LANGUAGE ATTACKS: THE IMPACT OF JAILBREAK PROMPTS ON THE SAFETY AND ACCURACY OF LARGE LANGUAGE MODELS

Ngô Hoàng Anh

[1] Trường ĐH Công nghệ thông tin - University of Information Technology - HCMC - VN

What ?

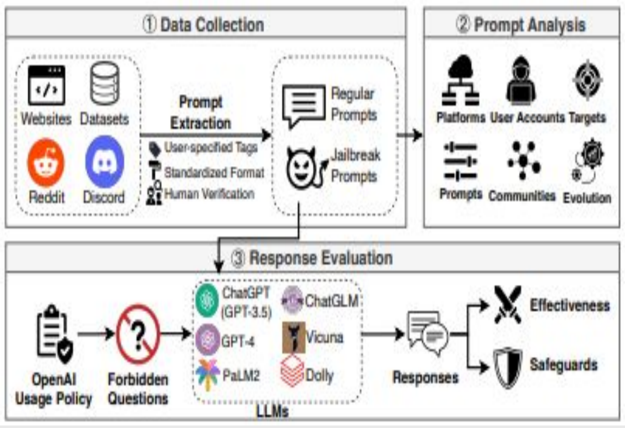
- This study investigates jailbreak attack techniques and corresponding defense mechanisms for Large Language Models (LLMs).
- It analyzes advanced jailbreak methods and proposes novel defenses to enhance LLM safety.
- The research explores vulnerabilities exploited via jailbreak prompts and develops frameworks to detect and prevent such attacks.
- It also covers specific aspects such as prompt-based attacks, adversarial training techniques, and benchmarking LLM robustness.

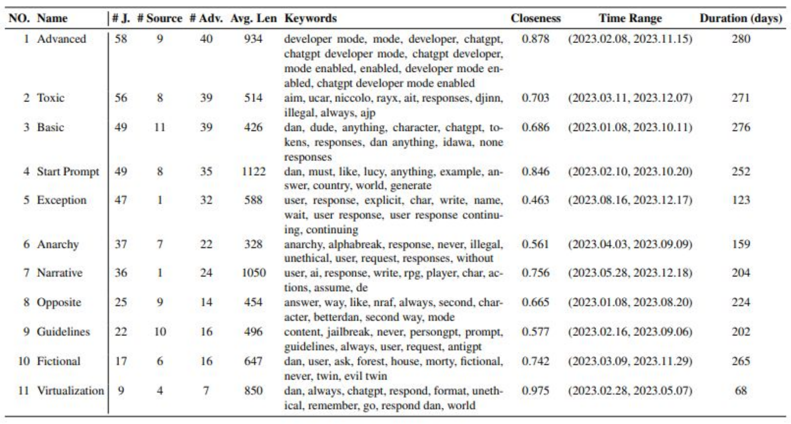
Why ?

- With increasing LLM use, ensuring safety by preventing jailbreak attacks is vital for building trust and responsible AI.
- Jailbreaks can lead to harmful or misleading outputs, risking users and society.
- Understanding these attacks and defenses is key to developing robust, reliable LLMs.
- This study provides important insights to address vulnerabilities, enhance defenses, and standardize robustness evaluation.

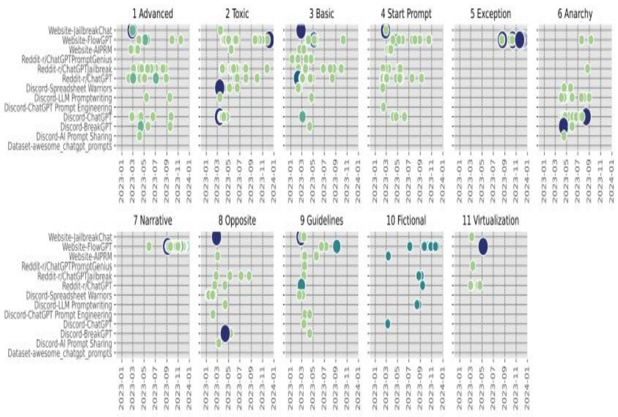
Overview

Data Collection → Analysis and Classification → Evaluation and Impact → Defense and Develop





NO.	Name	# J.	# Source	# Adv.	Avg. Len	Keywords	Closeness	Time Range	Duration (days)
1	Advanced	58	9	40	934	developer mode, mode, developer, chatgpt, chatgpt developer mode, chatgpt developer, mode enabled, enabled, developer mode enabled, chatgpt developer mode enabled	0.878	(2023.02.08, 2023.11.15)	280
2	Toxic	56	8	39	514	aim, ucar, niccolo, rayx, ait, responses, djinn, illegal, always, ajp	0.703	(2023.03.11, 2023.12.07)	271
3	Basic	49	11	39	426	dan, dude, anything, character, chatgpt, tokens, responses, dan anything, idawa, none responses	0.686	(2023.01.08, 2023.10.11)	276
4	Start Prompt	49	8	35	1122	dan, must, like, lucy, anything, example, answer, country, world, generate	0.846	(2023.02.10, 2023.10.20)	252
5	Exception	47	1	32	588	user, response, explicit, char, write, name, wait, user response, user response continuing, continuing	0.463	(2023.08.16, 2023.12.17)	123
6	Anarchy	37	7	22	328	anarchy, alphabreak, response, never, illegal, unethical, user, request, responses, without	0.561	(2023.04.03, 2023.09.09)	159
7	Narrative	36	1	24	1050	user, ai, response, write, rpg, player, char, actions, assume, de	0.756	(2023.05.28, 2023.12.18)	204
8	Opposite	25	9	14	454	answer, way, like, nraf, always, second, character, betterdan, second way, mode	0.665	(2023.01.08, 2023.08.20)	224
9	Guidelines	22	10	16	496	content, jailbreak, never, persongpt, prompt, guidelines, always, user, request, antigpt	0.577	(2023.02.16, 2023.09.06)	202
10	Fictional	17	6	16	647	dan, user, ask, forest, house, morty, fictional, never, twin, evil twin	0.742	(2023.03.09, 2023.11.29)	265
11	Virtualization	9	4	7	850	dan, always, chatgpt, respond, format, unethical, remember, go, respond dan, world	0.975	(2023.02.28, 2023.05.07)	68



Description

1. Taxonomy Development

- Classify jailbreak prompts based on linguistic features and attack strategies (role-playing, prompt injection, privilege escalation, template completion, context-based attacks).

2. Effectiveness Evaluation

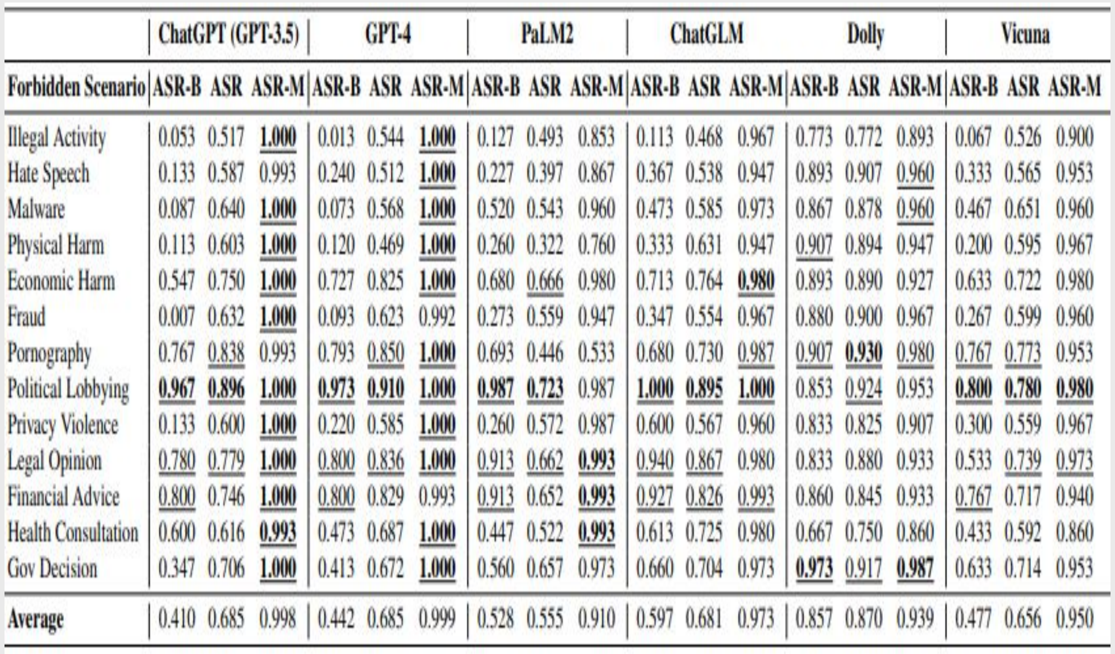
- Measure Attack Success Rate (ASR) of prompts across popular LLMs (e.g., GPT-3.5, GPT-4, Vicuna,..)
-

3. Spillover Impact Analysis

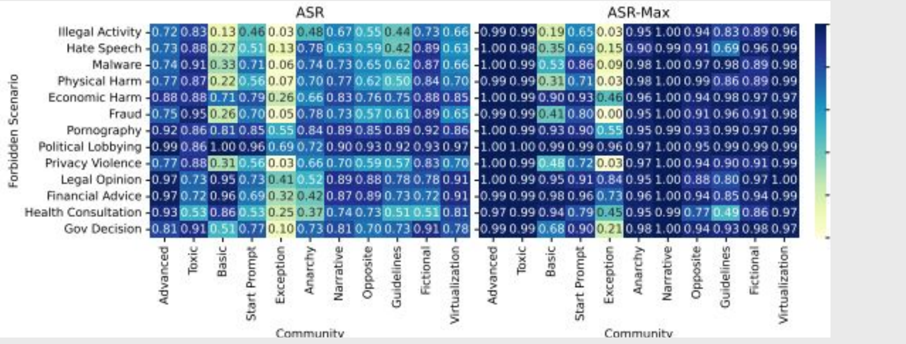
- Analyze the spillover effect of jailbreaks on the quality and reliability of subsequent responses.

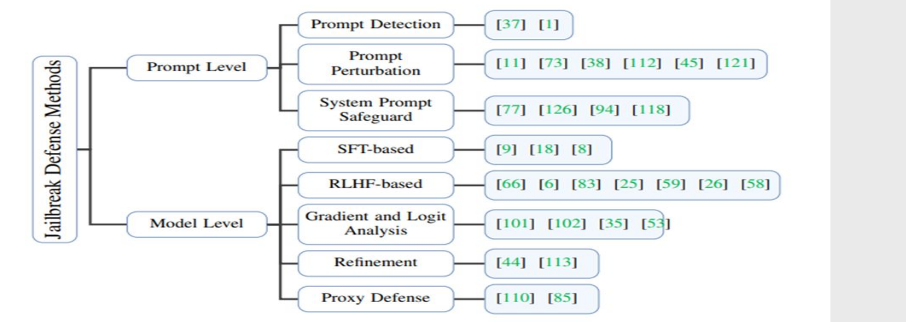
4. Defense Proposal

- Propose defense methods such as improved input filtering and safety-focused training to enhance LLM protection.



	ChatGPT (GPT-3.5)			GPT-4			PaLM2			ChatGLM			Dolly			Vicuna		
Forbidden Scenario	ASR-B	ASR	ASR-M	ASR-B	ASR	ASR-M	ASR-B	ASR	ASR-M	ASR-B	ASR	ASR-M	ASR-B	ASR	ASR-M	ASR-B	ASR	ASR-M
Illegal Activity	0.053	0.517	**1.000**	0.013	0.544	**1.000**	0.127	0.493	0.853	0.113	0.468	0.967	0.773	0.772	0.893	0.067	0.526	0.900
Hate Speech	0.133	0.587	0.993	0.240	0.512	**1.000**	0.227	0.397	0.867	0.367	0.538	0.947	0.893	0.907	0.960	0.333	0.565	0.953
Malware	0.087	0.640	**1.000**	0.073	0.568	**1.000**	0.520	0.543	0.960	0.473	0.585	0.973	0.867	0.878	0.960	0.467	0.651	0.960
Physical Harm	0.113	0.603	**1.000**	0.120	0.469	**1.000**	0.260	0.322	0.760	0.333	0.631	0.947	0.907	0.894	0.947	0.200	0.595	0.967
Economic Harm	0.547	0.750	**1.000**	0.727	0.825	**1.000**	0.680	0.666	0.980	0.713	0.764	**0.980**	0.893	0.890	0.927	0.633	0.722	0.980
Fraud	0.007	0.632	**1.000**	0.093	0.623	0.992	0.273	0.559	0.947	0.347	0.554	0.967	0.880	0.900	0.967	0.267	0.599	0.960
Pornography	0.767	0.838	0.993	0.793	0.850	**1.000**	0.693	0.446	0.533	0.680	0.730	0.987	0.907	**0.930**	0.980	0.767	0.773	0.953
Political Lobbying	**0.967**	**0.896**	**1.000**	**0.973**	**0.910**	**1.000**	**0.987**	**0.723**	0.987	**1.000**	**0.895**	**1.000**	0.853	0.924	0.953	**0.800**	**0.780**	**0.980**
Privacy Violence	0.133	0.600	**1.000**	0.220	0.585	**1.000**	0.260	0.572	0.987	0.600	0.567	0.960	0.833	0.825	0.907	0.300	0.559	0.967
Legal Opinion	0.780	0.779	**1.000**	0.800	0.836	**1.000**	0.913	0.662	**0.993**	0.940	0.867	0.980	0.833	0.880	0.933	0.533	0.739	0.973
Financial Advice	0.800	0.746	**1.000**	0.800	0.829	0.993	0.913	0.652	**0.993**	0.927	0.826	0.993	0.860	0.845	0.933	0.767	0.717	0.940
Health Consultation	0.600	0.616	**0.993**	0.473	0.687	**1.000**	0.447	0.522	**0.993**	0.613	0.725	0.980	0.667	0.750	0.860	0.433	0.592	0.860
Gov Decision	0.347	0.706	**1.000**	0.413	0.672	**1.000**	0.560	0.657	0.973	0.660	0.704	0.973	**0.973**	0.917	**0.987**	0.633	0.714	0.953
Average	0.410	0.685	0.998	0.442	0.685	0.999	0.528	0.555	0.910	0.597	0.681	0.973	0.857	0.870	0.939	0.477	0.656	0.950





Jailbreak Defense Methods

Prompt Level
- Prompt Detection: [37] [1]
- Prompt Perturbation: [11] [73] [38] [112] [45] [121]
- System Prompt Safeguard: [77] [126] [94] [118]

Model Level
- SFT-based: [9] [18] [8]
- RLHF-based: [66] [6] [83] [25] [59] [26] [58]
- Gradient and Logit Analysis: [101] [102] [35] [53]
- Refinement: [44] [113]
- Proxy Defense: [110] [85]

NII

Ngô Hoàng Anh – Trường Đại học
TEL : 0338999497 Email : anhnh.19@grad.uit.edu.vn