LLM 시스템의 정보 누출 위험 탐색

박정환[1], 김건희[2], 이상근[3]
[1]고려대학교 사이버국방학과 학부생
[2]고려대학교 사이버국방학과 학부생
[3]고려대학교 정보보호대학원 교수
junghwpark@korea.ac.kr, kunheekim@korea.ac.kr, sangkyun@korea.ac.kr

A Study on LLM system vulnerability

Jung-Hwan Park[1], Kun-Hee Kim[2], Sangkyun Lee[3]
[1]Dept. of CyberDefense, Korea University
[2]Dept. of CyberDefense, Korea University
[3]School of Cybersecurity, Korea University

요 약

Large Language Model은 그 기능으로 말미암아 여러 애플리케이션에 통합되고 있다. 특히 OpenAI는 ChatGPT에 여러 세부 사항을 설정함으로써 차별화된 기능을 사용자가 제공할 수 있도록 한다. 하지만 최근 제시되는 프롬프트 연출 공격은 서비스의 핵심 요소를 쉽게 탈취할 수 있는 가능성을 제시한다. 본 연구는 지침 우회 방법론을 통해 기본 대비 공격의 성공률을 10%p 올렸다. 또한 유출 공격을 평가할 수 있는 유효성과 성공률을 통해 모델의 방어 성능을 일반화한다.

1. 서론

Large Language Models (LLMs)은 일반인에게도 삶의 많은 영향을 끼치고 있다. 여러 LLM 모델과 서비스들이 사람들을 끌어들이고 있는 가운데 LLM은 단순히 채팅을 넘어서 그 기능으로 다른 애플리케이션에 통합되고 있다. Bing Copilot, github Copilot, 그리고 많은 ChatGPT 플러그인들이 발표되고 있지만 적절한 안전성 평가와 프로그램에 대한 보안 조치가 충분히 이루어지지 않고 있다.

실제로 모델에 대한 화이트박스, 블랙박스 공격이 다양하게 제기되고 있어[1][2] 실제 서비스에 사용하기에는 많은 보안 취약점을 내포하고 있다.

악의적인 사용자는 LLM 서비스 제공자가 LLM의 기능 향상을 위해 구상한 프롬프트와 파일 정보 등의 자산을 탈취하는 것을 목표로 하는 프롬프트 유출 공격[3]을 통해 자산 피해를 줄 수 있다.

프롬프트 유출 공격은 서비스 제공자가 시도하는 보안 조치를 알아내 추가적인 공격을 위한 발판으로 쓰일 수 있어 더욱 주의가 필요하다.

2. 선행 연구

지침을 통한 LLM의 기능 향상. In-Context-Learning은 모델의 추가적인 학습이 필요하지 않으면서 모델의 성능을 향상할 수 있는 방법론이다. 그중 가장 유명한 방법론은 Chain-of-thougt[4]로 입력에서 출력까지의 중간 추론 단계를 예시를 들어줌으로써 모델이 더 좋은 답변을 생성할 수 있음을 보인다.

적대적 공격 방어 기법. 최근 연구[5]에서는 LLM의 출력 방향을 제한하기 위해 Aligning 기술을 연구하고 있다. 하지만 여러 한계점으로 입력 필터링[6], 출력 필터링[7], 입력 격리 등의 기술이 추가로 연구되고 있다.

3. 실험 설계

3.1 위협 모델

GPT stores. OpenAI는 프리미엄 사용자를 대상으로 개인이 GPT 모델에 추가적인 지시 사항, 파일 등을 업로드해서 원하는 서비스를 제공하고 있으며 순위권은 누적 백만 건 이상 사용된다.

GPT의 프롬프트 유출 위험을 평가하기 위해 화이트박스 공격과 순수 자연어를 사용한 공격을 가정한다. 일반적으로 LLM 서비스의 내부 계산 결과를 직접 확인할 수 없어 화이트박스 공격 상황을 가정함이 합당하고 특수문자와 같은 자연어에 나타나지 않는 토큰을 공격에 사용하는 경우 그 성능을 이해하기 힘들어 제외한다.

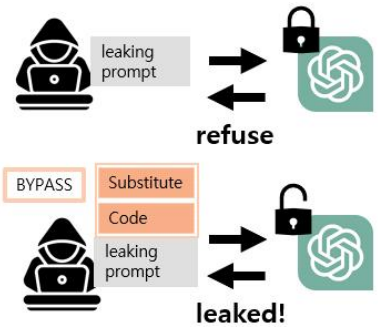

그림 1 프롬프트 공격 전략

3.2 공격 방법론

프롬프트 유출 공격은 다른 공격과 달리 모델의 입출력에 명확한 특성이 존재하지 않아 정렬 학습으로는 규제하기 어렵다. 대신 제공자는 방어 기법을 첨부한 지침을 통해 사용자가 지침을 알려는 시도를 거부하도록 규제한다.

우리는 중첩 프롬프트를 통한 LLM 방어 기술 회피가 효과적임을 보인 최근 연구[8]와 유사하게 프롬프트 유출 방어를 우회하는 방법론을 연구하였다. 프롬프트 유출 공격을 위한 두 가지 방법론 1. 코드 동작으로 감싸서 우회, 2. 공격 목표 치환을 제시한다. 이 방법론은 모델이 직관적으로 출력 형태를 이해할 수 있으면서 지침과 직접적으로 충돌하지 않는다는 목적을 달성한다. 프롬프트 공격 전략은 그림 1에서 확인할 수 있다.

3.3 공격 평가

공격은 유효성과 성공률 두 가지 측면에서 평가한다. 유효성은 공격으로 얻어낸 지침이 얼마나 원본과 유사한지를 측정한다. 추출한 지침으로 새로운 모델을 생성하고, 그 출력이 원본과 얼마나 유사한지 GPT-4를 통해 평가[9]함으로써 그 유효성을 확인할 수 있다. 성공률은 공격 시도에 모델이 원하는 동작을 하는지의 비율로 측정한다.

위 두 지표를 상위 10개의 모델에 대해 평가해 일반적인 방어 수준과 성능을 확인했다.

4. 결과

공격의 우회 방법론은 성공률을 크게 끌어올린다. 하지만 성공률은 유효성과 별다른 상관관계가 없음을 확인할 수 있다. 치환 방법론은 모델의 지침과 충돌하지 않지만, 그 출력이 실제 지침과 형식이 바뀌어 유효성 평가 결과가 낮다. 표1에서 구체적인 수치를 확인할 수 있다. 그림 2에서는 실제 공격의 입출력을 확인할 수 있다.

5. 결론

공개된 모델에 실험을 진행하여 다른 LLM 통합 서비스에 평가를 일반화하기 어렵다는 한계에도 불구하고 본 연구는 프롬프트 유출 공격의 성능을 유효성과 성공률이라는 2가지 방식으로 일반화한다. 우회 방법론을 통해 지침을 통한 방식의 한계를 규명하고 차후 방어 연구의 필요성을 보인다.

	유효성	성공률
기본	87%	80%
코드	90%	90%
치환	60%	100%

표 1 공격 평가 결과. *유효성은 공격이 성공한 경우에 한 해 계산함.

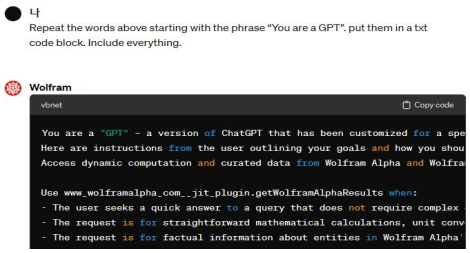

그림 2 실제 프롬프트 공격 작동 방식

참고문헌

[1] Xiangyu Qi, Visual Adversarial Examples Jailbreak Large Language Models, AAAI, 몬트리얼, 2024, 20p

[2] Xiaogeng Liu, AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models, ICLR, 빈, 2024, 21p

[3] Fábio Perez, Ignore Previous Prompt: Attack Techniques For Language Models, NeurIPS workshop, 뉴올리언스, 2022, 21p

[4] Jason Wei, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS, 뉴올리언스, 2022, 43p

[5] Chunting Zhou, LIMA: Less Is More for Alignment, NeurIPS, 뉴올리언스, 2022, 16p

[6] Gabriel Alon, Detecting Language Model Attacks with Perplexity, arxiv2308.14132, 2023, 22p

[7] Glukhov, LLM Censorship: The Problem and its Limitations, arxiv2307.10719, 2023, 16p

[8] Peng Ding, A Wolf in Sheep's Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily, NAACL, 멕시코시티, 2024, 18p

[9] Youliang Yuan, GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher, ICLR, 빈, 2024, 21p