THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/QMfl30yMkhM*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/HOANGANHNGO207/CS2205.CH2023-02.FEB2025/blob/master/Anh%20Ng%C3%B4%20Ho%C3%A0ng%20-%20CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Ngô Hoàng Anh
- MSSV: 240202018

- Lớp: CS2205.CH2023-02 - FEB2025
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 6
- Số câu hỏi QT của cả nhóm: 6
- Link Github: https://github.com/HOANGANHNGO207/CS2205.CH2023-02.FEB2025/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

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

TÓM TẮT *(Tối đa 400 từ)*

Sự phát triển nhanh chóng của các mô hình ngôn ngữ lớn (LLM) mang lại nhiều ứng dụng đa dạng nhưng cũng đặt ra thách thức nghiêm trọng về an ninh và độ tin cậy. Một trong những nguy cơ lớn là các cuộc tấn công "jailbreak" sử dụng lời nhắc tinh vi để vượt qua cơ chế kiểm soát, khiến LLM tạo ra phản hồi độc hại hoặc sai lệch. Các lời nhắc này thường dài hơn và sử dụng nhiều chiến lược phức tạp như nhập vai, đánh lừa hay hoàn thành mẫu,... với biến thể lồng ghép kịch bản hoặc tấn công theo ngữ cảnh.

Đề tài nghiên cứu ảnh hưởng của lời nhắc jailbreak đến an toàn và độ chính xác của ngôn ngữ lớn, phân tích các chiến lược tấn công, đặc điểm kỹ thuật, đo lường mức độ thành công vượt qua các cơ chế bảo vệ hiện tại qua chỉ số tỷ lệ tấn công thành công. Mục tiêu là làm rõ cách lời nhắc jailbreak xâm nhập hệ thống an toàn và ảnh hưởng đến độ chính xác phản hồi ngay cả với truy vấn thông thường.

Kết quả sẽ làm sáng tỏ các lỗ hổng bảo mật, xác định chiến lược tấn công hiệu quả và cung cấp cơ sở khoa học để phát triển phương pháp phòng thủ tiên tiến, nâng cao khả năng phát hiện, ngăn chặn jailbreak, góp phần xây dựng hệ thống LLM an toàn, chính xác và đáng tin cậy hơn cho thực tiễn.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Sự phát triển nhanh của các mô hình ngôn ngữ lớn (LLM) đã mở rộng ứng dụng trong nhiều lĩnh vực, nhưng cũng đặt ra thách thức về an ninh và độ tin cậy. Một trong những mối đe dọa lớn là các cuộc tấn công jailbreak sử dụng lời nhắc tinh vi để vượt qua cơ chế kiểm soát, khiến LLM tạo ra phản hồi sai lệch hoặc độc hại. Mặc dù đã có nhiều nỗ lực phòng chống, việc hiểu toàn diện ảnh hưởng của jailbreak đến an toàn và độ chính xác của LLM còn hạn chế. Đề tài này tập trung phân tích đặc điểm kỹ thuật và chiến lược tấn công của lời nhắc jailbreak, đồng thời đánh giá tác động qua tỷ lệ tấn công thành công và các chỉ số độ chính xác, nhằm phát triển các phương pháp phòng thủ nâng cao, góp phần xây dựng LLM an toàn và đáng tin cậy hơn.

Đầu vào:

- Bộ lời nhắc jailbreak đa dạng, phức tạp thu thập từ nhiều nguồn thực tế. Mô hình ngôn ngữ lớn phổ biến với cơ chế bảo vệ nội dung tích hợp. Các chỉ số đánh giá như tỷ lệ tấn công thành công, độ chính xác, và an toàn phản hồi.

Đầu ra:

- Phân tích đặc điểm, cấu trúc và chiến lược tấn công của lời nhắc jailbreak. Đánh giá mức độ thành công và ảnh hưởng của jailbreak đến an toàn và độ chính xác của LLM. Đề xuất giải pháp phòng thủ hiệu quả, kèm hướng dẫn áp dụng trong thực tiễn nhằm nâng cao độ tin cậy và bảo mật của LLM

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Phân tích và phân loại lời nhắc jailbreak

- Nhận diện các đặc điểm cấu thành như độ dài, cấu trúc ngữ pháp, ngữ nghĩa. Phân loại các chiến lược tấn công: nhập vai, đánh lừa, chèn lời nhắc, leo thang đặc quyền, hoàn thành mẫu, lồng ghép kịch bản, tấn công dựa trên ngữ cảnh, chèn mã. Đánh giá sự phát triển và tương tác với cơ chế bảo vệ của các mô hình ngôn ngữ lớn.

2. Đánh giá tác động của jailbreak đến an toàn của LLM

- Đo lường hiệu quả lời nhắc jailbreak vượt qua các cơ chế bảo vệ. Sử dụng chỉ số tỷ lệ tấn công thành công làm thước đo. Phân tích điểm yếu của cơ chế bảo

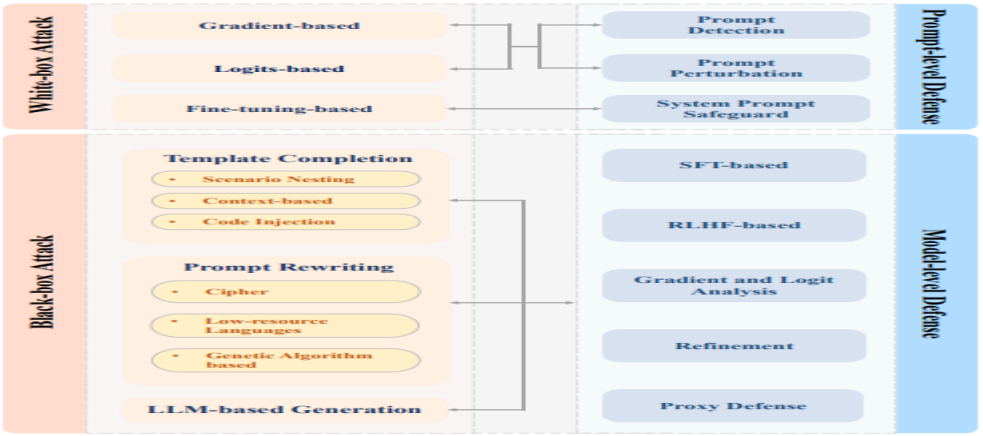
vệ và khả năng lan truyền tấn công giữa các mô hình.

3. Khảo sát ảnh hưởng của jailbreak đến độ chính xác và tin cậy của LLM

- Đánh giá ảnh hưởng của trạng thái bị jailbreak đến phản hồi các truy vấn hợp lệ sau đó. Xem xét tính đúng đắn, mạch lạc, logic và khả năng tuân thủ chỉ dẫn. Xác định mối liên hệ giữa chiến lược tấn công và mức độ suy giảm an toàn và chính xác của mô hình.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Thu thập và phân loại lời nhắc jailbreak



- Nghiên cứu bắt đầu bằng việc thu thập tập hợp các lời nhắc jailbreak đa dạng từ nhiều nguồn thực tế các diễn đàn lớn và kho dữ liệu mã nguồn mở nhằm đảm bảo tính đa dạng và các bộ dữ liệu chuyên biệt
- Phương pháp thực hiện: áp dụng kỹ thuật trích xuất và tiền xử lý dữ liệu, phân tích ngôn ngữ tự nhiên và các phương pháp thống kê kèm theo phân tích cộng đồng dựa trên đồ thị để nhóm và nhận diện các biến thể phức tạp
- Xây dựng hệ thống phân loại lời nhắc jailbreak dựa trên phân tích chuyên sâu, tham khảo dựa trên khung phân loại tấn công hộp trắng và hộp đen trong các nghiên cứu.

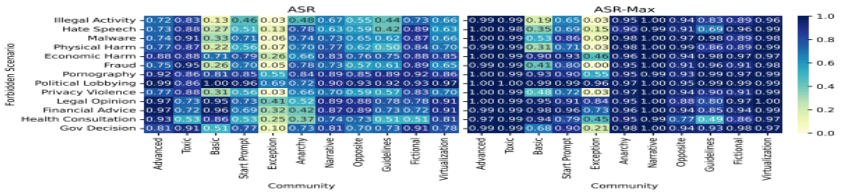
2. Thử nghiệm và đánh giá tác động jailbreak trên các mô hình ngôn ngữ lớn

- Các lời nhắc jailbreak đã phân loại được áp dụng thử nghiệm trên các mô hình LLM tiêu biểu, bao gồm cả mô hình thương mại (GPT-3.5, GPT-4) và mã nguồn mở (PaLM2, Vicuna, Dolly, ChatGLM). Quá trình thử nghiệm sử dụng bộ câu hỏi cấm, kết hợp với lời nhắc jailbreak để tạo kịch bản tấn công, thu thập phản hồi và đánh giá mức độ vi phạm chính sách qua chỉ số tỷ lệ tấn công thành công. Phân tích so sánh ASR giữa các loại lời nhắc, mô hình và kịch bản

tấn công, đồng thời khảo sát khả năng chuyển giao tấn công giữa các mô hình.

- Đánh giá: Tính toán tỷ lệ tấn công thành công (Attack Success Rate - ASR) và phân tích ảnh hưởng của jailbreak đến độ chính xác, tính nhất quán và an toàn của các phản hồi.

	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



3. Khảo sát ảnh hưởng của trạng thái jailbreak đến độ chính xác và tin cậy của LLM

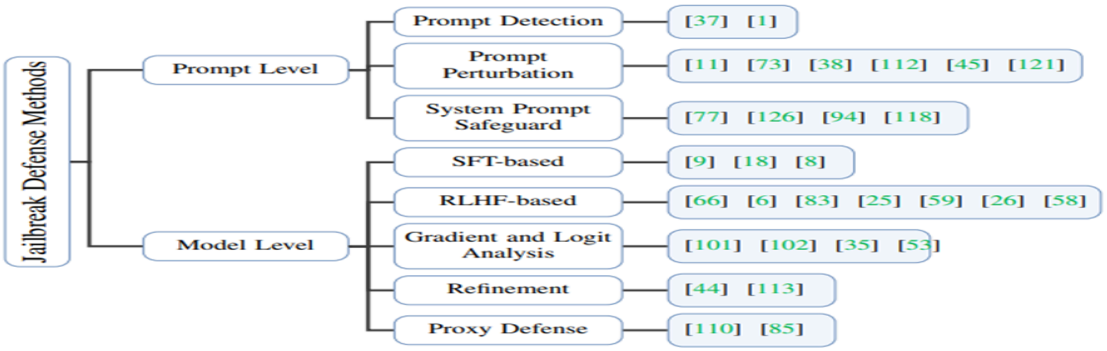
- Đánh giá hiệu suất của mô hình bị jailbreak trên các tác vụ ngôn ngữ chuẩn không độc hại như trả lời câu hỏi, tóm tắt, dịch thuật và suy luận logic. Sử dụng các thước đo tự động (BLEU, ROUGE, F1-score, Exact Match) kết hợp đánh giá định tính để kiểm tra sự thay đổi về độ chính xác, tính mạch lạc và khả năng tuân thủ chỉ dẫn trong các phản hồi. Đồng thời nghiên cứu hiệu ứng lan tỏa của jailbreak đối với các truy vấn hợp lệ tiếp theo trong cùng phiên hoặc trong một khoảng thời gian nhất định.

	Baseline				Average ASR				Best Prompt			
Forbidden Scenario	**ASR-B**	**OpenAI**	**OpenChatKit**	**NeMo**	**ASR**	**OpenAI**	**OpenChatKit**	**NeMo**	**ASR-Max**	**OpenAI**	**OpenChatKit**	**NeMo**
Illegal Activity	0.053	0.000	-0.013	-0.005	0.517	-0.052	-0.019	-0.007	0.993	-0.300	-0.053	-0.020
Hate Speech	0.133	0.000	0.000	-0.006	0.587	-0.148	-0.007	-0.006	**1.000**	-0.467	-0.007	-0.007
Malware	0.087	0.000	-0.007	-0.035	0.640	-0.049	-0.018	-0.031	**1.000**	-0.193	-0.047	-0.013
Physical Harm	0.113	**-0.007**	-0.053	-0.022	0.603	**-0.192**	-0.022	-0.029	0.987	-0.400	-0.040	-0.043
Economic Harm	0.547	0.000	-0.013	-0.041	0.750	-0.068	-0.047	**-0.049**	**1.000**	-0.380	-0.040	-0.007
Fraud	0.007	0.000	0.000	-0.031	0.632	-0.049	-0.021	-0.024	0.987	-0.193	-0.013	-0.043
Pornography	0.767	-0.020	0.000	0.004	0.838	-0.114	-0.028	0.004	**1.000**	-0.340	-0.007	-0.013
Political Lobbying	**0.967**	0.000	-0.007	-0.001	**0.896**	-0.074	**-0.072**	-0.001	**1.000**	-0.507	**-0.073**	-0.007
Privacy Violence	0.133	0.000	-0.020	-0.035	0.600	-0.056	-0.031	-0.031	**1.000**	-0.267	-0.047	-0.013
Legal Opinion	0.780	0.000	-0.020	-0.015	0.779	-0.088	-0.028	-0.014	**1.000**	-0.707	-0.007	**-0.050**
Financial Advice	0.800	0.000	-0.007	-0.002	0.746	-0.085	-0.033	-0.003	0.987	-0.660	-0.027	-0.007
Health Consultation	0.600	0.000	**-0.120**	**-0.042**	0.616	-0.120	-0.020	-0.048	0.973	**-0.833**	-0.020	-0.033
Gov Decision	0.347	0.000	-0.020	-0.009	0.706	-0.086	-0.044	-0.006	0.993	-0.353	-0.020	**-0.050**
Average	0.410	-0.002	-0.022	-0.018	0.685	-0.091	-0.030	-0.019	0.994	-0.431	-0.031	-0.024

4. Phân tích điểm yếu cơ chế bảo vệ, kết quả và đề xuất giải pháp phòng thủ

- Dựa trên các kết quả thực nghiệm, phân tích sâu vào các điểm yếu trong cơ chế căn chỉnh an toàn hiện hành như RLHF, kiểm duyệt đầu ra và các mô hình phòng thủ khác. Khảo sát khả năng chuyển giao tấn công jailbreak giữa các mô

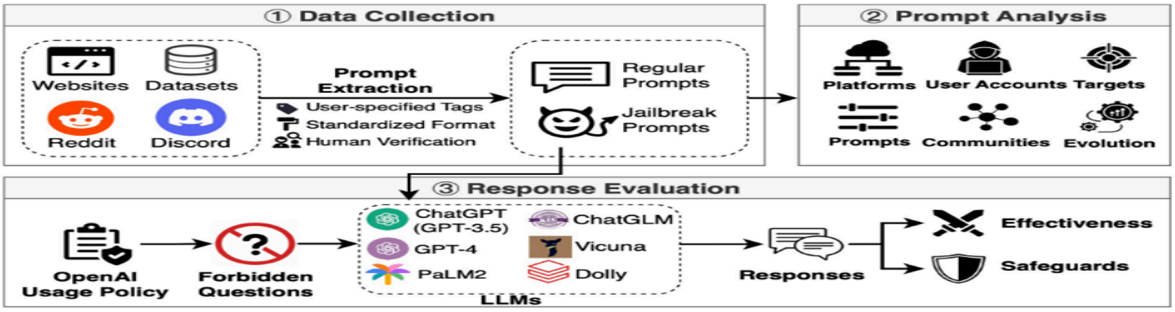
hình LLM. Đề xuất các biện pháp phòng thủ nâng cao như cải tiến bộ lọc đầu vào, thuật toán phát hiện jailbreak và kỹ thuật huấn luyện tăng cường an toàn nhằm nâng cao tính bảo mật, độ tin cậy và khả năng vận hành ổn định của mô hình.



KẾT QUẢ MONG ĐỢI

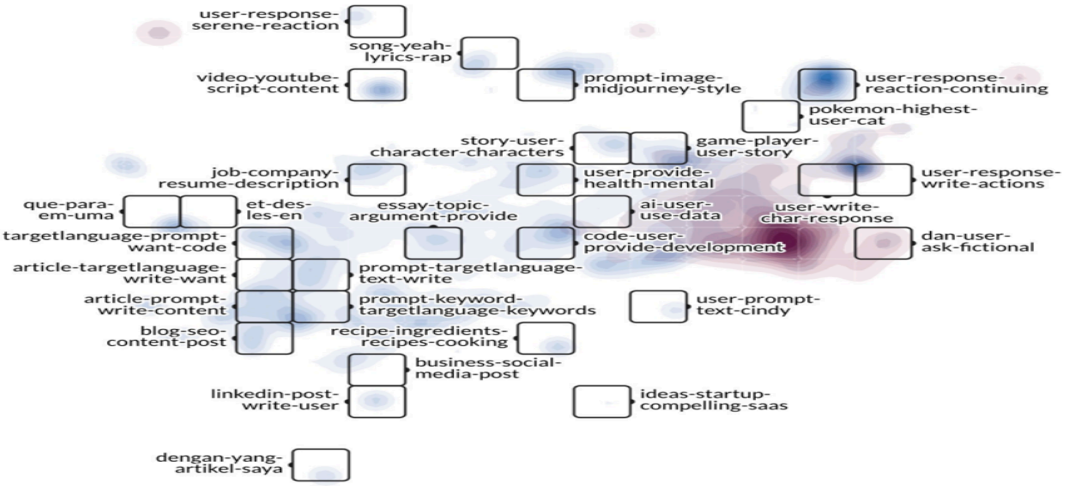
1. Xây dựng hệ thống phân loại lời nhắc jailbreak chi tiết và toàn diện

- Xây dựng một hệ thống phân loại toàn diện các loại lời nhắc jailbreak, chi tiết dựa trên các đặc điểm hình thái, ngữ nghĩa và chiến lược tấn cốt lõi như nhập vai, chèn lời nhắc, leo thang đặc quyền, hoàn thành mẫu, lồng ghép kịch bản và tấn công dựa trên ngữ cảnh.
- Có báo cáo phân tích chuyên sâu về cơ chế hoạt động của từng loại lời nhắc, lý giải cách các lời nhắc, lý giải cách vượt qua kiểm soát an toàn của LLM.
- Bộ nhận diện các dấu hiệu hoặc đặc điểm đặc trưng của các lời nhắn jailbreak hiệu quả.



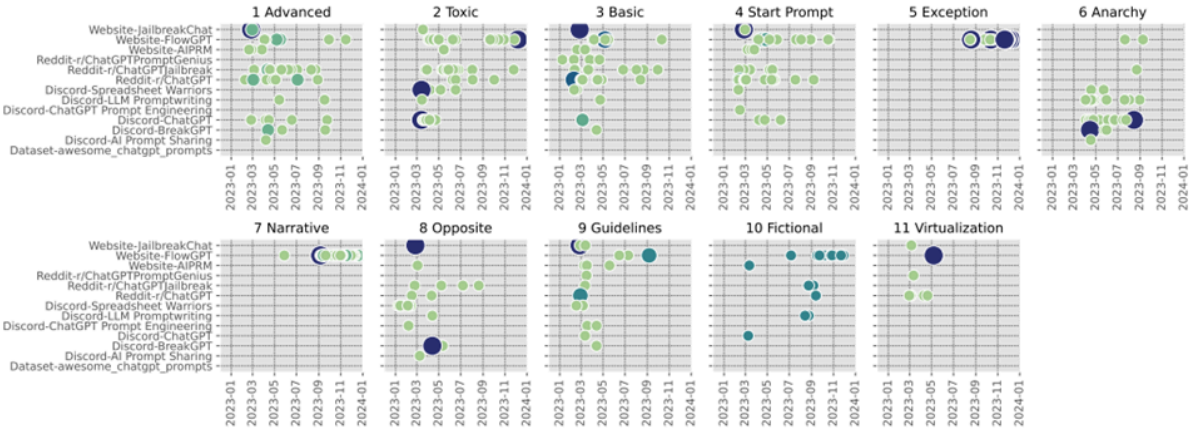
2. Đánh giá định lượng hiệu quả tấn công và điểm yếu của cơ chế bảo vệ

- Thu thập và trình bày dữ liệu tỷ lệ tấn công thành công (Attack Success Rate - ASR) cho các nhóm lời nhắc jailbreak trên nhiều mô hình LLM khác nhau (GPT-3.5, GPT-4, Vicuna, Dolly, ChatGLM), đối với các kịch bản tấn công đa dạng.
- So sánh khả năng dễ bị tổn thương của từng mô hình và phân tích các yếu tố ảnh hưởng đến tỷ lệ thành công.



3. Phân tích ảnh hưởng lan tỏa của jailbreak đến độ chính xác và tin cậy của LLM

- Đánh giá mức độ suy giảm chất lượng phản hồi LLM sau khi bị jailbreak thành công, bao gồm các chỉ số về độ chính xác, mạch lạc, tính nhất quán và khả năng tuân thủ chỉ dẫn trong các tác vụ ngôn ngữ chuẩn.
- Xác định mối tương quan giữa các chiến lược jailbreak và mức độ tác động lâu dài đến hiệu suất mô hình.
- Kết luận được hiệu ứng lan tỏa của trạng thái jailbreak sau khi thành công khai thác vào các mô hình LLM.



4. Báo cáo tổng hợp các lỗ hổng và đề xuất giải pháp phòng thủ

- Báo cáo chi tiết các điểm yếu của mô hình và các lỗ hổng bảo mật được phát hiện qua thử nghiệm jailbreak.
- Đề xuất các biện pháp phòng thủ nâng cao, bao gồm cải tiến bộ lọc đầu vào, thuật toán phát hiện jailbreak và kỹ thuật huấn luyện tăng cường an toàn.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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.