

※ 본 자료의 무단 복제 및 배포를 금합니다 ※

생성형 AI 시대

기업 LLM 보안 위협에 대응하는 필수 전략

이스트시큐리티 솔루션컨설팅팀

이규원 팀장

[본 자료는 웨비나 발표자료를 재구성한 배포용으로, 실제 발표 자료와 다를 수 있습니다]

기업 LLM 보안 위협에 대응하는 필수 전략
알약 xLLM

Contents

1 LLM 활용 시 취약점과 대응 전략

2 차세대 LLM보안 솔루션, 알약 xLLM

01.

LLM 활용 시 취약점과 대응 전략

주요 행정기관 금지 조치

美하원, 직원들 MS AI 챗봇 사용 금지... "데이터 유출 우려"

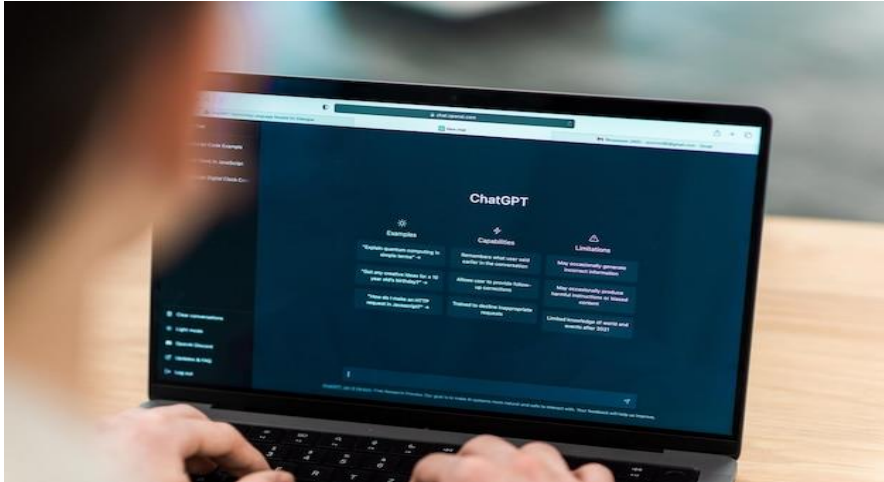


미국 하원 코파일럿 사용 금지

- 승인되지 않은 클라우드 서비스에 하원 데이터 유출 가능성 존재
- 사이버보안국의 “하원에서 사용할 수 없는 앱” 판단
- 모든 하원 MS 윈도우 디바이스 상에서 코파일럿 제거 및 차단 조치
- AI 서비스에 대한 안전장치 마련 촉구

주요 글로벌 금융기관 금지 조치

(서울=연합인포맥스) 정선미 기자 = 씨티그룹(NYS:C)과 골드만삭스(NYS:GS)가 인공지능(AI) 챗봇 챗GPT 사용을 제한했다고 런던 파이낸셜뉴스가 24일(영국시간) 보도했다.



소식통에 따르면 씨티와 골드만삭스는 직원들의 AI 소프트웨어 사용을 차단했다. 씨티는 전사적으로 챗GPT 접근을 허용하지 않았으며, 골드만삭스는 트레이딩 플로어에서의 사용을 제한했다.

소식통은 월가 은행들이 제삼자 소프트웨어의 사용을 일반적으로 제한하는 데 따른 것이며 챗GPT를 특정한 것은 아니라고 지적했다. 이들은 챗GPT 사용을 제한하는 이같은 조치에 대해 직원들에게 따로 알려진 것은 없다고 덧붙였다.

앞서 텔레그래프 보도에 따르면 JP모건 체이스(NYS:JPM)가 챗GPT 사용을 금지했다.

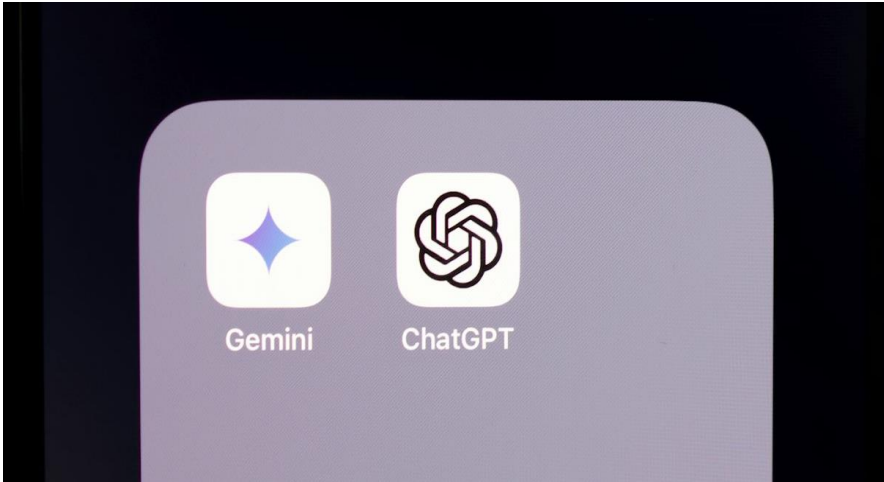
고강도 규제를 받는 은행과 금융기관은 직원들의 제삼자 소프트웨어와 웹사이트 접근을 허용하는 것에 극도로 경계심을 갖고 있다. 대부분 기관은 사무실에서 일할 때나 회사 기기를 사용할 때 페이스북이나 인스타그램과 같은 소셜미디어 플랫폼의 접근을 허용하지 않고 있다.

비공식 비즈니스 채널을 통한 중요/기밀 정보의 공유 금지

- 씨티: 전사적으로 ChatGPT 접근 차단.
- 골드만삭스: 트레이딩 플로어에서 ChatGPT 사용 제한.
- M&G: ChatGPT를 비즈니스용으로 승인하지 않음.
- 도이체뱅크: ChatGPT 접근 차단.
- 밴크오브아메리카: 금지 앱 목록에 ChatGPT 포함.

[출처: 연합인포맥스]

주요 국내외 기업 대응 사례



1 데일리포스트=김정은 기자 | 애플이 일부 직원을 대상으로 챗GPT와 같은 외부 대화형 인공지능 (AI) 이용을 제한하고 있다는 보도가 나왔다.

애플 사내 기밀 데이터 유출 우려로 타사 생성형 AI 사용 금지

- 3월 ChatGPT 버그로 사용자 대화 제목 유출 사건 발생
- 보안을 이유로 타사 생성형 AI 사용 제한
- 자체적인 보안 강화 조치를 통해 기밀 정보 보호 집중

국내 A사 사내에서 ChatGPT 등 생성형 AI 사용 전면 금지

- ChatGPT 사용 후 최소 세 차례 정보 유출 발생
- 보안 지침 위반 시 최대 해고를 포함한 징계 조치 경고
- 내부 AI 도구 개발 및 보안 강화 중

국내 B사 사내 챗봇 서비스에 기업용 챗GPT 오픈

- ChatGPT, Bard, Bing AI 등 외부 대화형 AI 차단
- 생성형 AI 서비스의 보안 이슈 심각성 인식 및 사내 보안 강화
- 보안 리스크를 최소화하는 방안 모색 중

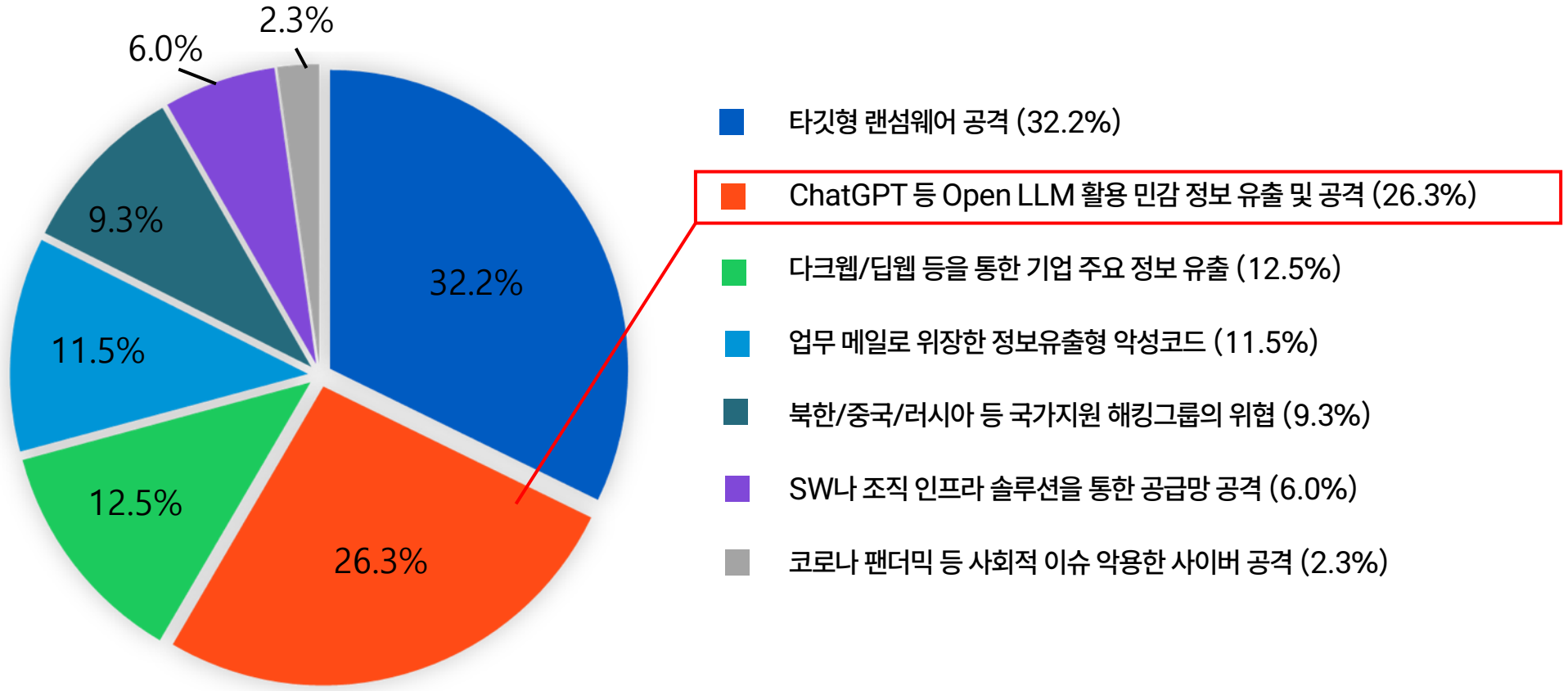
NIS, NSR, GenAI 위협 및 보안 가이드라인 발표 (2023.6)

대표 보안 위협	주요 원인	가능한 보안 위협
잘못된 정보	<ul style="list-style-type: none"> 편향 최신 데이터 학습 부족 환각 현상 	<ul style="list-style-type: none"> 사회적 혼란 조장 고위험 의사 결정 잘못된 의사 결정 유도
AI 모델 악용	<ul style="list-style-type: none"> 적대적 시스템 메시지 	<ul style="list-style-type: none"> 피싱 이메일 및 인몰 도용 사이버 보안 위협 코드 작성 대화형 서비스를 악용한 사이버 범죄 커뮤니티 활성화 사회 공학적 영향 가짜 뉴스 생성
유사 AI 모델 서비스 빙자	<ul style="list-style-type: none"> 유사 악성 서비스 접근 유도 	<ul style="list-style-type: none"> 스퀴팅 URL 및 확장 프로그램 가짜 애플리케이션
데이터 유출	<ul style="list-style-type: none"> 데이터 합성 과정의 문제 과도한 훈련 데이터 암기 문제 대화 과정에서 개인정보 및 민감정보 작성 	<ul style="list-style-type: none"> 훈련 데이터 유출 데이터 불법 처리 우려 기밀 유출 대화 기록 유출 데이터베이스 해킹 및 회원 추론 공격
플러그인 취약점	<ul style="list-style-type: none"> AI 모델의 적용 범위 확장 안정성 확인 미흡 해커 공격 범위 확장 취약점이 있는 서비스와 연결 	<ul style="list-style-type: none"> 새로운 도메인에서의 모델 오작동 '에이전트'화 된 AI 모델의 악용 멀티모달 악용
확장 프로그램 취약점	<ul style="list-style-type: none"> 확장 프로그램 내부의 악성 서비스 설치 서비스 제공 업체의 보안 조치 미흡 	<ul style="list-style-type: none"> 개인정보 수집 시스템 공격 호스팅 서버 및 스토리지 시스템 위협
API 취약점	<ul style="list-style-type: none"> 미흡한 API 키 관리 데이터와 명령 사이의 불분명한 경계 	<ul style="list-style-type: none"> API 키 탈취 악의적인 프롬프트 주입

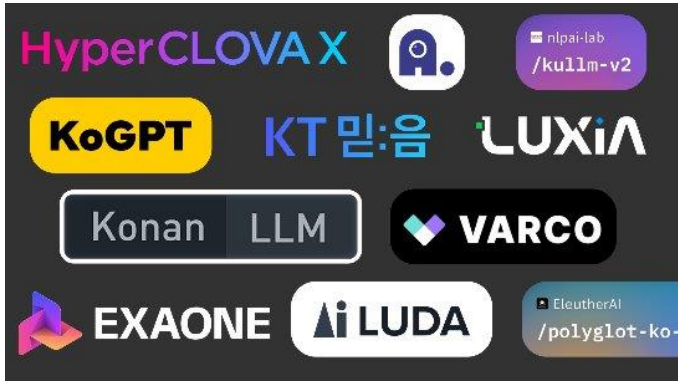


많은 실무 담당자들의 고민

공공/민간 보안 담당자 1791명 설문 : 현재 주요 보안위협 가운데 가장 큰 위협이 될 것으로 전망되는 트렌드는?



그렇다면 대응 전략은?



점차 확산되는 LLM 사용과 보안 위협, 충분히 대비하고 있습니까?

데이터 프라이버시 및 보안

LLM 대규모 데이터셋 내 포함된 민감 정보 유출은
개인정보 침해와 기업 기밀 정보 노출, 법적 문제 등을 야기

악성 콘텐츠 생성

악성 콘텐츠 생성으로 사용자 경험 악화 및 법적, 윤리적 문제 유발 가능

유해정보 생성

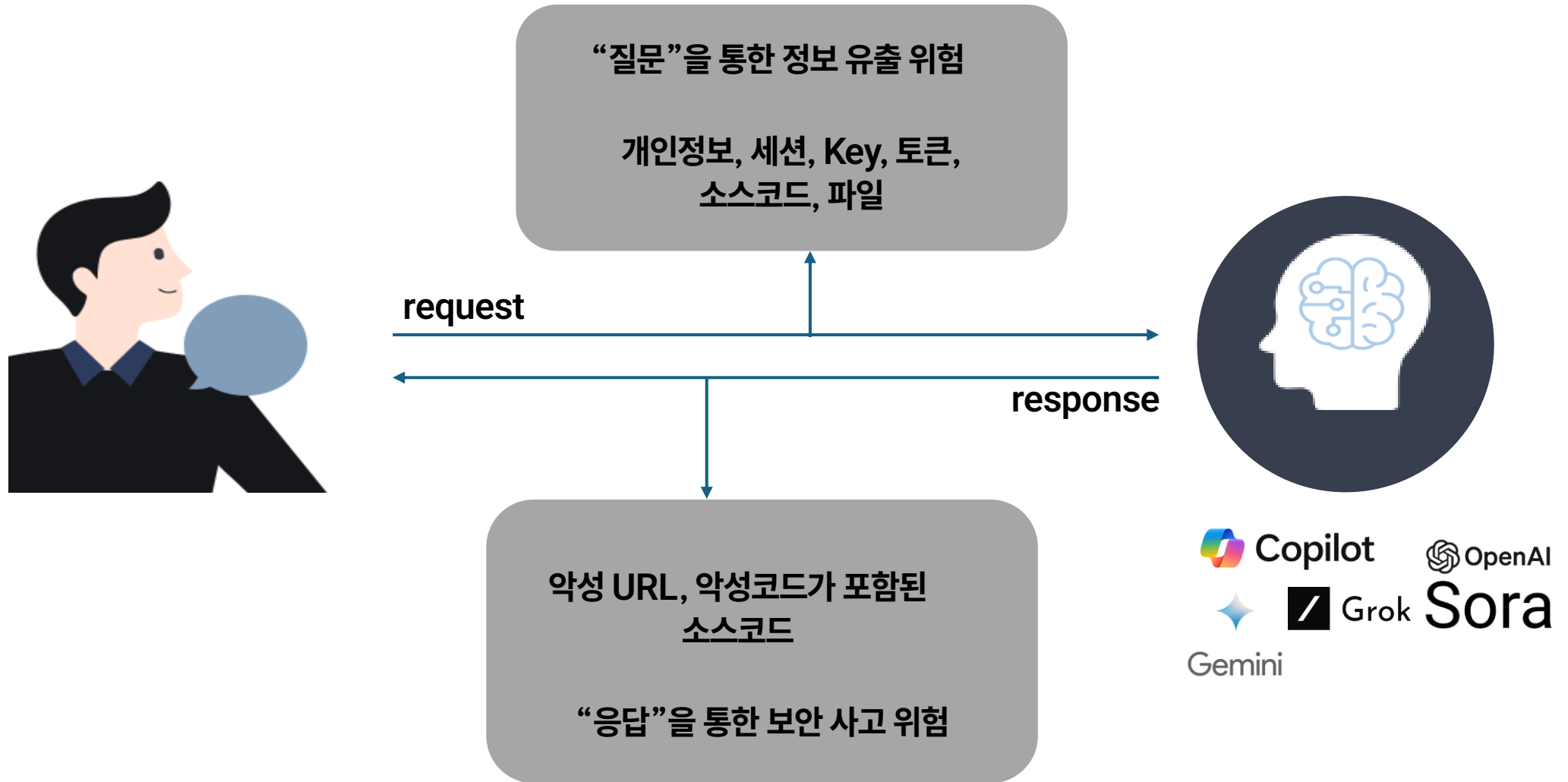
비속어, 욕설 등 부적절한 정보 생성으로
잘못된 신념과 지식을 제공해 사회에 부정적 영향 제공

- 1 늘어나는 LLM 서비스로 인한 개인정보 및 사내 기밀정보 유출 가능성 인지
- 2 LLM 서비스 내 보안 기능이 있으나 고객사 자체적인 보안 및 감사 관리 필요
- 3 개인정보보호위원회 등 LLM 관련 개인정보 보호 취약점 개선 권고사항 준수
- 4 공개 데이터 학습 시 중요 개인 식별정보 제거 및 개인정보 침해 최소화 노력
- 5 보안 담당자는 Open LLM을 활용한 민감 정보 유출 및 공격을 큰 위협으로 인식
- 6 실시간 모니터링, 액세스 제어 등 다양한 AI 관련 보안 솔루션을 검토 필요

02.

차세대 LLM보안 솔루션, 알약 xLLM

02. 차세대 LLM보안 솔루션, 알약 xLLM



“ 정보가 외부로 전달되기 이전, 원천적이고 안전한 보안 조치 ”



개인정보 탐지 및 익명화

- 주민등록번호, 여권번호 등 고유식별 정보 탐지 및 익명 처리
- 주소, 전화번호, 이메일 등 개인식별 정보(PII) 탐지 및 익명 처리



악성 콘텐츠 탐지

- 스팸, 피싱, 악성코드/링크 등의 악성 콘텐츠를 탐지하여 시스템 보안 강화 및 안전한 서비스 환경 제공



유해 표현 탐지

- 혐오와 같은 유해 표현을 감지하고, AI에 무해 정보를 보장하여 서비스 신뢰성을 강화



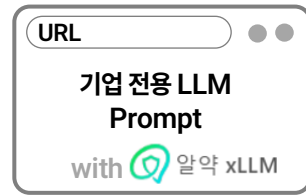
자격 증명(Credential) 탐지

- 프롬프트상 온라인에서의 사용자 이름, 비밀번호와 같은 자격증명(e.g: API key, access key, jwt)이 노출되었는지 탐지하여 자산 정보를 보호
- 유출로 인해 발생할 수 있는 계정 탈취, 무단 접근과 같은 2, 3차 보안 사고 예방

ALYac XLLM (개념도)



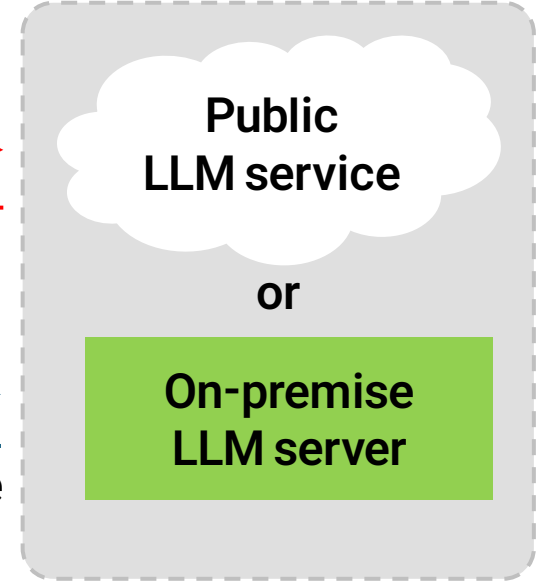
“상용 서비스”로 개인적인 접근과 사용 차단



request

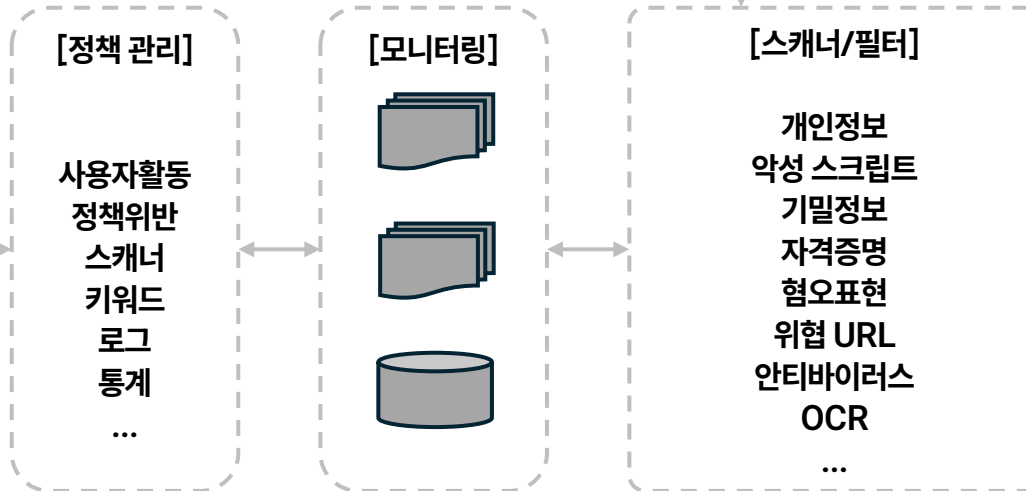


response



보안/관제 담당자

Control



알약 xLLM 위협 대응 흐름



알약 xLLM 주요 특징

시장 상황 및 LLM 보안 취약점에 따른 범위 설정

보안 목표	방어 범위	제공 기능
데이터 보호 Data Protection	데이터 유출/탈취 방어 Data Leakage/Exfiltration Prevention	개인 식별 정보(PII) 탐지 및 처리 예) 이름, 주소, 전화번호, 여권번호 익명화 처리
	기밀 정보 탐지 Confidential Information Detection	자산 정보/보안 정보 탐지 및 처리 예) private key, access token, api key 정보 마스킹처리
	기밀 정보 탐지 Confidential Information Detection	사전 정의된 기밀 정보 탐지 및 처리 예) 기업내 기밀 키워드 마스킹 처리
위협 방어 Threat Prevention	악성 콘텐츠 방어 Malicious Content Prevention	AI 시스템 또는 사용자의 자산 침해를 방어하기 위해 악성 URL/악성 코드 탐지 및 제거 예) 콘텐츠 내 피싱 링크/악성 스크립트 제작 요청/악성 스크립트 실행 요청
공급망 보안 Supply Chain Attack Prevention	취약점 관리 Vulnerabilities Management	공급망 공격을 방어하기 위해 AI 시스템, 배포 플랫폼, 라이브러리에 대한 취약점 관리 예) SBOM 기반 취약점 체크
위험 관리 Risk Management	유해 콘텐츠 탐지 Toxic Contents Detection	불법/사기/약물/폭력 등의 콘텐츠 탐지
	개인 민감 정보 탐지 Personal Sensitive Information Context Detection	금융정보/생체정보/위치정보/종교/성적지향/정치성향 등 민감 콘텐츠 탐지

주요 기능 및 특징

개인 식별 정보 및 기밀 정보 탐지

You 오후 02:29
 아래 내용을 참고해서 부드러운 느낌의 이력서 초안을 작성해줘. 이름을 김이스트입니다. 홈페이지는 estsecurity.com이며, 현재 알약 LLM Security 제품을 담당하고 있습니다. 제 핸드폰번호는 010-2321-5642이며, 주민등록번호는 950721-3844521입니다. 현재 서울특별시 서초구 반포대로 3에 거주하고 있습니다.

Took 1 steps ^

개인정보가 삽입된 프롬프트 입력

E ESTSecurity_input_filtering 오후 02:28

검사 결과

분석 시간	0.2297520637512207
약성 콘텐츠 탐지	음성
약성 URL 탐지	음성
기밀 정보 탐지	양성
혐오 표현 탐지	음성
온라인 자격 증명 노출 탐지	음성
개인 식별 정보 탐지	양성

알약 xLLM 서버에 의해 기밀 정보 탐지

무해화 결과

아래 내용을 참고해서 부드러운 느낌의 이력서 초안을 작성해줘. 이름을 [PERSON]입니다. 홈페이지는 estsecurity.com이며, 현재 [REDACTED-Confidential] 제품을 담당하고 있습니다. 제 핸드폰번호는 [PHONE_NUMBER]이며, 주민등록번호는 [KR_SSN]입니다. 현재 [LOCATION]에 거주하고 있습니다.

기밀 정보 무해화 후, ChatGPT등 LLM 서비스에 전달

D dolphin-mistral:latest 오후 02:29

안녕하세요! 저는 [PERSON]입니다. 홈페이지를 estsecurity.com으로 방문할 수 있습니다. 현재, [REDACTED-Confidential] 제품을 담당합니다. 연락처는 [PHONE_NUMBER]입니다. 주민등록번호는 [KR_SSN]입니다. 나는 현재 [LOCATION]에 거주하고 있습니다.

LLM서비스는 무해화 정보 기반으로 답변을 수행

저는 이력서의 초안을 작성해드리겠습니다. 제가 작성한 이력서의 본문은 어떤 내용으로 작성할지 알려주세요.

주요 기능 및 특징

악성 콘텐츠 및 URL 탐지

검사 결과

소모 시간	0.710440158849941
악성 콘텐츠 탐지	양성
기밀 정보 탐지	음성
혐오 표현 탐지	음성
온라인 자격 증명 노출 탐지	음성
개인 식별 정보 탐지	음성
악성 URL 탐지	음성

응답에서 악성 콘텐츠 감지되어

무해화 결과

악성 콘텐츠가 [MALICIOUS_CONTENT]로 필터되어 노출

Windows PowerShell 스크립트는 Windows에서 사용되는 명령줄 구문 및 인터페이스를 제공하며, 여기서는 TCP/IP 클라이언트가 서버와 통신할 수 있는 간단한 예시입니다. PowerShell은 명령줄 인터페이스를 사용하거나 그래픽 사용자 인터페이스에서도 실행할 수 있습니다.

PowerShell 코드는 서버와 TCP/IP 통신을 수행하고, 전송된 데이터를 수신하여 해당 데이터에 대한 명령어를 실행합니다. 그런 다음 결과를 서버로 보내며 루프가 반복됩니다.

아래는 PowerShell 코드의 한국어로 번역된 예시입니다:

[MALICIOUS_CONTENT]

이 코드는 PowerShell을 실행하여 TCP/IP 서버와 통신하며, 데이터를 수신한 다음 해당 데이터에 대한 명령어를 실행하고 결과를 전송합니다. 이러한 방식으로 TCP/IP 클라이언트와 서

알약 xLLM 서버에 의해 악성 콘텐츠 탐지

악성 정보 무해화 후, 사용자에게 전달

ESTsoft AI Service

PERSO, ALAN

the most universal interface for Human



PERSO.ai

AI Video Translator

Transcend Languages, Transform Communication

EST

최고의 상호작용을 위해 설계된 대화형 AI 에이전트, 앨런

사람과 말하듯 자연스러운 대화

“오늘 날씨에 맞는 옷차림은 뭐야?”
 “가볍게 들고 다닐 태블릿 거치대 추천해줘!”
 무슨 말이든 이해하고 알아듣는 앨런과 편하게 대화하세요.

기업 LLM 보안 위협에 대응하는 필수 전략, 알약 xLLM

감사합니다.

브로셔 다운로드

문의 및 컨설팅 요청