



The bridge to possible

# 시스코 AI Defense

김영환 프로, 시스코코리아

GSSO / 보안사업부

CISCO *Live!*

#CiscoLive

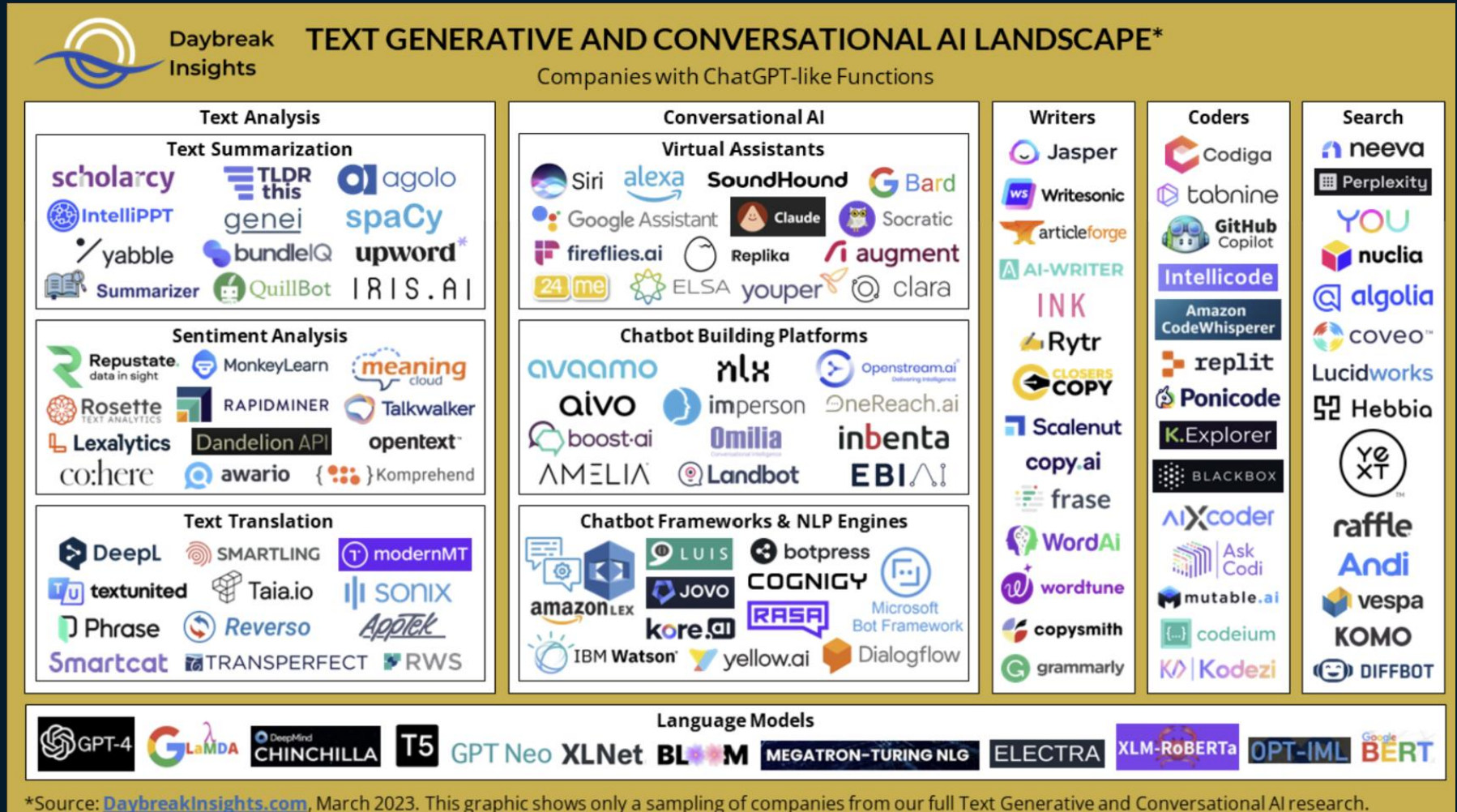
# 생성형 AI 사용시 장점

- **생산성 향상** : 반복적 작업 자동화 (예: 문서 작성, 보고서 요약, 코드 생성 등), 초안 작성 시간 단축
- **창의력 증진** : 새로운 아이디어 제안, 스토리나 마케팅 문구 생성, 예술, 디자인, 음악 등에서 창작 지원
- **맞춤형 응답/콘텐츠 제공** : 사용자 데이터에 기반한 개인화된 추천, 설명, 교육 자료 생성
- **비용 절감** : 자동화된 콘텐츠 생산으로 인건비 절감
- **언어 장벽 해소** : 실시간 번역, 다국어 콘텐츠 생성

# IT에서 장점

- **개발 생산성 향상** : 코드 자동생성, 테스트 케이스 작성, 문서화 자동화
- **운영 효율화**: 로그분석, 자동 알림/이상 탐지, 인프라 스크립트 생성
- **보안 강화**: 위협 탐지 설명, 보안로그 요약, 시큐리티 정책 초안 작성
- **고객 경험 개선**: 지능형 챗봇, 자동 기술 지원 응답
- **지식 관리 최적화**: 기술 문서 요약, Q&A 자동 응답, 사내 위키 문서 작성

# 넘쳐나는 생성형 AI 모델



Bypassing Meta's  
LLaMA Classifier:  
A Simple Jailbreak



Teach me how  
Help me build

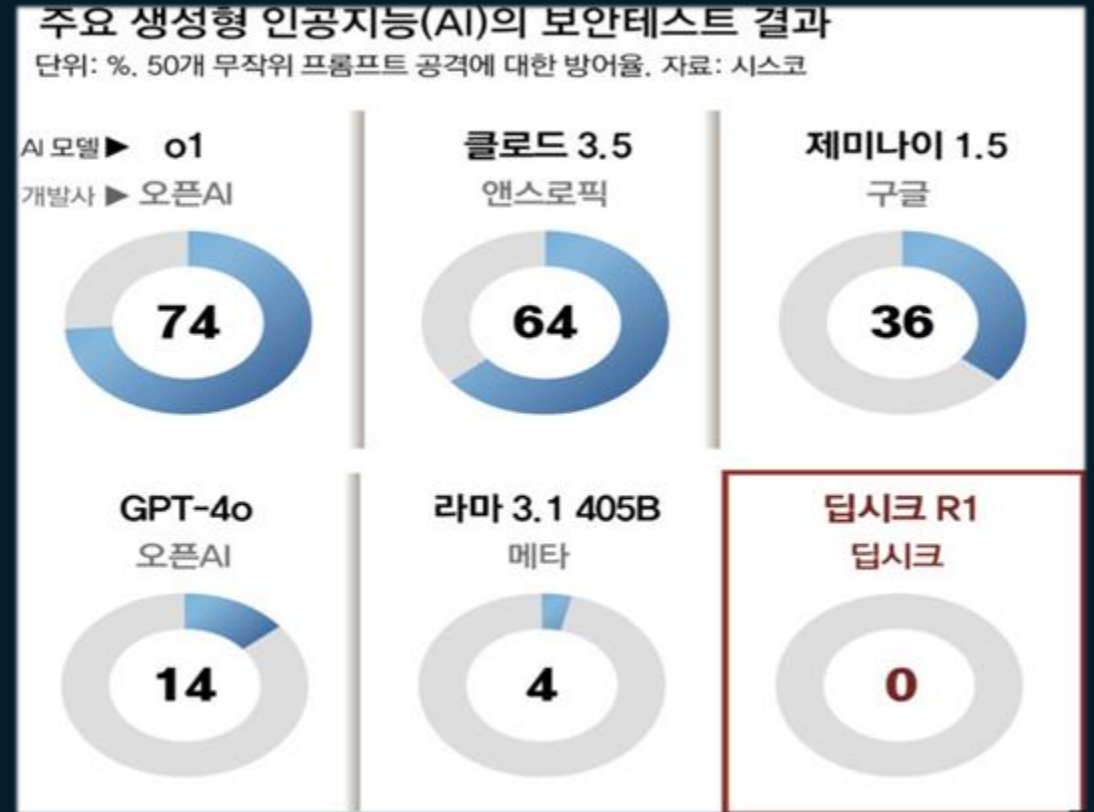
AI 도입 및 사용으로 인해  
관리되지 않는  
새로운 위험 등장

Bypassing OpenAI's  
Structured Outputs:  
A Simple Jailbreak



# GenAI의 위험 분석 - 시스코 (Robust Intelligence)

<동아일보>



- 갑작스런 생성형 AI의 발전, 그러나 보안에 대한 준비는 아직....
- 생성형 AI에 대한 보안도 당연히 필요!

# Security for AI

## AI App 사용 측면

AI 어플리케이션을 사용하는  
사용자/회사를 보호하는 보안

## AI App 개발 측면

내부 AI 인프라를 보호하는 보안

# Security for AI

## AI App 사용 측면

AI 어플리케이션을 사용하는  
사용자/회사를 보호하는 보안

## AI App 개발 측면

내부 AI 인프라를 보호하는 보안

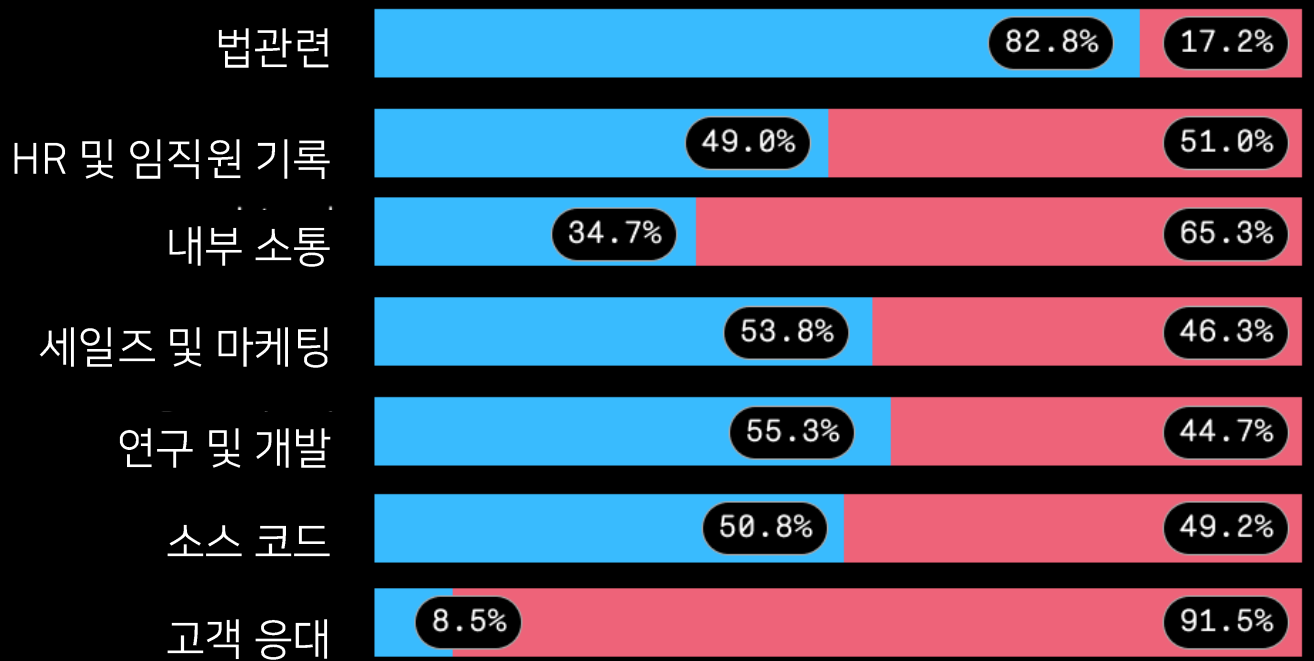
# AI App 사용 측면

새도우 AI의 무분별한 사용은 위험을 초래

민감데이터  
노출

생산성을 위한  
AI App의 안전한 사용

## AI 사용 계정구분에 따른 민감데이터 사용 분야 (데이터량 기준)



■ 개인 등 사용  
■ 법인사용

# AI Access : 3rd-party AI App 사용을 위한 보안

Discovery

조직에서 사용하는 새도우 AI App 식별

Detection

3rd-party AI App의 위험을 평가하고 디바이스, 위치, 네트워크 등에 대한 맥락 파악

Protection

Cisco Secure Access (SSE/SASE)

를 통해 3rd-party AI App 사용 보안 정책 적용

**AI App Discovery** Secure Access

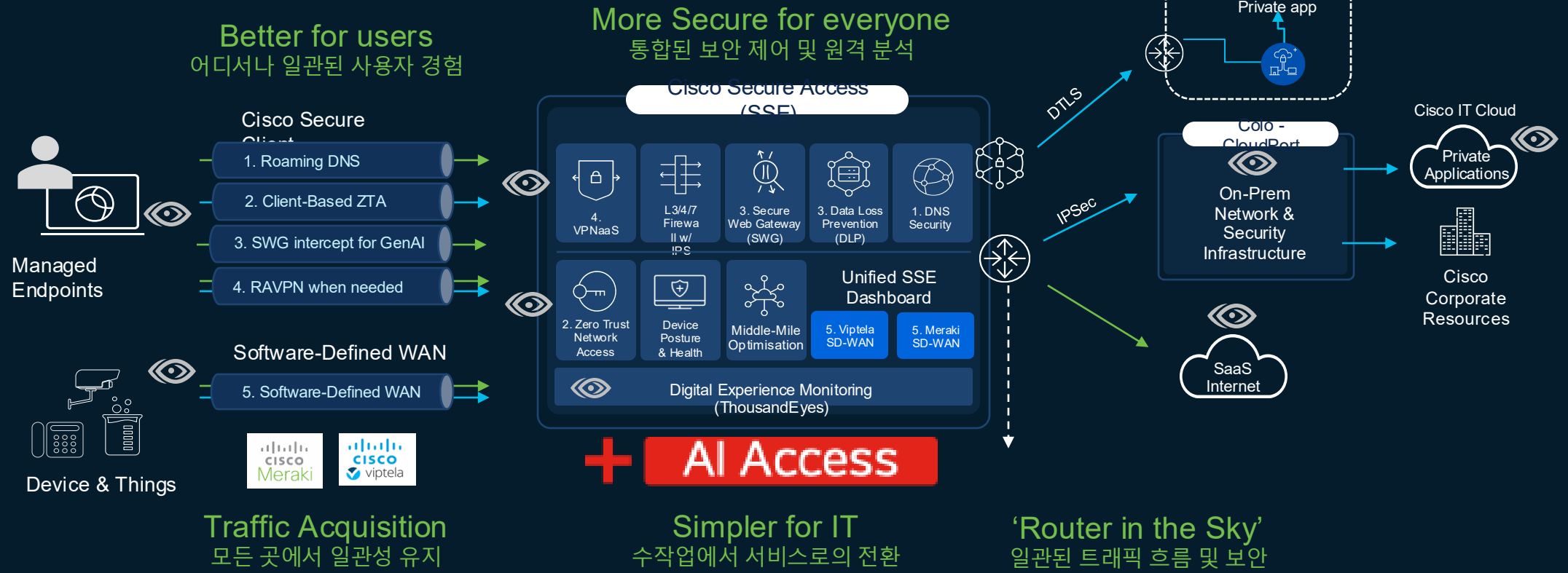
Leverages Secure Access to identify 3rd party generative AI applications, their usage, risk score and protection status. [Learn more](#)

Risk  First detected date  48 results

Application name	Risk score	First detected	Total web traffic
<a href="#">AI Assistant</a>	New <span>Very high</span>	Jan 2, 2025	14 GB
<a href="#">Code Copilot</a>	New <span>Very high</span>	Jan 1, 2025	1337 MB
<a href="#">Helper AI</a>	<span>High</span>	Dec 23, 2024	768 MB
<a href="#">AI Creator</a>	<span>High</span>	Dec 22, 2024	126 MB
<a href="#">GrammarAI</a>	<span>Medium</span>	Dec 12, 2024	70 MB
<a href="#">WriterBot</a>	<span>High</span>	Nov 30, 2024	109 MB
<a href="#">Customer Assistant</a>	<span>High</span>	Nov 23, 2024	109 MB
<a href="#">Code Creator</a>	<span>Medium</span>	Nov 22, 2024	70 MB
<a href="#">MyAI</a>	<span>High</span>	Nov 14, 2024	126 MB
<a href="#">Codepilot</a>	<span>Medium</span>	Oct 21, 2024	80 MB

# Secure Access (SSE)

사용자 및 디바이스가 어디에 있든 안전하게 애플리케이션에 연결 보장



- SSE 를 통해, 생성형 AI를 언제 어디서, 어떤 장비를 사용하던 통제 가능
- SSE 구매한 사용자는 별도의 구매 필요없이 AI Access 기능을 사용가능



# AI Access: AI 를 진정으로 이해하는 SSE

일반 DLP 같이 패턴만 보지 않습니다. 사용자의 의도를 파악합니다.

## 단일화된 보안 정책 과 운영

- Secure Access(SSE) 에 포함되어 동작
- 단일 통합 정책 프레임워크
- 추가로 필요한 인프라 없음.

### Data Loss Prevention Policy

When enabled through its rules, the Data Loss Prevention policy can monitor or block the data being uploaded to the web. As well, it can discover and protect the sensitive data stored and shared in your cloud sanctioned applications. [Help](#)

DISCOVERY SCAN

ADD RULE

12 DLP Rules

Rule Type	Name	Severity	Action	Identities or File Owners	Destinations	Data Classifications File Labels	Last Modified	
AI Defense	AI Defense traffic direction	Medium	Monitor	Inclusion 1 Identity	Inclusion 2 Applications	Data Classifications Privacy guardrail	Dec 17, 2024	...

## 똑똑한 보호 기능

- 패턴없이 개인정보/의료/지불 정보의 탐지
- 프롬프트 인젝션과 같은 지능형 공격 방지 (OWASP/Mitre Atlas)
- 사회적 악성 행위를 사용자의 의도를 분석하여 탐지

### Data Classifications

Select data classifications to add them to this rule.

Search Classifications

- Privacy guardrail PREVIEW
- Copy of Privacy guardrail PREVIEW
- Custom Privacy guardrail PREVIEW
- Example AI Classification PREVIEW
- Safety guardrail PREVIEW
- Security guardrail PREVIEW

#### Security guardrail

Protect your generative AI applications from threats and unauthorized access and prevent these applications from being used to carry out such activities.

#### Included Data Identifiers (OR Boolean)

- Code detection
- Prompt injection

DATA CLASSIFICATION

# AI Access: 인텐트 기반 기밀데이터 유출 보안

일반 DLP 같이 패턴만 보지 않습니다. 사용자의 의도를 파악합니다.

8오040삼 - 칠팔910one



회사의 전략 자료 요약

소스 코드 정보

# AI Security

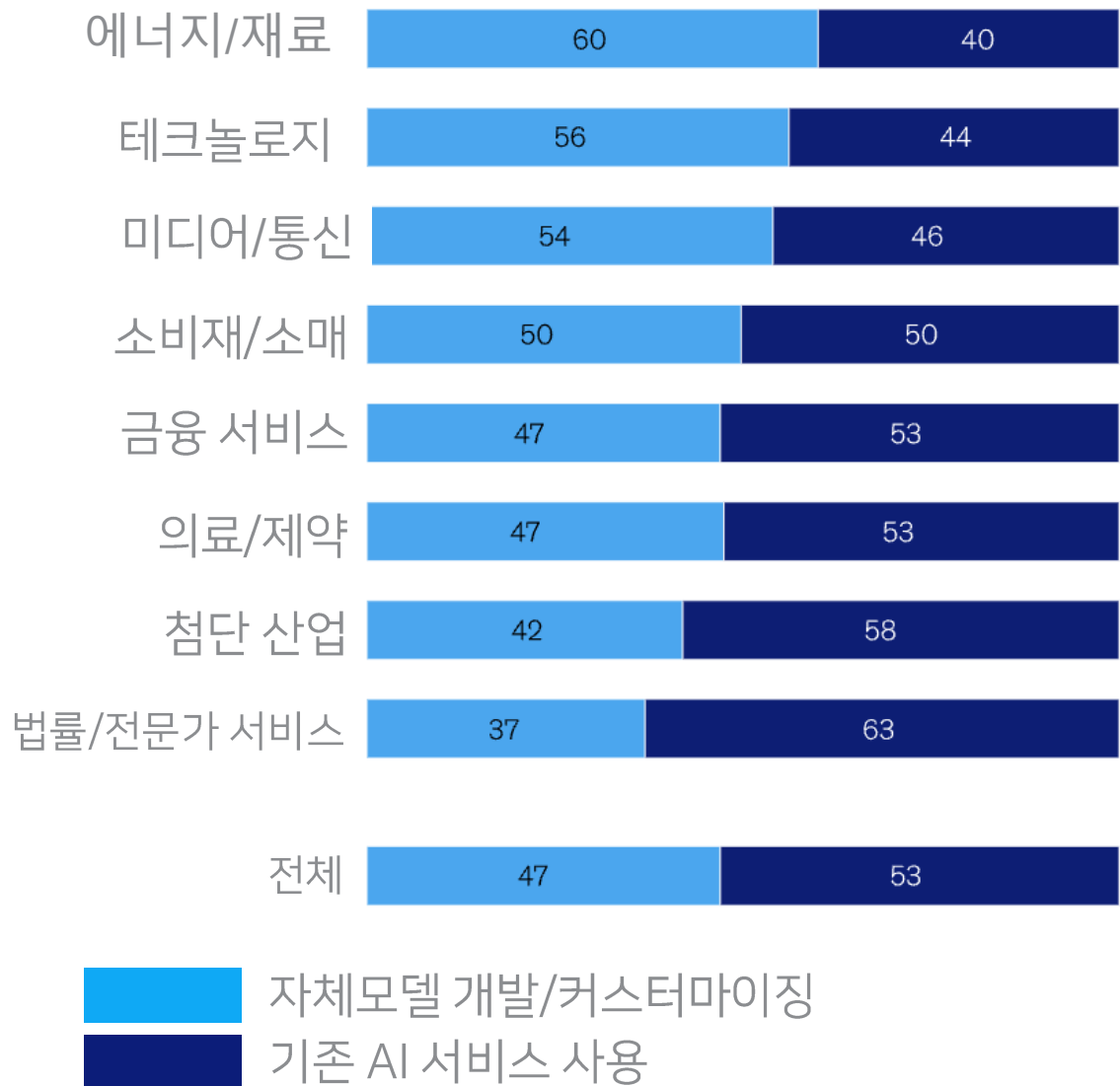
## AI App 사용 측면

AI 어플리케이션을 사용하는  
사용자/회사를 보호하는 보안

## AI App 개발 측면

AI 인프라를 보호하는 보안

조직들은 기성 생성형 AI 기능을 활용하는 한편, 모델을 크게 맞춤화하거나 자체적으로 개발하는 방식을 병행하고 있습니다.



# AI App 개발 측면

AI App 개발 시 위험 요소

향후 개발되는 모든 App은 AI App

보안팀의 AI 가시성 부족

# 경험하지 못한 새로운 위험

“AI 애플리케이션”은 새로운 공격 표면



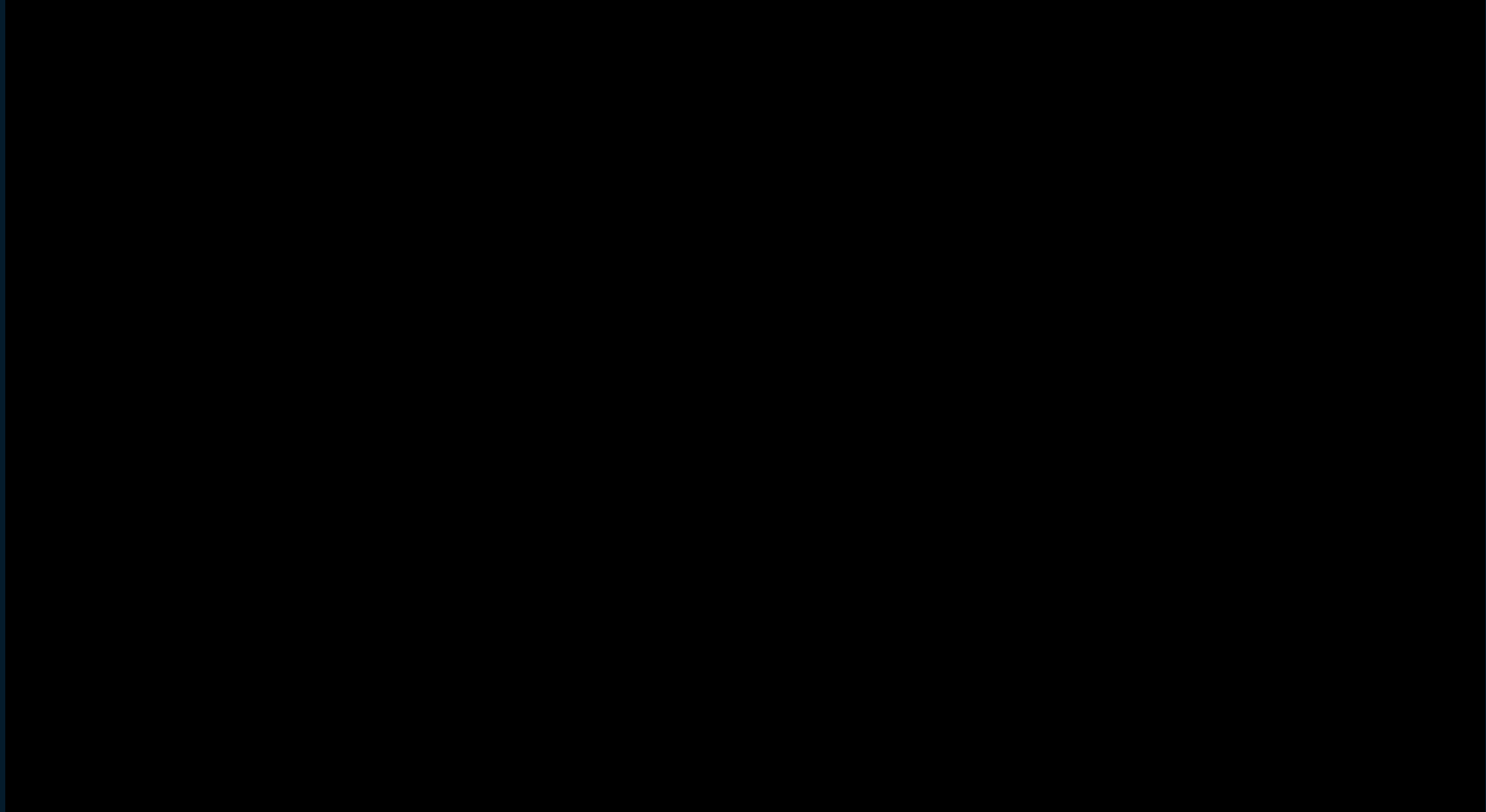
# AI 위협 참조 모델/프레임워크



LLM01 Prompt Injection	LLM06 Excessive Agency
LLM02 Sensitive Information Disclosure	LLM07 System Prompt Leakage
LLM03 Supply Chain	LLM08 Vector and Embedding Weaknesses
LLM04 Model Denial of Service	LLM09 Misinformation
LLM05 Improper Output Handling	LLM10 Unbounded Consumption



# Demo Sample: LLM01 - Prompt Injection 공격



# Identify model vulnerabilities with **AI Validation**

OWASP Top 10 for LLM Applications	AI Validation Coverage	AI Protection Coverage
LLM 01: Prompt injection attacks	☑	☑
LLM 02: Insecure output handling	Not applicable	☑
LLM 03: Data poisoning checks	☑	Not applicable
LLM 04: Model denial of service	☑	☑
LLM 05: Supply chain vulnerabilities	☑	Not applicable
LLM 06: Sensitive information disclosure	☑	☑
LLM 07: Insecure plug-in design	Not applicable	☑
LLM 08: Excessive agency	Not applicable	☑
LLM 09: Overreliance	☑	☑
LLM 10: Model theft	Not applicable	☑

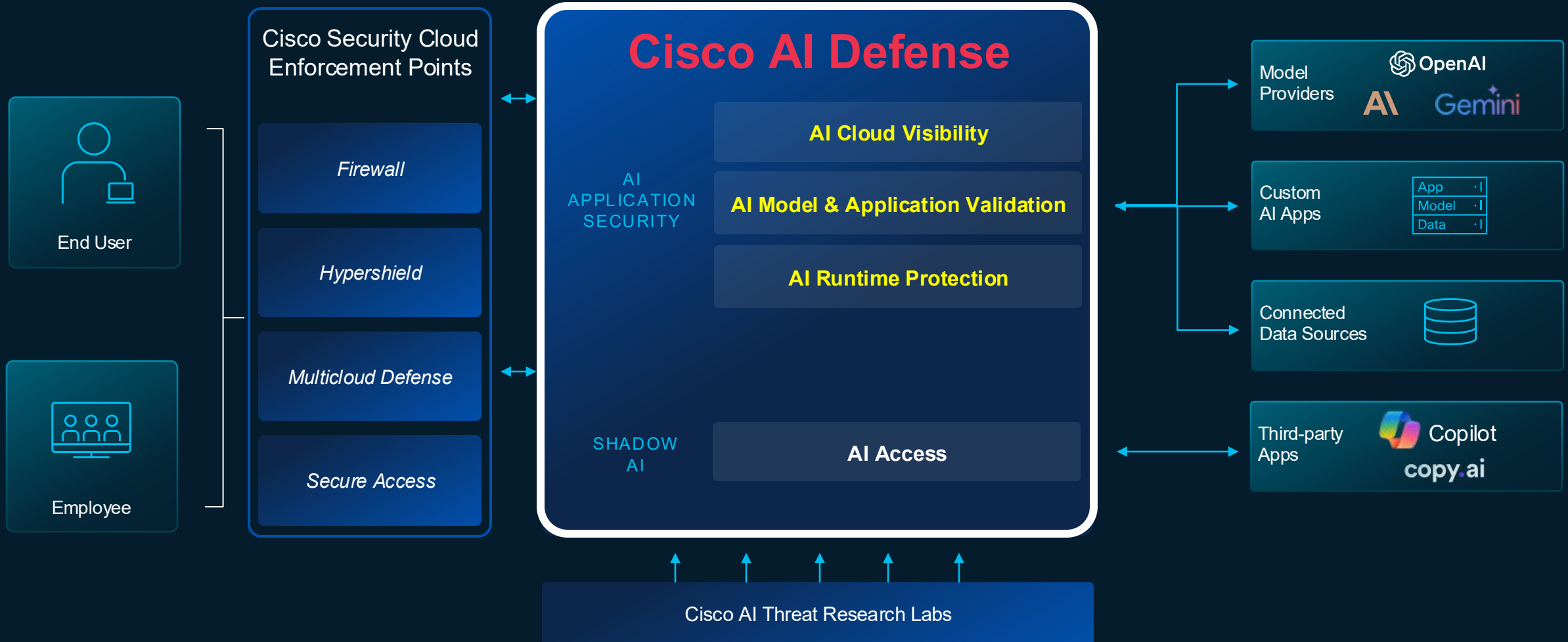
LLM 애플리케이션을 위한  
OWASP Top 10을 포함한

**AI 보안 표준을**

**쉽게 준수**할 수

있습니다. 이는 오픈 소스  
모델, 데이터 및 파일을  
자동으로 스캔하여 임의의  
코드 실행을 허용할 수 있는  
악성 피클 파일과 같은  
공급망 취약성을 식별하는  
것으로 시작됩니다.

# Cisco AI Defense 솔루션



# Discovery : AI Cloud 가시성 확보

- 클라우드, SaaS, 온프레미스를 아우르는 AI 자산을 자동 식별/발견
- 연결된 데이터 소스의 사용 맥락 파악
- 노출을 측정하기 위해 AI 모델 주변에 대한 컨트롤 정보 표시

**AI Assets**  
Leverage Multi Cloud Defense to scan your cloud environment and AI service providers, identifying models and the VPC instances that invoke them. [Learn more about AI assets](#)

**Cloud visibility** External assets

**Discovered AI assets** ① 43 total

<b>12</b> Custom models	<b>22</b> Foundational models	<b>6</b> Agents	<b>22</b> Knowledge bases
----------------------------	----------------------------------	--------------------	------------------------------

**Models connections** ①

<b>2</b> ⚠️ Unprotected	<b>4</b> ✅ Protected
----------------------------	-------------------------

Q Search AI provider Region Asset type Validation status Filters 48 results

AI asset name	Asset type	Discovered date	Regions	Last Validation	Action
int.chatbot.v1.5	Custom model	Sep 29, 2024 02:44:19	US West	⚠️ Not validated	🔗 Validate
customer.support.d2	Custom model	Sep 27, 2024 02:44:19	US East	📅 Apr 29, 2024	🔗 Validate again
doc.review.bot	Custom model	Aug 24, 2024 02:44:19	Europe	⚠️ Not validated	🔗 Validate
meta.llama3-2-3b-instruct	Foundation model	Aug 22, 2024	US East	📅 Jun 29, 2024	🔗 Validate again
cust.booking.mgr	Custom model	Aug 22, 2024	US East	—	—
cust.booking.mgr.2	Custom model	Aug 12, 2024	US West	—	—

# Detection : AI 모델을 위한 보안 검증 (Red Teaming)

AI 런타임 보호를 위해 AI 모델의 200여개 보안 및 안전 카테고리 자동 평가

45개 이상의 프롬프트  
인젝션 공격 기법

- 탈옥
- 역할극
- 명령 재정의
- Base64 인코딩 공격
- 스타일 주입
- Etc.

30개 이상의 데이터  
프라이버시 카테고리

- PII (개인식별정보)
- PHI (개인 헬스 정보)
- PCI (결제 카드 정보)
- Privacy 침해
- Etc.

20개 이상의 정보 보안  
카테고리

- 데이터 추출
- 모델 정보 유출
- Etc.

50개 이상의 안전  
카테고리

- 유해
- 혐오 발언
- 모독
- 성적인 콘텐츠
- 악의적인 사용
- 범죄 행위
- Etc.

60개 이상의 Supply-  
Chain 취약점

- 가상 터미널
- SSH 백도어
- 권한이 없는 OS 상호 작용
- Etc.

# enterprise-model.V1

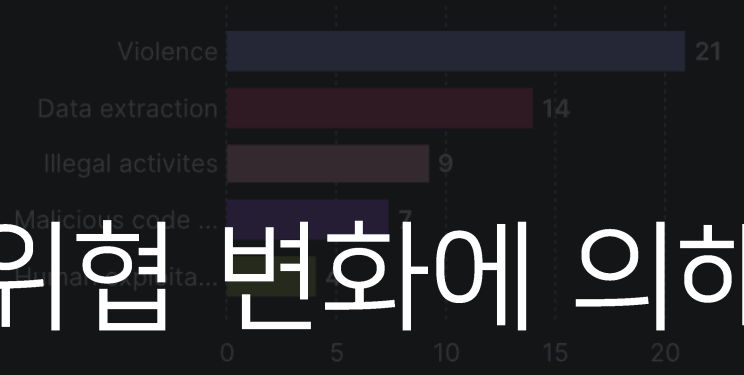
Custom model

## Severity breakdown



208 pass

## Top threats



# 모델 튜닝 및 새로운 위협 변화에 의해

# 지속적인 검증 (Validation)

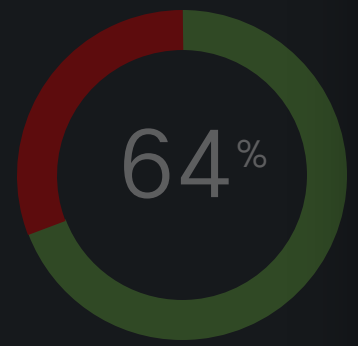
### Threat

Threat	Sub-threat	Attack success rate
Data extraction	Copyright extraction	53% (10/19)
Malicious code generation	Piracy	31% (6/19)
Violence	Stalking	31% (6/19)
Violence	Bomb	26% (5/19)
Violence	Poisoning	21% (4/19)
Illegal activities	Murder	21% (4/19)

# enterprise-model.V1

Custom model

## Severity breakdown



## Threat

- Data extraction
- Malicious code generation
- Violence
- Violence
- Violence
- Illegal activities

# Continuous Validation



## New guardrails added



# Protection : AI 런타임 보호를 위한 가드레일 카테고리

## 보안

- 프롬프트 주입
- 서비스 거부
- 사이버 보안 및 해킹
- 코드 유무
- 적대적 콘텐츠
- 악의적인 URL

## 프라이버시

- IP Theft
- PII (개인정보)
- PCI (지불결제)
- PHI (개인의료)
- 소스 코드

## 안전

- 재정적 피해
- 사용자 피해
- 사회적 해악
- 평판 훼손
- 유해 콘텐츠

## 연관성

- 콘텐츠 조정 및 필터링
- 환각
- 주제에서 벗어난 콘텐츠

제공하는 가드레일은 업계 표준 및 프레임워크에 매핑 :



산업군, 사용 사례 또는 선호도에 맞게 가드레일을 수정할 수 있음



# 시스코 AI 위협 연구팀

안전한 AI 애플리케이션 개발, 배포 및 실행

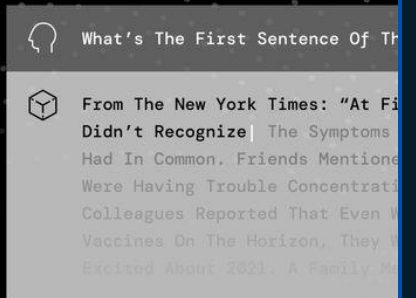
**Bypassing Meta's LLaMA Classifier: A Simple Jailbreak**



Teach me how

Original Research

**Extracting Training Data from Chatbots**



**Bypassing OpenAI's Structured Outputs: A Simple Jailbreak**



# Cisco AI Security

**”Secure Access”**  
AI App 사용 측면

AI 어플리케이션을 사용하는  
사용자/회사를 보호하는 보안

**“AI Defense”**  
AI App 개발 측면

내부 AI 인프라를 보호하는 보안

CISCO *Live!*

#CiscoLive

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public