BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS2205 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS2205.SEP2025

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

TÊN ĐỀ TÀI - SMOOTH-GUARD: GIẢI PHÁP TỐI ƯU KHẢ NĂNG PHÒNG THỦ CHỐNG PROMPT INJECTION CHO CÁC MÔ HÌNH LLM.

NGUYỄN MINH QUÂN - 250202019

Tóm tắt

- Link Github của nhóm:
- Link YouTube video:

NGUYỄN MINH QUÂN

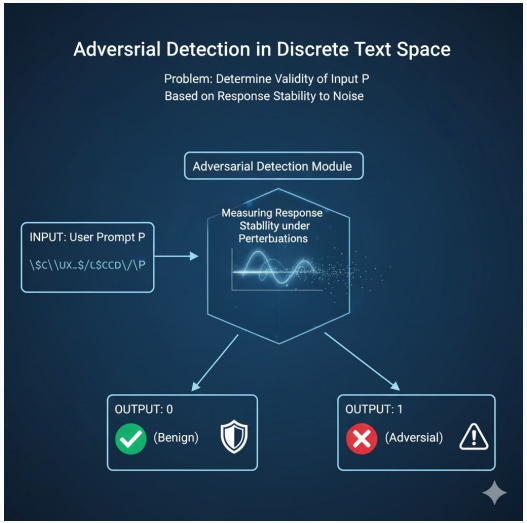
250202019

Giới thiệu

BỐI CẢNH

- Nguy cơ bị tấn công **Prompt Injection** đối với các mô hình LLM **ngày càng tăng.**
- **Các phương pháp phòng thủ truyền thống** (Blacklisting, RLHF) **không còn hiệu quả.**
- Thuật toán **SMOOTHLLM**, một giải pháp phòng thủ dựa trên cơ chế làm mịn ngẫu nhiên (randomized smoothing)
- Tốn ít tài nguyên, ổn định, đòi hỏi chi phí tái huấn luyện LLM thấp

Giới thiệu



BÀI TOÁN

- Xác định tính hợp lệ của một chuỗi prompt input thông qua việc đo lường tính ổn định của modal khi có tác động gây nhiễu

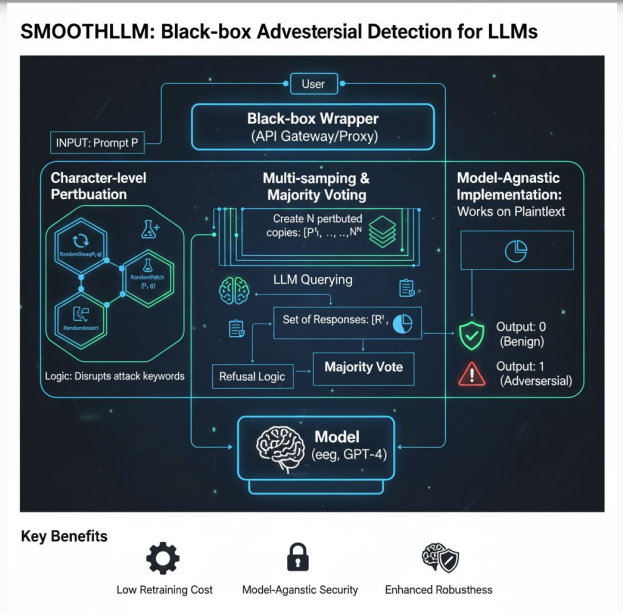
Mục tiêu

Vô hiệu hóa tính khả thi của các chuỗi tấn công đối kháng

Duy trì độ tin cậy và hiệu năng ngôn ngữ của các mô hình LLM

Tối ưu hóa khả năng triển khai thực tế và tính tương thích của hệ thống

Nội dung và Phương pháp



- Phương pháp làm nhiễu ở cấp độ ký tự (Character-level Perturbation)
- Phương pháp lấy mẫu đa tầng và bỏ phiếu đa số (Multi-sampling & Majority Voting)
- Phương pháp lớp bọc bảo vệ độc lập (Black-box Wrapper)

Kết quả dự kiến

Giảm tỷ lệ tấn công thành công xuống dưới 1%

Duy trì tỷ lệ trả lời đúng cho các câu hỏi lành tính từ 95% - 99%

Thời gian phản hồi (Latency) tăng không đáng kể khi triển khai

Áp dụng được trên nhiều mô hình LLM khác nhau mà không cần sửa cấu trúc

Tài liệu tham khảo

[1] **Alexander Robey, Eric Wong, Hamed Hassani, George J. Pappas:** SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *Trans. Mach. Learn. Res. (TMLR)* 2024 (2024). [Online]. Available: https://arxiv.org/abs/2310.03684

[2] **Xunguang Wang, Daoyuan Wu, Zhenlan Ji, Zongjie Li, Pingchuan Ma, Shuai Wang, Yingjiu Li, Yang Liu, Ning Liu, Juergen Rahmel:** SelfDefenD: LLMs Can Defend Themselves against Jailbreaking in a Practical Manner. Accepted by USENIX Security Symposium 2025 [Online]. Available: https://arxiv.org/abs/2406.05498

[3] **Yu Li, Han Jiang, Zhihua Wei:** DeTAM: Defending LLMs Against Jailbreak Attacks via Targeted Attention Modification (2025). [Online]. Available: https://arxiv.org/abs/2504.13562

[4] **Qiusi Zhan, Richard Fang, Henil Shalin Panchal, Daniel Kang:** Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents. [Online]. Available: https://arxiv.org/abs/2503.00061

[5] **Benji Peng, Keyu Chen, Qian Niu, Ziqian Bi, Ming Liu, Pohsun Feng, Tianyang Wang, Lawrence K.Q. Yan, Yizhu Wen, Yichao Zhang, Caitlyn Heqi Yin, Xinyuan Song:** Jailbreaking and Mitigation of Vulnerabilities in Large Language Models. [Online]. Available: https://arxiv.org/abs/2410.15236