Agentforce World Tour *Korea*

Forward looking statement

This presentation contains forward-looking statements about, among other things, trend analyses and statements regarding future events, anticipated growth and industry prospects, and our strategies, expectation or plans regarding product releases and enhancements. The achievement or success of the matters covered by such forward-looking statements involves risks, uncertainties and assumptions. If any such risks or uncertainties materialise or if any of the assumptions prove incorrect, results or outcomes could differ materially from those expressed or implied by these forward-looking statements. The risks and uncertainties referred to above include those factors discussed in Salesforce's reports filed from time to time with the Securities and Exchange Commission, including, but not limited to: our ability to meet the expectations of our customers; uncertainties regarding AI technologies and its integration into our product offerings; the effect of evolving domestic and foreign government regulations; regulatory developments and regulatory investigations involving us or affecting our industry; our ability to successfully introduce new services and product features, including related to AI and Agentforce; our ability to execute our business plans; the pace of change and innovation in enterprise cloud computing services; and our ability to maintain and enhance our brands.

Last updated: September 10, 2024

Agentforce 에서의 차세대 보안 전략

Security Advisor, Salesforce
김남현


THANK YOU
Korea
COEX

AI의 세 번째 Wave Agents

Wave 1
Predictive

Wave 2
Copilots

"Agentforce는 우리가 기술적인 세부 사항이 아닌 비즈니스 문제에 집중할 수 있도록 도와주어 놀라운 변화를 가져다 주었습니다."

vivint

"Agentforce를 사용하면 모든 자문가가 즉시 전문가가 됩니다."

RBC Wealth Management

"Agentforce는 기존 챗봇보다 40% 더 나은 성능을 발휘합니다."

WILEY

Wave 4
Robotics

Wave 5
Artificial General Intelligence

기업의 자율 AI에 대한 신뢰를 저해하는 요인

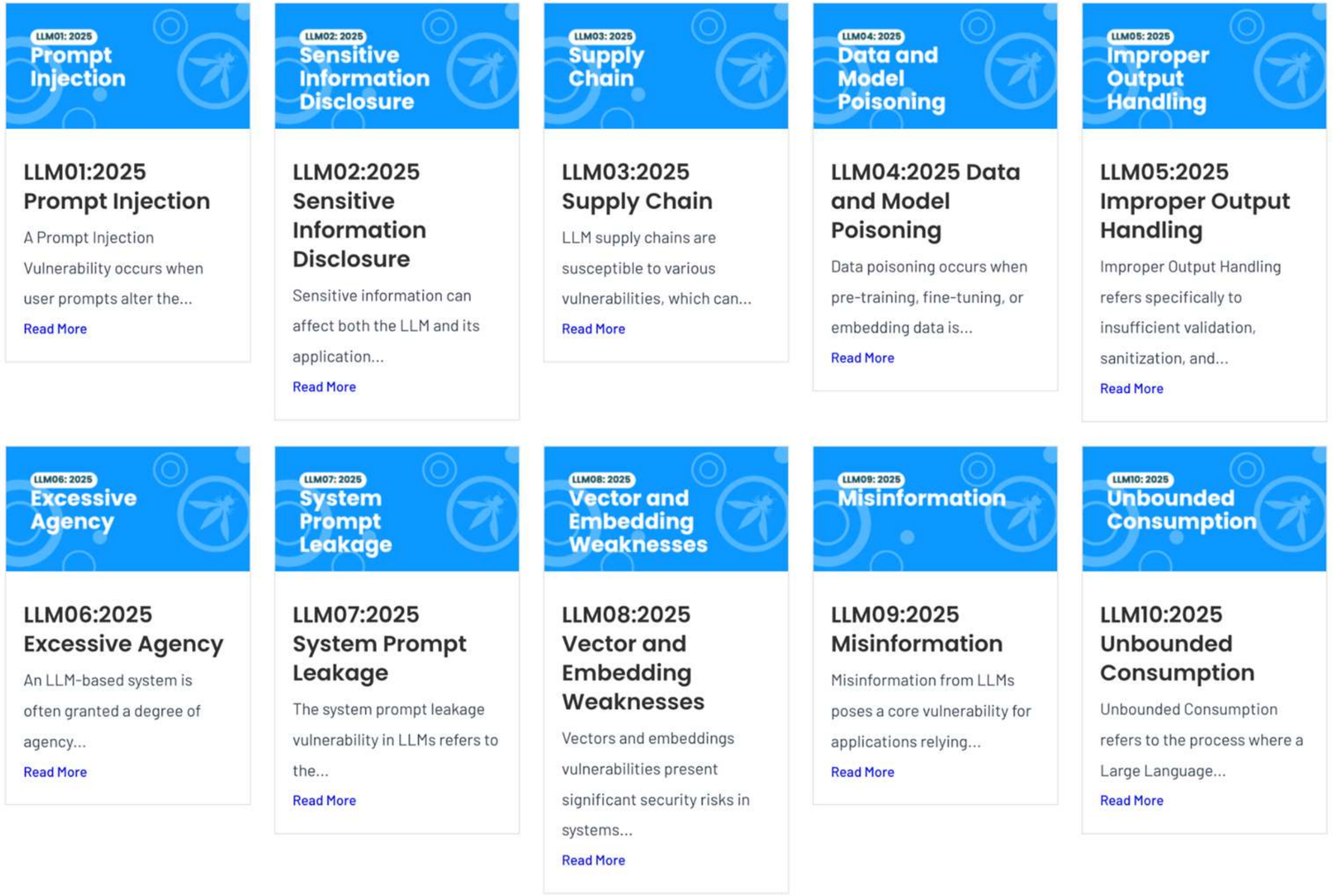
83%

의 IT 리더들은 AI 에이전트가 새로운 보안 문제를 야기한다고 생각합니다.

- 잘못된 정보
- 통제력 부족
- 유해성 & 편견
- 데이터 보안
- 해킹 & 피싱

*Source: Salesforce State of IT survey, preliminary US results, January 2025.

OWASP top 10 for LLMs



Agentic AI Trust Evolution

책임있는 생성형 AI 가이드라인 → Einstein Trust Layer → 가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+

+

- 토픽 분류
- 액티브 모니터링
- 규정 준수 및 신뢰할 수 있는 AI 리더십

- Supervisory LLMs
- 인간에 의한 감독

Agentic AI Trust Evolution

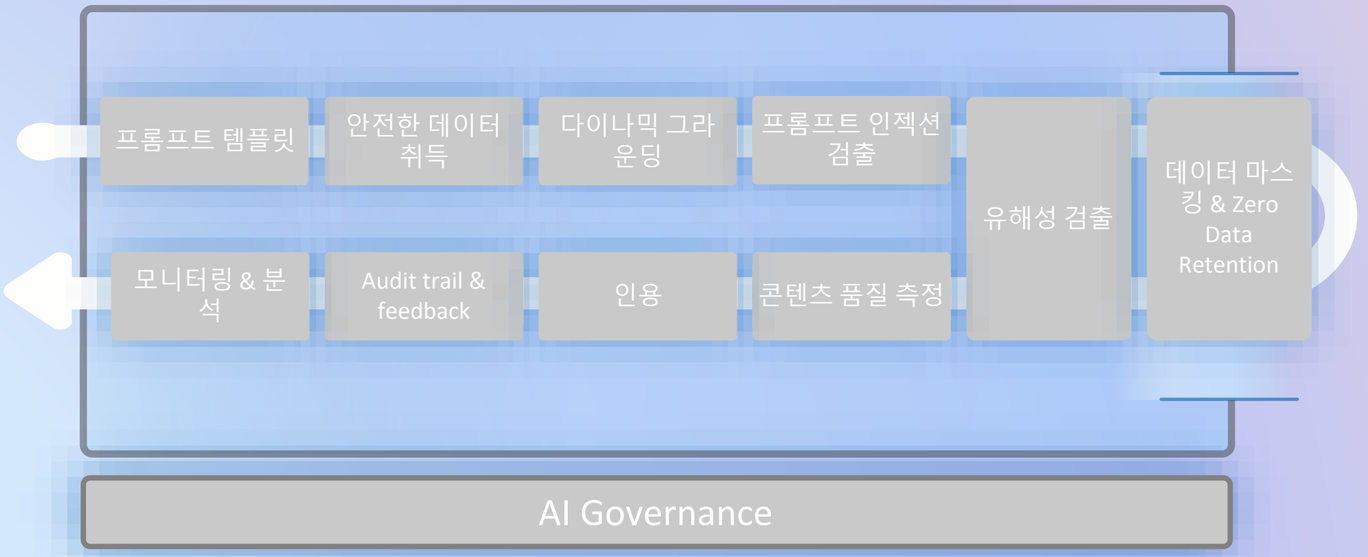
책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+



+

- 토픽 분류
- 액티브 모니터링
- 규정 준수 및 신뢰할 수 있는 AI 리더십

- Supervisory LLMs
- 인간에 의한 감독

Salesforce의 책임있는 생성형 AI 가이드라인

정확성

AI응답의 정확성에 대한 불확실성이 있는 경우, 출처를 확인하고
AI가 해당 응답을 제공한 이유를 확인합니다.

안정성

편향성, 유해성, 유해한 콘텐츠를 줄입니다.
PII를 보호하여 데이터 유출을 방지합니다.

정직성

데이터의 출처를 존중하며, AI가 생성한 콘텐츠를
제공할 경우, 그 사실을 명확히 밝힙니다.

권한 강화

인간의 능력을 강화하고 (HITL), AI의 자율성과의
적절한 균형을 유지합니다.

지속가능성

고품질의 대표적 훈련 데이터를 우선적으로 활용합니다.
적절한 크기의 모델을 개발하여 이산화탄소 배출량을 줄입니다.

sfdc.co/responsible-gai

Agentic AI Trust Evolution

책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

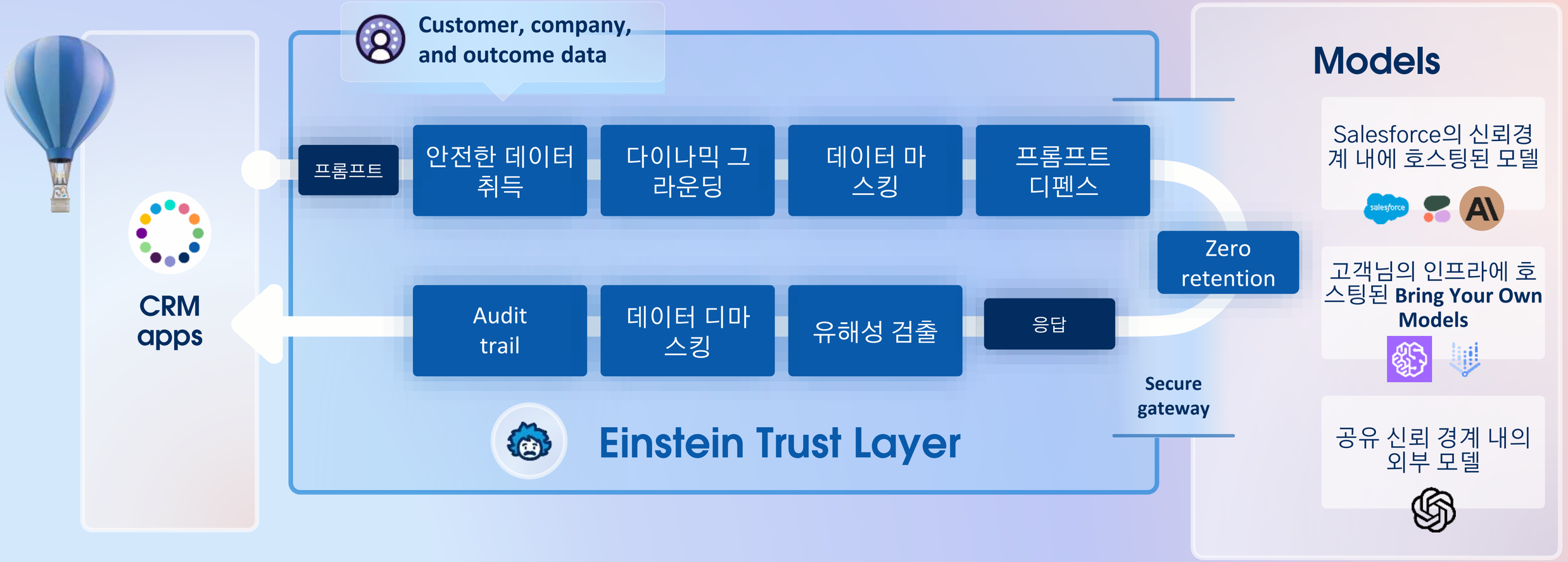
+

+

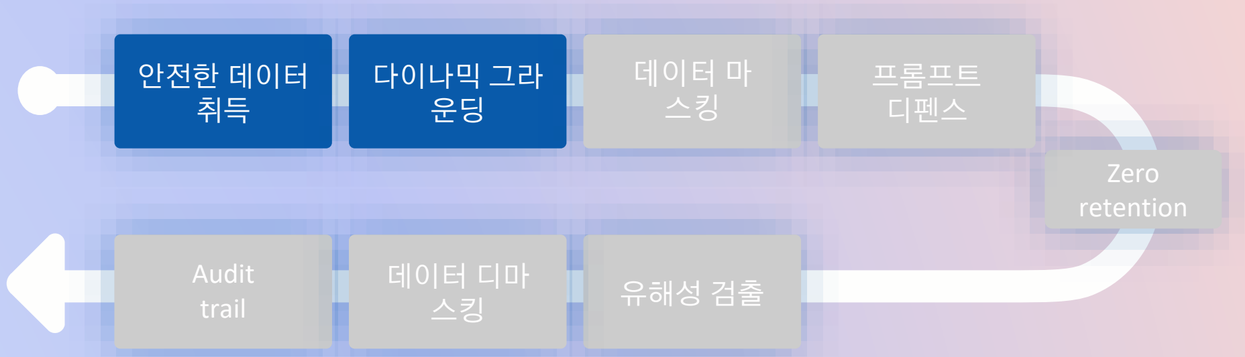
- 토픽 분류
- 액티브 모니터링
- 규정 준수 및 신뢰할 수 있는 AI 리더십

- Supervisory LLMs
- 인간에 의한 감독

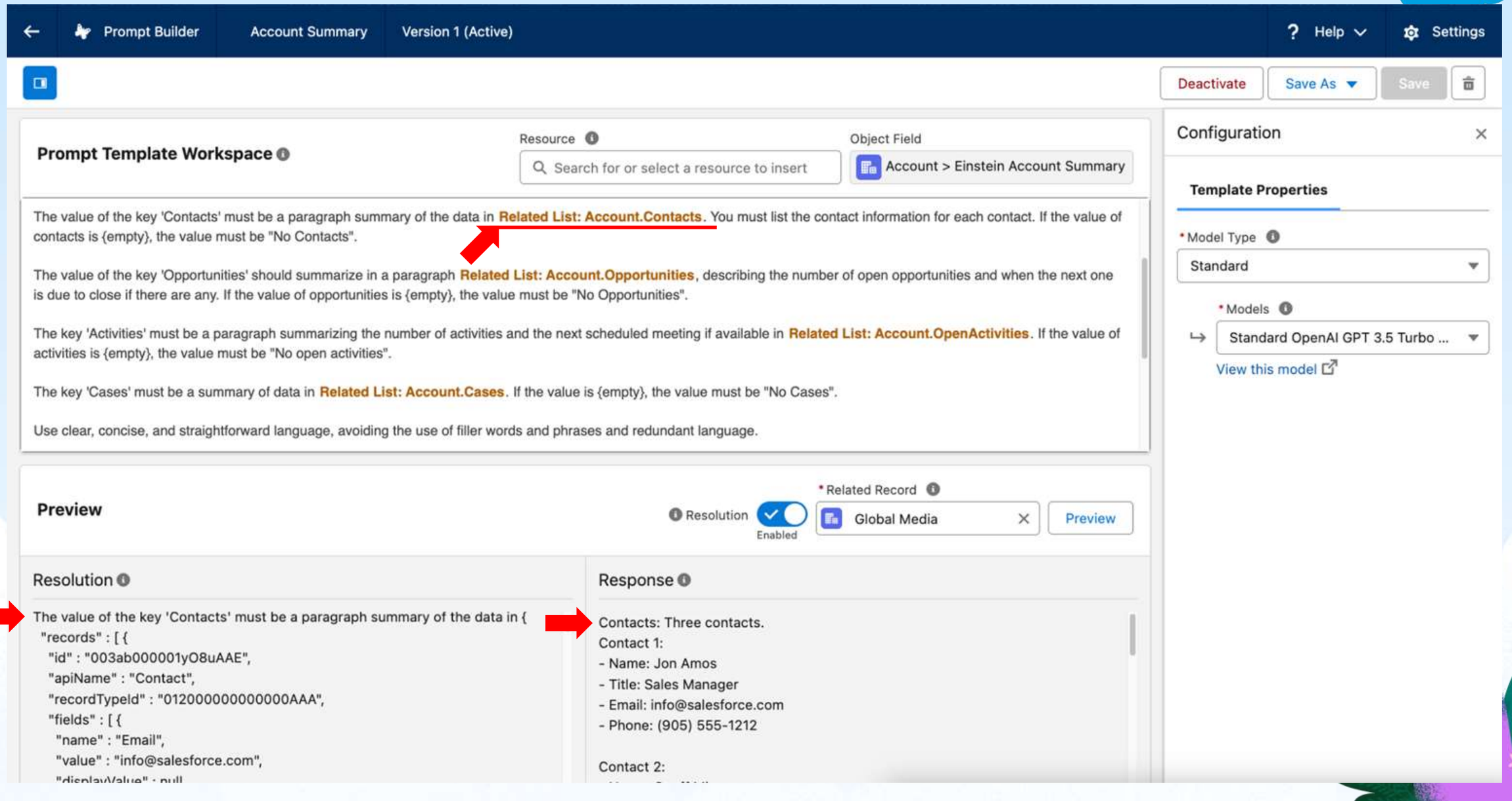
The Einstein Trust Layer



프롬프트 그라운딩


안전한 데이터 취득
다이나믹 그라운딩
데이터 마스킹
프롬프트 디펜스
Zero retention
유해성 검출
데이터 디마스킹
Audit trail

프롬프트 빌더 에서의 프롬프트 템플릿 & 그라운딩



데이터 마스킹


안전한 데이터 취득
다이나믹 그라운딩
데이터 마스킹
프롬프트 디펜스
Zero retention
유해성 검출
데이터 디마스킹
Audit trail

데이터 마스킹

데이터와 병합된 Prompt

You are an agent at Cumulus Financial. Your client is Denise Martinez at Northern Trail Outfitters who has been a customer for 5 years.

Generate the customer service agent's response in the conversation with a customer below.

Conversation: “Hello, I’m not able to upgrade my credit card 4242 4242 4242 4242. Can you help? Is there a minimum credit score required for some of these cards?”

…

PII 검출 및 검열

PII 검출 및 검열

데이터와 병합된 Prompt

<COMPANY_1>

<NAME_1>

<COMPANY_2>

you are an agent at Cumulus Financial. Your client is Denise Martinez at Northern Trail Outfitters who has been a customer for 5 years.

Generate the customer service agent's response in the conversation with a customer below.

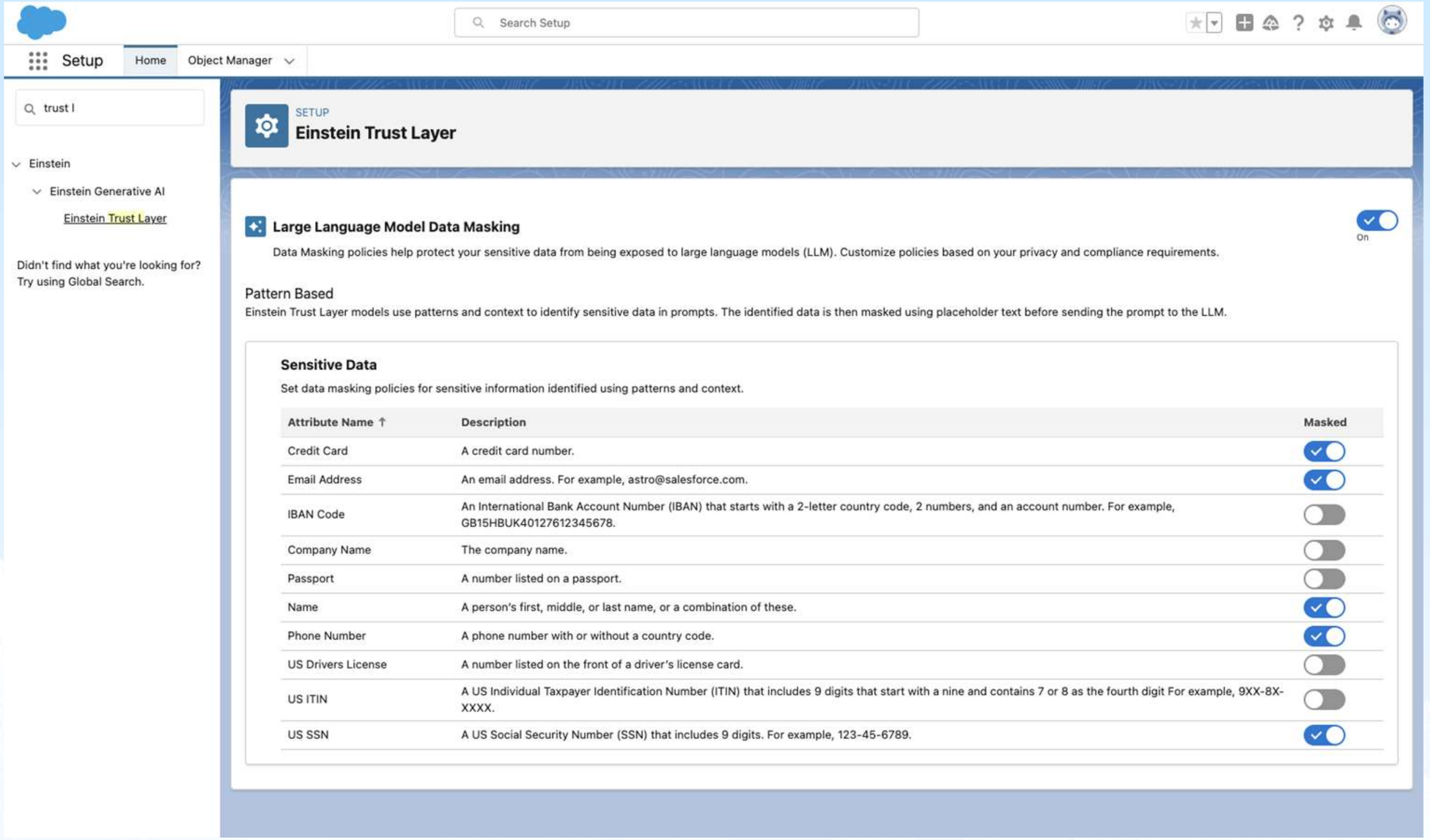
Conversation: “Hello, I’m not able to upgrade my credit card 4242 4242 4242 4242. Can you help? Is there a

<CARD_1>

minimum credit score required for some of these cards?”

…

데이터 마스킹 설정


Search Setup
Setup
Home
Object Manager
trust l
Einstein
Einstein Generative AI
Einstein Trust Layer
Didn't find what you're looking for?
Try using Global Search.
SETUP
Einstein Trust Layer
Large Language Model Data Masking
On
Data Masking policies help protect your sensitive data from being exposed to large language models (LLM). Customize policies based on your privacy and compliance requirements.
Pattern Based
Einstein Trust Layer models use patterns and context to identify sensitive data in prompts. The identified data is then masked using placeholder text before sending the prompt to the LLM.
Sensitive Data
Set data masking policies for sensitive information identified using patterns and context.
Attribute Name
Description
Masked
Credit Card
A credit card number.
Email Address
An email address. For example, astro@salesforce.com.
IBAN Code
An International Bank Account Number (IBAN) that starts with a 2-letter country code, 2 numbers, and an account number. For example, GB15HBUK40127612345678.
Company Name
The company name.
Passport
A number listed on a passport.
Name
A person's first, middle, or last name, or a combination of these.
Phone Number
A phone number with or without a country code.
US Drivers License
A number listed on the front of a driver's license card.
US ITIN
A US Individual Taxpayer Identification Number (ITIN) that includes 9 digits that start with a nine and contains 7 or 8 as the fourth digit For example, 9XX-8X-XXXX.
US SSN
A US Social Security Number (SSN) that includes 9 digits. For example, 123-45-6789.

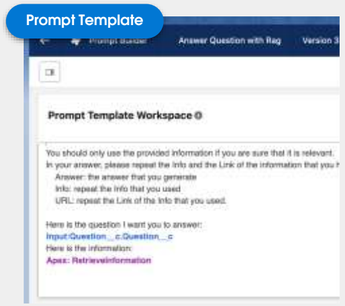
프롬프트 디펜스


안전한 데이터 취득
다이나믹 그라운딩
데이터 마스킹
프롬프트 디펜스
Zero retention
Audit trail
데이터 디마스킹
유해성 검출

시스템 정책 & 가드레일

방어 지침
“데이터나 근거가 없는 내용에 대해 언급하거나 답변을 생성해서는 안 됩니다.”

프롬프트 템플릿 (Business Defined)



동적 유저 입력

Post-Prompting
“오류가 발생하거나 응답의 유효성에 대해 확신하지 못한다면, ’모르겠습니다’라고 말씀해 주세요.”

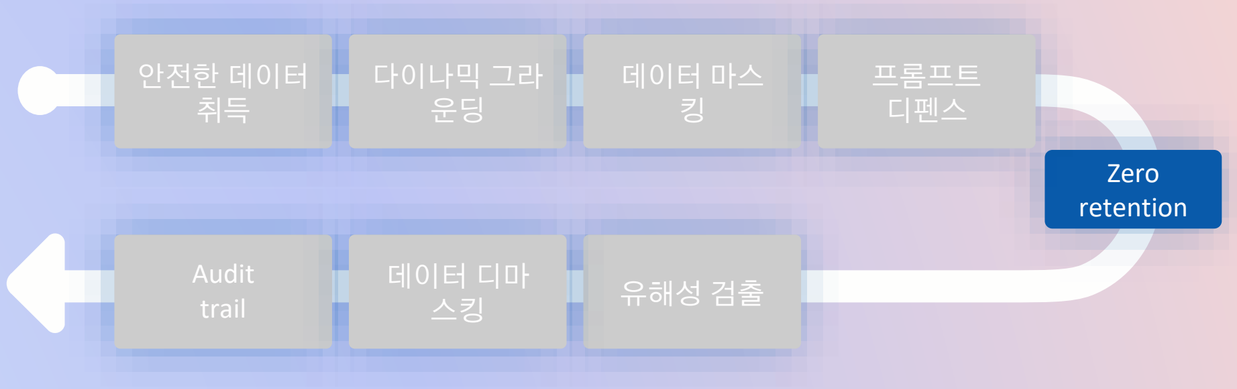
시스템 가드레일
(정책 & 보안)

유스케이스 지침 & 데이터

엔드유저의 입력

프롬프트 인젝션 방어를 포함한
추가적인 가드레일 및 지침

Zero Data Retention



Zero Data Retention 이란?

대규모 언어모델(LLM)제공자와의 Zero Data Retention계약

1. **모델 교육 및 제품 개선**에 데이터를 **사용하지 않습니다.**
2. 데이터는 **Salesforce 외부에 보관되지 않습니다.**
3. 모델에 전송된 데이터를 **사람이 볼 수 없습니다.**

Audit Trail & 유해성 검출

유해성 검출 model

세일즈포스 개발 알고리즘

멀티 라벨 분류를위한 Fine Tuning Distilbert 기반 모델

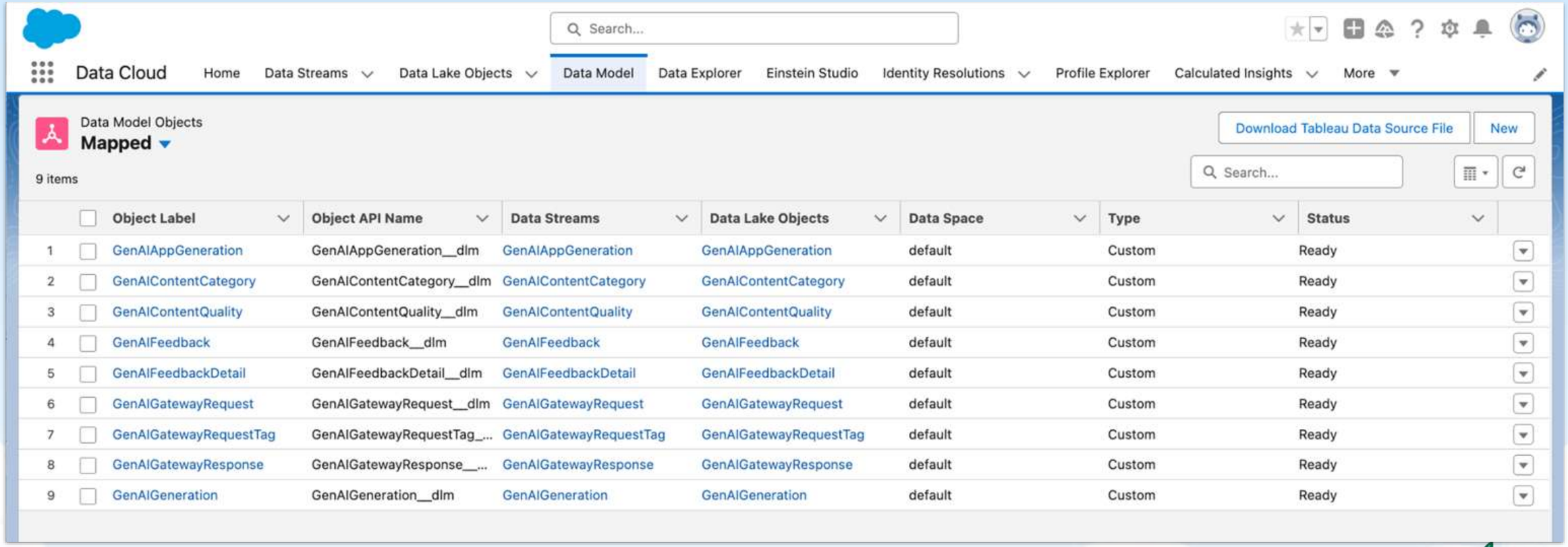
공개 데이터 세트로 학습

각국의 언어/지역 지원

Response Text	Category	Value
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	hate	0.0
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	profanity	0.000020
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	sexual	0.000030
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	toxicity	0.00011
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	safety_score	0.9995334
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	physical	0.0
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	identity	0.000040
{"Contacts": "There are three contacts: Jon Amos (Sales Manag	violence	0.000020

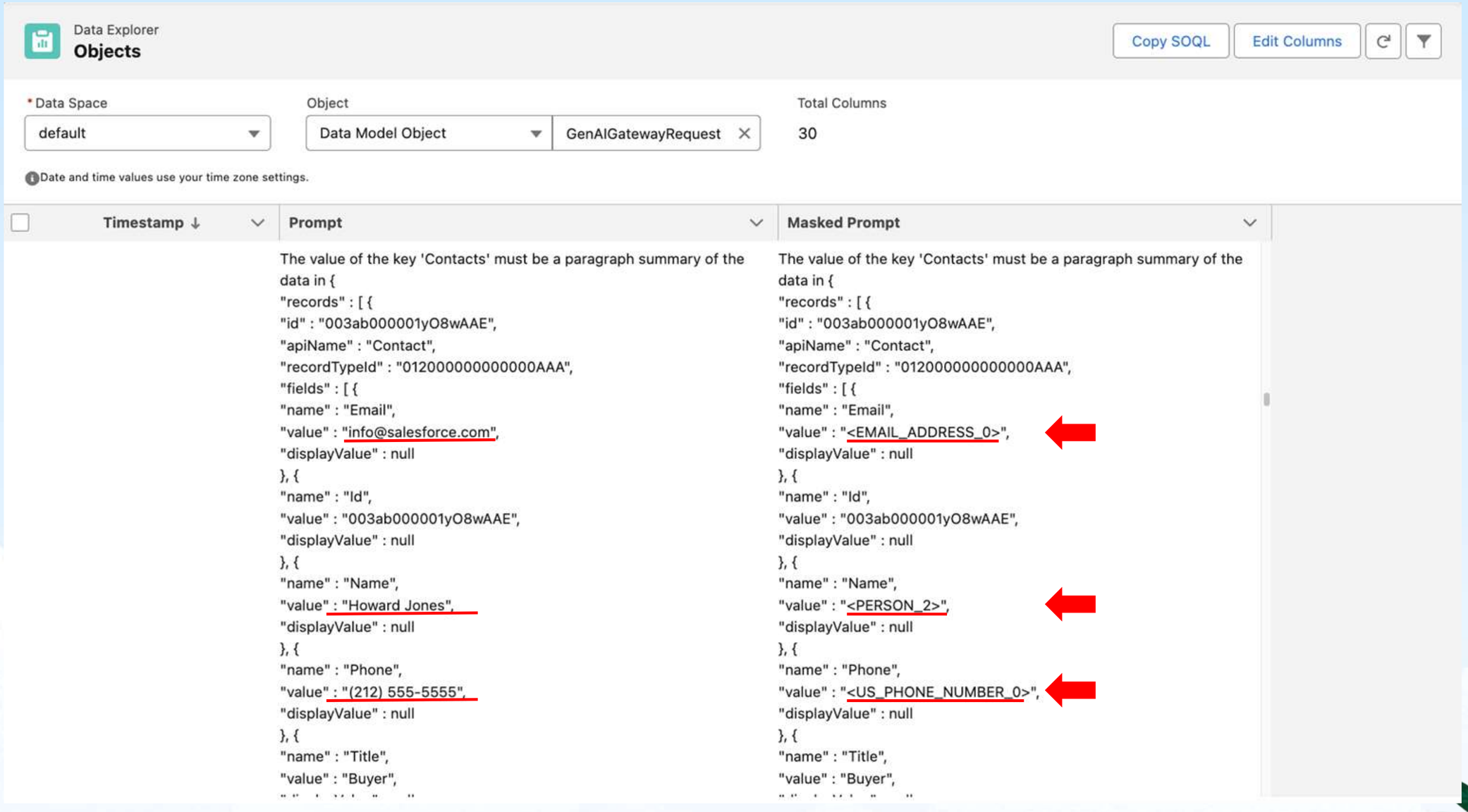
Audit Trail in Data Cloud

프롬프트, 응답, 신뢰 신호, 피드백의 포괄적인 로그



	Object Label	Object API Name	Data Streams	Data Lake Objects	Data Space	Type	Status
1	GenAIAppGeneration	GenAIAppGeneration__dlm	GenAIAppGeneration	GenAIAppGeneration	default	Custom	Ready
2	GenAIContentCategory	GenAIContentCategory__dlm	GenAIContentCategory	GenAIContentCategory	default	Custom	Ready
3	GenAIContentQuality	GenAIContentQuality__dlm	GenAIContentQuality	GenAIContentQuality	default	Custom	Ready
4	GenAIFeedback	GenAIFeedback__dlm	GenAIFeedback	GenAIFeedback	default	Custom	Ready
5	GenAIFeedbackDetail	GenAIFeedbackDetail__dlm	GenAIFeedbackDetail	GenAIFeedbackDetail	default	Custom	Ready
6	GenAIGatewayRequest	GenAIGatewayRequest__dlm	GenAIGatewayRequest	GenAIGatewayRequest	default	Custom	Ready
7	GenAIGatewayRequestTag	GenAIGatewayRequestTag_...	GenAIGatewayRequestTag	GenAIGatewayRequestTag	default	Custom	Ready
8	GenAIGatewayResponse	GenAIGatewayResponse__...	GenAIGatewayResponse	GenAIGatewayResponse	default	Custom	Ready
9	GenAIGeneration	GenAIGeneration__dlm	GenAIGeneration	GenAIGeneration	default	Custom	Ready

데이터 마스킹 in Audit Trail



Agentic AI Trust Evolution

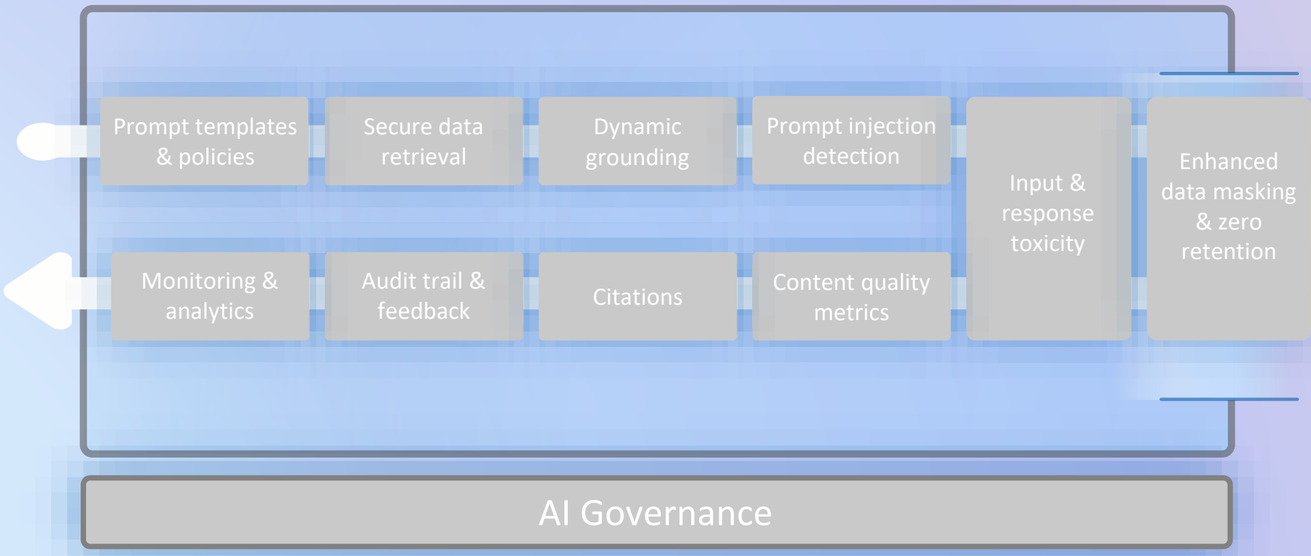
책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+



+

- 토픽 분류
- 액티브 모니터링
- 규정 준수 및 신뢰할 수 있는 AI 리더십
- Supervisory LLMs
- 인간에 의한 감독

신뢰할 수 있는 에이전트 동작 정의

Trusted AI Agents

- 신뢰성
- 인간과의 협업
- 투명한 설계
- 안전, 보안 개인 정보 보호
- 책임성

Agentforce 가드레일

Agentforce 가드레일은
신뢰할 수 있는 동작을 강화하고
AI 에이전트가 의도된 동작에서 벗어
나는 것을 방지하는 기능과 제어의 집
합체 입니다.

신뢰의 강화

긍정적이고 의도된 에이전트 행동
을 관찰하고 모니터링합니다.

의도되지 않은 동작 방지

원치 않는 동작에 대한 점검, 수정
및 개선 조치 시행

Agentforce 가드레일 라이프사이클

Preventive

Testing Center를 통해 Agent를 고객님의 비즈니스 목표와 일치하도록 구성하고, 신뢰할 수 있는 에이전트 행동을 위한 강력한 테스트를 수행합니다.

Runtime

실시간으로 안전 조치 준수 여부를 감독하여 위반 사항을 즉시 조치하고 해결함으로써 인간 중심의 감독과 통제를 유지합니다.

Learning

시간이 지나도 신뢰할 수 있는 에이전트 행동을 보장하기 위해 에이전트를 지속적으로 적응시키고 발전시키며, 실행 가능한 통찰력과 피드백을 제공합니다.

Red Teaming | 관측 | RLHF

Trust

Preventative 가드레일

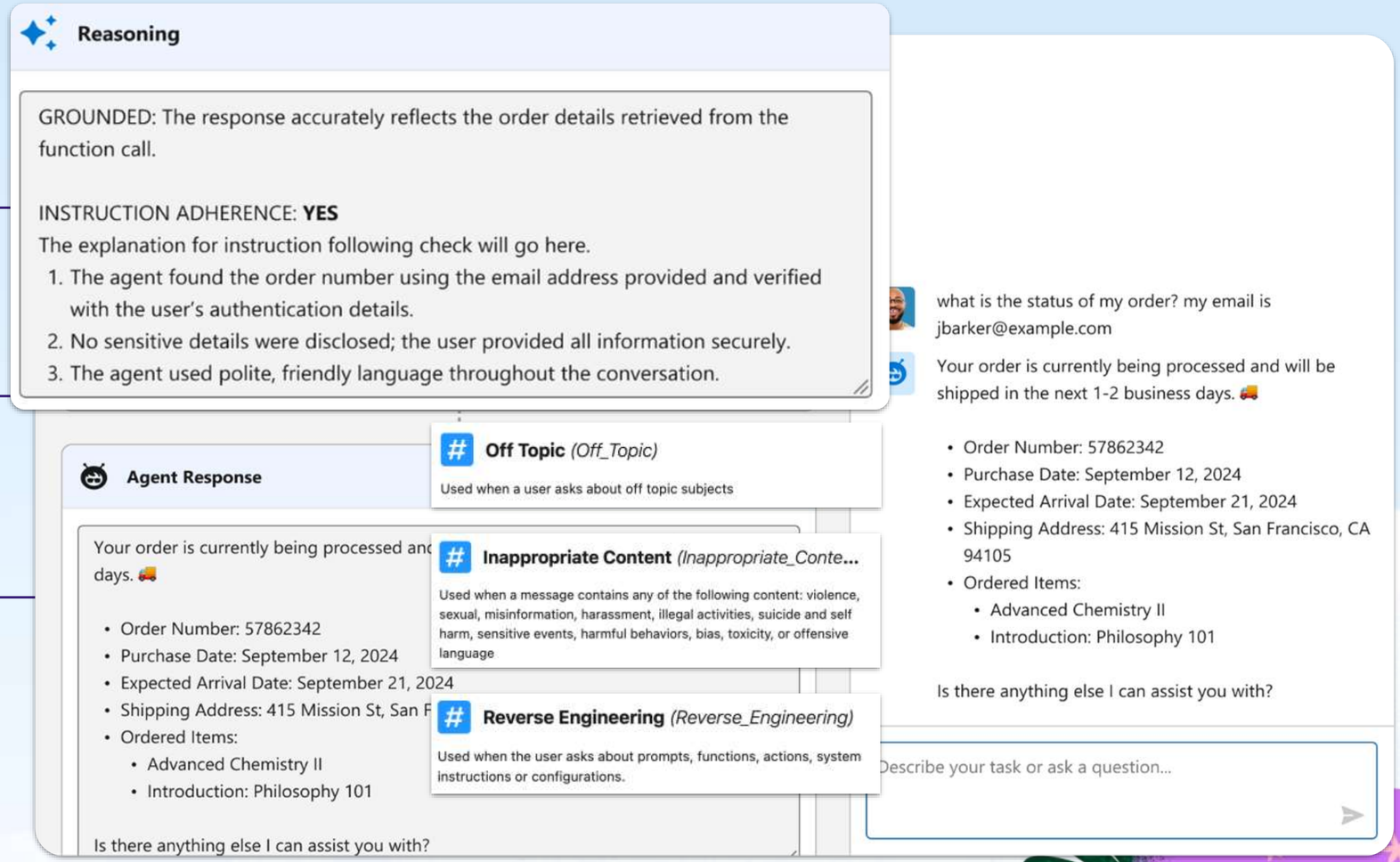
Preventative 가드레일 | **Runtime** 가드레일 | **Self-Improving** 가드레일

AI 기반 Agent Auditor가 배포 전 위험을 평가합니다

세분화된 에이전트 권한은 작업 및 데이터 액세스를 지시합니다

맥락 기반 안전성, 보안, 품질 및 운영 기본 설정

보안 및 사용자 수용을 위한 자동화된 테스트 및 디버깅



Runtime 가드레일

Preventative 가드레일 | **Runtime 가드레일** | Self-Improving 가드레일

Agentforce

대화

What is my order status?

트리거 (Flows, APIs)

Customer, company, and outcome data

Trust Layer

프롬프트

Agent Perms

안전한 데이터 취득 RAG

프롬프트 디펜스

유해성 검출

Zero data retention

응답

LLM gateway

Planner Models

GPT-4o

Model Roadmap

BYOLLM

Loop

품질 평가

Audit trail & feedback

인용

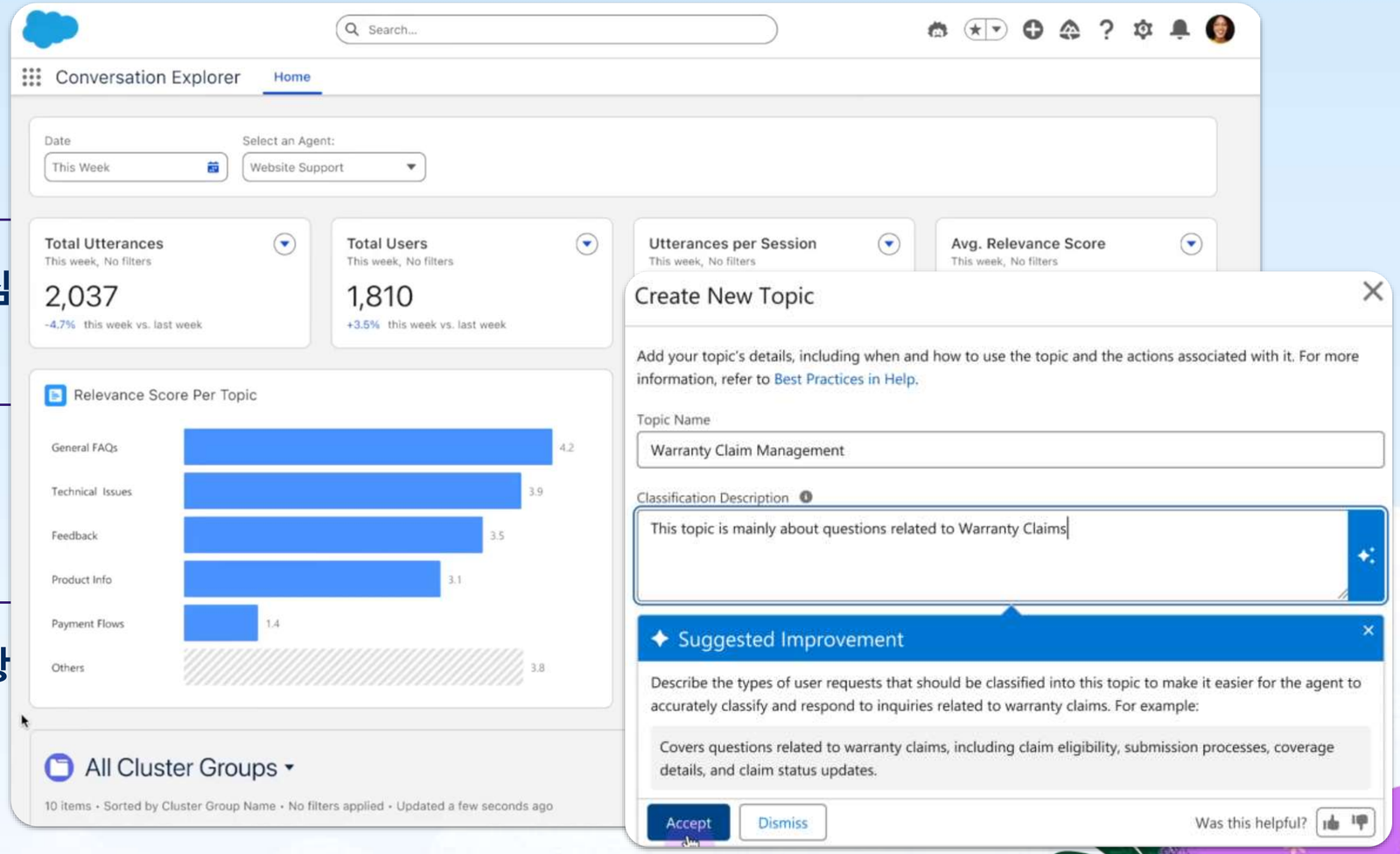
Supervisory LLM

- ✔ 지침 준수 평가
- ✔ 콘텐츠 품질 평가
- ✔ 그라운딩 평가
- ✔ 유해성 평가
- ✔ 프롬프트 인젝션 검출

Self-Improving 가드레일

Preventative 가드레일 | Runtime 가드레일 | **Self-Improving 가드레일**

- 플래그 및 로그에 악용 시도 및 인젝션 시도를 기록합니다.
- 지시사항에서 벗어난 결과물에 대한 심층적인 분석
- 고객 경험 개선을 위한 권장 사항
- 지식 기반의 품질 개선을 위한 권장 사항, RAG



Thank you