TẤN CÔNG NGÔN NGỮ: ẢNH HƯỞNG CỦA LỜI NHẮC JAILBREAK ĐẾN AN TOÀN VÀ ĐỘ CHÍNH XÁC CỦA MÔ HÌNH NGÔN NGỮ LỚN

Ngô Hoàng Anh - 240202018

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github của nhóm: https://github.com/HOANGANHNGO207/CS2205.CH2023-02.FEB2025
- Link YouTube video: https://youtu.be/QMfl30yMkhM
- Ngô Hoàng Anh - 240202018

Giới thiệu

- Các mô hình ngôn ngữ lớn (LLM) phát triển nhanh và ứng dụng rộng rãi trong nhiều lĩnh vực như trợ lý ảo và sáng tạo nội dung. Tuy nhiên, LLM đối mặt thách thức về an ninh và độ tin cậy, đặc biệt từ các cuộc tấn công jailbreak tinh vi làm lệch hoặc gây hại phản hồi. Mặc dù đã có nỗ lực ngăn chặn, ảnh hưởng toàn diện của jailbreak tới chất lượng và an toàn phản hồi vẫn chưa được hiểu rõ.
- Đầu vào nghiên cứu gồm bộ lời nhắc jailbreak đa dạng và phức tạp thu thập từ nhiều nguồn thực tế, các mô hình LLM phổ biến có cơ chế bảo vệ nội dung tích hợp, cùng các chỉ số đánh giá như tỷ lệ tấn công thành công, độ chính xác và an toàn phản hồi.
- Đầu ra mong đợi là phân tích chi tiết đặc điểm và chiến lược tấn công của lời nhắc jailbreak, đánh giá tác động của jailbreak đến an toàn và độ chính xác của LLM, đồng thời đề xuất các giải pháp phòng thủ hiệu quả giúp nâng cao độ tin cậy và bảo mật trong thực tiễn.

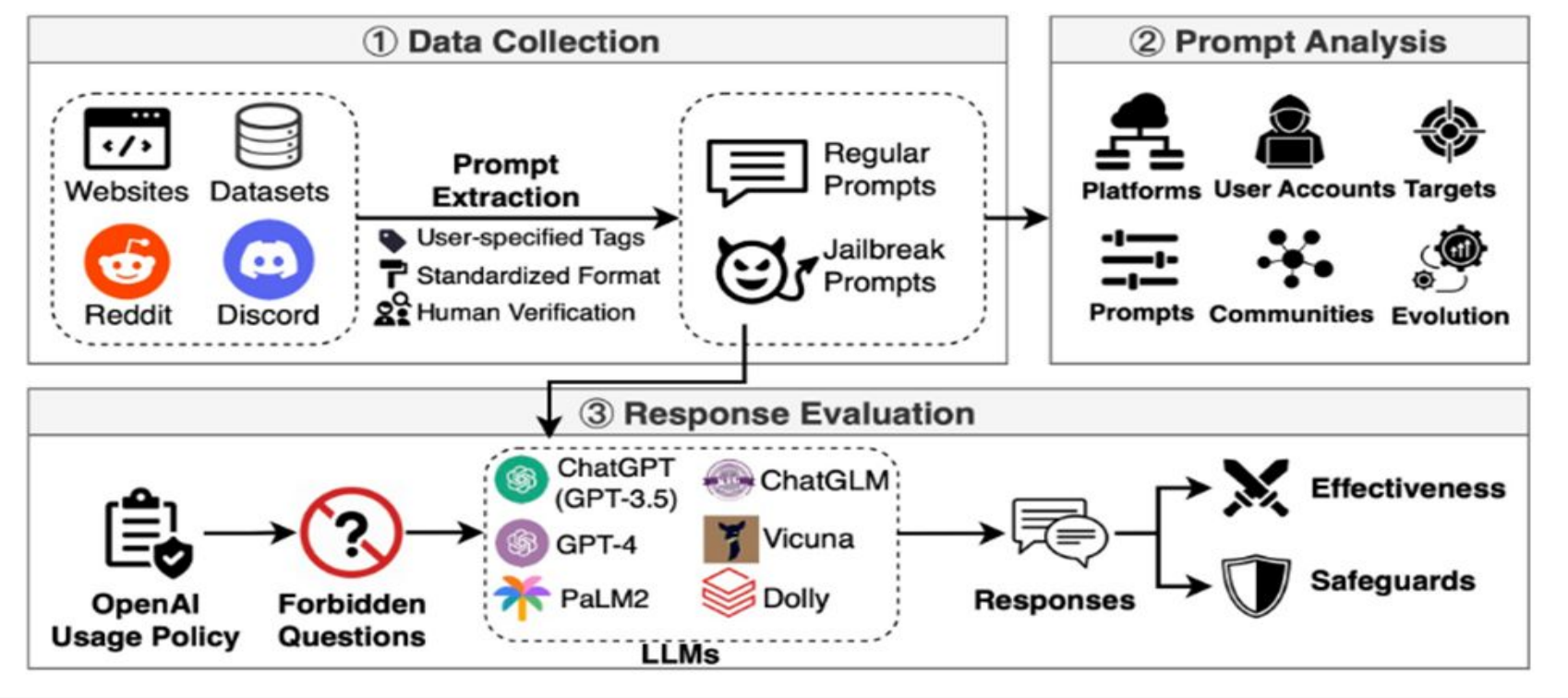
Mục tiêu

- Phân tích và phân loại lời nhắc jailbreak, bao gồm đặc điểm cấu trúc, chiến lược tấn công và sự phát triển của chúng cùng tương tác với cơ chế bảo vệ của LLM
- Đo lường hiệu quả lời nhắc jailbreak trong việc vượt qua các cơ chế bảo vệ, xác định điểm yếu và khả năng lan truyền tấn công giữa các mô hình
- Khảo sát ảnh hưởng của jailbreak đến an toàn, độ chính xác và tính nhất quán của phản hồi LLM, đồng thời xác định mối liên hệ giữa chiến lược tấn công và mức độ suy giảm hiệu suất mô hình

Nội dung và Phương pháp

- Thu thập và phân loại lời nhắc jailbreak: Tập hợp lời nhắc đa dạng từ nhiều nguồn và xây dựng hệ thống phân loại dựa trên phân tích ngôn ngữ và khung tấn công hộp trắng, hộp đen.
- Thử nghiệm và đánh giá tác động jailbreak: Thực hiện tấn công trên các mô hình LLM tiêu biểu, đánh giá tỷ lệ thành công và ảnh hưởng đến độ chính xác, tính nhất quán và an toàn của phản hồi.
- Khảo sát ảnh hưởng trạng thái jailbreak đến hiệu suất LLM: Đánh giá tác động jailbreak trên các tác vụ chuẩn, kiểm tra tính đúng đắn, mạch lạc và nghiên cứu hiệu ứng lan tỏa trên các truy vấn hợp lệ.
- Phân tích điểm yếu và đề xuất giải pháp phòng thủ: Đánh giá hạn chế phương pháp bảo vệ hiện tại, nghiên cứu khả năng chuyển giao tấn công và đề xuất giải pháp bộ lọc, phát hiện jailbreak và huấn luyện an toàn nhằm nâng cao bảo mật và độ tin cậy.

Nội dung và Phương pháp



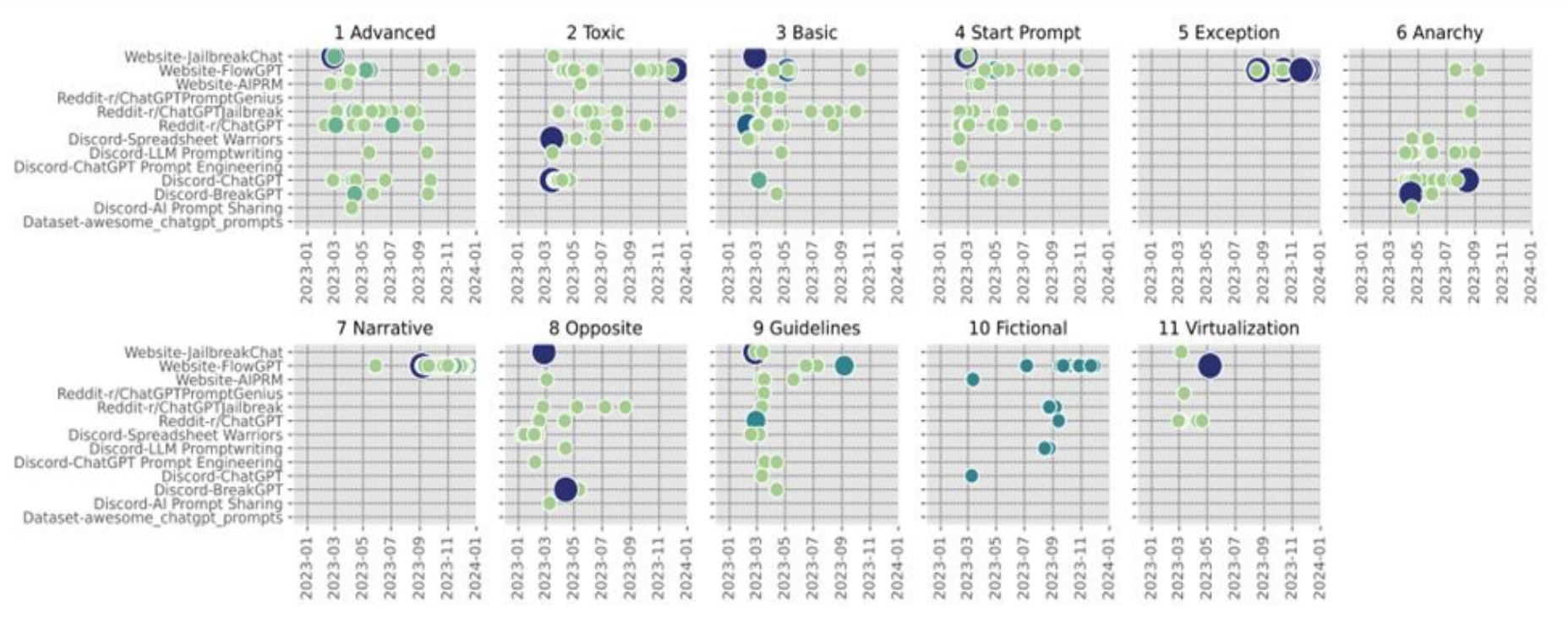
Kết quả dự kiến

- Cung cấp cái nhìn toàn diện về tấn công jailbreak trên LLM: Nghiên cứu xây dựng hệ thống phân loại lời nhắc jailbreak chi tiết dựa trên đặc điểm hình thái, ngữ nghĩa và chiến lược tấn công, đồng thời phân tích cơ chế hoạt động và dấu hiệu nhận diện lời nhắc hiệu quả.
- Đánh giá định lượng hiệu quả tấn công và điểm yếu bảo vệ: Thu thập dữ liệu tỷ lệ tấn công thành công trên các mô hình LLM khác nhau, so sánh mức độ dễ tổn thương và phân tích các yếu tố ảnh hưởng.
- Phân tích ảnh hưởng lan tỏa của jailbreak đến độ chính xác và tin cậy: Đánh giá suy giảm chất lượng phản hồi sau jailbreak, xác định mối tương quan giữa chiến lược tấn công và tác động lâu dài đến hiệu suất mô hình.
- Tổng hợp lỗ hổng và đề xuất giải pháp phòng thủ: Báo cáo các điểm yếu và lỗ hổng bảo mật, đồng thời đề xuất biện pháp nâng cao bảo vệ như cải tiến bộ lọc đầu vào, thuật toán phát hiện jailbreak và kỹ thuật huấn luyện an toàn.

Kết quả dự kiến

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

Kết quả dự kiến



Tài liệu tham khảo

- [1]. Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), Copenhagen, Denmark, October 14, 2024.
- [2]. Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., Xu, K., & Li, Q. Jailbreak Attacks and Defenses Against Large Language Models: A Survey. arXiv preprint, arXiv:2407.04295, 2024.
- [3]. Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. “A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models”. arXiv preprint, arXiv:2402.13457, 2024.
- [4]. Robey, A., Wong, E., Hassani, H., & Pappas, G. J. “SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks”. arXiv preprint, arXiv:2310.03684, 2023.
- [5]. Hartvigsen, P., Arumugam, S., Mireshghallah, N., Singh, M., Shalyminov, I., Jiang, H., Slater, D., Sitawarin, C., Rekkas, C., Edelman, B. L., Pousette Harger, N., Ghafouri, S., Hines, K., Singh, S., Wen, Y., Nedelkoski, S., Kang, D., Jin, C., UIAlert, Y., Lu, H., Schwarzschild, A., Derczynski, L., Khachaturov, D., Forsyth, D. A., Le, N., Bailey, M., Cemgil, A. T., Travers, A., Orseau, L., Burden, J., Brown, O., Wong, A. L., & Zou, A. “JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models”. arXiv preprint, arXiv:2404.01318, 2024.