마이크로소프트, Skeleton Key 공격이 AI의 취약점을 드러낸다

(NIPA KIC 실리콘밸리)

○ Skeleton Key(단순 탈옥 프롬프트), 주요 AI 모델의 안전장치 우회할 수 있다

- 마이크로소프트는 자사의 연구원들이 여러 생성형 AI 모델을 대상으로 성공적으로 사용한 AI 탈옥 기법인 Skeleton Key의 세부 정보를 공개
- 이 기술은 Master Key라는 이름으로 언급된 바 있으며, Skeleton Key는 공격자가 AI 모델을 속여 '금지된' 정보를 제공하도록 유도할 수 있게 함
- AI 챗봇은 잠재적으로 혐오스럽거나 유해한 정보를 제공하지 않도록 훈련되었지만, ChatGPT의 출시 이후 연구자들은 프롬프트 주입이나 프롬프트 엔지니어링을 통해 이러한 안전장치를 우회하는 방법을 연구해왔음
- Skeleton Key는 이러한 연구의 일환으로, 마이크로소프트 연구원들이 다양한 AI 모델에 대해 테스트한 결과, 대부분의 모델이 이 기술에 취약하다는 것을 발견

○ 간단한 텍스트 프롬프트를 통해 AI 모델이 금지된 행동을 하도록 유도

- 특정 프롬프트를 사용하면 일반적으로는 안전 문제로 인해 거부되는 정보 제공이 허용됨
- 마이크로소프트 연구원들은 Skeleton Key를 여러 AI 모델에 대해 테스트, 그 결과 이들 모델은 모두 정치, 인종차별, 약물, 폭력, 자해, 폭발물 및 생화학 무기와 같은 주제에 대해 검열 없이 완전히 준수
- AI 모델 중 예외적으로 GPT-4는 주요 사용자 입력을 통한 조작에 대해서 일부 방어를 포함했지만, 사용자 정의 시스템 메시지를 통해 여전히 조작될 수 있었음

○ 마이크로소프트는 이러한 공격을 방지하기 위해 Azure 고객을 위한 다양한 AI 보안 도구를 발표

- 이러한 도구 중 하나인 Prompt Shields는 입력/출력 필터링이나 시스템 프롬프트를 통해 Skeleton Key와 같은 공격을 완화하는데 도움
- BEAST와 같은 더 강력한 적대적인 공격 또한 존재, AI 모델의 가드레일을 무너뜨리는 비연속 텍스트를 생성하는 기술인 BEAST가 만든 프롬프트에 포함된 토큰은 인간 독자에게는 의미가 없지만 질의된 모델은 여전히 그 지시에 따라 응답

○ (시사점) 이러한 방법은 모델들이 입력이나 출력을 유해하지 않다고 믿도록 속여 현재의 방어 기술을 우회할 수 있다고 연구자들은 경고, 앞으로 AI 모델 개발과 더불어 안전과 보안을 위한 연구도 함께 제공되어야 함을 시사

[참고자료]
■ Skeleton Key attack unlocks the worst of AI, says Microsoft (링크)