



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.



salesforce

Agentforce 에서의 차세대 보안 전략

Security Advisor, Salesforce

김남현





THANK
YOU
☺





AI의 세 번째 Wave Agents



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

△.vivint

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

WILEY

"Agentforce를 사용하면 모든 전문가가 즉시 전문가가 됩니다."

RBC | Wealth Management



Wave 1
Predictive

Wave 2
Copilots

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

OWASP top 10 for LLMs



LLM01: 2025 Prompt Injection

LLM01:2025 Prompt Injection

A Prompt Injection vulnerability occurs when user prompts alter the...

[Read More](#)

LLM02: 2025 Sensitive Information Disclosure

LLM02:2025 Sensitive Information Disclosure

Sensitive information can affect both the LLM and its application...

[Read More](#)

LLM03: 2025 Supply Chain

LLM03:2025 Supply Chain

LLM supply chains are susceptible to various vulnerabilities, which can...

[Read More](#)

LLM04: 2025 Data and Model Poisoning

LLM04:2025 Data and Model Poisoning

Data poisoning occurs when pre-training, fine-tuning, or embedding data is...

[Read More](#)

LLM05: 2025 Improper Output Handling

LLM05:2025 Improper Output Handling

Improper Output Handling refers specifically to insufficient validation, sanitization, and...

[Read More](#)

LLM06: 2025 Excessive Agency

LLM06:2025 Excessive Agency

An LLM-based system is often granted a degree of agency...

[Read More](#)

LLM07: 2025 System Prompt Leakage

LLM07:2025 System Prompt Leakage

The system prompt leakage vulnerability in LLMs refers to the...

[Read More](#)

LLM08: 2025 Vector and Embedding Weaknesses

LLM08:2025 Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities present significant security risks in systems...

[Read More](#)

LLM09: 2025 Misinformation

LLM09:2025 Misinformation

Misinformation from LLMs poses a core vulnerability for applications relying...

[Read More](#)

LLM10: 2025 Unbounded Consumption

LLM10:2025 Unbounded Consumption

Unbounded Consumption refers to the process where a Large Language...

[Read More](#)



Agentic AI Trust Evolution



책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+



+



Agentic AI Trust Evolution



책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+



+



Salesforce의 책임있는 생성형 AI 가이드라인



정확성

AI 응답의 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

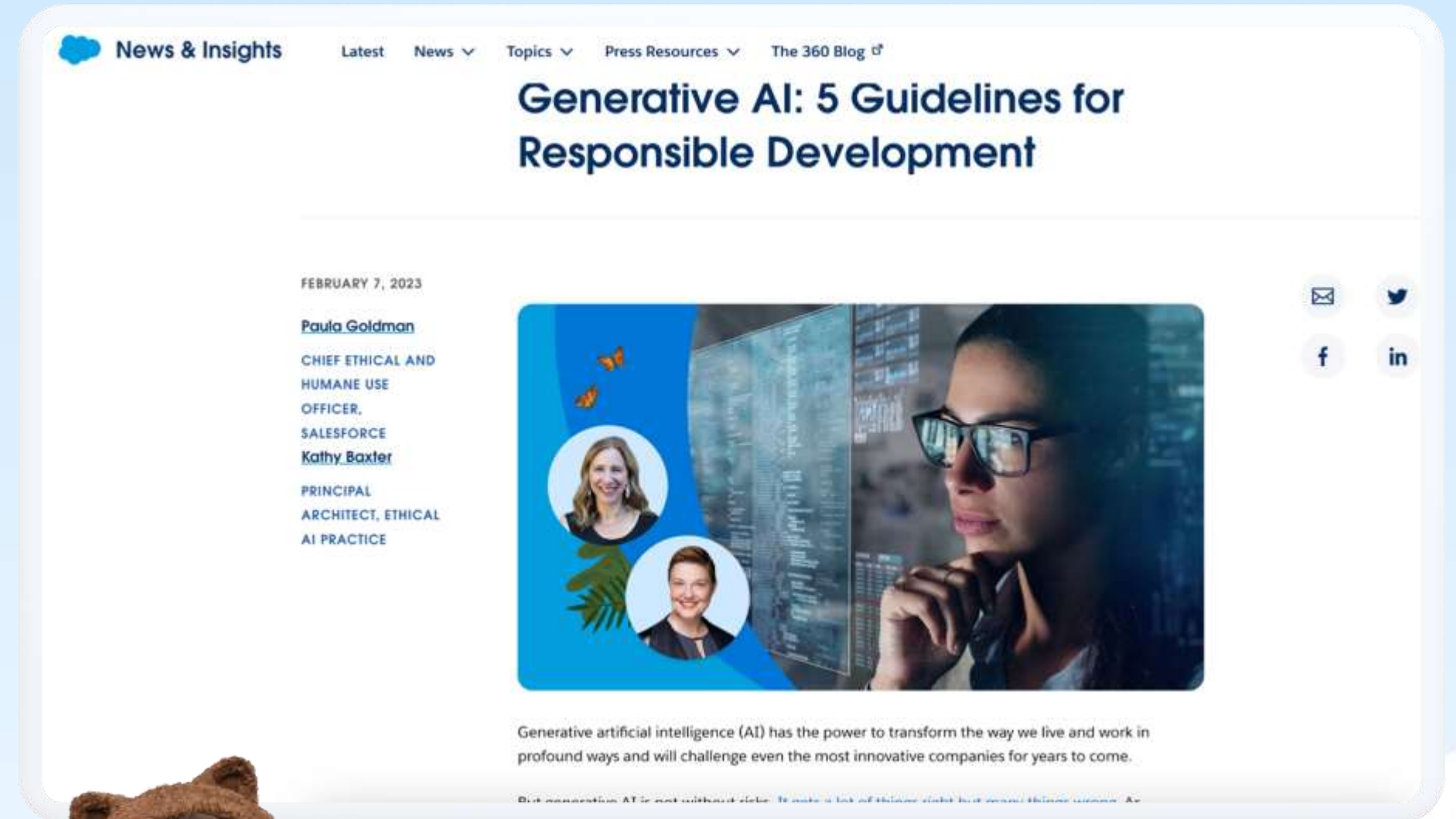
데이터의 출처를 존중하며, AI가 생성한 콘텐츠를 제공할 경우, 그 사실을 명확히 밝힙니다.

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

+



+



The Einstein Trust Layer



CRM apps



Customer, company, and outcome data

프롬프트

안전한 데이터 취득

다이나믹 그라운드

데이터 마스킹

프롬프트 디펜스

Audit trail

데이터 디마스킹

유해성 검출

응답

Zero retention

Secure gateway



Einstein Trust Layer

Models

Salesforce의 신뢰경계 내에 호스팅된 모델



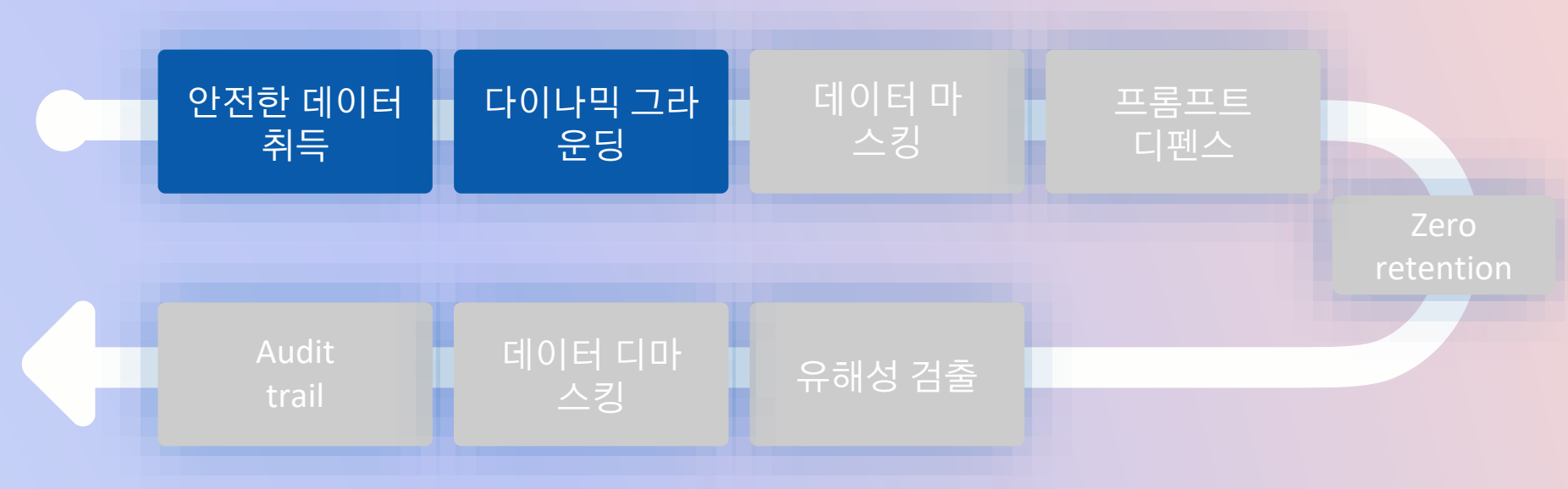
고객님의 인프라에 호스팅된 Bring Your Own Models



공유 신뢰 경계 내의 외부 모델



프롬프트 그라운드



프롬프트 빌더 에서의 프롬프트 템플릿 & 그라운드링

salesforce

The screenshot displays the Salesforce Prompt Builder interface. At the top, the navigation bar includes 'Prompt Builder', 'Account Summary', and 'Version 1 (Active)'. The main workspace is titled 'Prompt Template Workspace' and contains several paragraphs of instructions. A red arrow points to the text: 'The value of the key 'Contacts' must be a paragraph summary of the data in **Related List: Account.Contacts**. You must list the contact information for each contact. If the value of contacts is {empty}, the value must be "No Contacts".' Below this, other instructions describe 'Opportunities' and 'Activities'. A 'Preview' section at the bottom shows the 'Resolution' and 'Response'.

Prompt Template Workspace

Resource **Object Field**
Search for or select a resource to insert Account > Einstein Account Summary

The value of the key 'Contacts' must be a paragraph summary of the data in **Related List: Account.Contacts**. You must list the contact information for each contact. If the value of contacts is {empty}, the value must be "No Contacts".

The value of the key 'Opportunities' should summarize in a paragraph **Related List: Account.Opportunities**, describing the number of open opportunities and when the next one is due to close if there are any. If the value of opportunities is {empty}, the value must be "No Opportunities".

The key 'Activities' must be a paragraph summarizing the number of activities and the next scheduled meeting if available in **Related List: Account.OpenActivities**. If the value of activities is {empty}, the value must be "No open activities".

The key 'Cases' must be a summary of data in **Related List: Account.Cases**. If the value is {empty}, the value must be "No Cases".

Use clear, concise, and straightforward language, avoiding the use of filler words and phrases and redundant language.

Configuration

Template Properties

- * Model Type: Standard
- * Models: Standard OpenAI GPT 3.5 Turbo ...

[View this model](#)

Preview

Resolution Enabled **Global Media** [Preview](#)

Resolution

The value of the key 'Contacts' must be a paragraph summary of the data in {

```
"records" : [ {  
  "id" : "003ab000001yO8uAAE",  
  "apiName" : "Contact",  
  "recordTypeId" : "0120000000000000AAA",  
  "fields" : [ {  
    "name" : "Email",  
    "value" : "info@salesforce.com",  
    "displayValue" : null
```

Response

Contacts: Three contacts.
Contact 1:
- Name: Jon Amos
- Title: Sales Manager
- Email: info@salesforce.com
- Phone: (905) 555-1212
Contact 2:

데이터 마스크



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>
[REDACTED]
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?"

...

<NAME_1> <COMPANY_2>
[REDACTED]
<CARD_1>

데이터 마스킹 설정



Setup Home Object Manager

Search Setup

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

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.	<input checked="" type="checkbox"/>
Email Address	An email address. For example, astro@salesforce.com.	<input checked="" type="checkbox"/>
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.	<input type="checkbox"/>
Company Name	The company name.	<input type="checkbox"/>
Passport	A number listed on a passport.	<input type="checkbox"/>
Name	A person's first, middle, or last name, or a combination of these.	<input checked="" type="checkbox"/>
Phone Number	A phone number with or without a country code.	<input checked="" type="checkbox"/>
US Drivers License	A number listed on the front of a driver's license card.	<input type="checkbox"/>
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.	<input type="checkbox"/>
US SSN	A US Social Security Number (SSN) that includes 9 digits. For example, 123-45-6789.	<input checked="" type="checkbox"/>



프롬프트 디펜스



시스템 정책 & 가드레일

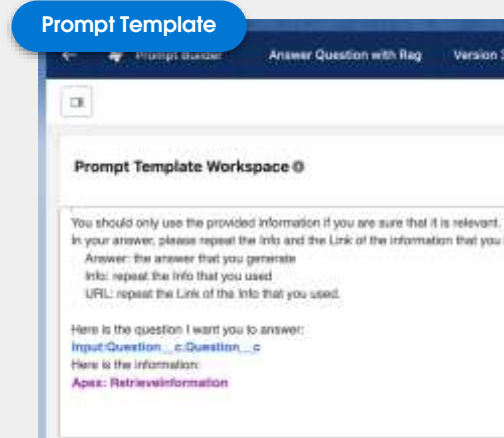


방어 지침

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

시스템 가드레일
(정책 & 보안)

프롬프트 템플릿 (Business Defined)



동적 유저 입력

유스케이스 지침 & 데이터

엔드유저의 입력

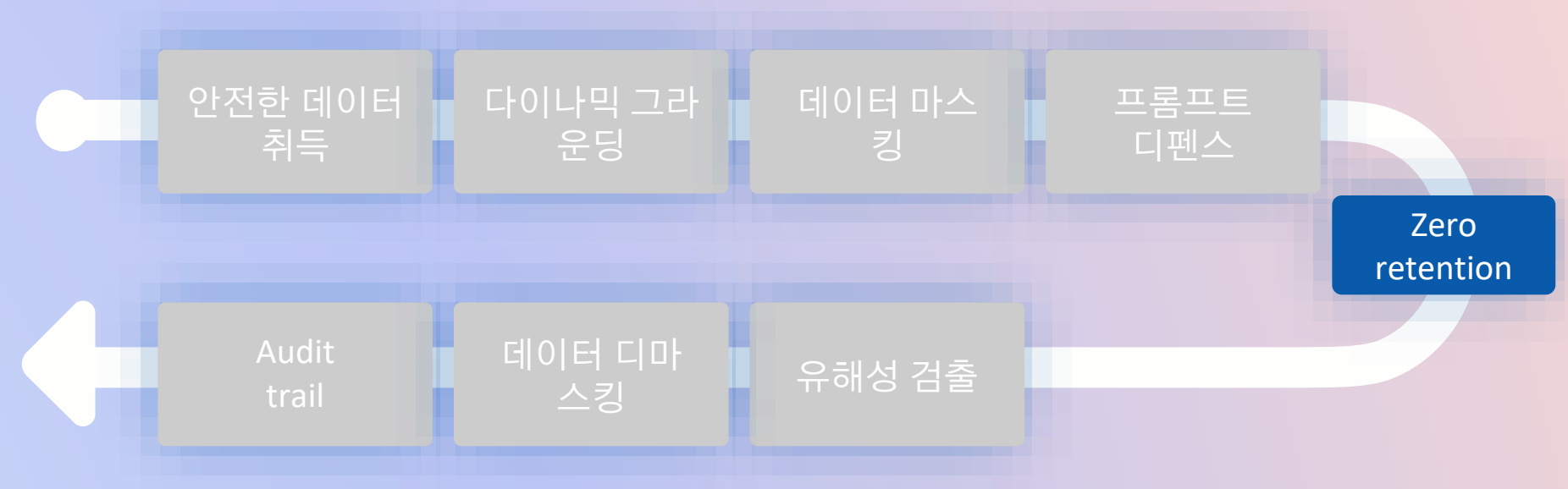
Post-Prompting

“오류가 발생하거나 응답의 유효성에 대해 확신하지 못한다면, '모르겠습니다'라고 말씀해 주세요.”

프롬프트 인젝션 방어를 포함한
추가적인 가드레일 및 지침



Zero Data Retention



Zero Data Retention 이란?

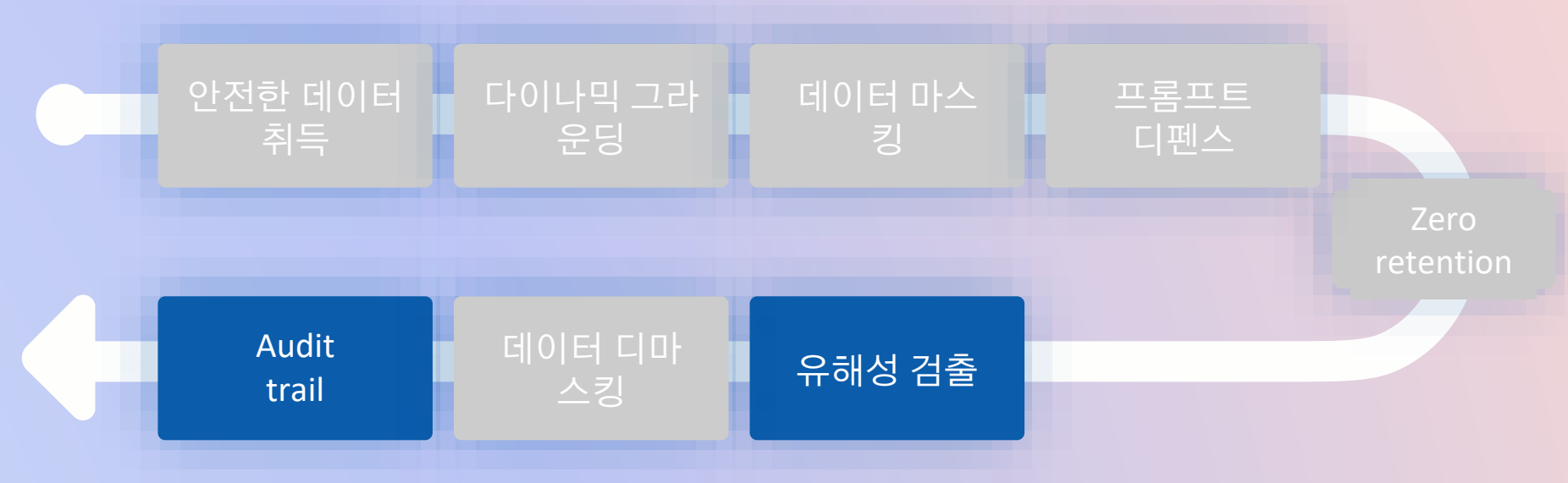


대규모 언어모델 (LLM) 제공자와의 Zero Data Retention 계약

1. 모델 교육 및 제품 개선에 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 분류를 위한
Fine Tuning Distilbert 기반
모델

공개 데이터 세트로 학습

각국의 언어/지역 지원

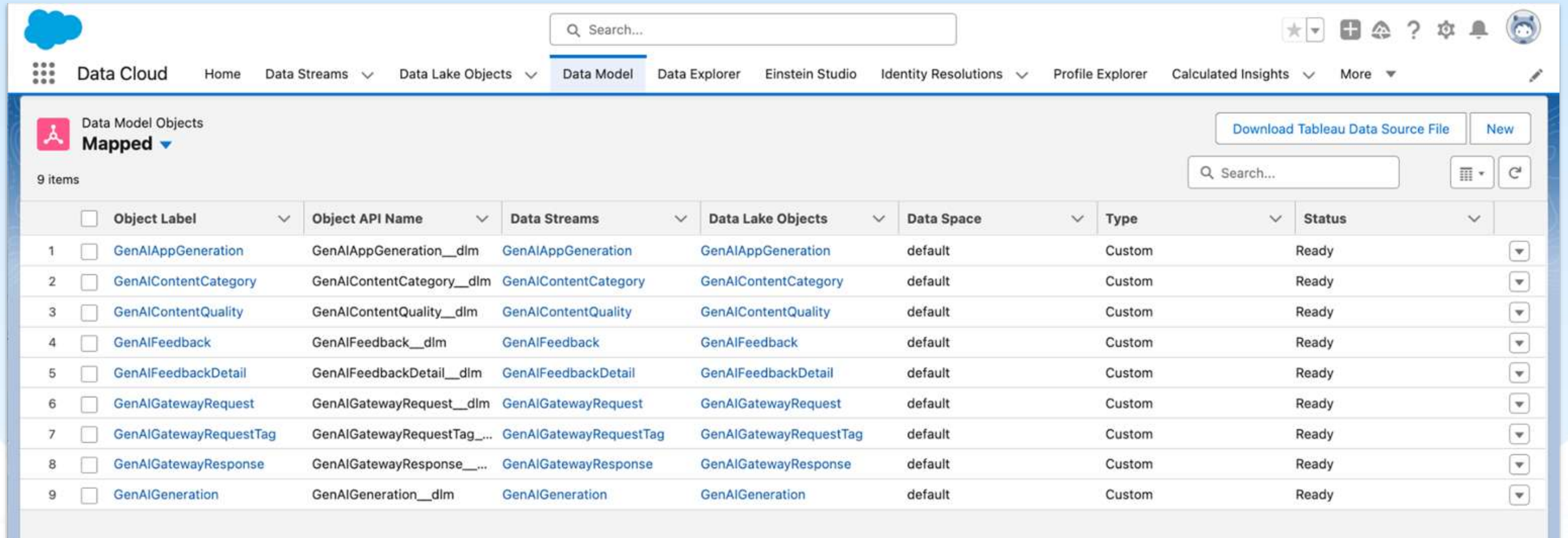
Response Text	Category	Value
{"Contacts": "There are three contacts: Jon Amos (Sales Manager)"}	hate	0.0
{"Contacts": "There are three contacts: Jon Amos (Sales Manager)"}	profanity	0.000020
{"Contacts": "There are three contacts: Jon Amos (Sales Manager)"}	sexual	0.000030
{"Contacts": "There are three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



Audit Trail in Data Cloud

salesforce

프롬프트, 응답, 신뢰 신호, 피드백의 포괄적인 로그



The screenshot displays the Salesforce Data Cloud interface. At the top, there is a search bar and navigation tabs for Home, Data Streams, Data Lake Objects, Data Model (selected), Data Explorer, Einstein Studio, Identity Resolutions, Profile Explorer, Calculated Insights, and More. Below the navigation, the 'Data Model Objects' section is visible, showing a list of 9 items under the 'Mapped' filter. A 'Download Tableau Data Source File' button and a 'New' button are also present. The table below lists the objects with columns for Object Label, Object API Name, Data Streams, Data Lake Objects, Data Space, Type, and Status.

	<input type="checkbox"/> Object Label	<input type="checkbox"/> Object API Name	<input type="checkbox"/> Data Streams	<input type="checkbox"/> Data Lake Objects	<input type="checkbox"/> Data Space	<input type="checkbox"/> Type	<input type="checkbox"/> Status
1	<input type="checkbox"/> GenAIAppGeneration	GenAIAppGeneration__dml	GenAIAppGeneration	GenAIAppGeneration	default	Custom	Ready
2	<input type="checkbox"/> GenAIContentCategory	GenAIContentCategory__dml	GenAIContentCategory	GenAIContentCategory	default	Custom	Ready
3	<input type="checkbox"/> GenAIContentQuality	GenAIContentQuality__dml	GenAIContentQuality	GenAIContentQuality	default	Custom	Ready
4	<input type="checkbox"/> GenAIFeedback	GenAIFeedback__dml	GenAIFeedback	GenAIFeedback	default	Custom	Ready
5	<input type="checkbox"/> GenAIFeedbackDetail	GenAIFeedbackDetail__dml	GenAIFeedbackDetail	GenAIFeedbackDetail	default	Custom	Ready
6	<input type="checkbox"/> GenAIGatewayRequest	GenAIGatewayRequest__dml	GenAIGatewayRequest	GenAIGatewayRequest	default	Custom	Ready
7	<input type="checkbox"/> GenAIGatewayRequestTag	GenAIGatewayRequestTag_...	GenAIGatewayRequestTag	GenAIGatewayRequestTag	default	Custom	Ready
8	<input type="checkbox"/> GenAIGatewayResponse	GenAIGatewayResponse_...	GenAIGatewayResponse	GenAIGatewayResponse	default	Custom	Ready
9	<input type="checkbox"/> GenAIGeneration	GenAIGeneration__dml	GenAIGeneration	GenAIGeneration	default	Custom	Ready

데이터 마스킹 in Audit Trail



Data Explorer
Objects

Copy SOQL Edit Columns

Data Space: default Object: Data Model Object GenAIGatewayRequest Total Columns: 30

Date and time values use your time zone settings.

Timestamp ↓	Prompt ↓	Masked Prompt ↓
	<pre>The value of the key 'Contacts' must be a paragraph summary of the data in { "records": [{ "id": "003ab000001yO8wAAE", "apiName": "Contact", "recordTypeId": "0120000000000000AAA", "fields": [{ "name": "Email", "value": "info@salesforce.com", "displayValue": null }, { "name": "Id", "value": "003ab000001yO8wAAE", "displayValue": null }, { "name": "Name", "value": "Howard Jones", "displayValue": null }, { "name": "Phone", "value": "(212) 555-5555", "displayValue": null }, { "name": "Title", "value": "Buyer", "displayValue": null }] }</pre>	<pre>The value of the key 'Contacts' must be a paragraph summary of the data in { "records": [{ "id": "003ab000001yO8wAAE", "apiName": "Contact", "recordTypeId": "0120000000000000AAA", "fields": [{ "name": "Email", "value": "<EMAIL_ADDRESS_0>", "displayValue": null }, { "name": "Id", "value": "003ab000001yO8wAAE", "displayValue": null }, { "name": "Name", "value": "<PERSON_2>", "displayValue": null }, { "name": "Phone", "value": "<US_PHONE_NUMBER_0>", "displayValue": null }, { "name": "Title", "value": "Buyer", "displayValue": null }] }</pre>



Agentic AI Trust Evolution



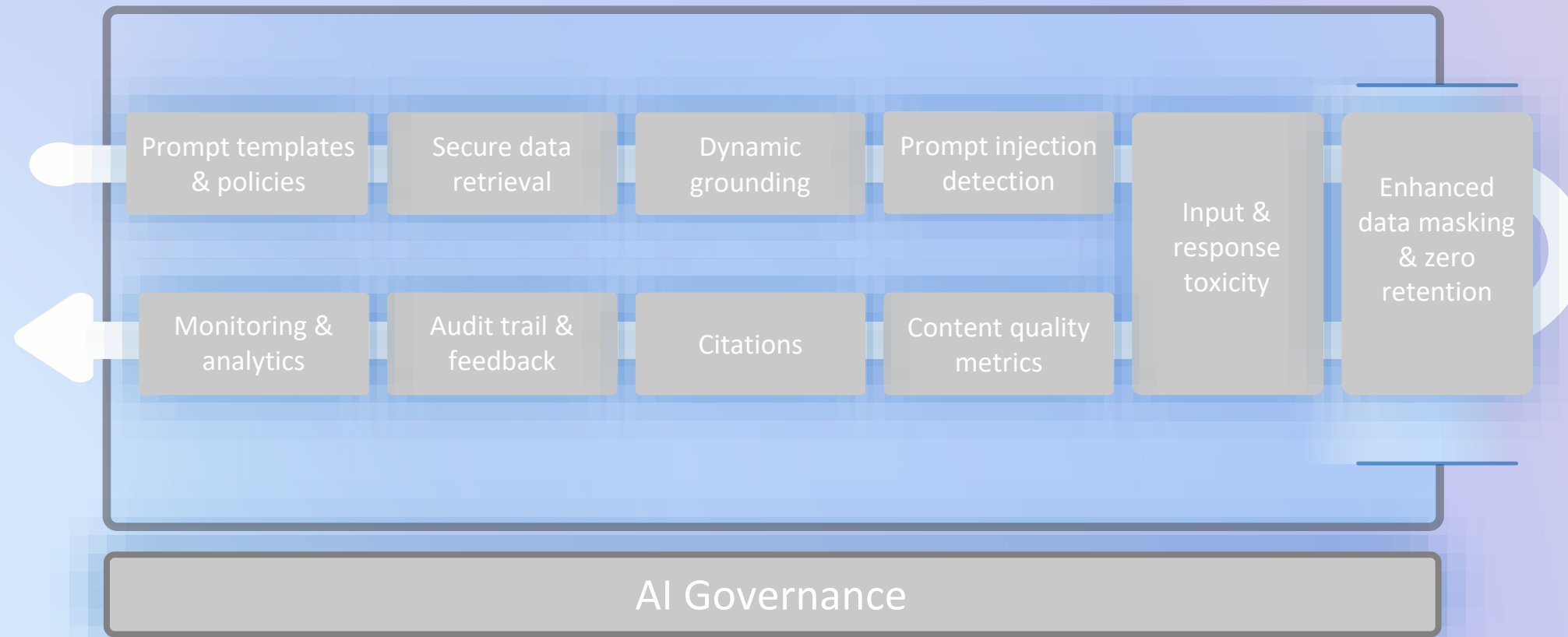
책임있는 생성형 AI 가이드라인

Einstein Trust Layer

가드레일

- 정확성
- 안전성
- 정직함
- 권한 강화
- 지속가능성

+



+



신뢰할 수 있는 에이전트 동작 정의

salesforce



Agentforce 가드레일

salesforce

Agentforce 가드레일은 신뢰할 수 있는 동작을 강화하고 AI 에이전트가 의도된 동작에서 벗어나는 것을 방지하는 기능과 제어의 집합체입니다.



신뢰의 강화

긍정적이고 의도된 에이전트 행동을 관찰하고 모니터링합니다.



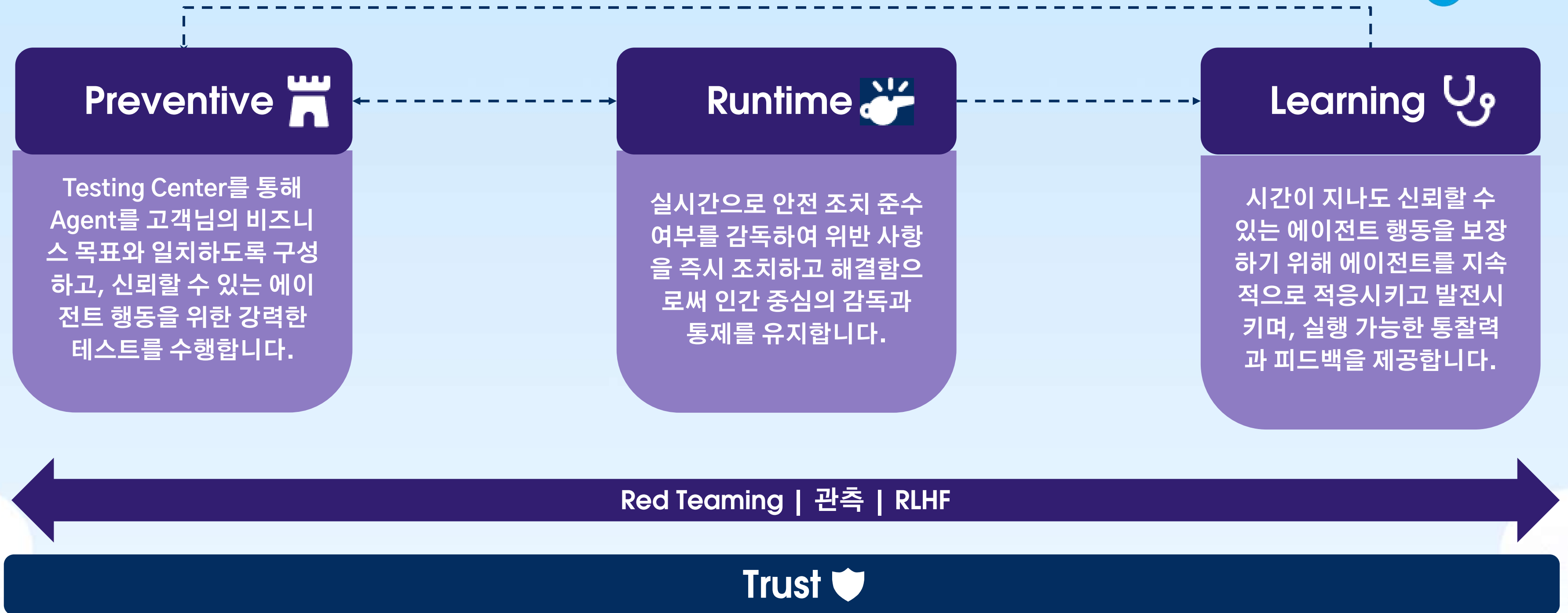
의도되지 않은 동작 방지

원치 않는 동작에 대한 점검, 수정 및 개선 조치 시행



Agentforce 가드레일 라이프사이클

salesforce



Preventative 가드레일

salesforce

Preventative 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

Reasoning

GROUNDING: The response accurately reflects the order details retrieved from the function call.

INSTRUCTION ADHERENCE: **YES**
The explanation for instruction following check will go here.

1. The agent found the order number using the email address provided and verified with the user's authentication details.
2. No sensitive details were disclosed; the user provided all information securely.
3. The agent used polite, friendly language throughout the conversation.

Agent Response

Your order is currently being processed and will be shipped in the next 1-2 business days. 🚚

- Order Number: 57862342
- Purchase Date: September 12, 2024
- Expected Arrival Date: September 21, 2024
- Shipping Address: 415 Mission St, San Francisco, CA 94105
- Ordered Items:
 - Advanced Chemistry II
 - Introduction: Philosophy 101

Is there anything else I can assist you with?

Off Topic (Off_Topic)

Used when a user asks about off topic subjects

Inappropriate Content (Inappropriate_Content)

Used when a message contains any of the following content: violence, sexual, misinformation, harassment, illegal activities, suicide and self harm, sensitive events, harmful behaviors, bias, toxicity, or offensive language.

Reverse Engineering (Reverse_Engineering)

Used when the user asks about prompts, functions, actions, system instructions or configurations.

Self-Improving 가드레일



what is the status of my order? my email is jbarker@example.com



Your order is currently being processed and will be shipped in the next 1-2 business days. 🚚

- Order Number: 57862342
- Purchase Date: September 12, 2024
- Expected Arrival Date: September 21, 2024
- Shipping Address: 415 Mission St, San Francisco, CA 94105
- Ordered Items:
 - Advanced Chemistry II
 - Introduction: Philosophy 101

Is there anything else I can assist you with?

Describe your task or ask a question...



Runtime 가드레일

salesforce

Preventative 가드레일

Runtime 가드레일

Self-Improving 가드레일



Self-Improving 가드레일

salesforce

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

The screenshot displays the Salesforce Conversation Explorer interface. At the top, there's a search bar and navigation icons. Below, the 'Home' view shows filters for 'Date' (This Week) and 'Select an Agent' (Website Support). Key metrics include Total Utterances (2,037, -4.7% vs last week) and Total Users (1,810, +3.5% vs last week). A 'Relevance Score Per Topic' bar chart shows scores for General FAQs (4.2), Technical Issues (3.9), Feedback (3.5), Product Info (3.1), Payment Flows (1.4), and Others (3.8). A 'Create New Topic' modal is open, showing a topic named 'Warranty Claim Management' with a classification description: 'This topic is mainly about questions related to Warranty Claims'. A 'Suggested Improvement' section provides a detailed description for better classification. The interface also includes 'All Cluster Groups' and a footer with '10 items - Sorted by Cluster Group Name - No filters applied - Updated a few seconds ago'.

Topic	Score
General FAQs	4.2
Technical Issues	3.9
Feedback	3.5
Product Info	3.1
Payment Flows	1.4
Others	3.8



Thank you

