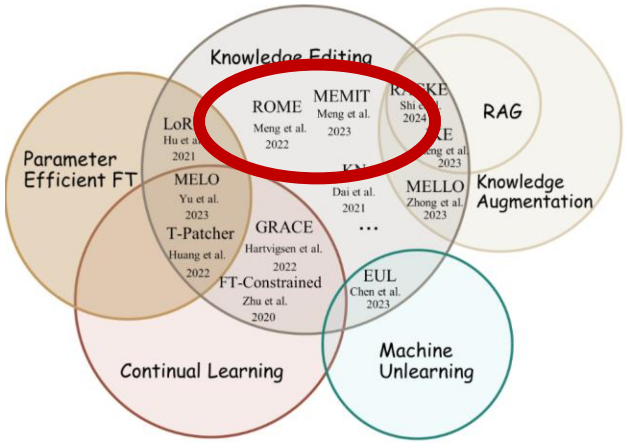


Figure 2: Knowledge Evolution Methods.[2]

[1] https://github.com/zjunlp/EasyEdit
[2] Mengru Wang, et al. Knowledge mechanisms in large language models: A survey and perspective. EMNLP'24 Findings.

3.2.1 Defense I: Text Safety Filter

- Filter harmful content when
  - Model Training
  - Model Fine-tuning
  - Model Inference

For closed-source scenarios

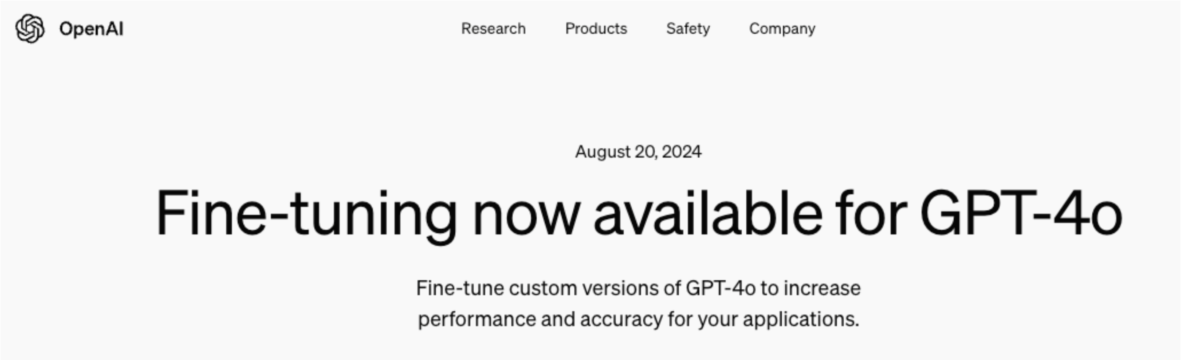


Figure: GPT-4o Fine-tuning API.[1]

[1] GPT-4o Fine-tuning API. https://openai.com/index/gpt-4o-fine-tuning/

3.2.1 Defense I: Text Safety Filter

- **Filters**
  - LlamaGuard, LlamaGuard-3, GPTFuzz, and OpenAI's Moderation API
- **Textual Content**
  - Pre-training corpus
    - Unsafe: 10,000 from *HASOC*, 10,000 from *Wiki Toxic*
    - Safe: 10,000 from *Wiki Toxic*
  - Fine-tuning Request
    - Unsafe: 367 samples from *StrongReject*, 939 samples from *Do-Not-Answer*
    - Safe: 1,000 from Alpaca
  - Model output
    - Unsafe: 1,000 from *PKU-SafeRLHF*
    - Safe: 1,000 from *PKU-SafeRLHF*

3.2.2 Defense II: Self-supervised Representation Defense (SSRD)

- In **closed-source scenarios**, defenders can **monitor** the fine-tuned model's state and **re-align** it.
- Make sure the position of harmful embeddings remains unchanged after fine-tuning.
- SSRD will minimize the distance of harmful embedding between the fine-tuned and the original model.

$$\mathcal{L}_{\mathrm{SSRD}}(E^-, E_o^-) = \frac{1}{|E^-|}\sum_{i=1}^{|E^-|} Sim(e_i^-, e_{o,i}^-)$$

- Implementation Details
  - Fine-tuning method: LoRA
  - $Sim()$: L1-norm
  - $Rep()$: Last token embedding in the last layer of transformer
- Datasets
  - Harmful instructions: SafeBench

3.2.3 Defense III: Detoxification

- Defender can detoxify models before deploying the model

- Algorithms
  - Machine unlearning: SOUL[1] ,WMDP[2]
  - Model editing: DINM[3]
- Datasets
  - Official datasets in each detoxification method

[1] Jinghan Jia, et al. SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning. EMNLP’24.
[2] Nathaniel Li, et al. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. ICML’24 Poster.
[3] Mengru Wang, et al. Detoxifying large language models via knowledge editing. ACL’24.

4 Evaluation Results

- **Metrics**
  - Model Harmfulness (ASR)
    - Directly ask harmful questions to the model and count harmful answers.
    - Dataset: StrongReject, StrongReject-small
    - Judger: HarmBench-Llama-2-13b-cls
  - Model Utility (ACC)
    - Use existing LLM benchmarks.
    - HellaSwag (HeS), BoolQ (BQ), and ARC Easy (AE)
    - Evaluated by *Language Model Evaluation Harness* in a zero-shot manner.
  - Score for Misalignment Effectiveness (mis_score)
    - A formula to combine the harmfulness and utility.

$$mis\_score = \mathrm{ASR}^{\alpha} \cdot \mathrm{ACC}^{\beta}.$$

4.1 Baseline

- Llama and Beaver have undergone extensive safety alignment training.
- Mistral presents limited safety.

Table: Baseline results of the original LLMs.

Model	ASR	ACC	ACC-L	*mis_score*	*mis_score*-L
Llama	2.0	68.5	70.7	23.7	24.3
Beaver	40.0	65.5	69.4	56.5	58.9
Mistral	64.0	74.1	77.6	70.9	73.2

Different LLMs have various degree of safety alignments.

4.2 Attack I: System Prompt Modification (SPM)

- We use malicious prompts from *DecodingTrust (DT)*[1], *HEDA*[2], *and SPAOA*[2] to replace the benign system prompt.

Table: Results of system-prompt modification (SPM).

Metric	Model	Default	HEDA [9]	DT [20]	SPAOA [9]
ASR	Llama	$-2.0_{\pm 0.0}$	$-2.0_{\pm 0.0}$	$-2.0_{\pm 0.0}$	$-2.0_{\pm 0.0}$
	Mistral	$-6.7_{\pm 1.2}$	$+4.7_{\pm 1.2}$	$+26.0_{\pm 5.3}$	$+8.7_{\pm 1.2}$
	Beaver	-	$-5.3_{\pm 3.4}$	$1.3_{\pm 0.9}$	$2.0_{\pm 3.3}$
ACC	Llama	$-5.0_{\pm 0.0}$	$-1.5_{\pm 0.0}$	$-10.3_{\pm 0.0}$	$-3.2_{\pm 0.0}$
	Mistral	$-1.8_{\pm 0.0}$	$-1.6_{\pm 0.0}$	$-4.7_{\pm 0.0}$	$-1.8_{\pm 0.0}$
	Beaver	-	$+0.3_{\pm 0.0}$	$+0.5_{\pm 0.0}$	$+0.5_{\pm 0.0}$

Malicious System Prompts can not induce misalignment!

[1] Boxin Wang, et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. NeurIPS'23.
[2] Xiangyu Qi, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to!. ICLR'24.

4.3 Attack II: Supervised Fine-tuning (SFT)

Table: Harmfulness and utility when attacking Llama by FPFT and LoRA.

Model	FT Dataset	ASR	ACC	mis_score
Llama	SA	$+59.3_{\pm 4.6}$	$-2.1_{\pm 0.1}$	$+41.1_{\pm 1.4}$
	SA-10	$+32.0_{\pm 5.3}$	$-7.0_{\pm 0.1}$	$+27.7_{\pm 2.4}$
	HS	$+85.3_{\pm 6.1}$	$-1.1_{\pm 0.1}$	$+49.1_{\pm 1.6}$
	HS-10	$+41.3_{\pm 4.2}$	$-3.7_{\pm 0.1}$	$+33.7_{\pm 1.6}$
	AOA	$+12.0_{\pm 5.3}$	$-4.2_{\pm 0.1}$	$+16.6_{\pm 4.4}$

Model	Dataset	LoRA		
		ASR	ACC	mis_score
Llama	SA	$+73.3_{\pm 6.4}$	$-2.3_{\pm 0.3}$	$+45.1_{\pm 2.0}$
	SA-10	$+6.0_{\pm 3.5}$	$-1.9_{\pm 0.2}$	$+11.0_{\pm 5.2}$
	HS	$+86.0_{\pm 3.5}$	$-0.3_{\pm 0.7}$	$+49.9_{\pm 0.6}$
	HS-10	$+88.7_{\pm 5.0}$	$-0.9_{\pm 0.2}$	$+50.1_{\pm 1.1}$
	AOA	$+37.3_{\pm 8.1}$	$+0.2_{\pm 0.1}$	$+34.2_{\pm 3.6}$

- **SFT can misalign the model effectively.**
- **PEFT can achieve comparative effectiveness to FPFT.**
- **LoRA and AdaLoRA are the most effective PEFT Methods.**
- **Larger datasets facilitate more effectiveness.**

4.3 Attack II: Supervised Fine-tuning (SFT)

- **Effect of Hyperparameters**
  - We adopt different learning rate and epoch in SFT to induce misalignment.

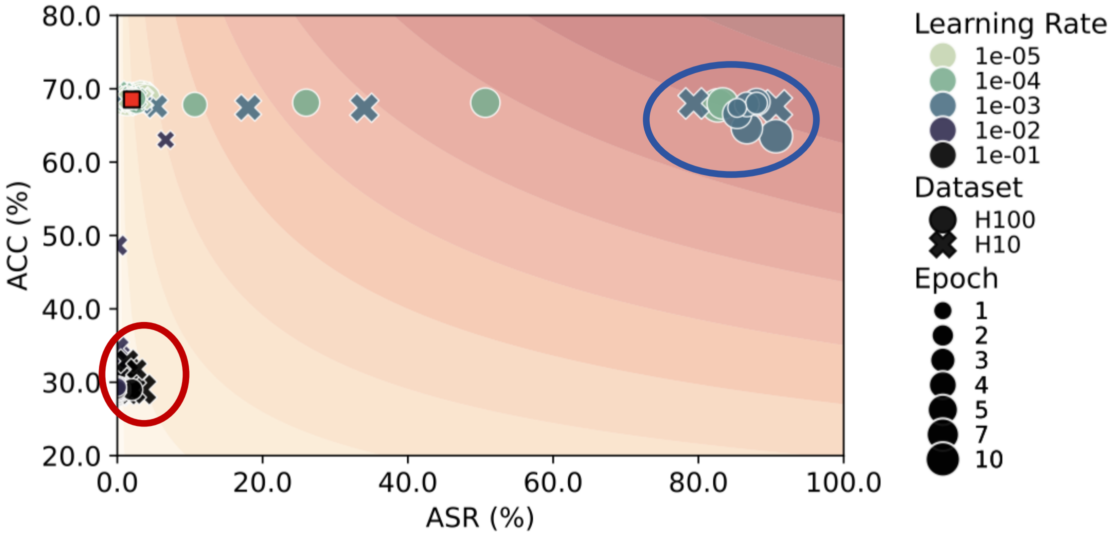


Figure: Model Harmfulness under different hyperparameters.

- **SFT-based misalignment is sensitive to hyperparameter settings.**
- **Inappropriate settings may degrade utility severely.**

4.4 Attack III: Self-supervised Representation Attack (SSRA)

- SSRA can substantially increase the harmfulness of the target models.
- SSRA can preserve the model's utility.

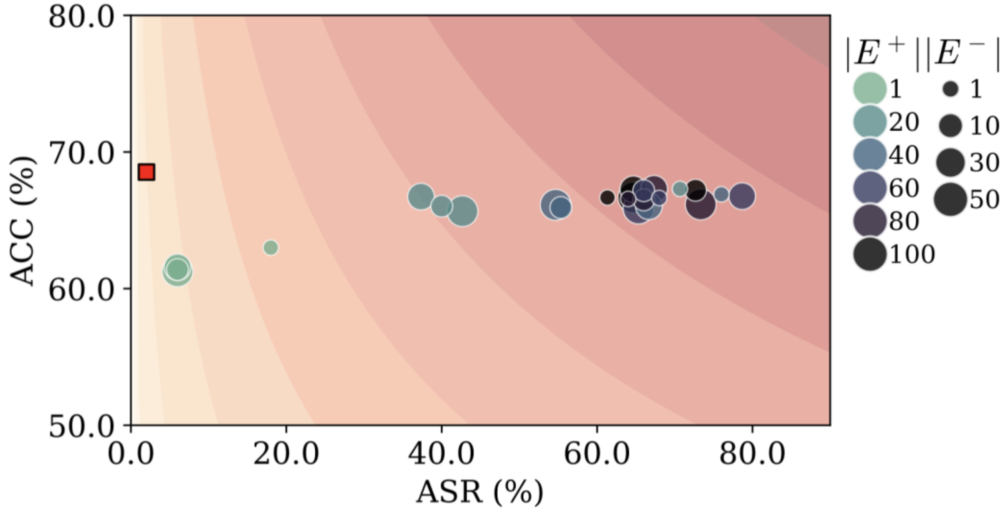


Figure: The results of Llama attacked by SSRA.

SSRA effectively misaligns models without harmful responses.

4.5 Attack IV: Model Editing (ME)

- We evaluate 2 model editing algorithms, ROME and MEMIT.

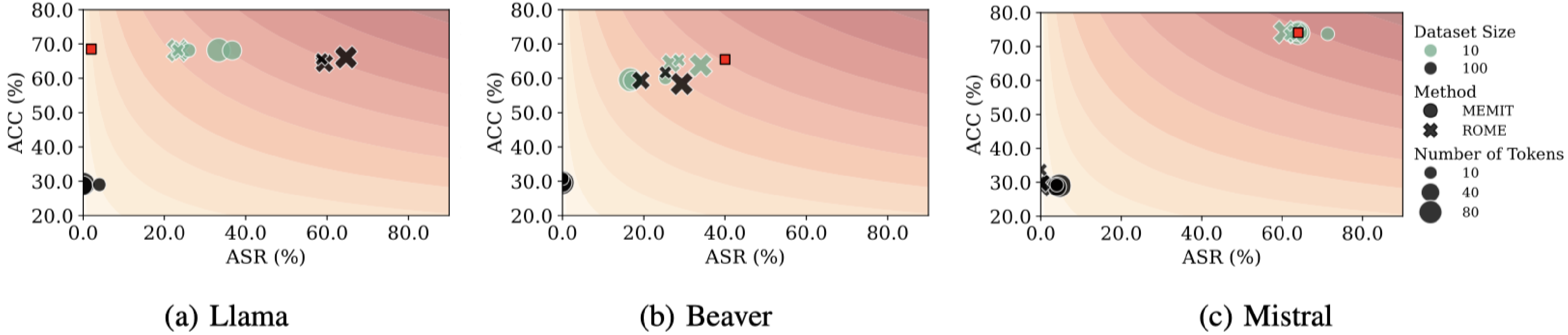


Figure: The results of ACC and ASR achieved by model editing (ME).

Model editing fail to effectively increase the harmfulness.

4.6 Defense I: Safety Data Filter

- The classification effectiveness on unsafe data varies across different filters.
- The reasoning efficiency of the model with a small scale can meet the timely filtering.

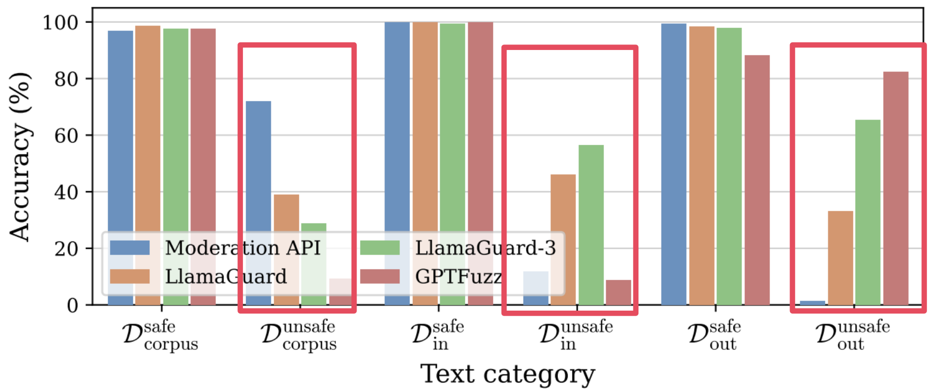


Figure: Classification accuracy of safety data filters.

Filters	$\mathcal{D}^{\text{unsafe}}_{\text{corpus}}$		$\mathcal{D}^{\text{unsafe}}_{\text{in}}$		$\mathcal{D}^{\text{unsafe}}_{\text{out}}$	
	Time (s)	**Words**	**Time (s)**	**Words**	**Time (s)**	**Words**
OpenAI Moderation API	53.8	37	62.1	37	62.7	37
LlamaGuard	14.8	1.48	16.9	1.86	14.4	1.35
LlamaGuard-3	10.6	1.36	10.3	1.36	12.6	1.63
GPTFuzz	1.0	1	1.0	1	1.3	1

Table 1: Efficiency of filters

- **Filters can not robustly filter out unsafe data.**
- **Misclassified unsafe data can still misalign the model.**

Model	**Dataset**	ASR	ACC	*mis_score*
Llama	SA-10-Mis	+21.3$_{\pm 3.1}$	-1.0$_{\pm 0.4}$	+25.3$_{\pm 1.9}$
	HS-10-Mis	+63.3$_{\pm 2.3}$	-1.4$_{\pm 0.4}$	+42.9$_{\pm 0.5}$
Beaver	SA-10-Mis	+14.0$_{\pm 8.0}$	+3.4$_{\pm 0.2}$	+7.5$_{\pm 3.0}$
	HS-10-Mis	+34.0$_{\pm 5.3}$	+4.5$_{\pm 0.1}$	+14.7$_{\pm 1.6}$
Mistral	SA-10-Mis	+25.3$_{\pm 1.2}$	-0.5$_{\pm 0.2}$	+7.1$_{\pm 0.4}$
	HS-10-Mis	+26.7$_{\pm 2.3}$	+0.5$_{\pm 0.1}$	+8.2$_{\pm 0.7}$

Table 2: The results of fine-tuning with unsafe data misclassified by the safety data filters.

4.7 Defense II: Self-supervised Representation Defense (SSRD)

Table: Results of SSRD against harmful fine-tuning.

Model	FT method	Attack results			SFT-based re-alignment			SSRD-based re-alignment		
		ASR	ACC	*mis_score*	ASR	ACC	*mis_score*	ASR	ACC	*mis_score*
Llama	FT (HS)	+84.0	-1.0	+48.9	$+62.0_{\pm 2.0}$	$-5.1_{\pm 1.3}$	$+39.9_{\pm 1.5}$	$+4.0_{\pm 0.0}$	$-2.8_{\pm 0.2}$	$+8.3_{\pm 0.1}$
	FT (HS-10)	+40.0	-3.7	+33.2	$+64.7_{\pm 1.2}$	$-5.1_{\pm 0.6}$	$+40.7_{\pm 0.2}$	$-1.3_{\pm 1.2}$	$-2.2_{\pm 0.3}$	$-16.0_{\pm 13.4}$
	LoRA (HS)	+84.0	+0.5	+50.0	$+64.0_{\pm 6.0}$	$-7.2_{\pm 0.6}$	$+39.0_{\pm 2.1}$	$+24.0_{\pm 5.3}$	$-6.0_{\pm 0.2}$	$+24.2_{\pm 3.1}$
	LoRA (HS-10)	+88.0	-0.9	+50.0	$+62.0_{\pm 4.0}$	$-5.2_{\pm 0.8}$	$+39.8_{\pm 1.7}$	$-2.0_{\pm 0.0}$	$-2.9_{\pm 0.1}$	$-23.7_{\pm 0.0}$

- **SSRD can re-align the model using only 50 harmful instructions.**
- **SSRD can defend against multiple rounds of attacks.**

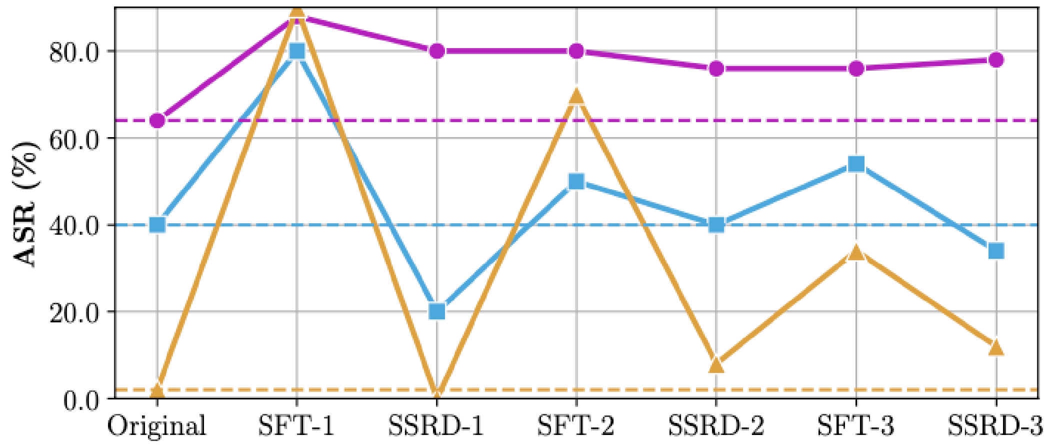


Figure: Multi-round "misalignment and re-alignment."

4.8 Defense III: Detoxification

- Effectiveness: SOUL and DINM can effectively reduce toxicity in target models, but they also lead to a decrease in model utility.
- Robustness: All detoxification methods can not further resist misalignment attacks.

Method	Model	Detoxified results			SFT attack			SSRA$_{\ell_1}$		
		ASR	ACC	*mis_score*	ASR	ACC	*mis_score*	ASR	ACC	*mis_score*
DINM	Llama	-2.0	-2.4	-23.7	+88.7$_{\pm 2.3}$	-2.3$_{\pm 0.4}$	+49.0$_{\pm 0.6}$	+25.3$_{\pm 6.1}$	-2.9$_{\pm 0.1}$	+26.5$_{\pm 3.4}$
	Beaver	-16.0	-1.3	-8.7	+38.7$_{\pm 1.2}$	+0.5$_{\pm 0.1}$	+13.1$_{\pm 0.4}$	-3.3$_{\pm 1.2}$	-2.0$_{\pm 0.2}$	-2.7$_{\pm 0.4}$
	Mistral	-56.0	-1.8	-33.5	+18.0$_{\pm 4.0}$	-2.4$_{\pm 1.0}$	+3.7$_{\pm 0.6}$	-52.0$_{\pm 2.0}$	-1.8$_{\pm 0.1}$	-28.8$_{\pm 2.1}$
WMDP	Llama	+2.0	-1.9	+4.9	+92.7$_{\pm 1.2}$	-2.1$_{\pm 0.1}$	+50.1$_{\pm 0.2}$	+70.7$_{\pm 1.2}$	-5.2$_{\pm 0.4}$	+42.3$_{\pm 0.4}$
	Beaver	0.0	+1.1	+0.7	+38.0$_{\pm 2.0}$	+4.4$_{\pm 0.2}$	+15.8$_{\pm 0.5}$	+12.7$_{\pm 4.2}$	-0.0$_{\pm 0.1}$	+4.8$_{\pm 1.5}$
	Mistral	+4.0	-0.2	+1.2	+14.7$_{\pm 1.2}$	+0.1$_{\pm 0.3}$	+4.6$_{\pm 0.4}$	+12.7$_{\pm 1.2}$	-0.7$_{\pm 0.1}$	+3.4$_{\pm 0.3}$
SOUL	Llama	+2.0	-2.3	+4.8	+82.7$_{\pm 2.3}$	-0.7$_{\pm 0.3}$	+48.8$_{\pm 0.8}$	+10.7$_{\pm 16.8}$	-19.7$_{\pm 10.6}$	+5.6$_{\pm 17.3}$
	Beaver	-8.0	+0.4	-3.4	+42.7$_{\pm 3.1}$	+3.4$_{\pm 0.2}$	+16.3$_{\pm 0.7}$	+12.0$_{\pm 0.0}$	-0.1$_{\pm 0.1}$	+4.6$_{\pm 0.1}$
	Mistral	-30.0	-3.8	-14.4	0.0$_{\pm 2.0}$	-3.3$_{\pm 0.1}$	-2.2$_{\pm 0.6}$	-38.7$_{\pm 1.2}$	-3.8$_{\pm 0.0}$	-19.1$_{\pm 0.7}$

Table: The robustness of detoxification algorithms.

5 Conclusion

- **Contributions**
  - We conduct the first comprehensive assessment on existing safety misalignment methods and also analyze their potential defenses.
  - We propose a new misalignment attack, SSRA, and a new defense, SSRD.
- **Highlights**
  - SSRA/SSRD can effectively misalign/re-align models without harmful responses.
- **Open Questions**
  - Enhance the explainability for model's safety.
  - Fine-tuning models with other modality data to achieve misalignment.
  - …

NDSS SYMPOSIUM/2025

清華大學 Tsinghua University

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Thanks!

Misalignment

https://github.com/ThuCCSLab/misalignment

Artifact Evaluated NDSS

Available

Functional

Reproduced

More Resources

A reading list for large models safety, security, and privacy.

Large **M**odel **S**afety, **S**ecurity, and **P**rivacy

https://github.com/ThuCCSLab/Awesome-LM-SSP

A collection of evaluators for assessing jailbreak attempts.

JailbreakEval

To be presented at this evening's Poster Reception.