AI 보안의 새로운 패러다임

ALYac LLM Guard 1.0

알약 LLM Guard 1.0

이스트시큐리티는 LLM(대규모 언어 모델) 사용 시 발생할 수 있는 데이터 유출과 악성 콘텐츠 수신 위협을 탐지하고 차단합니다. 안전성을 강화하여 개인과 기업이 효과적으로 AI를 사용할 수 있도록 지원하며, AI 서비스 기업이 고객에게 신뢰성 높은 솔루션을 제공하는 데 기여합니다.

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

알약 LLM Guard 위협 대응 흐름

위협 모니터링

사용자/AI의 상호작용에서 기밀 정보, 악성 콘텐츠 등 위험 요소를 감지합니다.

경고 및 알림

위협 탐지시 탐지 안내 알림을 통해 추가조치 필요 여부를 판단할 수 있습니다.

위협 대응

탐지된 위협에 대해 사전에 정의된 방법으로 대응합니다. (맞춤화 가능)

대응 내역 수집/보고

제공되는 대시보드를 통해 탐지된 위협, 대응 조치, 위험도, 성과 등을 관리자에게 보고합니다.

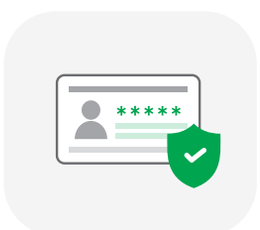
ESTSECURITY

기대효과

LLM 보안 위협에 대한 대응 능력 강화

기밀 정보/자산 정보 탐지 및 악성 컨텐츠 탐지로 사용자/AI 서비스 환경 보호	AI 활용에 대한 정보보안 거버넌스 달성 지원	탐지 위협, 대응 조치 등 AI 관련 위협 가시성 제공

핵심 기능



개인정보 탐지 및 익명화

주민등록번호, 여권번호와 같은 고유식별 정보와 주소, 전화번호, 이메일 같은 개인식별 정보(PII) 등을 탐지 및 익명 처리하여 민감 정보가 외부로 유출되는 위험을 줄일 수 있습니다.



악성 콘텐츠 탐지

스팸, 피싱, 악성코드/링크 등의 악성 콘텐츠를 탐지하여 사용자에게 전달되지 않도록 보호하여 시스템 보안을 강화하고 안전한 서비스 환경을 제공합니다.



유해 표현 탐지

혐오와 같은 유해 표현을 감지하여, 사용자의 서비스 경험을 좋게하고,
AI에 무해 정보를 보장하여 서비스 신뢰성을 높일 수 있습니다.



자격 증명 (Credential) 탐지

프롬프트상 온라인에서의 사용자 이름, 비밀번호와 같은 자격증명(e.g: API key, access key, jwt)이 노출되있는지 탐지하여 자산 정보를 보호하여, 그 유출로 인해 발생 할 수 있는 계정 탈취나 무단 접근과 같은 2차, 3차 보안 사고를 사전에 예방할 수 있습니다.

상담 문의 정보

메일 : sales@estsecurity.com 홈페이지 : www.estsecurity.com 상담전화 : 02.583.4616

ESTSECURITY