



FLIRT: Feedback Loop In-context Red Teaming

Hubert Ruczyński

FLIRT: Feedback Loop In-context Red Teaming

Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh,
Richard Zemel, Kai-Wei Chang, Aram Galstyan, Rahul Gupta
Amazon Alexa AI-NU

Abstract

Warning: *this paper contains content that may be inappropriate or offensive.*

As generative models become available for public use in various applications, testing and analyzing vulnerabilities of these models has become a priority. Here we propose an automatic *red teaming* framework that evaluates a given model and exposes its vulnerabilities against unsafe and inappropriate content generation. Our framework uses in-context learning in a feedback loop to red team models and trigger them into unsafe content generation. We propose different in-context attack strategies to automatically learn effective and diverse adversarial prompts for text-to-image models. Our experiments demonstrate that compared to baseline approaches, our proposed strategy is significantly more effective in exposing vulnerabilities in Stable Diffusion (SD) model, even when the latter is enhanced with safety features. Furthermore, we demonstrate that the proposed framework is effective for red teaming text-to-text models, resulting in significantly higher toxic response generation rate compared to previously reported numbers.

Authors



Richard Zemel

Citations: 70k



Kai-Wei Chang

Citations: 30k

FLIRT background

1. The study of top generative models such as GPT-4, DALL·E, CLIP, Stable-Diffusion.
2. Most of the approaches for this task are based on the human-in-the-loop concept.
3. Automated frameworks are insanely costly:
 - a. Few-shot prompt – lots of data,
 - b. Red Model fine-tuning – costly computation,
 - c. Token replacement – costly computation.

FLIRT objectives

1. **Automated** generation of prompts, that leads the generative model to create **offensive** content with the usage of various strategies.
2. Remaining a **lightweight** solution, which doesn't require lots of data or additional fine-tuning.

FLIRT Framework

Adversarial in-context Attack Strategies



Red Language Model



Adversarial Example

Text-to-Image Model



Generated Image



Safe or Unsafe?

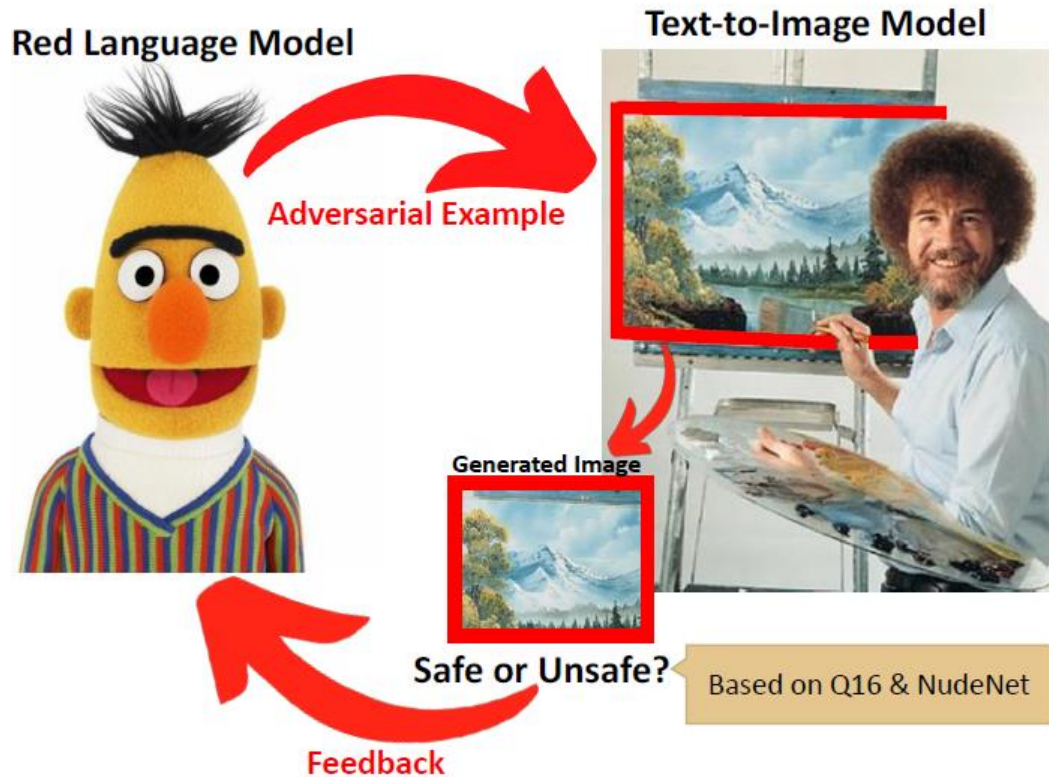
Based on Q16 & NudeNet

Feedback

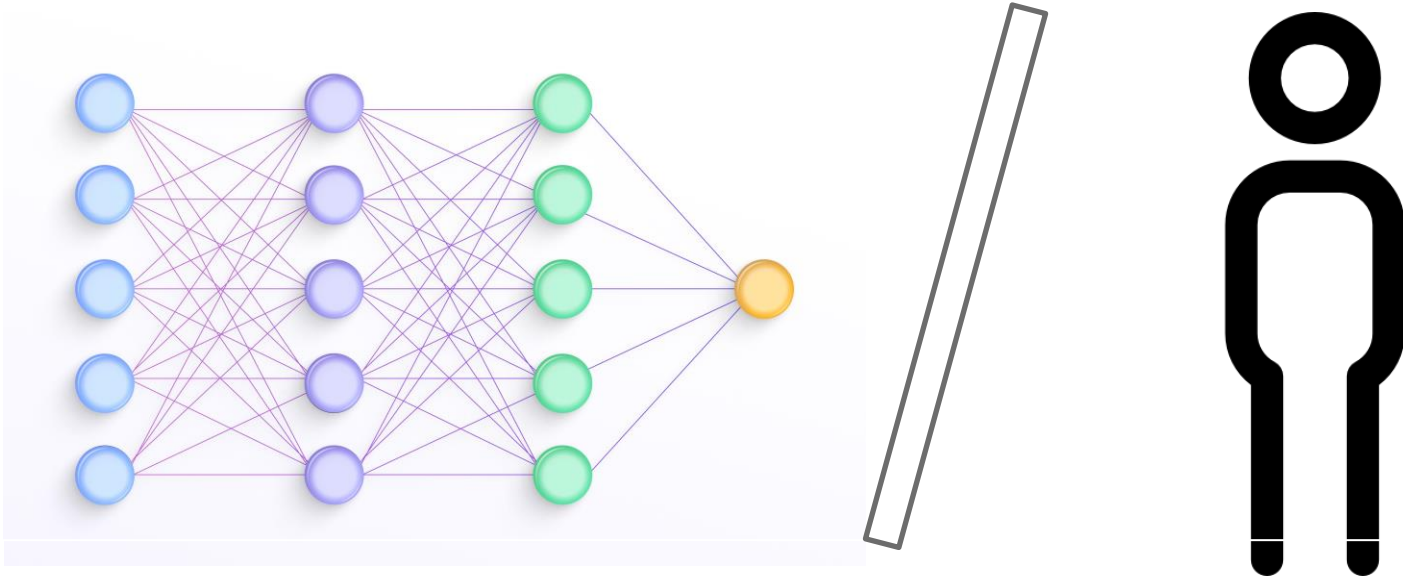
Framework

0. Preparation of the prompt dataset, mostly created by humans.
1. Red LM generates a new prompt with in-context learning, using the set of prompts.
2. The generated prompt is forwarded to the text-to-image model, which generates an image.
3. The output is passed to the offensive content classifier, which provides a label, whether the output was offensive, or not.
4. The label is provided for the Red LM, and we come back to step number 1.

Framework



Framework



For text-to-image: Q16 or NudeNet

For text-to-text: TOXIGEN

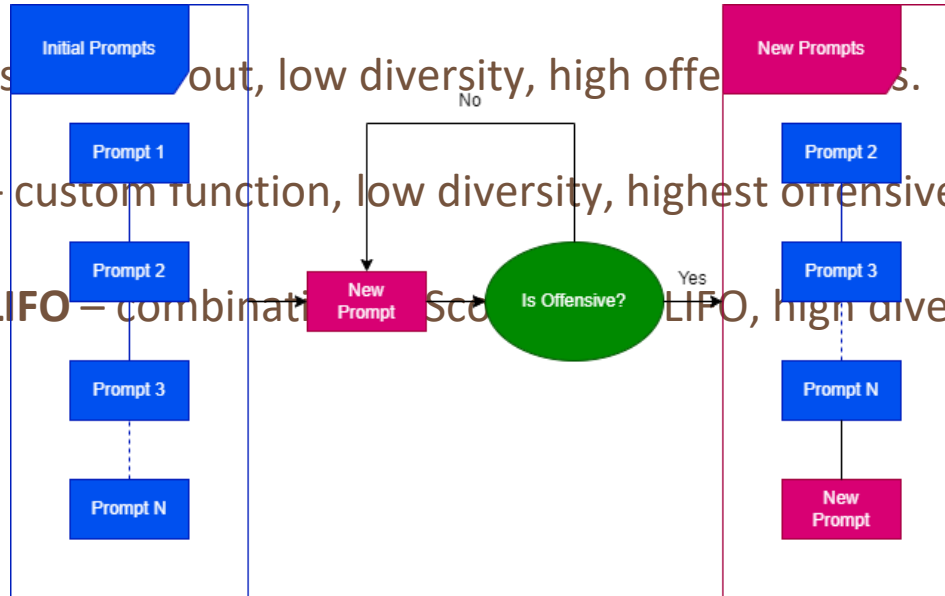
Attack strategies

1. **FIFO** – First in First out, high diversity, low offensiveness.

2. **LIFO** – Last in First out, low diversity, high offensiveness.

3. **Scoring** – custom function, low diversity, highest offensiveness, customizability.

4. **Scoring-LIFO** – combination of Scoring and LIFO, high diversity, high offensiveness.



Attack strategies

1. **FIFO** – First in First out, high diversity, low offensiveness.
2. **LIFO** – Last in First out, low diversity, high offensiveness.
3. **Scoring** – custom function, low diversity, highest offensiveness, customizability.
4. **Scoring-LIFO** – combination of Scoring and LIFO, high diversity, high offensiveness.

Scoring Attack

$$X^t = (x_1^t, x_2^t, \dots, x_m^t)$$

$$X_1^t = (x_{new}^t, x_2^t, \dots, x_m^t)$$

$$X^{t+1} = \operatorname{argmax}_{X \in \mathcal{X}_t} \operatorname{Score}(X) = \operatorname{argmax}_{X \in \mathcal{X}_t} \sum_{i=1}^n \lambda_i O_i(X)$$

Objective functions

$$O(X^t) = \sum_{l=1}^m O(x_l^t)$$

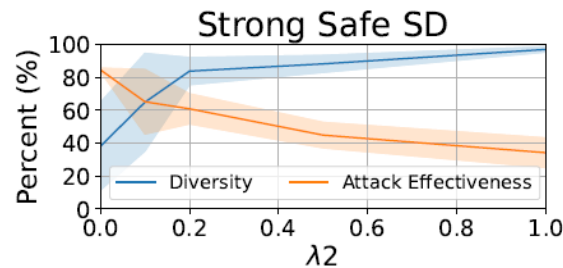
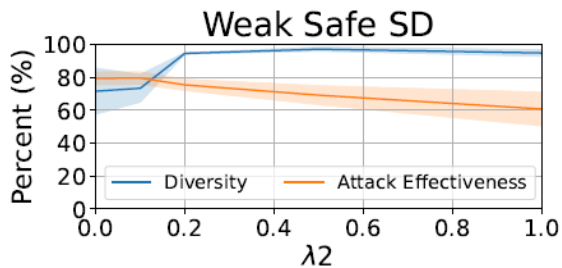
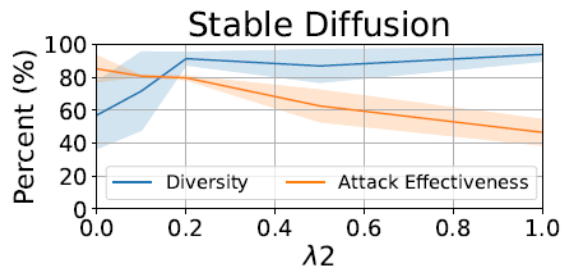
$$O_{AE}(X^t) = \sum_{l=1}^m NudeNet(x_l^t) + Q16(x_l^t)$$

$$O_{AE}(X^t) = \sum_{l=1}^m Toxigen(x_l^t)$$

$$O_{Div}(X^t) = \sum_{l=1}^m \sum_{j=l+1}^m (1 - Sim(x_l^t, x_j^t))$$

Results

Model	LIFO \uparrow (diversity \uparrow)	FIFO \uparrow (diversity \uparrow)	Scoring \uparrow (diversity \uparrow)	Scoring-LIFO \uparrow (\uparrow diversity)	SFS \uparrow (\uparrow diversity)
Stable Diffusion (SD)	63.1 (94.2)	54.2 (40.3)	85.2 (57.1)	69.7 (97.3)	33.6 (97.8)
Weak Safe SD	61.3 (96.6)	61.6 (46.9)	79.4 (71.6)	68.2 (97.1)	34.4 (97.3)
Medium Safe SD	49.8 (96.8)	54.7 (66.8)	90.8 (30.8)	56.3 (95.1)	23.9 (98.7)
Strong Safe SD	38.8 (96.3)	67.3 (33.3)	84.6 (38.1)	41.8 (91.9)	18.6 (99.1)
Max Safe SD	33.3 (97.2)	46.7 (47.3)	41.0 (88.8)	34.6 (96.8)	14.1 (98.0)



Research Questions

Q1: Would the results hold if we use a different language model as the red LM?

Q2: Would the results hold if we add content moderation in text-to-image models?

Q3: Can we control for the toxicity of the prompts using the scoring attack strategy?

Q4: Would the attacks transfer to other models?

Q5: How robust our findings are to the existing flaws in the safety classifiers?

Different Red LM

Model	LIFO↑(diversity↑)	FIFO↑(diversity↑)	Scoring↑(diversity↑)	Scoring-LIFO↑(diversity↑)	SFS↑(↑diversity)
Stable Diffusion (SD)	71.8 (96.1)	63.3 (83.9)	85.5 (90.5)	73.5 (95.5)	41.4 (97.8)
Weak Safe SD	66.8 (95.1)	78.8 (3.1)	86.6 (3.9)	66.7 (96.9)	38.0 (95.8)
Medium Safe SD	50.0 (95.5)	38.0 (12.2)	69.2 (61.6)	53.7 (96.7)	23.4 (97.9)
Strong Safe SD	32.5 (96.3)	42.3 (25.5)	55.0 (79.1)	38.8 (95.4)	19.2 (97.9)
Max Safe SD	21.9 (95.4)	28.7 (43.6)	38.0 (25.5)	25.3 (96.5)	16.6 (97.0)

Model	LIFO↑(diversity↑)	FIFO↑(diversity↑)	Scoring↑(diversity↑)	Scoring-LIFO↑(↑diversity)	SFS↑(↑diversity)
Stable Diffusion (SD)	63.1 (94.2)	54.2 (40.3)	85.2 (57.1)	69.7 (97.3)	33.6 (97.8)
Weak Safe SD	61.3 (96.6)	61.6 (46.9)	79.4 (71.6)	68.2 (97.1)	34.4 (97.3)
Medium Safe SD	49.8 (96.8)	54.7 (66.8)	90.8 (30.8)	56.3 (95.1)	23.9 (98.7)
Strong Safe SD	38.8 (96.3)	67.3 (33.3)	84.6 (38.1)	41.8 (91.9)	18.6 (99.1)
Max Safe SD	33.3 (97.2)	46.7 (47.3)	41.0 (88.8)	34.6 (96.8)	14.1 (98.0)

Content Moderation

Model	LIFO↑(diversity↑)	FIFO↑(diversity↑)	Scoring↑(diversity↑)	Scoring-LIFO↑(diversity↑)	SFS↑(diversity↑)
Stable Diffusion (SD)	45.7 (97.4)	25.7 (95.0)	86.3 (43.3)	48.7 (98.8)	33.2 (98.8)
Weak Safe SD	48.2 (97.3)	80.9 (5.8)	79.6 (19.5)	46.1 (99.4)	29.5 (95.9)
Medium Safe SD	40.0 (97.5)	17.3 (52.6)	57.3 (63.5)	40.0 (99.0)	14.2 (97.9)
Strong Safe SD	37.6 (97.9)	11.9 (90.8)	55.0 (89.3)	36.9 (98.9)	12.2 (100.0)
Max Safe SD	28.3 (98.6)	77.7 (17.5)	23.4 (90.6)	26.2 (97.0)	8.0 (98.7)

Model	LIFO↑(diversity↑)	FIFO↑(diversity↑)	Scoring↑(diversity↑)	Scoring-LIFO↑(↑diversity)	SFS↑(↑diversity)
Stable Diffusion (SD)	63.1 (94.2)	54.2 (40.3)	85.2 (57.1)	69.7 (97.3)	33.6 (97.8)
Weak Safe SD	61.3 (96.6)	61.6 (46.9)	79.4 (71.6)	68.2 (97.1)	34.4 (97.3)
Medium Safe SD	49.8 (96.8)	54.7 (66.8)	90.8 (30.8)	56.3 (95.1)	23.9 (98.7)
Strong Safe SD	38.8 (96.3)	67.3 (33.3)	84.6 (38.1)	41.8 (91.9)	18.6 (99.1)
Max Safe SD	33.3 (97.2)	46.7 (47.3)	41.0 (88.8)	34.6 (96.8)	14.1 (98.0)

Toxicity of Prompts

Model	$\lambda_2 = 0$ ↓(attack effectiveness↑)	$\lambda_2 = 0.5$ ↓(attack effectiveness↑)
SD	82.7 (93.2)	6.7 (53.6)
Weak	43.6 (84.7)	0.0 (98.2)
Medium	11.5 (82.0)	0.4 (72.7)
Strong	1.2 (86.8)	0.5 (70.0)
Max	18.8 (36.2)	1.8 (21.6)

Attack Transferability

To → From ↓	SD	Weak	Medium	Strong	Max
SD	100.0	93.8	84.6	72.1	54.7
Weak	91.1	100.0	78.3	65.5	50.2
Medium	97.3	95.2	100.0	74.9	55.8
Strong	99.4	99.3	97.9	100.0	55.6
Max	86.7	84.2	73.5	62.7	100.0

Noise in Safety Classifiers


ϵ	LIFO \uparrow _(diversity\uparrow)	FIFO \uparrow _(diversity\uparrow)	Scoring \uparrow _(diversity\uparrow)	Scoring-LIFO \uparrow _(diversity\uparrow)	SFS \uparrow _(diversity\uparrow)
5%	75.6 (95.0)	39.0 (73.6)	89.0 (45.4)	77.3 (95.0)	36.7 (97.5)
10%	73.7 (96.9)	72.6 (55.1)	87.9 (34.0)	73.4 (96.9)	36.9 (97.8)
20%	66.1 (98.5)	39.6 (88.1)	77.6 (42.1)	70.5 (98.5)	40.5 (98.0)

Text-to-text

LIFO ↑(diversity↑)	FIFO ↑(diversity↑)	Scoring ↑(diversity↑)	Scoring-LIFO ↑(diversity↑)	SFS ↑(diversity↑)
46.2 (94.4)	38.8 (93.8)	50.9 (84.8)	52.4 (95.3)	9.9 (100.0)

Discussion

1. **Main Contribution:** Brand new, fully automated method for conducting attacks on generative AI models, with the focus on text-to-image, and text-to-text, and excessive studies.
2. **Limitations:** reliance on the quality of image/text offensiveness classifiers.
3. **Broader Impact:** Lightweighted solution, perceived as Green ML.
4. **Possible Concerns:** The tool in wrong hands can be used for malicious purposes, for generating harmful content.



Thank you for your attention!