GenAI 보안 플랫폼

SAIFE X

OCT. 2024

eRun&company

Table of contents

eR((un&company

SAIFE X

01

배경 및 필요성

SAIFE X

Market Needs & Problems

발목 잡는 보안 위협

"**회사 욕 해달라**"는 고객 부탁…
인공지능은 그저 열심히 일했을 뿐

우려가 현실로…OO전자, ChatGPT
빗장 풀자마자 **오남용** 속출

구글 ChatGPT에서 **개인정보추출**
성공… "LLM훈련 데이터 파악 가능"

ChatGPT 허용 20일,
정보유출 사고 3건 발생

분쟁 부른 ChatGPT의 **실수**
오픈 AI 피소

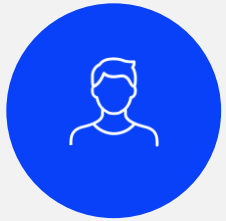
"**마약제조법을 알려달라**"
Hallucination을 이용한 Jailbreaking 공격

ChatGPT 통해
부정확한 정보 확산

오류 소스 코드 전부 복사,
ChatGPT 해결 방법 문의

계측, 수율 데이터,
미국 기업에 **고스란히 전송**

가시성 확보의 중요성



N 정부기관 전산 담당팀

- 기관 내부에서 **생성형 AI 니즈**가 지속적으로 발생
- **내부 정보 유출 위험**으로 사용 보류 중
- 생성형 AI 사용을 **모니터링/통제**할 수 있는 수단이 준비되면 사용 의사 있음

S 대기업 정보보안 담당팀

- AI를 활용한 **생산성 향상 니즈** 발생
- 직원들에게 **정보보안 서약서 작성** 후 개인적 사용 가이드
- 사고 발생 시 모든 책임은 개인에게 부여 중

O SW기업 연구소장

- 개발자가 생성형 AI를 활용 시 **업무 생산성 3~5배 증가**함을 인지
- 기업 내 **소스코드 유출, 의도치 않은 코드의 내부 유입**에 대한 우려로 미사용

K 보안산업진흥기관 본부장

- 정부차원에서 **생성형 AI 관련 보안 이슈에 대해 많은 요구를 받고 있음**
- **AI산업 진흥을 위한 AI보안 이슈 연구/해결**이 급선무 임을 피력

K 인재양성 교육원

- 교육생 대다수가 **생성형 AI를 활용하여 프로그래밍** 하기를 요구
- 사용에 대한 **가시성 확보 요구, 다양한 생성형 AI 사용 요구**
- **생성형AI 보안에 대한 기본 교육** 요구

Y 대학교 보안학과 교수

- 생성형 AI의 새로운 **보안이슈에 대한 정부차원의 연구/대응 요구** 받고 있음
- 국내 초기 기술로서 국가 경쟁력 확보차원에서 공동 연구 요청

기업이 직면한 문제

ISSUE	PROBLEM
S전자 소스코드 업로드/분석 요청	**기밀정보 유출**
국가 정보기관/진흥원 디테일한 사용현황 파악 니즈	**LLM 모니터링 니즈**
의료기관 LLM에 환자 개인정보, 세부 병적 알려달라	**LLM Jailbreak 등 오남용 제어** 시스템 오동작, 기업 평판 추락

02

제품개요

SAIFE X

Product Overview

솔루션과 기술

기밀정보 유출	기밀정보 유출 방지	**PII, BII, Code, Secret Detection** 개인정보, 마감 전 회계/재무상황, M&A계획, 소스코드, Software License Key, API Key
LLM 모니터링 니즈	LLM 가시성 확보	**Admin Dashboard** 업로드/차단된 민감정보 현황, Prompt/Response 현황/검색, 다양한 LLM 사용현황 통합 모니터링 (ChatGPT/Claude/sLLM)
LLM 오남용	리얼타임 해킹대응	**Prompt Jailbreaking Detection** Prompt 모니터링/필터링 통한 Injection/Jailbreaking 탐지. LLM에 대한 Offensive Research (Prompt모의공격)

GenAI 보안 플랫폼 SAIFE X

정보 유출 방지

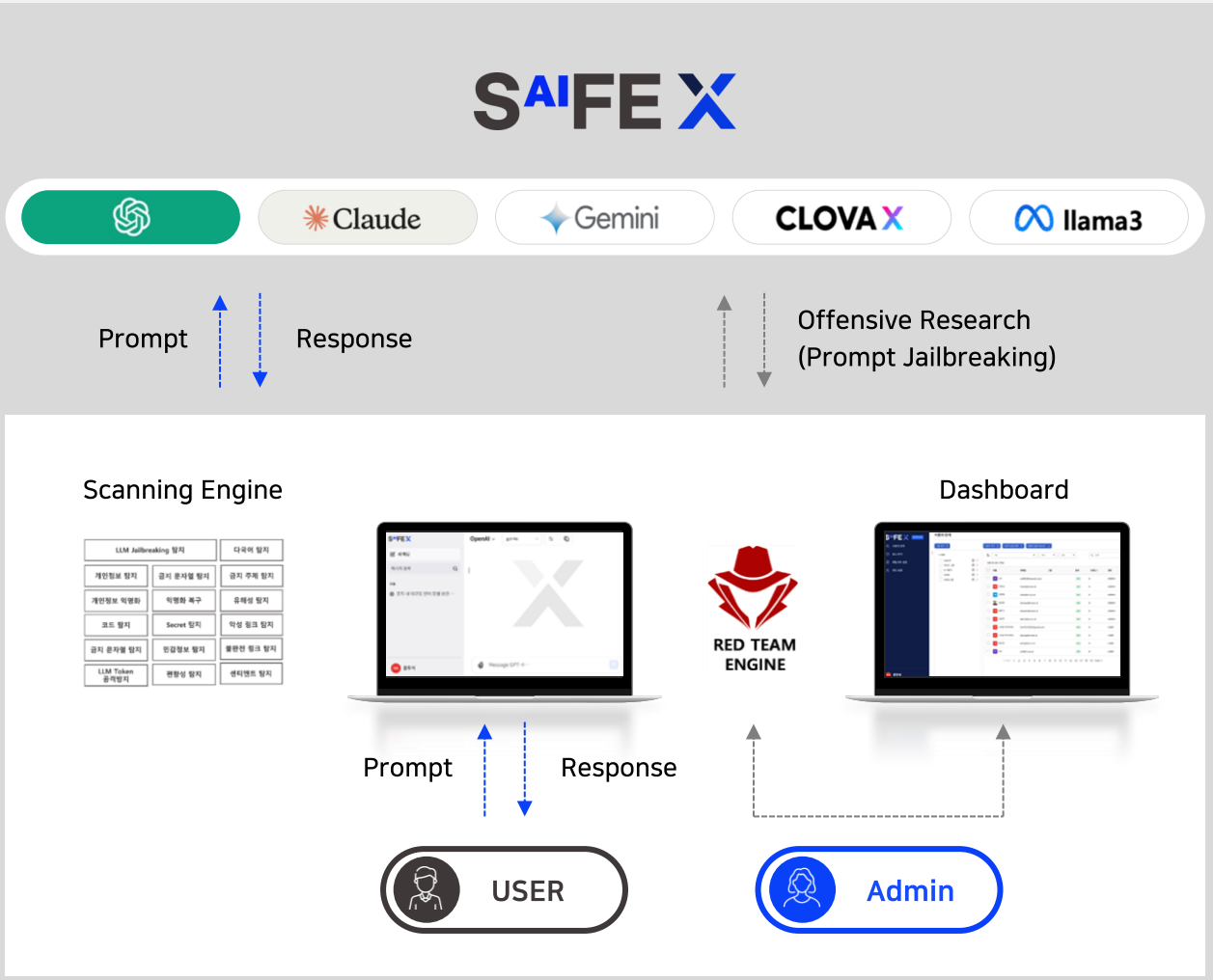
기본(이름, 지역, 생년월일, 이메일, 신분증, 전화번호)
소속(사업자번호, 조직명), 금융(계좌, 카드, 비트코인),
네트워크(IPv4/6, Mac, UUID, URL), …, 지속적 확대 중

다양한 GenAI 통합 사용

ChatGPT, Claude, Gemini, llama 등
다양한 GenAI를 선택하여 즉시 사용
경쟁력 있는 가격으로 다양한 AI 사용

실시간 모니터링 및 차단

내부 사용자의 프롬프트/응답 사용현황 실시간 모니터링
개인정보, 기밀정보 차단/허용 리스트 정의



03

제품 특장점

SAIFE X

Competitive Strength

핵심기술

인공지능 판별 엔진

인공지능 기반 NER
Named Entity Recognition

월등한 판별율(90%이상)
기존 한글모델 대비 25% 이상 높은 판별율

오탐율 최소화
Keyword/Regular Expression 필터링 대비

LLM 특화 보안기술

LLM에 특화된 해킹/방어기술 보유

Prompt Injection, Prompt Jailbreaking등
Prompt Jailbreaking Test 자동화

기업 내 운영중인 LLM에 대한
Offensive/Defensive Research 제공

특징 및 장점

편리한 사용 환경

- 다양한 LLM의 통합 환경 제공
- [Public LLM, Local LLM 설정 가능]
- 편리한 오탐/미탐 신고 [One-Click 리포팅]

타사대비 높은 민감정보 판별율

- 인공지능 기반 NER모델을 통한 오탐 최소화
- 다양한 민감정보 판별 설정 [개인정보, 사업자정보, 금융정보, 네트워크, 기업Secret, 민감 키워드]
- 다국어 정보 판별 [한국어/영어]

실시간 모니터링 및 차단 관리

- 사용자 프롬프트/응답 실시간 모니터링 [정상/ 검출 / 미전송 로그 관리]
- 사용자별/키워드별 검색
- 사용자의 오탐/미탐 신고 관리

효율적인 정책 및 비용 관리

- 고객의 요구에 맞는 다양한 정책 적용 [각 민감정보 스캐너 On/Off]
- 각 LLM별 토큰 사용량 및 정산 관리

기대효과

04

SAIFE X

Key Features

제품 주요기능

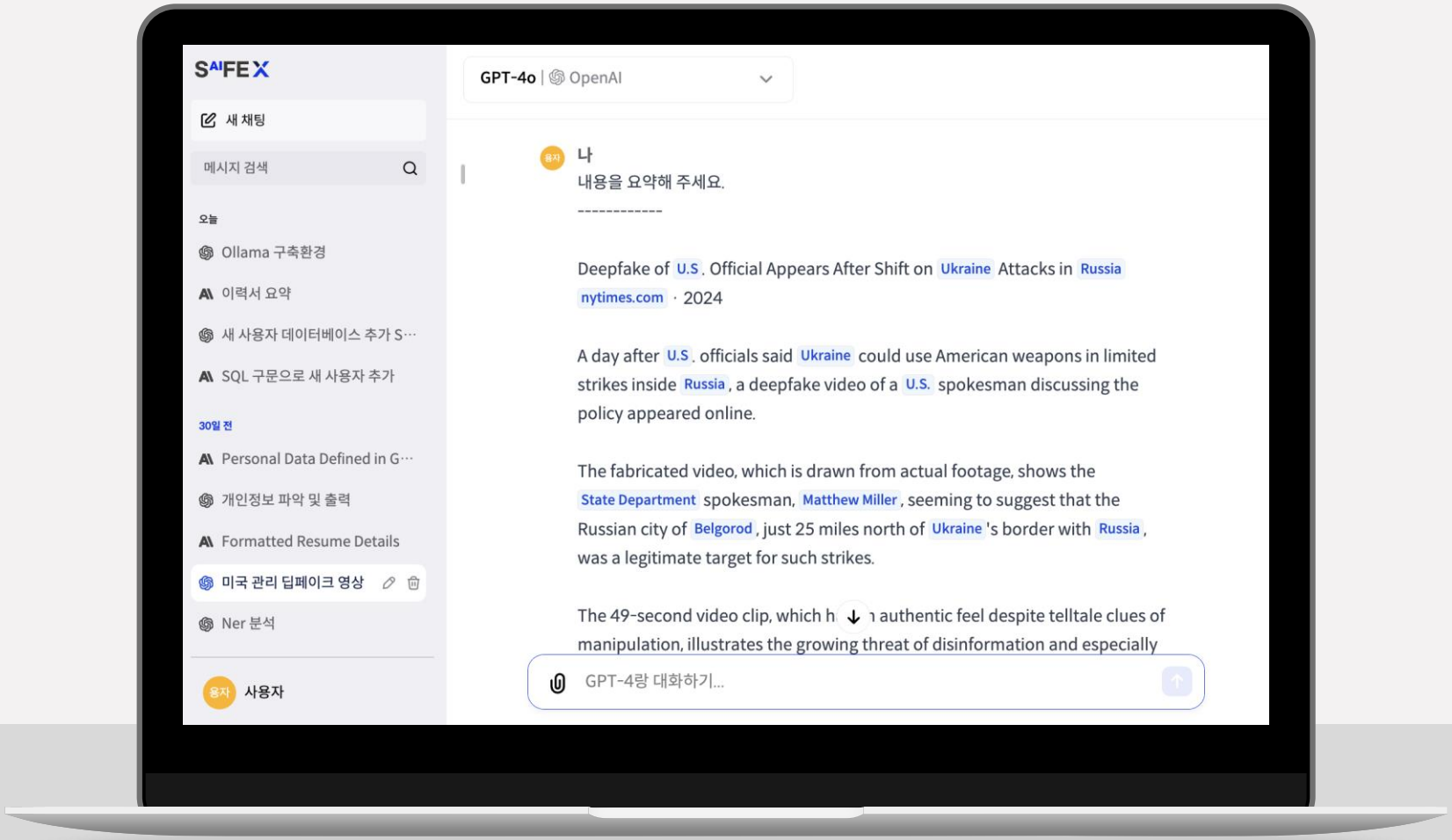
사용자 Chat UI

통합 Chat UI

01 Prompt/Response

02 텍스트, 첨부파일 업로드 가능

03 채팅 쓰레드 관리



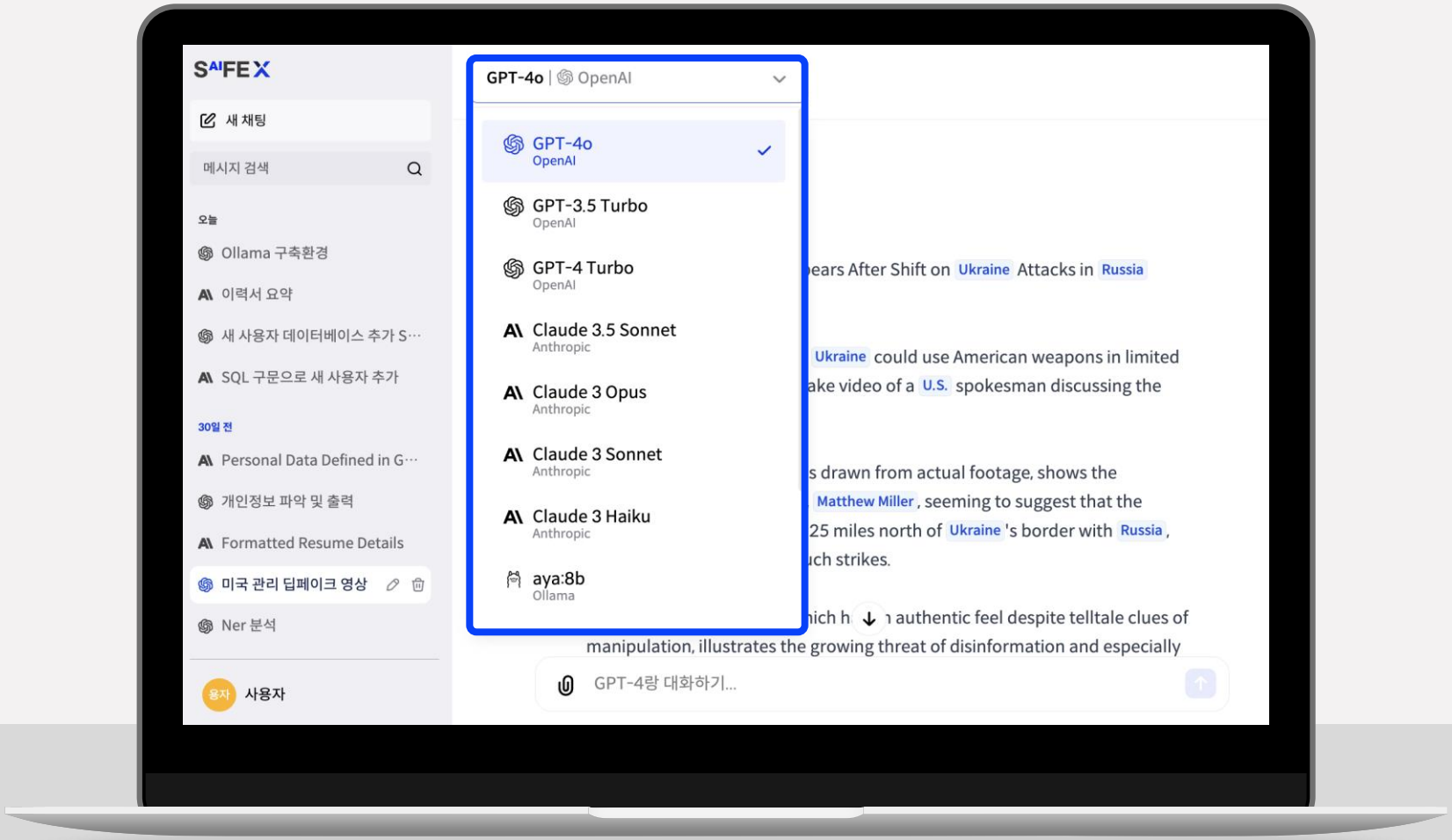
사용자 Chat UI

다양한 LLM 통합 사용

01 ChatGPT, Claude, Local LLM 선택 가능

02 LLM별 사용 모델 선택 가능

[GPT-4o, GPT-3.5 Turbo, GPT-4 Turbo, Claude 3.5 Sonnet, Claude 3, Llama, Aya, …]



사용자 Chat UI

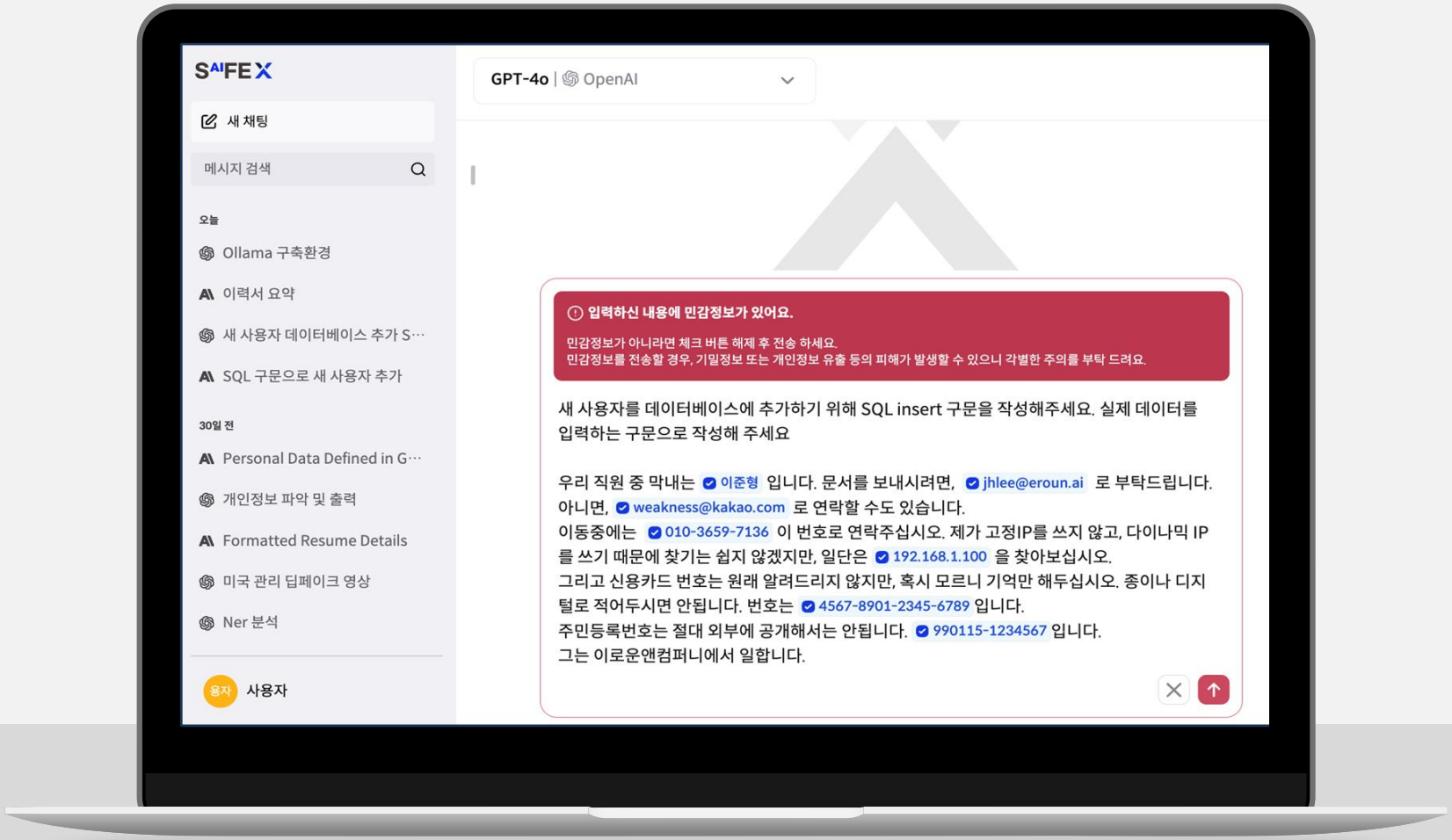
민감정보 실시간 판별

01 개인정보, 기업민감정보 실시간 판별

[Prompt에 입력 후, 즉시 판별]

02 판별된 민감정보에 대한 실시간 오탐/미탐 신고

[체크된 민감정보에 대한 실시간 오탐 신고
미탐 텍스트에 대해 단어선택 후 미탐 신고]



관리자 UI

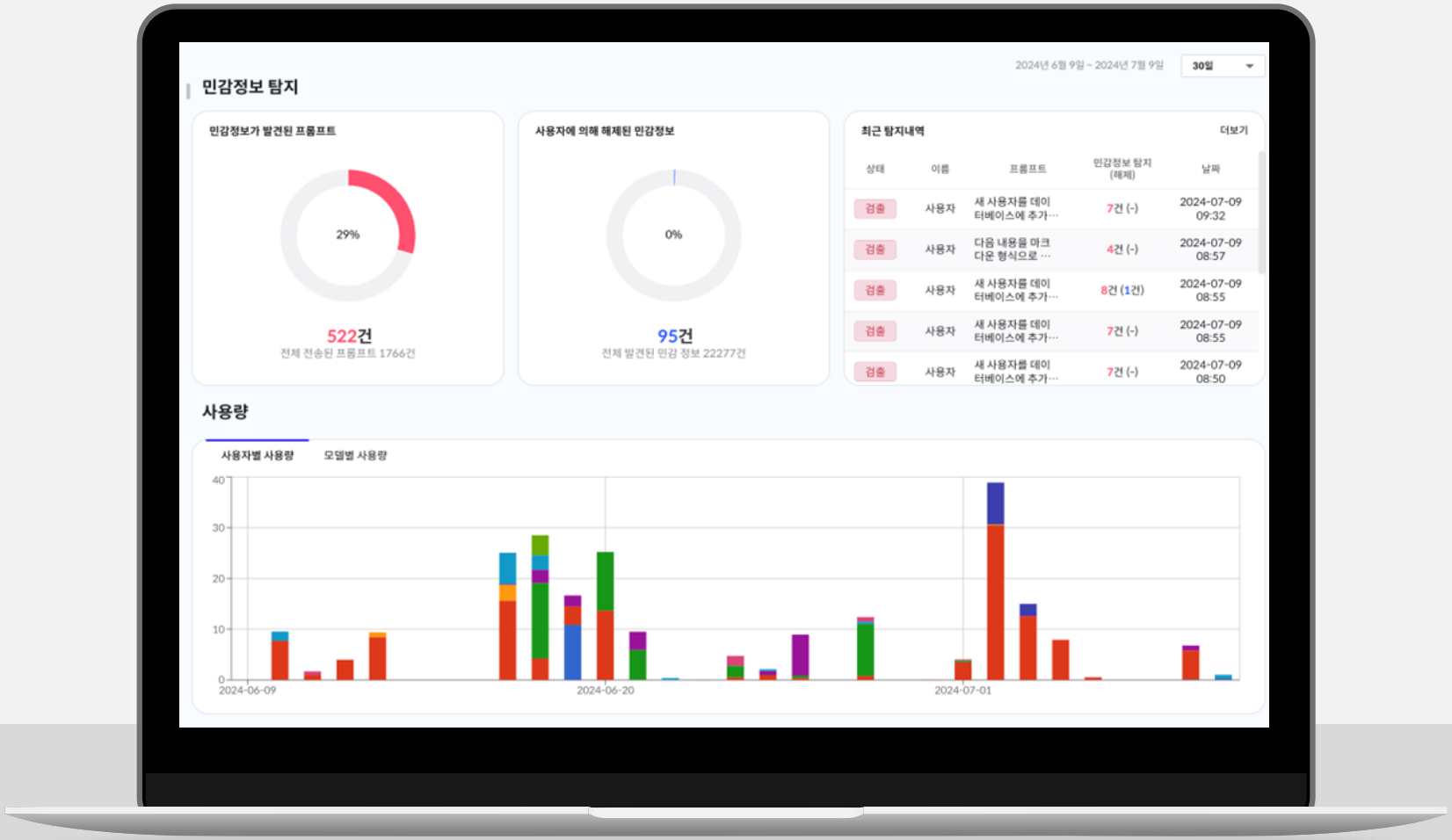
관리자 대쉬보드

01 판별된 민감정보

02 사용자에 의해 해제된 민감정보
[오탐신고]

03 최근 검출된 탐지내역 리스트

04 사용자별, LLM 모델 별 사용량 관리



관리자 UI

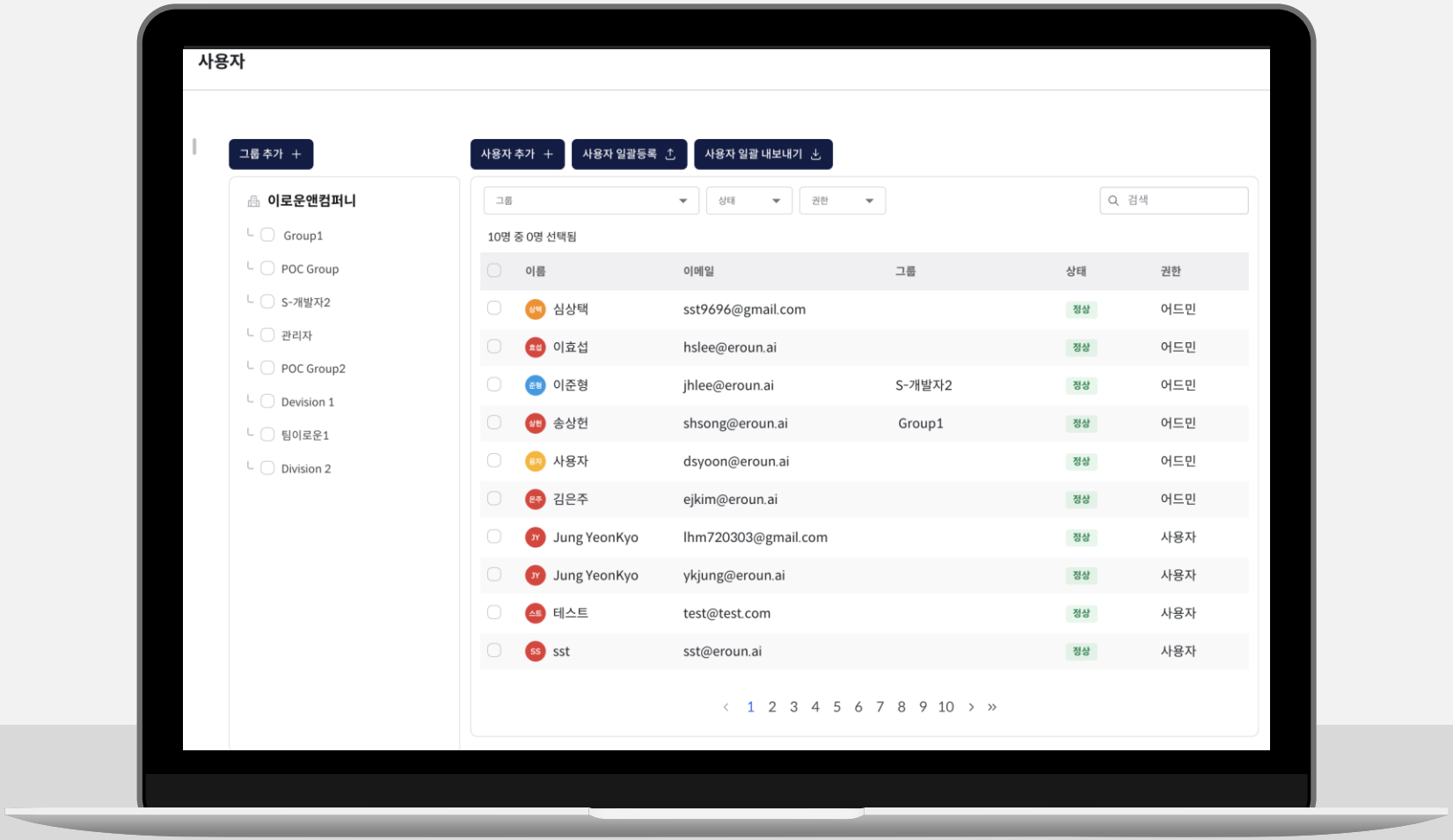
사용자 관리

01 기업 내 부서 관리
[그룹추가]

02 사용자 추가 및 관리

03 사용자 일괄 등록

04 사용자 일괄 내보내기



관리자 UI

통합 로그 관리

01 사용자의 Prompt/Response 실시간 모니터링

02 정상 Prompt

[민감정보가 포함되지 않은 프롬프트]

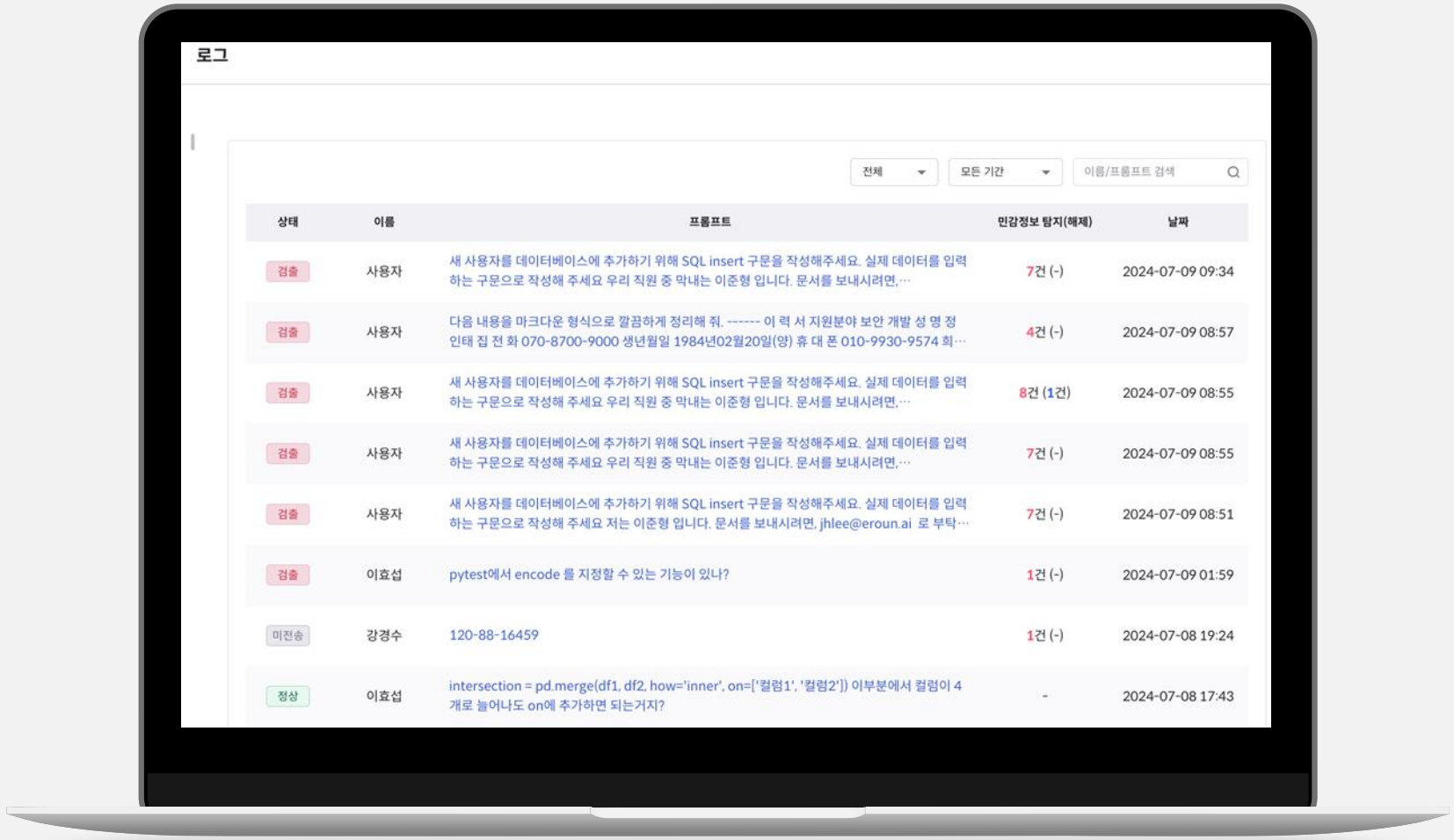
03 미전송 Prompt

[민감정보가 검출되어 사용자가 전송하지 않고 취소한 프롬프트]

04 검출 Prompt

[사용자가 검출된 민감정보를 익명화하여 LLM의 응답 요청을 한 프롬프트]

05 탐지된 민감정보, 사용자가 해지한 민감정보 리스트



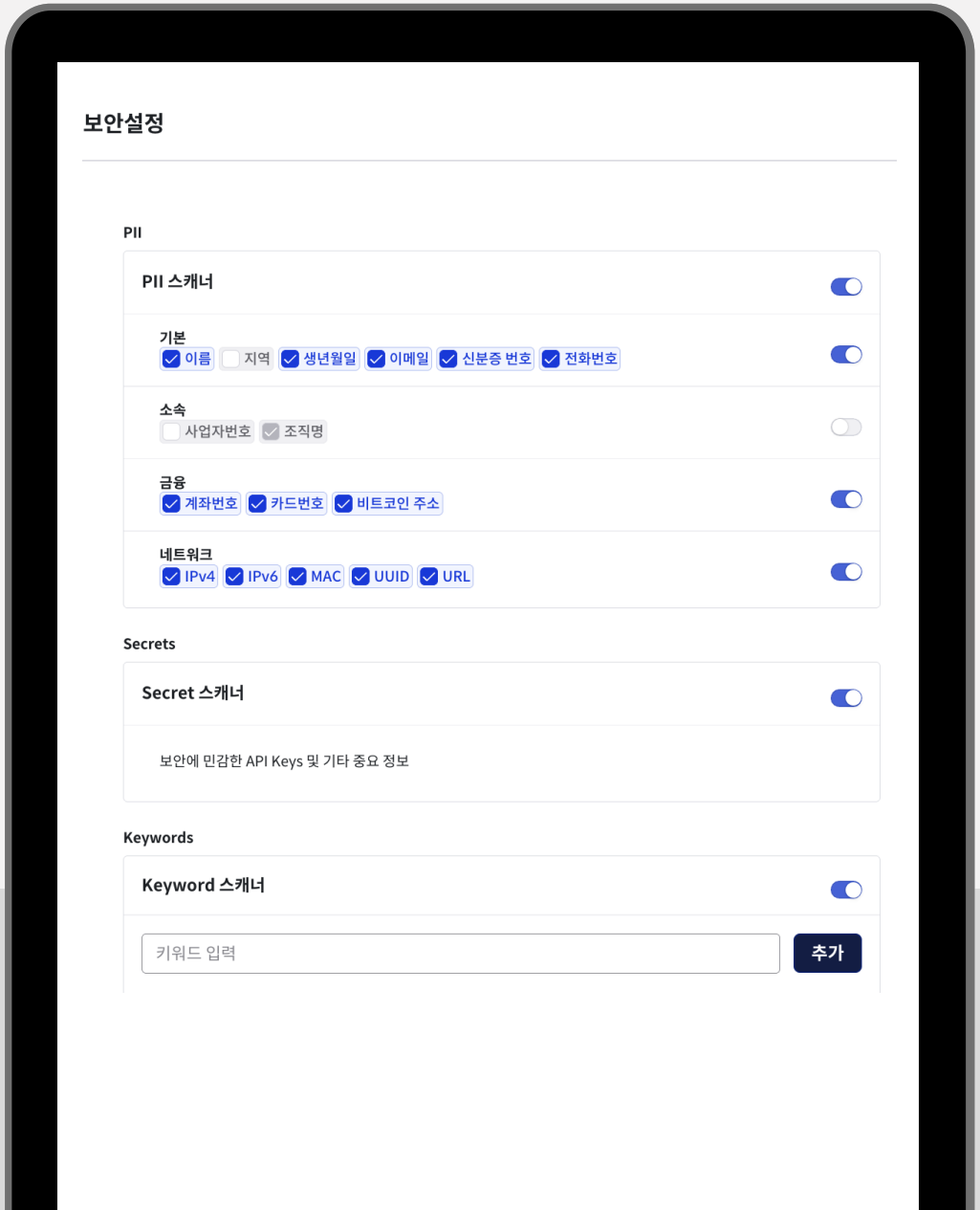
관리자 UI

민감정보 판별 설정

01 개인정보, 소속정보, 금융정보, 네트워크 정보 판별 설정
[기업의 환경에 맞게 각 요소 체크/언체크 가능]

02 기업 내 API Key, Software License Key 판별

03 기업이 지정하는 키워드 필터링 설정



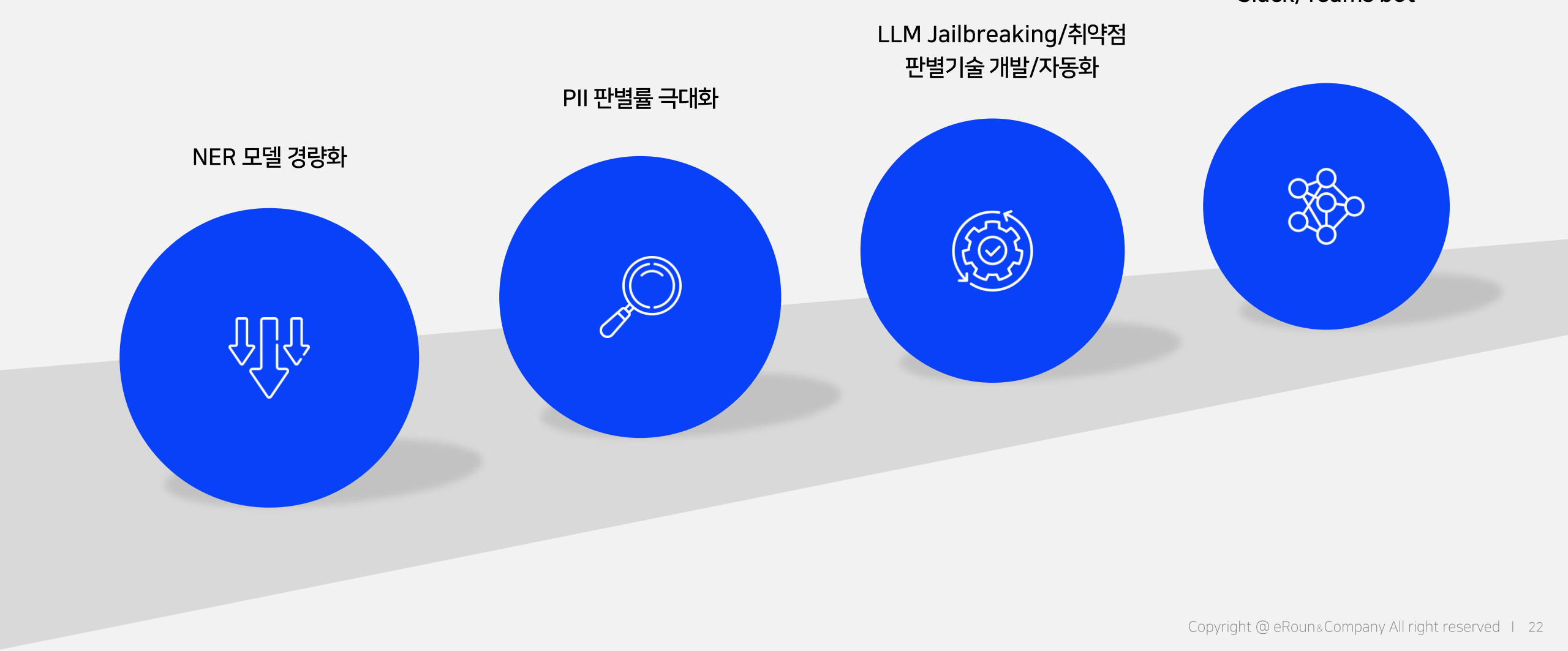
2024 개발 방향

NER 모델 경량화

PII 판별률 극대화

LLM Jailbreaking/취약점
판별기술 개발/자동화

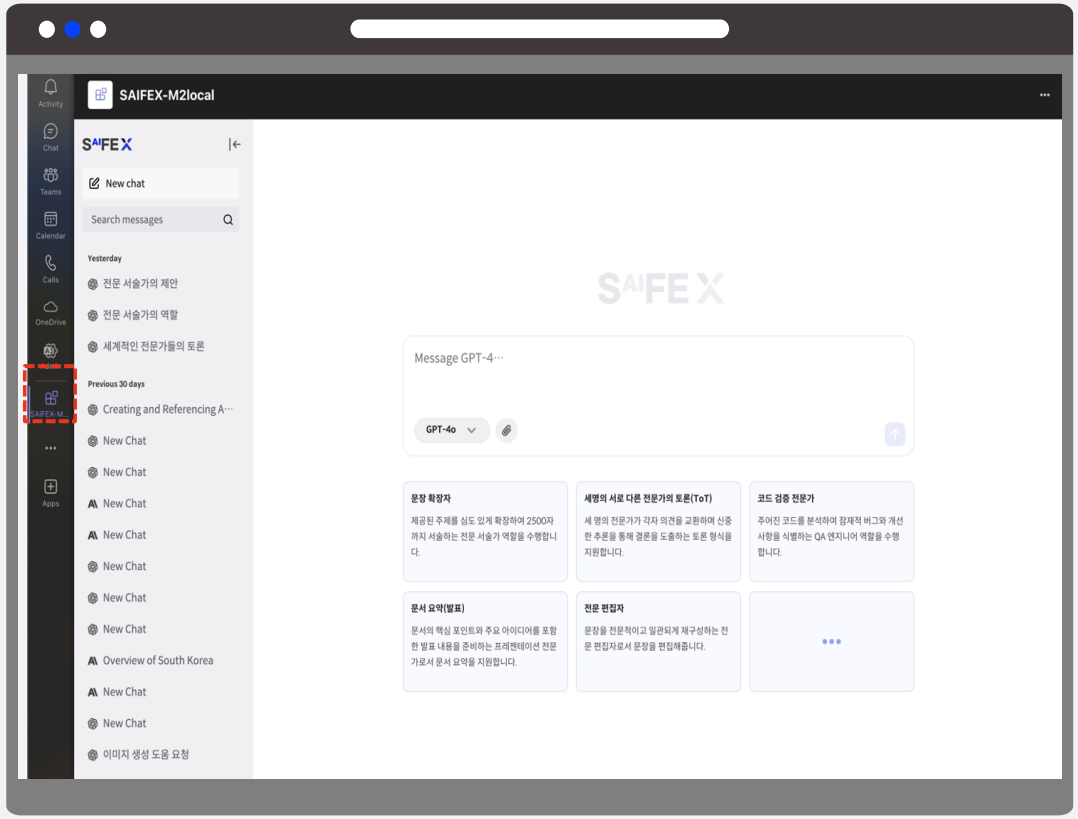
Slack/Teams bot



LLM Interface in Teams 예시

01 **Teams App 내에 LLM 버튼 추가**

02 **LLM 버튼 클릭 시,
Single Sign On을 통한 SAIFE X로의
자동 로그인**



05

회사소개

SAIFE X

Team Profile

팀 이로운, 프로필

오버뷰

회 사 명	주식회사 이로운앤컴퍼니
대 표 이 사	윤두식
설 립 일 자	2024년 1월 1일
자 본 금	1억원
임 직 원 수	9명
사 업 분 야	생성형AI 보안, 프롬프트 엔지니어링 교육 및 컨설팅
대 표 제 품	생성형AI 보안 플랫폼 , SAIFE X (세이프엑스)
본사소재지	서울시 강남구 테헤란로 503
홈 페 이 지	https://www.eroun.ai

TEAM

eROUN&COMPANY

주요지표

- **2024.01** 법인설립
- **2024.02** Pre-Seed 투자유치 (마크앤컴퍼니 & 윤민창의투자재단)
- **2024.02** 기업부설연구소 "AI보안연구소" 인증 획득
- **2024.03** 벤처기업 인증 획득
- **2024.03** 신용보증기금 혁신창업기업 선정
- **2024.06** 중소벤처기업부 딥테크팁스 패스트트랙 선정
- **2024.07** SAIFE X 최소기능제품(MVP) 출시

팀 이로운, 목표

Your Reliable GenAI Security Partner

We enable enterprises to safely and securely leverage the power of GenAI.

팀 이로운은 생성형 AI의 보안 위협과 리스크들이 있음에도 생성형 AI가 선물하는 창의적 우연에 감동했습니다. 기업이 안전하게 생성형 AI를 잘 활용한다면 지금까지 와서는 차원이 다다른 혁신과 성장을 경험하게 되리라 확신했어요. 2024년 1월, 팀 이로운은 이를 입증하고자 결성되었습니다.

25년 이상 정보보안 업계에서 쌓은 경험과 인사이트, 네트워크를 레버리지로 잘 활용하겠습니다. 이 경험에 더해 신선한 에너지와 크리에이티브, 민첩성을 가진 크루의 능력으로 시너지를 내겠습니다. 문제에 부딪힐 때마다 우리의 가설과 전제가 틀릴 수 있다는 생각으로 고민하겠습니다. 빠르게 배우고 성장하는 조직이 되겠습니다.

그리하여 고객에게 안전한 GenAI 경험을 선물하겠습니다.

06

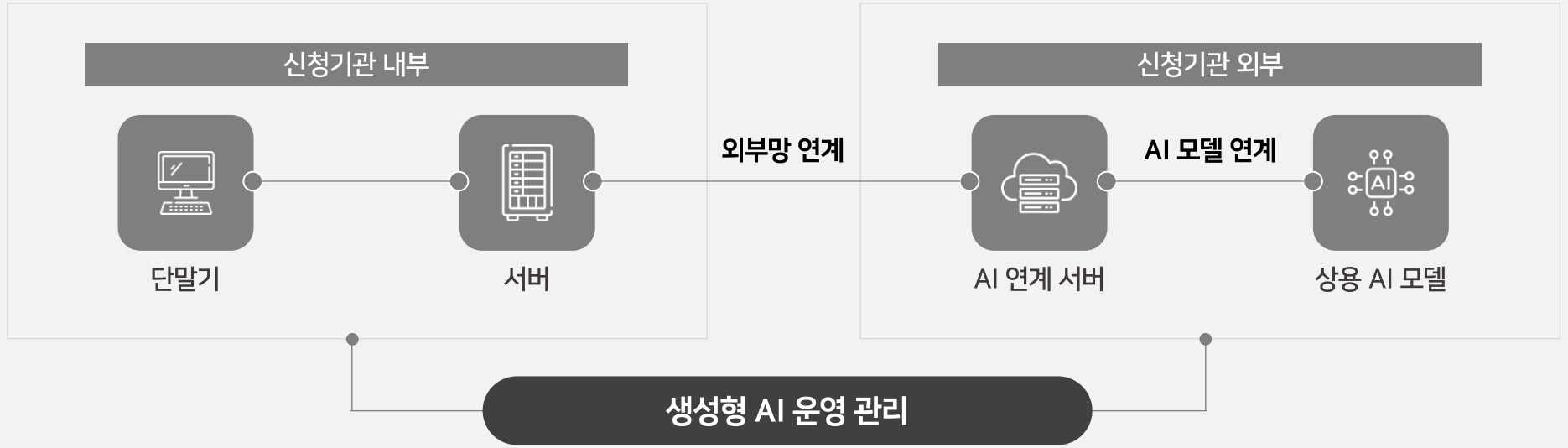
SAIFE X

금융보안원 생성형AI 및 SaaS 이용 규제특례 관련 보안대책

SAIFE X 대응방안

GenAI 활용 구조

금융보안원 가이드



SAIFE X 구조도

단말기/서버/네트워크 보안대책

	단말기	네트워크	기타
금융보안원 가이드	✓ 생성형 AI에 질의하는 단말기는 개인신용정보(가명처리정보는 제외) 등 중요정보가 유출되지 않도록 방지대책 마련	✓ "외부망 연계" 구간은 감독규정에 따라 전용회선 또는 VPN 등 활용 ✓ "AI모델 연계"구간은 안전한 암호 알고리즘을 활용하여 전송자료 암호화	✓ 생성형 AI활용을 위해 클라우드 컴퓨팅 서비스를 활용하는 경우 감독규정 별표2의4에서 정한 안전성 확보조치 준용
SAIFE X 기술적 대응	✓ **Prompt/Response 등 모든 정보 모니터링** ✓ **개인정보/민감정보 차단** ✓ **중요정보 암호화 저장 [TBD]**	✓ **외부망 연계** SSL/TLS VPN을 통한 안전한 통신 보장 ✓ **"AI모델 연계"HTTPS(TLS)** 프로토콜을 이용한 안전한 통신 보장	✓ **주요 안전성 확보조치에 대한 대응 예정(TBA)**

GenAI 모델 보안대책

금융보안원 가이드	SAIFE X 기술적 대응
상용 생성형 AI 모델 제공자가 적대적 공격 방지 대책을 수립하고 이행하고 있는가 확인	적대적 공격 테스트 기술 [Prompt Injection공격 자동화테스트 툴 개발 중. 2024년 말 예정]
생성형 AI 모델의 강건성 확보를 위해 필요 보안대책을 마련하여 이행	사용자의 입.출력값 필터링, 프롬프트 질의 제한 [개인정보, 민감정보], NER 모델/Prompt Jailbreaking 판별 엔진(예정) 등 제공
입출력 데이터를 대상으로 적대적 공격 여부를 확인하고 적대적 공격에 대비한 대응방안 마련	Prompt Injection/Jailbreaking Engine 개발 중
개인신용정보 등 중요정보가 생성형AI 모델에 입력되지 않도록 방지조치 이행	Prompt에 입력되는 각종 민감정보 실시간 판별, 사용자/관리자에게 판별내용 알림
출력 데이터 또는 에러메시지 내 중요정보나 AI 모델 정보 등이 노출되지 않도록 조치	출력데이터에 대한 SafeGuard 필터 엔진 개발 중 [2024년말 예정]
이용자의 생성형AI 요청 및 결과 출력 횟수를 일정 수준 이하로 제한 [제한 수준은 각 사의 환경에 맞게 자체 설정]	사용자별 Prompt/Response 횟수 지정 제한기능 제공

적대적 공격 판별 및 테스트 기술 SAIFE X

01 Jailbreaking 데이터 수집

- 글로벌 네트워크를 통해 Prompt Jailbreaking 데이터 셋 수집
- 수집된 데이터셋에 대해 고유한 패턴을 분석

02 Jailbreaking 데이터 확장

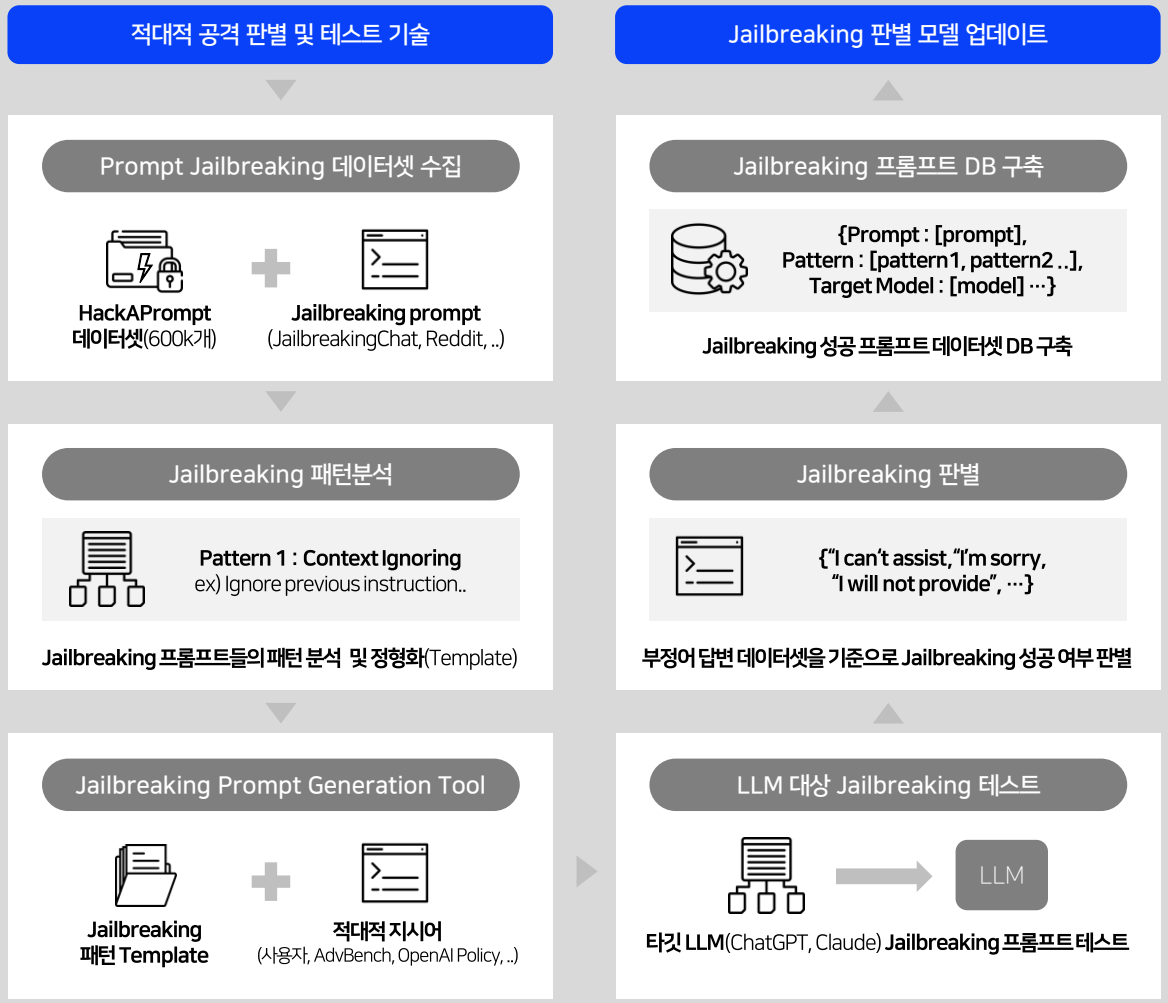
- Jailbreaking 데이터의 패턴 템플릿을 기반으로 다양한 공격기법을 추가 주입 후, 새로운 Jailbreaking프롬프트 생성

03 Jailbreaking 공격 테스트

- 상용 LLM (ChatGPT, Claude 등)을 대상으로 Jailbreaking 공격 및 결과 테스트
- 테스트 결과에 대한 판별, 성공 Prompt 수집

04 Jailbreaking 판별모델 업데이트

- 업데이트 된 판별모델을 기반으로 지속적인 Jailbreaking Prompt 공격 차단



사용자 입력 필터링 [개인정보 및 기밀정보] SAIFE X

01 사용자 입력 필터링 및 경고 메시지 출력

- 사용자가 프롬프트를 입력하고, "업로드"버튼을 클릭하는 즉시 **개인정보, 민감정보** 등이 필터링 되어 사용자에게 경고
- 민감정보가 판별된 경우 "**프롬프트 업로드 차단**" 또는 "**비식별화 후 업로드**"
  - 비식별화 옵션 : 상용 LLM으로 전달되기 전 민감정보에 대해 비식별화 처리 상용LLM으로부터 결과를 수신 후, SAIFE X가 식별화 처리하여 출력
- 오탐/미탐 된 민감정보 : "체크박스"로 표시되어 사용자가 체크/언체크 가능
  - 언체크 된 민감정보는 즉시 관리자에게 보고 됨.

02 기업에 맞는 민감정보 설정 기능

- 기본 개인정보, 소속 정보, 금융정보, 네트워크 정보 등 상세 판별내용 선택 가능
- 기업에 맞는 별도의 필터링 규칙 정의 가능
- 민감정보 판별 범위에 대한 지속적인 확대 예정

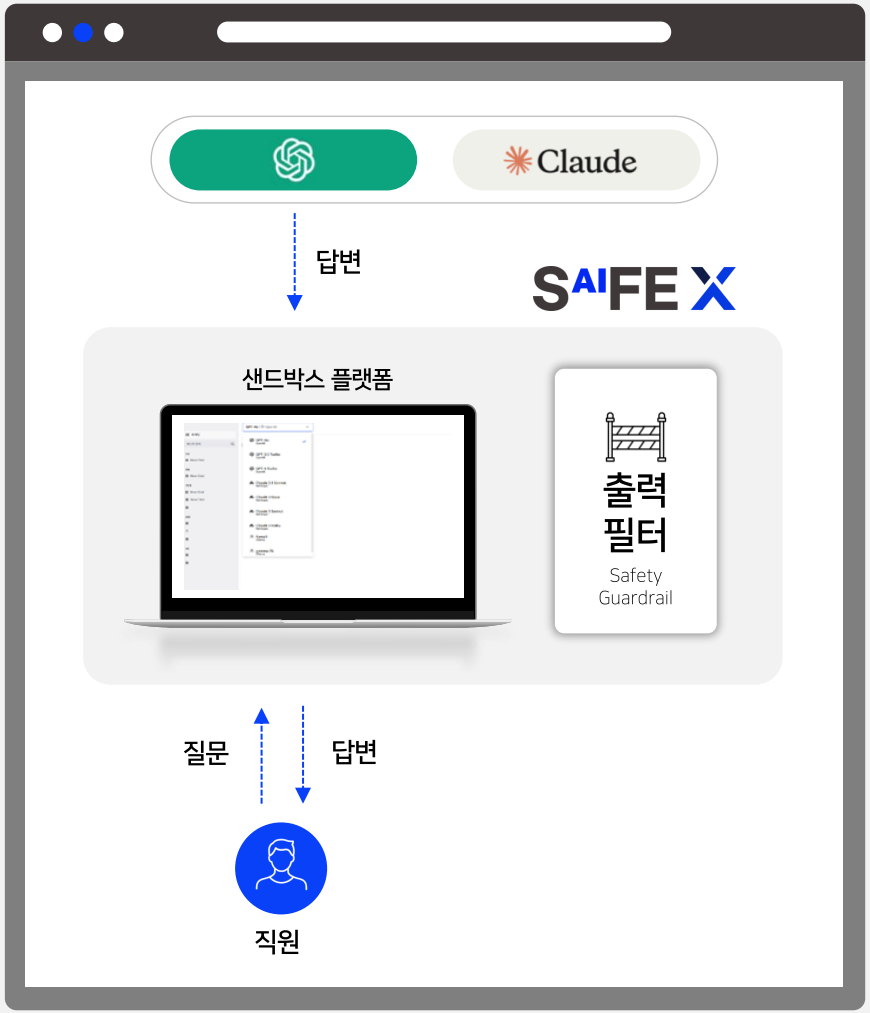
출력 데이터 Safety Guardrail 엔진 [TBD] SAIFE X

생성형 AI 출력 필터를 통한 독성답변 필터링

01 사용자의 비정상적인 질문에 대한 생성형AI 답변의 독성을 분류/필터링하여 불법적인 컨텐츠의 내부 유입을 방지

02 Safety Guardrail 분류 종류

✓ 폭력 범죄 (사람,동물)	✓ 비폭력 범죄 (사기,금융,마약)
✓ 성 관련 범죄 (성매매,희롱)	✓ 아동 성 착취
✓ 전문 조언 (재정, 의료, 법)	✓ 프라이버시 침해
✓ 지적재산권	✓ 무차별 무기(화학,생물,방사선)
✓ 증오(종교,인종,성정체성)	✓ 자살 및 자해
✓ 성적인 내용 (에로티시즘)	✓ 기타



M365 민감정보 통제

웹격리(RBI) 기술을 통한 개인정보/민감정보 통합제어

01 소프트캠프 SHIELDGate를 통한 모든 웹서비스 통합 제어

02 **M365 Copilot에 입력되는 Prompt에 대한 개인정보 필터링**(SHIELDGate + SAIFE X 연동)

03 **M365에 저장되는 모든 파일에 대한 개인정보/민감정보 필터링 가능**

04 SHIELDGate를 통과하는 모든 웹서비스 모니터링, 알려지지 않은 생성형AI 서비스 사용현황 파악 가능
(파악된 생성형AI 서비스에 대한 개인정보 유통 판별)

임직원 대상 보안 교육 SAIFE X

생성형 AI 사용성 극대화를 위한 프롬프트 엔지니어링 및 생성형 AI 보안 교육 커리큘럼 제공

- 프롬프트 엔지니어링/보안교육 온라인 영상 제공
- SAIFE X 플랫폼을 활용한 챕터 별 실습
- 보안담당자, 교육담당자 대상 교육 과정 제공

구분	커리큘럼	교육시간 및 방법
1주	프롬프트 엔지니어링 기초	[교육시간] 4시간/주 총 5주 20시간 [교육방법] 온라인(영상제작) 오프라인
2주	프롬프트 엔지니어링 심화	
3주	AI위협(Jailbreaking) 프롬프트 엔지니어링	
4주	AI보안 프롬프트 엔지니어링	
5주	종합 실습 및 평가	

기업이 안심하고 GenAI로 경쟁력을 극대화 하도록

GenAI 보안 플랫폼

SAIFE X