

ORACLE

Oracle RAG 아키텍처

시멘틱 검색과 믿음직한 LLM 구축을 위한 Oracle RAG 아키텍처

김태완

Cloud Engineering, Oracle Korea

Jan 25, 2024



발표 내용 요약

OCI Generative AI의 주요 특징을 살펴보고,
LLM 프롬프트 엔지니어링을 통해서 답변의 신뢰성을 높이는 RAG 기법에 대하여 소개합니다.
마지막으로 오라클이 RAG 기법을 지원하기 위해 개발한 OCI Generative AI Agent 아키텍처를 소개합니다.

- 목차
 - ✓ OCI Generative AI 소개
 - ✓ 기업용 LLM 신뢰성 확보: RAG
 - ✓ 오라클 RAG 지원 아키텍처: OCI Generative AI Agent

Safe harbor statement



The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

OCI Generative AI 서비스

엔터프라이즈 기업을 위한 LLM 모델 서비스

GA of OCI Generative AI
2024.01.23

The screenshot shows the 'Generative AI overview' page in the Oracle Cloud console. The page is divided into several sections:

- Generative AI overview:** A dark banner with the text 'Power your apps with large language models and generative AI' and a 'Watch service tour' button.
- Metrics in my compartment:** Three colored boxes showing metrics: 'Dedicated AI clusters' (7), 'Custom models' (3), and 'Active endpoints' (12).
- Get started:** A section with a 'Go to playground' button and a 'Playground' link.
- Resources:** A list of links including 'All documentation', 'Rest API reference', 'Workshops', 'Tutorials', and 'Pricing'.
- Scope:** A dropdown menu for 'Compartment' set to 'My compartment'.

The screenshot shows the 'Create model' page in the Oracle Cloud console. The page is divided into several sections:

- Create model:** A section with a 'Fine-tuning configuration' sub-section.
- Fine-tuning configuration:** A section with a 'Base model' dropdown set to 'cohere.command-light.15.6', a 'Fine-tuning method' dropdown set to 'T-Few', and a 'Dedicated AI cluster in genaiusers (root)' dropdown set to 'generativeaidedicatedcluster20240115214815'.
- Advanced options:** A section with various configuration options: 'Total training epochs' (3), 'Learning rate' (0.01), 'Training batch size' (16), 'Early stopping patience' (6), 'Early stopping threshold' (0.01), and 'Log model metrics interval in steps' (10).

OCI Generative AI 서비스

엔터프라이즈 기업을 위한 LLM 서비스 지원 모델

GA of OCI Generative AI
2024.01.23

The screenshot shows the Oracle Cloud Generative AI overview page. The top navigation bar includes the Oracle Cloud logo, a search bar, and the region 'US West (Phoenix)'. The left sidebar lists navigation options: Overview, Playground, Dedicated AI clusters, Custom models, Endpoints, and Scope. The main content area is titled 'Generative AI overview' and features a description of the service, a 'Watch service tour' button, and a 'Metrics in my compartment' section with three cards: 'Dedicated AI clusters' (7), 'Custom models' (3), and 'Active endpoints' (12). Below this is a 'Get started' section with a 'Go to playground' button and three cards: 'Playground', 'Dedicated AI clusters', and 'Custom models'. The 'Playground' card describes it as a visual interface for exploring models. The 'Dedicated AI clusters' card describes spinning up hardware units. The 'Custom models' card describes creating custom models by fine-tuning base models.



+



오라클의 AI 서비스 스택

Applications

Fusion Applications

NetSuite


Fusion Analytics

Industry Applications

3rd Party Applications

AI Services

NEW



OCI Generative AI

NEW




OCI GenAI Agents



Digital Assistant



Speech



Language



Vision




Document Understanding


Data Platforms




OCI Data Science



AI Vector Search
in Oracle Database




MySQL HeatWave Vector Store




OCI Data Labeling

Data

AI Infrastructure



Compute bare metal instances and VMs
with NVIDIA GPUs



OCI Supercluster with RDMA networking




Block, object, and file storage; HPC filesystems

Oracle AI Partners

오라클의 AI 서비스 스택

Applications

NEW



OCI Generative AI

NEW



OCI GenAI Agents



Digital Assistant



Speech



Language



Vision




Document Understanding

ORACLE AI Demos

Search

Back



Speech Transcription

Speech can be used to transcribe batches of audio files.

Speech

Audio Files

Name	Duration
banking-1.wav	00:04:01
banking-2.wav	00:04:22
banking-3.wav	00:05:07
insurance-1.wav	00:04:25
insurance-2.wav	00:04:44
insurance-3.wav	00:06:03

banking-1.wav

Analyze

Results

00:00:01-00:00:02 Good afternoon.

00:00:02-00:00:03 Thank you for calling Wells Fargo.

00:00:03-00:00:04 My name is ready.

00:00:04-00:00:09 How can I help you? Yes, I would like to set up a joint account for me and my kid.

00:00:10-00:00:15 I'm so, um, if we do open this account, it would be linked to this account.

00:00:15-00:00:22 Is that okay? Asked, okay, perfect I am.

00:00:22-00:00:27 So what is your **forename**? My full name is Eric melton Jr. k.

00:00:27-00:00:33 Take me so much I'm in and I will be asking a series of information just to confirm your identity.

00:00:33-00:00:34 So, **Cosider** with me.

00:00:35-00:00:45 What are **your** national insurance number is X246374. K&Q.

00:00:45-00:00:56 **Im** a man can also have you verify your date of quarter Alexander May 26th 1990.

Text for analysis

The European sovereign debt crisis was a period when several European countries experienced the collapse of financial institutions, high government debt, and rapidly rising bond yield spreads in government securities. The debt crisis began in 2008 with the collapse of Iceland's banking system, then spread primarily to Portugal, Italy, Ireland, Greece, and Spain in 2009, leading to the popularisation of an offensive monetary (PMOS). It has led to a loss of confidence in European businesses and economies. The crisis was eventually controlled by the financial guarantees of European countries, who feared the collapse of the euro and financial contagion, and by the International Monetary Fund (IMF).

0007000 Characters remaining

Analyze

Language Detection

English

Text Classification

Finance/Banking

Named Entity Recognition

The **European sovereign debt crisis** was a period when several European countries experienced the collapse of financial institutions, high government debt, and rapidly rising **bond yield** spreads in government securities. The **debt crisis** began in 2008 with the collapse of **Iceland's** banking system, then spread primarily to **Portugal**, **Italy**, **Ireland**, **Greece**, and **Spain** in 2009, leading to the popularisation of an offensive monetary (PMOS). It has led to a loss of confidence in European businesses and economies. The **crisis** was eventually controlled by the financial guarantees of European countries, who feared the collapse of the **euro** and financial contagion, and by the **International Monetary Fund** (IMF).

Key phrase extraction

collapse european sovereign debt crisis several european countries collapse of financial institutions high government debt rising bond yield spreads government securities debt crisis Iceland's banking system loss of confidence european businesses and economies financial guarantees european countries euro and financial contagion international monetary fund portugal Italy Greece popularisation pmos crisis IMF Ireland Spain

Sentiment

Document sentiment Negative

No sentiment aspects returned


Sentence level sentiment

1 To E

ORACLE AI Demos

Search

Back



Translation

OCI Language Translate is a multilingual neural machine translation service to translate text from one language into another.

Language

Text Translation

Source language


English

Target language

오라클의 AI 서비스 스택

AI Services

NEW




OCI Generative AI

NEW




OCI GenAI Agents



Digital Assistant



Speech



Language



Vision



Document Understanding



Auto design	99.26%
Car	99.26%
Bumper	99.23%
Wheel	98.95%



Oracle Generative AI 전략 & 차별성



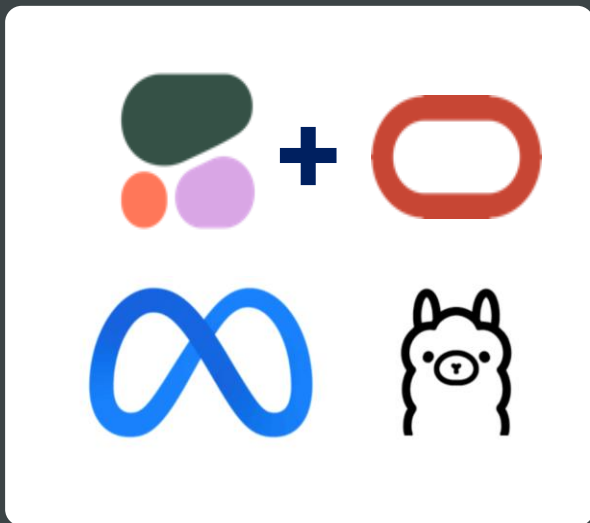
Model



DATA



INFRASTRUCTURE



Oracle
Database 23C



OpenSearch



Heatwave



ADW



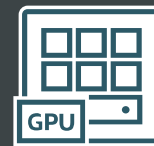
ATP



EXACI



EXACC



H100/A100/A10 GPU



Cloud Native
OKE



Super Cluster



Chatbot

Oracle Enterprise LLM/Gen AI Platform



Oracle Enterprise LLM Platform

엔터프라이즈 기업을 위한 SOTA LLM 모델 제공



Llama-2 70B

Meta에서 개발한 llama-2 70B 파라미터 텍스트 생성 모델
연구 및 상업적 용도로 무료로 사용 가능한 개방형 LLM



Command

Command는 Cohere의 고성능 생성형 LLM 모델
모델 크기: 6B (Medium), 52B 파라미터 (XL)

Command XL 모델이 Command M 모델 보다 더 높은 정확성, Command M은 비용 및 처리 속도에 강점

Summarize

Summarize 모델은 문서에서 가장 중요한 정보를 정확하게 포착하고 고품질 요약물을 생성하는 특화 모델

Embed

Embed는 텍스트를 벡터로 임베딩하는 모델로 영어 및 다국어 언어 임베딩 모델(V3) 지원.
더 작고 빠르지만 성능이 약간 떨어지는 '라이트' 버전의 모델도 제공 (영어전용).

생성

- ✓ cohere.command v15.6
- cohere.command-light v15.6
- meta.llama-2-70b-chat

요약

cohere.command v15.6

포함

cohere.embed-english-light-v2.0
cohere.embed-english-light-v3.0
cohere.embed-english-v3.0
cohere.embed-multilingual-light-v3.0
cohere.embed-multilingual-v3.0

Oracle Gen AI Partnership with Cohere

오라클은 LLM 선도 기업인 Cohere와 긴밀한 파트너십을 통해 엔터프라이즈를 위한 Generative AI 서비스 개발

스텐포드에서 수행한 자연어 LLM 모델의 성능을 측정하는 HELM(Holistic Evaluation of Language Models)의 결과에 따르면 Cohere 모델이 가장 뛰어난 성능을 제공

Cohere 모델은 파라미터가 52B(520억개 파라미터)로 구성되고, OpenAI의 다빈치 모델은 178B(1,780억개) 파라미터를 갖음

Cohere 모델이 OpenAI 모델에 비교하여 모델 크기가 작으면서 뛰어난 성능 제공

작은 LLM 모델이 갖는 강점

- 빠른 추론, 빠른 성능 제공
- 작은 모델을 Foundation AI 모델 (Base Model)로 추가학습 시키는데 투입되는 GPU 수량과 시간을 줄임
- 운영 비용 절감 (Production에 투입 GPU 수량 감소)

Cohere 모델의 특성은 엔터프라이즈를 위한 Generative AI에 적합한 가격, 비용, 성능 효율성 제공



Generative AI for enterprise— offered through Oracle

Company	Model	Model type	Mean win rate
cohere	Cohere Command (52B)	Command	93.0%
OpenAI	Davinci Instruct 002	Command	93.0%
OpenAI	Davinci Instruct 003	Command	89.8%
NVIDIA Microsoft	TNLG v2 (530B) <i>not publicly available or viable to serve given size</i>	Base	85.5%
ANTHROPIC	Anthropic v4 (52B)	Command	84.2%
AI21labs	J1 Grande v2 (17B)	Command	80.6%
AI21labs	Luminous Supreme (70B)	Command	78.3%
cohere	Cohere XL (52B)	Base	74%
Meta	OPT (175B)	Base	67.8%
OpenAI	GPT-3 Davinci (175B)	Base	62.8%
AI21labs	J1-Jumbo (178B)	Base	59.2%
AI21labs	Luminous Extended (30B)	Command	58.2%
Hugging Face	BLOOM (176B)	Base	52.9%

Cohere delivers top-tier LLM performance, outperforming peers in independent LLM benchmarks

(not included: GPT-4 from OpenAI)

Cohere's Command model is ranked very highly in HELM, **while being more efficient** (52B parameters compared to GPT-3 with 175B parameters) and **more easily customized**

OCI Generative AI 서비스: Playground

Generation: Document Transformation

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information." The "Model" dropdown is set to "cohere.command". The "Example" dropdown is set to "Rewrite instructions with steps". The "Input" field contains the text: "Rewrite the following steps as a numbered list: First, reboot the computer to make sure you have a clean memory footprint. Next, click the main menu, and select settings. In the search box of the settings app, type 'updates' and click search. Select the search results titled 'Software updates'. In the Software updates screen, click the 'Check for updates' link. If no updates are available you're done. If updates are available for your computer, click 'Proceed with this update' link. Read the update description and confirm the installation by clicking 'Confirm'. The installation might take several minutes and might require the computer to restart several times during the installation. If restart happens, you are asked to login to the computer after each restart. After the process completes, you get a summary of the installed software." The "Output" field shows the generated response: "1. Reboot the computer to make sure you have a clean memory footprint 2. Click the main menu, and select settings 3. In the search box of the settings app, type 'updates' and click search 4. Select the search results titled 'Software updates' 5. In the Software updates screen, click the 'Check for updates' link 6. If no updates are available, you're done 7. If updates are available for your computer, click 'Proceed with this update' link 8. Read the update description and confirm the installation by clicking 'Confirm' 9. The installation might take several minutes and might require the computer to restart several times during the installation 10. If restart happens, you are asked to login to the computer after each restart 11. After the process completes, you get a summary of the installed software".

Summarization

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information." The "Model" dropdown is set to "cohere.command". The "Example" dropdown is set to "Summarize a blog post". The "Input" field contains the text: "트랜스포머(Transformer)는 버트(BERT) 및 GPT-3와 같이 오늘날 널리 사용되는 언어 모델의 종류이다. 처음에는 자연어처리(NLP)를 위해 개발되었지만 컴퓨터 비전, 약물 발견 등 전 산업에서 활용에 혁신을 가져오고 있다. 예를 들어, 은행 금융 서비스 및 보험 회사에서는 고객을 유지하고 비용을 줄이기 위해 정제된 규칙과 규정 준수를 준수해야 한다. 이후 솔루션은 기업에 정보 검색, 지도 분석, 고객 서비스 등과 관련된 프로세스 자동화 및 응용 프로그램을 효율적으로 수행하고 그에 따른 위험을 완화하도록 돕는다. 트랜스포머 모델은 문장 속 단어와 같은 소문 키워드를 추적할 수 있는 능력을 높이고 변화하는 규칙과 규정 준수를 준수해야 한다. 이전 솔루션은 기업에 정보 검색, 지도 분석, 고객 서비스 등과 관련된 프로세스 자동화 및 응용 프로그램을 효율적으로 수행하고 그에 따른 위험을 완화하도록 돕는다. 구글(Google)의 2017년 논문(Attention is All You Need)에 처음 등장한 트랜스포머는 지금까지 개발된 모델 중 가장 사용되고 강력한 성능을 보여주며, '트랜스포머 AI'가 불리는 혁신적인 형태의 혁신을 주도하고 있다. 여기에 구글은 2020년에 대규모 이미지 인식용 트랜스포머(API) 모델을 발표(Transfomers for Image Recognition at Scale-다량)하기도 했다." The "Output" field shows the generated response: "Transformer is the core of today's popular language models, such as BERT and GPT-3. It was originally developed for natural language processing, but is now being used in various fields, such as computer vision and drug discovery. Transformers can be used to automate compliance processes, improve customer service, and reduce risk. It is a self-attention model that learns the relationship between sequential data, such as words in a sentence, and their meanings. Google's 2017 paper 'Attention is All You Need' introduced the first transformer, which has since become one of the most popular and powerful machine learning models. Google has also developed a transformer for image recognition (API) and announced it in 2020."

Document Embedding(D: 1024)

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information." The "Model" dropdown is set to "cohere.embed-english-light-v2.0". The "Example" dropdown is set to "Add a list of sentences or phrases to generate embeddings (maximum of 96 inputs)". The "Input" field contains a list of sentences: "1. 방탄소년단은 BTS라고도 불리고, HIVE 소속이다.", "2. 트세라핀은 4세대 대표 K-Pop 걸그룹이다.", "3. 로제카림도 HIVE 소속이다.", "4. 제왕은 프세라임 라이더이다.", "5. 수지는 뽕블루스 드라마 이후의 주연 배우다.", "6. 아이유는 여자애들 출연한 가수들 연기하지.", "7. 수지는 이태원에서 전직된 걸그룹 스타를 연기했다." The "Output" field shows the generated response: "Output vector projection: [1, 2, 3, 4, 5, 6, 7]".


Code Generation

The screenshot shows the OCI Generative AI Playground interface. The left sidebar has tabs for Overview, Playground (selected), Dedicated AI clusters, Custom models, Endpoints, and Scope. The main area is titled "Generative AI Playground" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information." The "Model" dropdown is set to "cohere.embed-english-light-v2.0". The "Example" dropdown is set to "Add a list of sentences or phrases to generate embeddings (maximum of 96 inputs)". The "Input" field contains a list of sentences: "1. 방탄소년단은 BTS라고도 불리고, HIVE 소속이다.", "2. 트세라핀은 4세대 대표 K-Pop 걸그룹이다.", "3. 로제카림도 HIVE 소속이다.", "4. 제왕은 프세라임 라이더이다.", "5. 수지는 뽕블루스 드라마 이후의 주연 배우다.", "6. 아이유는 여자애들 출연한 가수들 연기하지.", "7. 수지는 이태원에서 전직된 걸그룹 스타를 연기했다." The "Output" field shows the generated response: "Output vector projection: [1, 2, 3, 4, 5, 6, 7]".



OCI Generative AI 서비스: 전용 모델

Create dedicated AI cluster



Dedicated AI clusters can take a few minutes to create. After a cluster is in an active state, you can use it for fine-tuning or hosting workloads.

Compartment

genaiusers (root)

Name *Optional*

HR_Query

Description *Optional*

Chat agent to provide responses to all the organizational HR questions.

Cluster type

☐ Hosting - beta ☒ Fine-tuning - beta


☒ I agree that I will use this dedicated AI cluster only for beta testing. At the end of my beta testing, I will delete this dedicated AI cluster, or it will be deleted on my behalf if I do not take action.






Create


Cancel

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago) 



Bucket Information

Tags

General

Namespace: axk4z7krhqtx


Compartment: [genaiusers](#)


Created: Thu, Aug 31, 2023, 16:08:56 UTC


ETag: 093eab27-5c55-446e-80e1-0012d028819e


OCID: ...xycuh76q [Show](#) [Copy](#)

Usage

Approximate Object Count: 4 objects 

Approximate Size: 86.42 KiB 

Uncommitted Multipart Uploads Approximate Count: 0 uploads 


Uncommitted Multipart Uploads Approximate Size: 0 bytes 


Features


Default Storage Tier: Standard

Visibility: Private

Encryption Key: Oracle managed key [Assign](#)

Auto-Tiering: ☒ Disabled [Edit](#) 

Emit Object Events: ☒ Disabled [Edit](#) 

Object Versioning: ☒ Disabled [Edit](#) 

Resources

Objects

Metrics

Pre-Authenticated Requests


Work Requests







Lifecycle Policy Rules

Replication Policy

Objects

Upload

More Actions 

<input type="checkbox"/>	Name	Last Modified	Size	Storage Tier	
<input type="checkbox"/>	 small_data_dup_records.jsonl	Thu, Aug 31, 2023, 16:15:01 UTC	21.83 KiB	Standard	
<input type="checkbox"/>	 small_data_dup_records_1.jsonl	Thu, Aug 31, 2023, 16:27:55 UTC	21.83 KiB	Standard	
<input type="checkbox"/>	 small_data_wrongid_records.jsonl	Thu, Aug 31, 2023, 16:45:18 UTC	21.83 KiB	Standard	



OCI Generative AI 서비스: LLM 모델 Fine-Tuning

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Create model

1 Model definition

2 Fine-tuning configuration

3 Data selection

Fine-tuning configuration

Define the model type, dedicated AI cluster type and hyperparameters for this specific model.

1

Models of different categories have different cluster hardware requirements for fine-tuning. The dedicated AI cluster drop-down list is filtered to show clusters that are compatible in size with the requirements of the selected base model.

Base model

cohere.command-light.15.6

Fine-tuning method ⓘ

T-Few

Dedicated AI cluster in **genaiusers (root)** ⓘ [\(Change compartment\)](#)

generativeaidedicatedaicluster20240115214815

Create a new dedicated AI cluster

Advanced options

ⓘ Hide hyperparameters

Total training epochs ⓘ

3

Enter 1 or a higher integer.

Learning rate ⓘ

0.01

Enter a number that's between 0 and 1.0.

Training batch size ⓘ

16

Enter 1 or a higher integer. Higher numbers use more GPU to give better results.

Early stopping patience ⓘ

6

Enter 1 or a higher integer for the grace period of the evaluation cycle before the early stopping stops the training. Enter 0 to disable early stopping.

Early stopping threshold ⓘ

0.01

Enter 1 or a higher integer for the grace period of the evaluation cycle before the early stopping stops the training. Enter 0 to disable early stopping.

Log model metrics interval in steps ⓘ

10

Enter an integer between 1 and the total number of training steps to enable logging of the training loss, and enter 0 to disable logging.

Restore defaults

Previous

Next

Cancel

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Create model

1 Model definition

2 Fine-tuning configuration

3 Data selection

Data selection

1

Ensure that each file contains at least 32 examples. Learn about [training data requirements](#)

Choose bucket in **genaiusers (root)** ⓘ [\(Change compartment\)](#)

afshin-test

Training file

small_data_dup_records.jsonl

If your file is not listed, you might not have "read" permission to the selected bucket.

Data preview

This section shows how the data is ingested for the first few records of the training file. If the data records aren't parsing correctly, check the file format and ensure that the data meets the [requirements](#).

1

Warning: only found 7 examples (25 incorrectly encoded, 0 duplicated), at least 32 valid examples required.

Refresh Data

prompt	completions
I am a salesman at Oracle who wants to start with Oracle's previous featured success stories to find new business opportunities. Use Case: I plan to speak to a prospect in the IT consulting industry. This prospect is a growing IT consultancy firm based in Asia-Pacific, providing tailored solutions to clients across various industries. I am going to speak to their head of operations. Related Oracle Success Story: Inoapps turns to Oracle Cloud ERP and HCM to reimagine business operations. Description: Inoapps is a global IT consultancy and Oracle Platinum Partner that delivers innovative solutions to clients across various industries. Inoapps wanted to transform its own business operations and enhance its employee experience by moving to a cloud-based platform. Inoapps chose Oracle Cloud ERP and Oracle Cloud HCM to replace its legacy systems and streamline its finance, procurement, project management, and human resources processes. With Oracle Cloud, Inoapps has improved its visibility and control over its financial performance, reduced its operational costs and risks, increased its productivity and efficiency, and empowered its employees with self-service capabilities and career development tools. ##	Sales pitch: Subject: Streamline Your IT Consultancy's Operations with Oracle Cloud ERP and HCM Dear [Head of Operations], As a growing IT consultancy firm, the efficiency and effectiveness of your operations play a critical role in your success. I recently learned about Inoapps, a global IT consultancy and Oracle Platinum Partner, that transformed its business operations and enhanced its employee experience using Oracle Cloud ERP and Oracle Cloud HCM. By moving to Oracle Cloud, Inoapps replaced its legacy systems, streamlining finance, procurement, project management, and human resources processes. This resulted in improved visibility and control over financial performance, reduced operational costs and risks, increased productivity and efficiency, and empowered employees with self-service capabilities and career development tools. I believe that your IT consultancy firm can benefit from implementing Oracle Cloud ERP and HCM, just as Inoapps did. I would like to propose a meeting to discuss how these cloud-based solutions can help you optimize your operations and create a better working environment for your employees. Please let me know when you are available for a conversation, and I will arrange a meeting with our product specialists to provide more insights into the benefits of Oracle Cloud ERP and HCM for your organization. Looking forward to discussing this opportunity with you. Best regards, [Your Name] [Your Title] [Your Contact Information] END
I am a salesman at Oracle who wants to start with Oracle's previous featured success stories to find new business opportunities. Use Case: I plan to speak to a prospect in the food and beverage industry. This prospect is a rapidly expanding restaurant chain in Latin America with ambitious growth plans. I am going to speak to their head of IT. Related Oracle Success Story: Oracle Cloud provides key ingredients to sustaining growth at Madero Restaurants. Description: Madero Restaurants is a leading casual dining chain in Brazil, with more than 200 locations and 15,000 employees. Madero wanted to support its rapid growth and expansion plans by modernizing its IT infrastructure and applications. Madero selected Oracle Cloud Infrastructure (OCI), Oracle Cloud ERP, Oracle Cloud HCM, and Oracle Cloud CX to create a unified and scalable cloud platform. With Oracle Cloud, Madero has improved its operational efficiency and agility.	Sales pitch: Subject: Scale Your Restaurant Chain with Oracle Cloud: Enhanced Efficiency, Customer Experience, and Workforce Enablement Dear [Head of IT], I understand that your restaurant chain is experiencing rapid growth in Latin America, and I wanted to share a success story from Madero Restaurants, a leading casual dining chain in Brazil. Madero modernized its IT infrastructure and applications with Oracle Cloud to support its expansion plans and improve its business operations. By implementing Oracle Cloud Infrastructure (OCI), Oracle Cloud ERP, Oracle Cloud HCM, and Oracle Cloud CX, Madero created a unified and scalable cloud platform that improved operational efficiency and agility, optimized supply chain and inventory management, enhanced customer service and loyalty, and enabled its workforce with mobile and collaborative tools. I believe that your restaurant chain can achieve similar success by leveraging Oracle Cloud solutions. I would like to propose a meeting to discuss how these innovative technologies can help you streamline your operations.

Previous

Submit

Cancel

14 Oracle Cloud Summit | Copyright © 2024, Oracle and/or its affiliates

OCI Generative AI 서비스: LLM 모델 Fine-Tuning

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Overview

Endpoints

Acme_Company_HR_QA

View in playground

Add tags

Edit

Delete

Endpoint Information

Tags

Compartment OCID: ...jm3kucc7lotk2oqep47q

Model name: generativeaimodel20230911152851

OCID: ...ktexjey43qu2ficnd2wa

Model version: v20230911152851

Description:

Dedicated AI Cluster: ocid1.generativeaidedicatedcluster.oc1.us-chicago-1.amaaaaaacqy6p4qastho3j5jnhvxpqnv2b6hrcwzplwua2xguz7o7kqma

State: Creating

Lifecycle details:

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Edit

Add tags

Delete

Custom models can take several hours to fine-tune and become active. After a custom model is in an active state, you must create an endpoint for that model to host it for inference.

Model details

Fine-tuning configuration

Tags

Compartment: ...tk2oqep47q

Base model: ocid1.generativeaimodel.oc1.us-chicago-1.amaaaaapi24rzaalcbnpexqkxpwfdqskcupap27qtr7ew5ierjpbih2a

OCID: ...yqdp6i6zq

Capability: Text generation

Description: -

Category: Small

State: Active

Dedicated AI cluster: ocid1.generativeaidedicatedcluster.oc1.us-chicago-1.amaaaaaacqy6p4qastho3j5jnhvxpqnv2b6hrcwzplwua2xguz7o7kqma

Lifecycle details: Created Model

Dataset: sales_pitch_generation_train.jsonl

Created: Fri, 15 Sep 2023 17:42:24 UTC

Model performance

Accuracy

0.73396

Loss

1.09664

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Generative AI

Generative AI Playground

Overview

Playground

Dedicated AI clusters

Custom models

Endpoints

Scope

Compartment

genaiusers (root)

Model

cohere.command

Example

Generate an email

View code

As a corporate vice president, generate an email congratulating a team that has just shipped a new cloud service. Emphasize the great positive impact the new service will have on the productivity of their customers.

Hi Team,

I am so proud to announce that we have just shipped our new cloud service! This has been a huge undertaking and you all have worked so hard to make it happen. I want to thank you from the bottom of my heart for your dedication and hard work.

This new cloud service is going to have a huge positive impact on the productivity of our customers. It will allow them to work from anywhere in the world and access their files and data from any device. This is going to be a game changer for our company and for our customers.

Thank you again for all your hard work and dedication. I am so proud to be a part of such a great team.

Best,

[Your Name]

Parameters

Maximum output tokens

600

Temperature

0.5

Top p

0.75

Top k

0

Stop sequences

Enter sequence and press enter

Frequency penalty

1

Presence penalty

0

Show likelihoods

-15

0

Generate

Regenerate

Clear

Character count - 876 | Token limit - 4000

Terms of Use and Privacy

Cookie Preferences

Copyright © 2023, Oracle and/or its affiliates. All rights reserved.



OCI Generative AI 서비스: LLM 모델 관리

ORACLE Cloud

Search resources, services, documentation, and Marketplace

Generative AI

Overview

Playground

Dedicated AI clusters

Custom models

Endpoints

Scope

Compartment

genaiusers (root)

Generative AI overview

Power your apps with large language models and generative AI

OCI Generative AI is a fully managed service that provides a set of state-of-the-art, customizable LLMs that cover a wide range of use cases for text generation. Use the playground to try out the models out-of-the-box or create and host your own fine-tuned custom models based on your own data on dedicated AI clusters.

Metrics in genaiusers (root) Compartment

Dedicated AI clusters

7

Custom models

6

Endpoints

10

Get started

Playground

The playground is a visual interface for exploring the hosted pretrained and custom models without writing a single line of code. Use the playground to test your use cases and refine prompts and parameters. When you're happy with the results, you can view the code and integrate Generative AI into your applications.

Go to playground

Dedicated AI clusters

Spin up dedicated hardware units for fine-tuning custom models and hosting them.

Custom models

Create custom models by fine-tuning the base models with your own dataset.

Endpoints

Create and manage endpoints to host your custom models.

ORACLE Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

custom-cohere-light-endpoint

View in playground

Edit

Add tags

Move endpoint

Delete

Endpoint Information

Tags

Compartment OCID: ...jrn3kuc70k2oqep47q Show Copy

OCID: ...yluztfo7behwtoria Show Copy

Description:

State: Active

Lifecycle details: Created Endpoint

Dedicated AI Cluster: ooid1.generativeai.dedicatedcluster.oc1.us-chicago-1.amaaaaaaacyf6pkaentkars2ch3k2c4qabxtqum2457v2vf44c3a

Model name: cohere.command-light

Model version: 15.6

Created on: Mon, 15 Jan 2024 21:50:41 UTC

Created by: laleh.haghshenas@oracle.com

Content moderation: Enabled

Resources

Endpoint metrics

Work requests

Endpoint metrics

Start time: Jan 15, 2024 22:33:43 UTC

End time: Jan 15, 2024 23:33:43 UTC

Quick selects: Last hour

Reset charts

Total processing time

Interval: Auto

Statistic: Mean

No data for this time range.

Number of calls

Interval: Auto

Statistic: Count

No data for this time range.

Service Errors Count

Interval: Auto

Statistic: Count

No data for this time range.

Client Errors Count

Interval: Auto

Statistic: Count

No data for this time range.

Total number of input

Interval: Auto

Statistic: Sum

No data for this time range.

Total number of output

Interval: Auto

Statistic: Sum

No data for this time range.

16 Oracle Cloud Summit | Copyright © 2024, Oracle and/or its affiliates

OCI Generative AI 서비스

SOTA Open-Source & Commercial Model 지원 서비스

GA of OCI Generative AI
2024.01.23

The screenshot shows the OCI Generative AI overview page. The header includes the Oracle Cloud logo, a search bar, and the region 'US West (Phoenix)'. The left sidebar has a 'Generative AI' section with links to Overview, Playground, Dedicated AI clusters, Custom models, and Endpoints. The main content area is titled 'Generative AI overview' and includes a description of the service, a 'Watch service tour' button, and a 'Metrics in my compartment' section with three cards: 'Dedicated AI clusters' (7), 'Custom models' (3), and 'Active endpoints' (12). Below this is a 'Get started' section with a 'Go to playground' button and three cards: 'Playground', 'Dedicated AI clusters', and 'Custom models'. A 'Resources' section on the right lists links for documentation, API reference, workshops, tutorials, and pricing.



+



End-2-End Managed LLM Service

Oracle Enterprise LLM Platform

고려사항



**Designed for
enterprise**



Tailored to your data

OCI Generative AI 서비스를 통해 자신의 데이터를 활용해 모델을 개선할 수 있습니다.



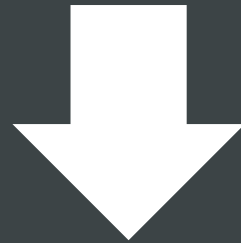
**Security and
privacy**



**Predictable performance
and pricing**

Oracle Enterprise LLM Platform

기업용 LLM의 도전 과제

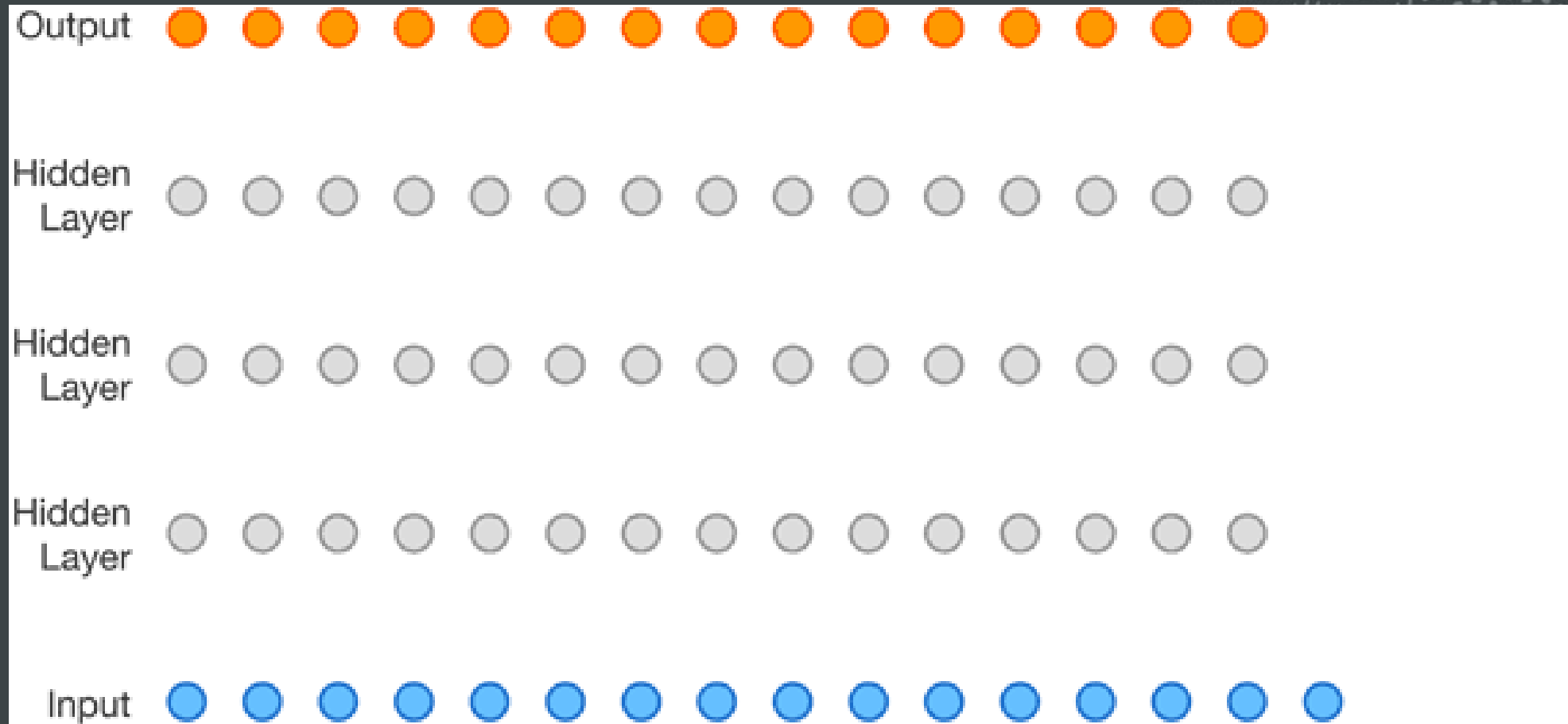


Prompt Engineering Based Solution



RAG Approach

Auto-Regressive LLM 모델의 한계: Hallucination, 신뢰성 이슈



Reference: <https://deepmind.google/>

RAG(Retrieval Augmented Generation)

Open Domain Question 해결 기법 (2021.04)

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡} Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com

Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.



Patrick Lewis

Natural Language Processing
Research Scientist

co:here

UCL NLP

Meta AI Research



Natural Language Patrick

I'm a London-based AI and Natural Language Processing Research Scientist, working at [co:here](#). Before this, I was a research scientist at the Fundamental AI Research Lab (FAIR) at Meta AI. I completed my PhD working with Sebastian Riedel and Pontus Stenetorp, splitting my time between FAIR and University College London.

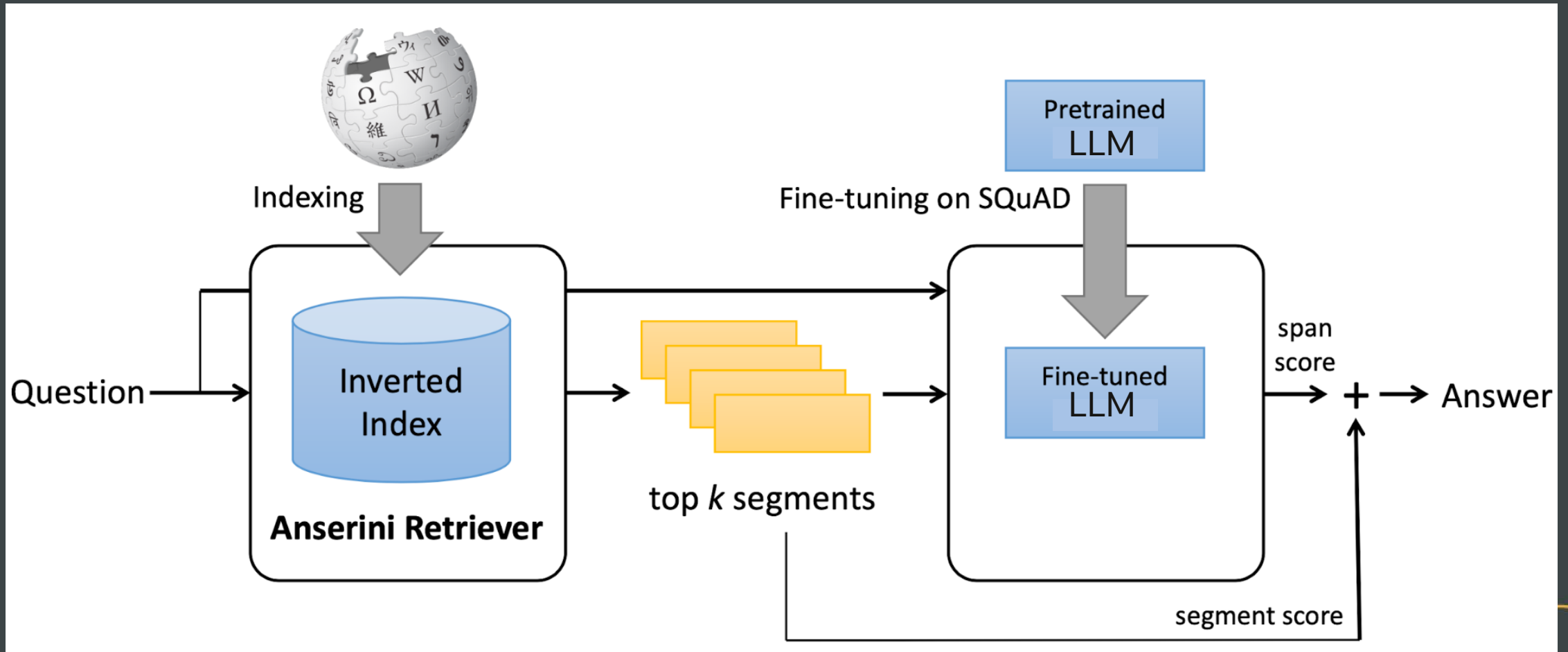
I work at the intersection of information retrieval techniques (IR) and large language models (LLMs), and have done a lot of work on *Retrieval-Augmented Language Models*.

I'm interested in how to represent, store and retrieve *knowledge* for use in large language models. My work focuses on building more powerful, efficient, robust and update-able models that can perform well on a wide range of NLP tasks, but also excel on knowledge-intensive NLP tasks such as Question Answering and Fact Checking.

I also occasionally tweet (🐦) [here](#) and toot (🐙) [here](#)

RAG(Retrieval Augmented Generation)

Open Domain Question 해결 기법 (2021.04)



인간 커뮤니케이션

W5H1, 육하원칙



Who	누가
When	언제
Where	어디서
Prompt	
How	어떻게
What	무엇을
Why	왜



LLM과 커뮤니케이션

Prompt Engineering



Role 역할

Instruction 지시

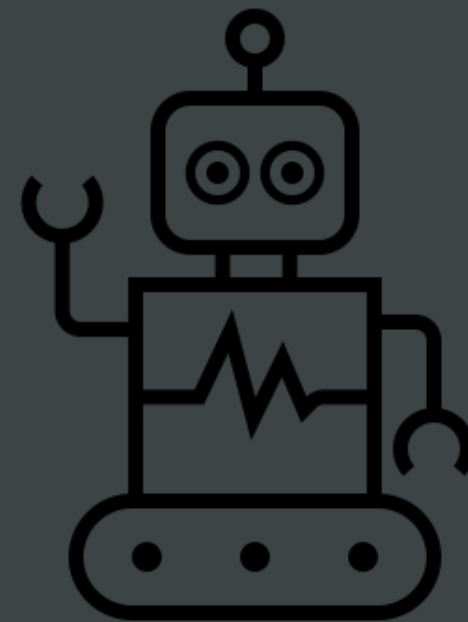
Context 문맥

Prompt

Question 질문

Few Shot 예시

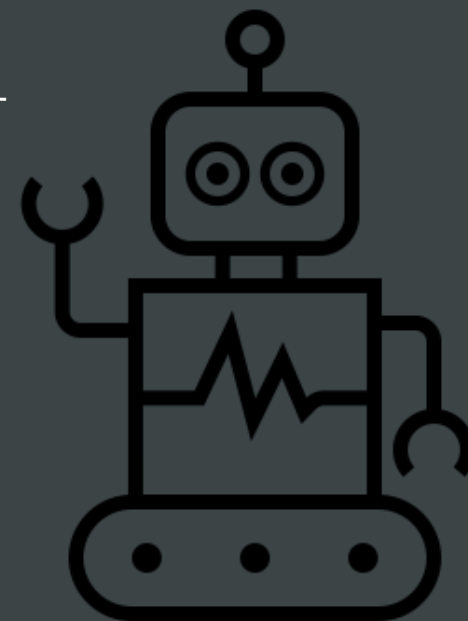
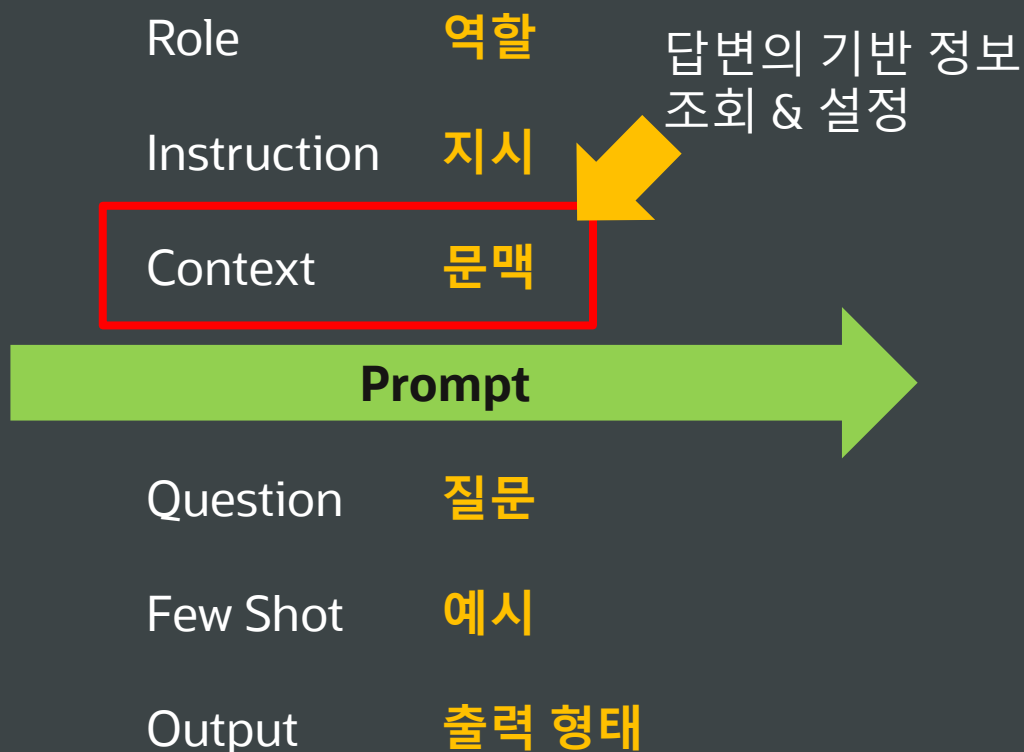
Output 출력 형태



Generative AI (LLM,
Large Language Model)

LLM과 커뮤니케이션

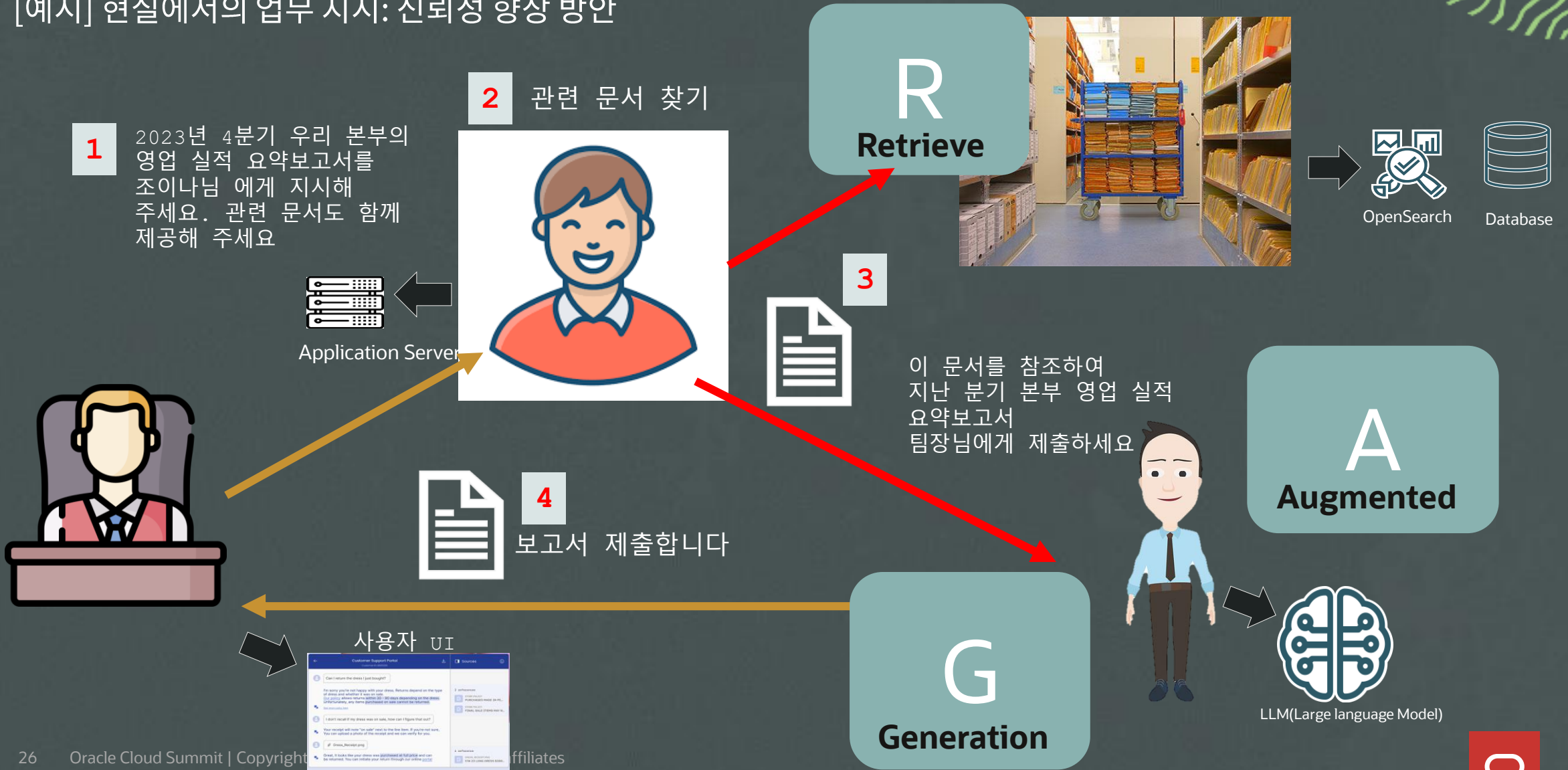
Prompt Engineering



Generative AI (LLM,
Large Language Model)

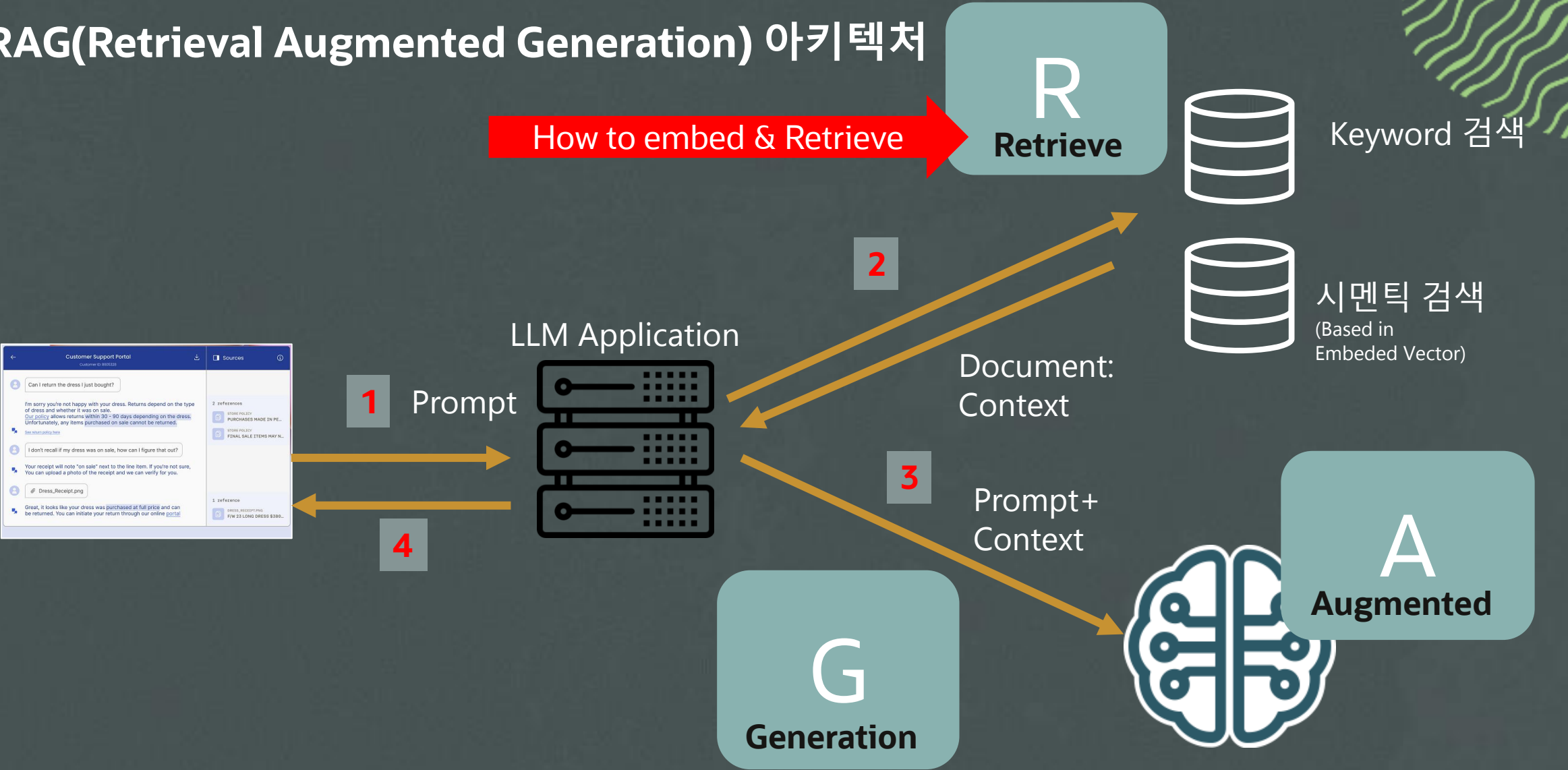
RAG (Retrieval Augmented Generation)

[예시] 현실에서의 업무 지시: 신뢰성 향상 방안



RAG(Retrieval Augmented Generation) 아키텍처

How to embed & Retrieve



RAG(Retrieval Augmented Generation) 아키텍처

Embedding model의 성능 지표 (Huggingface Leaderboard)

생성
✓ cohere.command v15.6
cohere.command-light v15.6
meta.llama-2-70b-chat

요약
cohere.command v15.6

포함
cohere.embed-english-light-v2.0
cohere.embed-english-light-v3.0
cohere.embed-english-v3.0
cohere.embed-multilingual-light-v3.0
cohere.embed-multilingual-v3.0

Overall	Bitext Mining	Classification	Clustering	Pair Classification	Reranking	Retrieval	STS	Summarization				
English	Chinese	Polish										
Overall MTEB English leaderboard												
Metric: Various, refer to task tabs												
Languages: English												
Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)	STS Average (10 datasets)	Average (10 datasets)
1	cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55	82.62	30.18
2	bge-large-en-v1.5	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29	83.11	31.61
3	cohere-embed-multilingual-v3.0		1024	512	64.01	76.01	46.6	86.15	57.86	53.84	83.15	30.99
4	bge-base-en-v1.5	0.44	768	512	63.55	75.53	45.77	86.55	58.86	53.25	82.4	31.07

Massive Text Embedding Benchmark (MTEB) Leaderboard: 2023.11.15 기준



RAG(Retrieval Augmented Generation) 아키텍처

VectorDBs in OCI



OpenSearch



Oracle 23C



MySQL Heatwave



PostgreSQL



OCI Cache
: Managed Redis

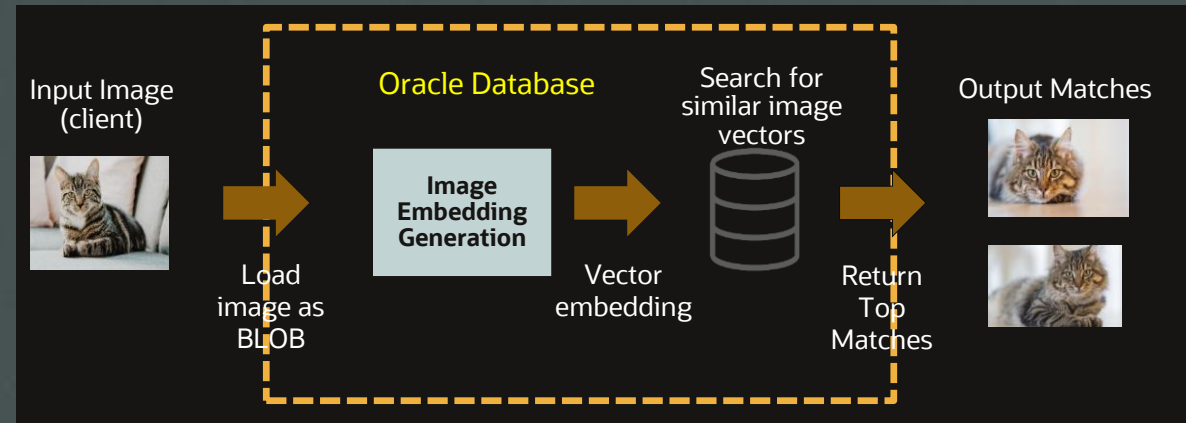
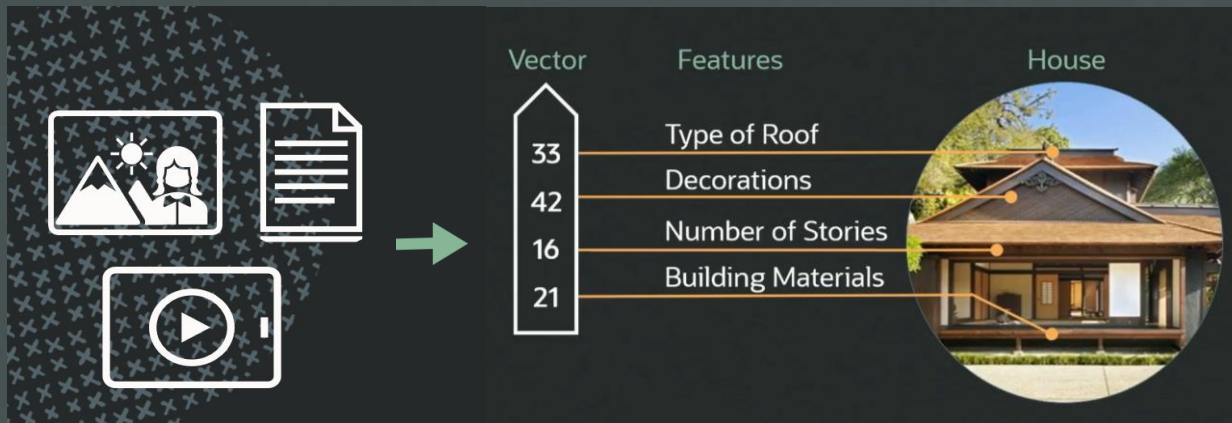
RAG 지원을 위한 Oracle DB의 전략 : Vector Type in DB 23c

Vector Search: 오라클 DB 내에서 Vector Type 제공

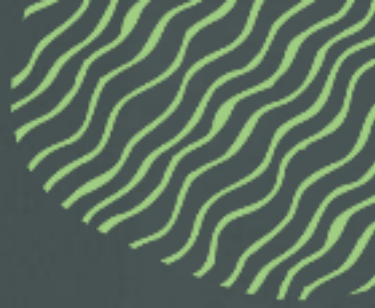
- 23c에서 Vector Type 지원
- AI vector는 비정형 데이터 자체가 아니라 semantic content를 나타내는 특정 feature들을 담고 있는 데이터
- 기존 테이블 구조에 vector type을 추가하여 keyword + semantic search 수행 가능
- 기존의 RAC/Exadata/Partition/Sharding과 결합되어 대규모의 Vector DB에 대해 성능/안정성 제공

Oracle DB 내에서의 Vector DB 구성의 장점

- 고객들은 이미 오라클 DB 내에 대규모의 **operational/analytical enterprise data**를 운용 중
- 별도의 Vector DB를 구성하는 것보다 이미 운영 중인 오라클 DB 안에 vector embedding을 추가하는 것이 (1) 구현의 용이성, (2) 비용 절감, (3) 성능/안정성 제공 측면에서 효율적



RAG 지원을 위한 Oracle DB의 전략 : Vector Type in DB 23c



```
create table galaxies (  
  id number,  
  name varchar2(50),  
  doc varchar2(500),  
  embedding vector  
);
```

```
insert into galaxies values (1, 'M31',  
  'Messier 31 is a barred spiral galaxy in  
  the Andromeda constellation which has a lot  
  of barred spiral galaxies.',  
  '[0,2,2,0,0]');
```

```
insert into galaxies values (2, 'M33',  
  'Messier 33 is a spiral galaxy in the  
  Triangulum constellation.',  
  '[0,0,1,0,0]');
```

```
insert into galaxies values (3, 'M58',  
  'Messier 58 is an intermediate barred  
  spiral galaxy in the Virgo constellation.',  
  '[1,1,1,0,0]');
```

RAG 지원을 위한 Oracle DB의 전략 : Vector Type in DB 23c

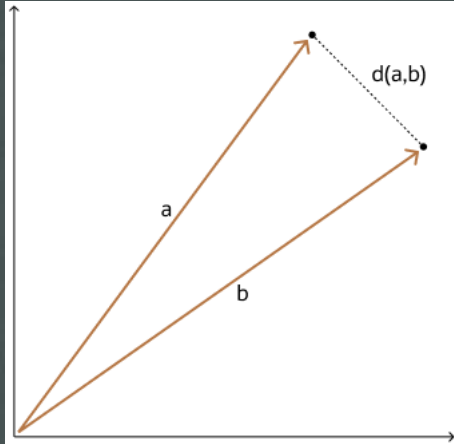
```
create table galaxies (  
    id number,  
    name varchar2(50),  
    doc varchar2(500),  
    embedding vector  
);
```

```
insert into galaxies values (1, 'M31',  
    'Messier 31 is a barred spiral galaxy in  
    the Andromeda constellation which has a lot  
    of barred spiral galaxies.',  
    '[0,2,2,0,0]');
```

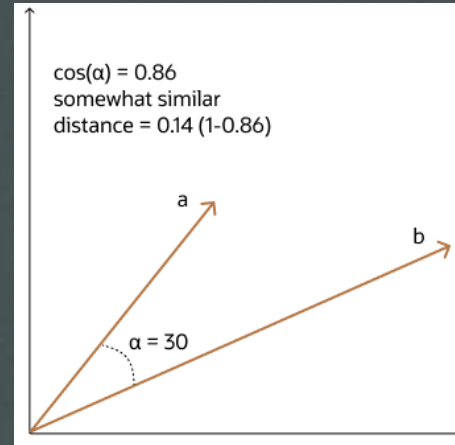
```
insert into galaxies values (2, 'M33',  
    'Messier 33 is a spiral galaxy in the  
    Triangulum constellation.',  
    '[0,0,1,0,0]');
```

```
insert into galaxies values (3, 'M58',  
    'Messier 58 is an intermediate barred  
    spiral galaxy in the Virgo constellation.',  
    '[1,1,1,0,0]');
```

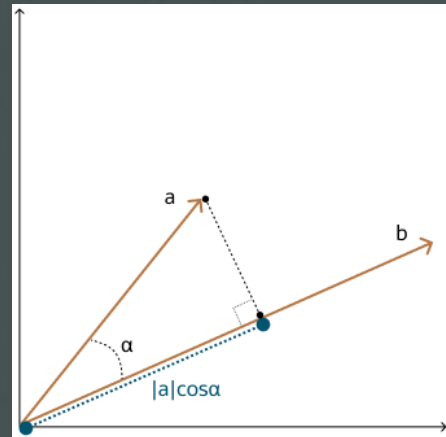
RAG 지원을 위한 Oracle DB의 전략 : Vector 유사도 연산



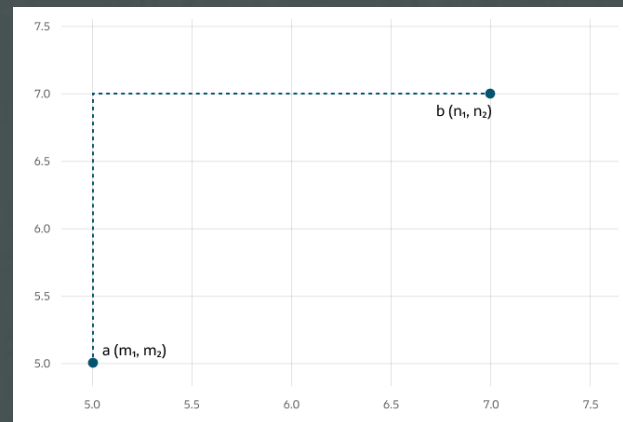
Euclidean Distance



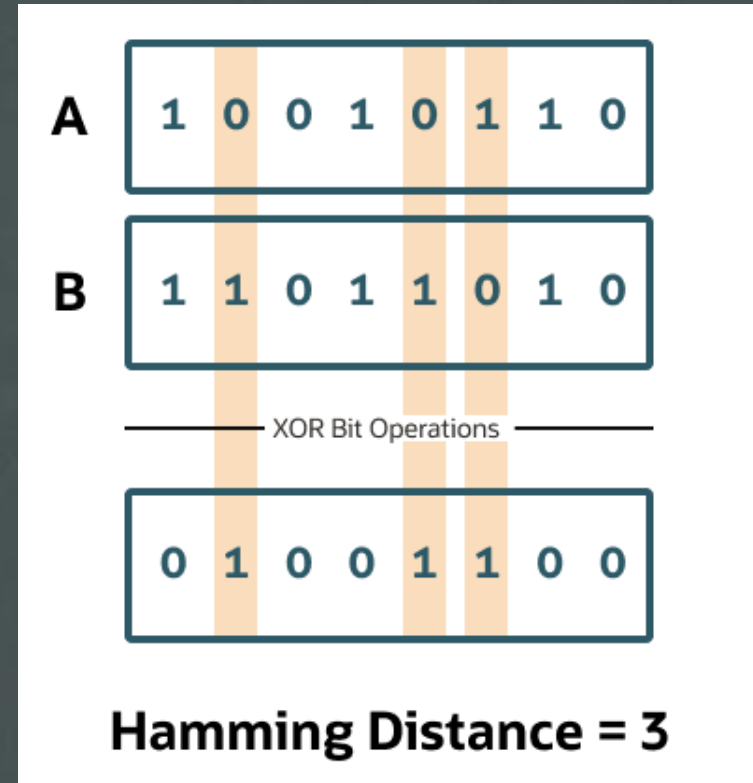
Cosine Similarity



Dot Product Similarity



Manhattan Distance



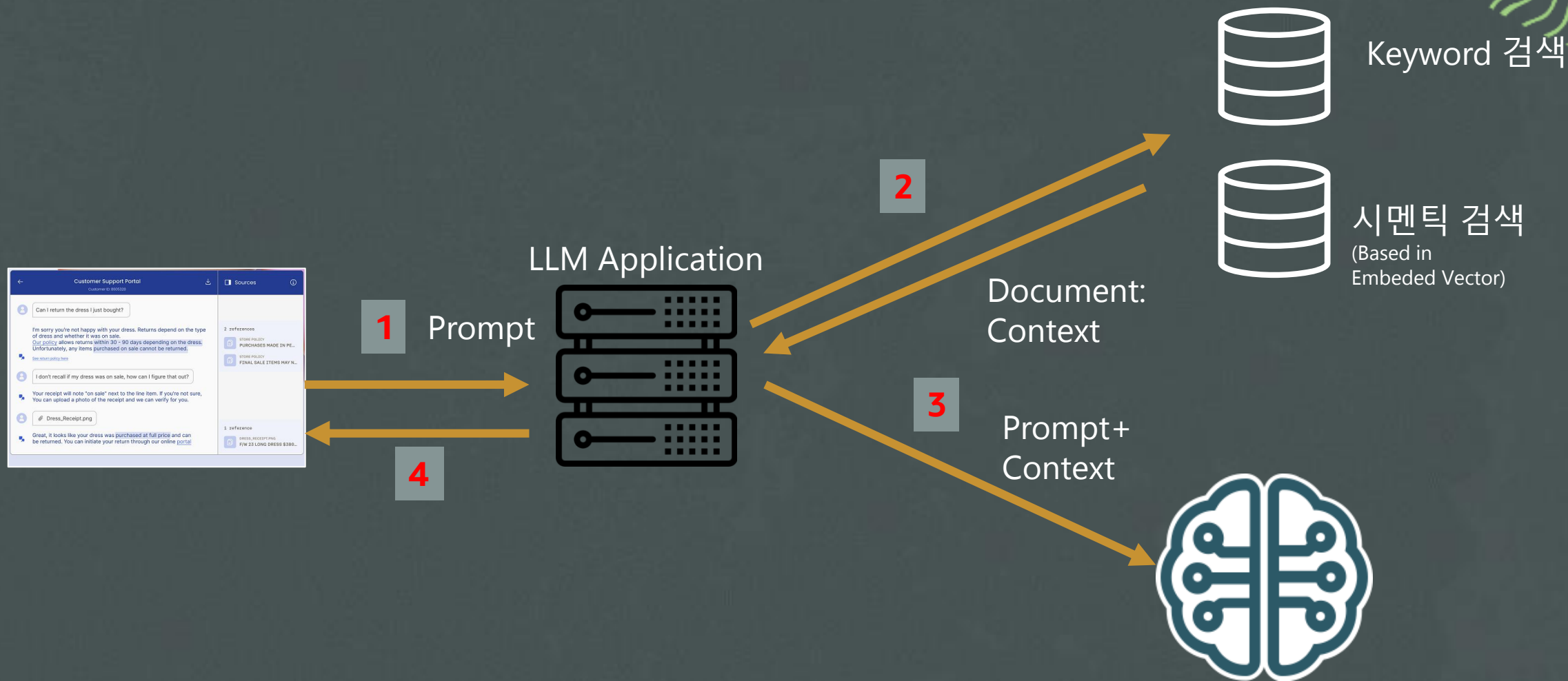
Hamming Distance

RAG 지원을 위한 Oracle DB의 전략 : Vector 유사도 쿼리

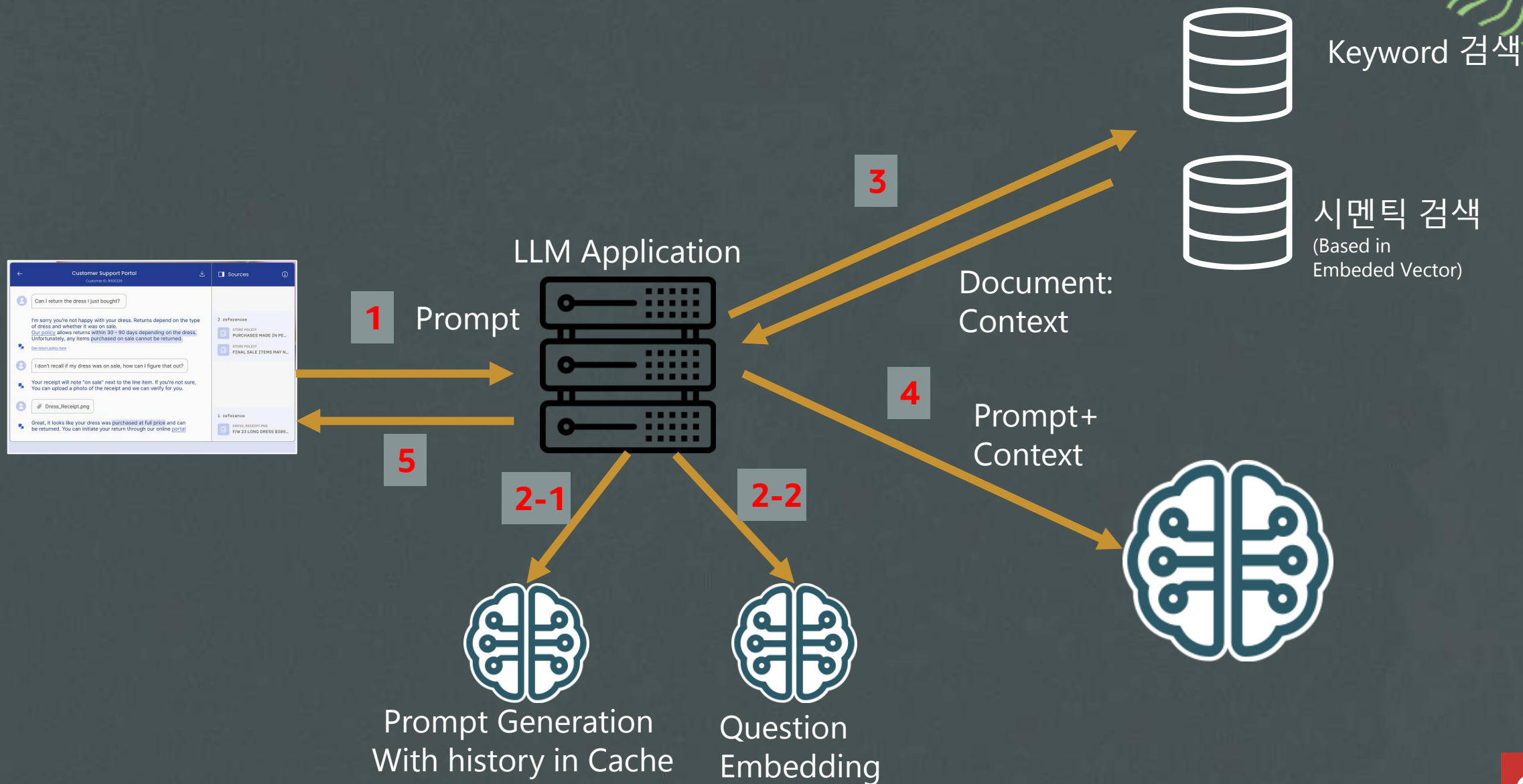
```
SELECT name,  
       VECTOR_DISTANCE( embedding,  
                        (SELECT embedding FROM galaxies WHERE name = 'NGC1073'),  
                        COSINE ) as vector_similarity  
FROM galaxies  
WHERE name <> 'NGC1073'  
ORDER BY VECTOR_DISTANCE( embedding,  
                          (SELECT embedding FROM galaxies WHERE name = 'NGC1073'),  
                          COSINE )  
FETCH FIRST 3 ROWS ONLY;
```

NAME	VECTOR_SIMILARITY
M31	0.000000000000000002220446049250313
M91	0.000000000000000002220446049250313
M77	0.000000000000000002220446049250313

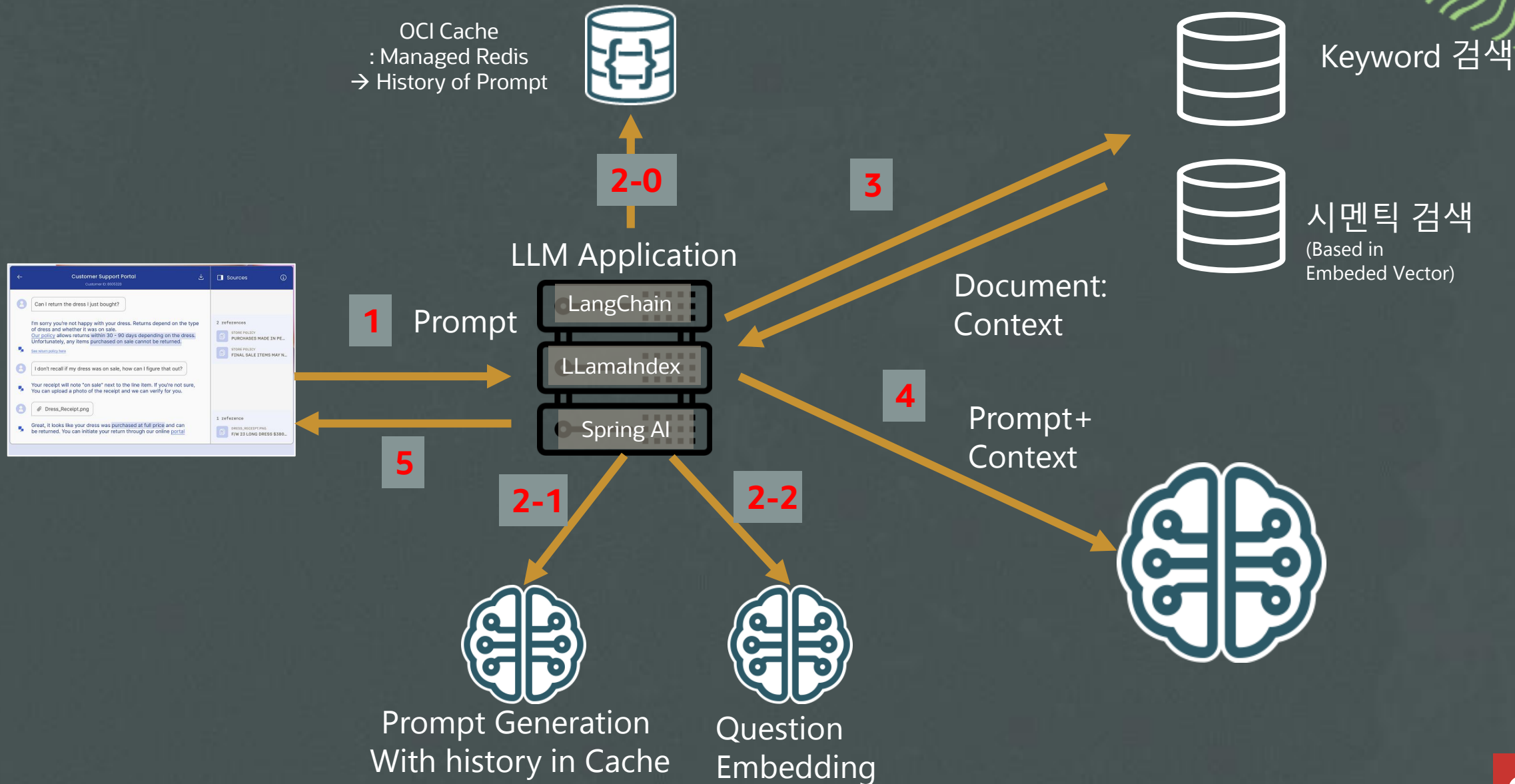
RAG(Retrieval Augmented Generation) 아키텍처



RAG(Retrieval Augmented Generation) 아키텍처



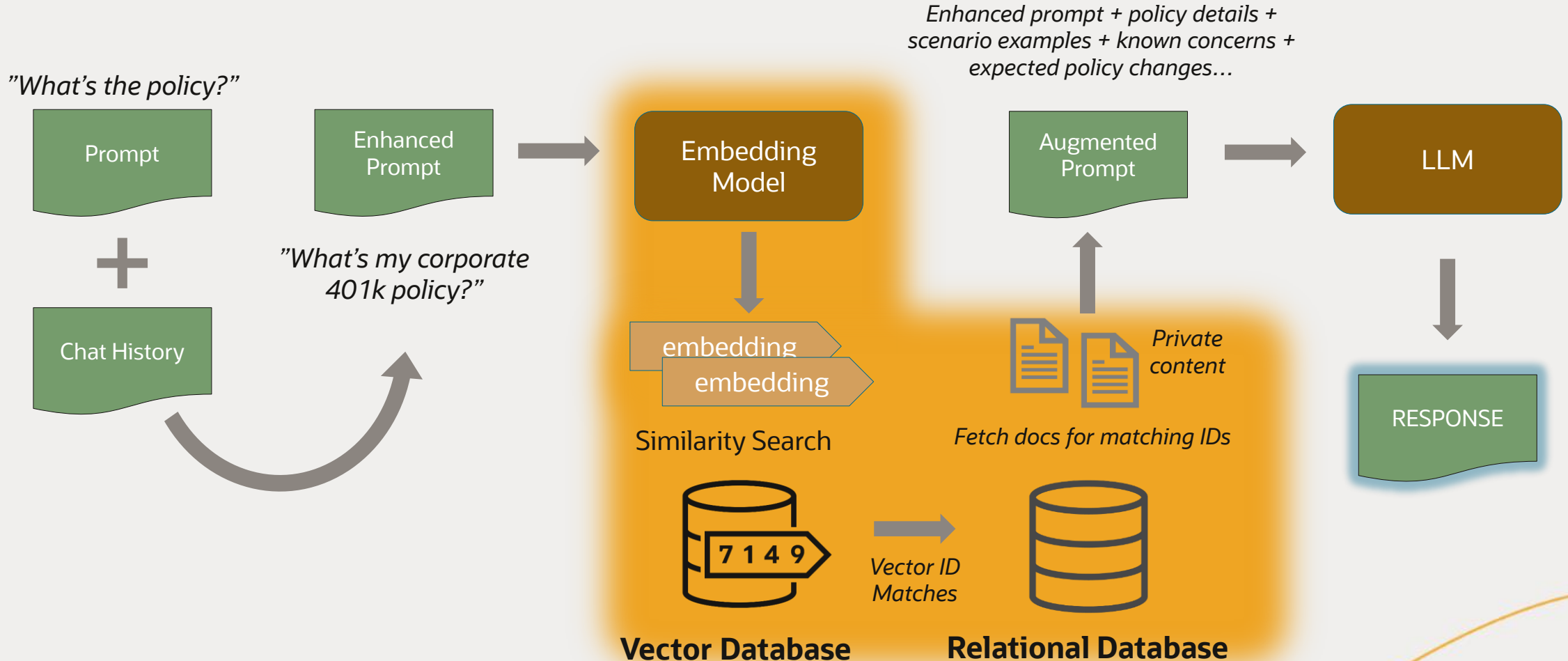
RAG(Retrieval Augmented Generation) 아키텍처



RAG(Retrieval Augmented Generation)

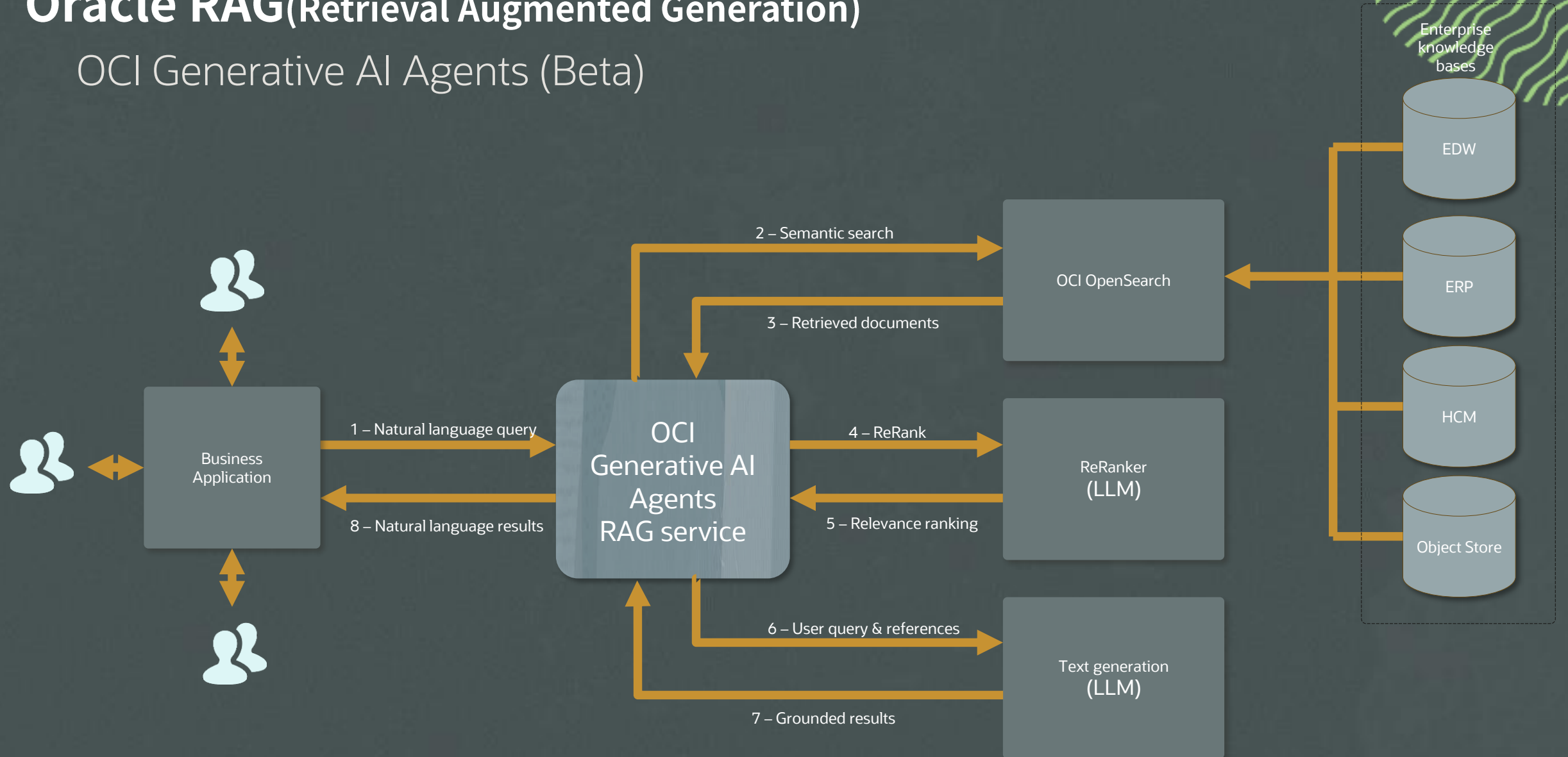
기업 지식/데이터를 활용하여 hallucination 없는 Q&A 지원

- ✓ Enterprise-level Gen AI Application 개발을 위해서는 RAG 지원은 필수
- ✓ 오라클은 Oracle DB 내에서 Vector DB의 구성 지원

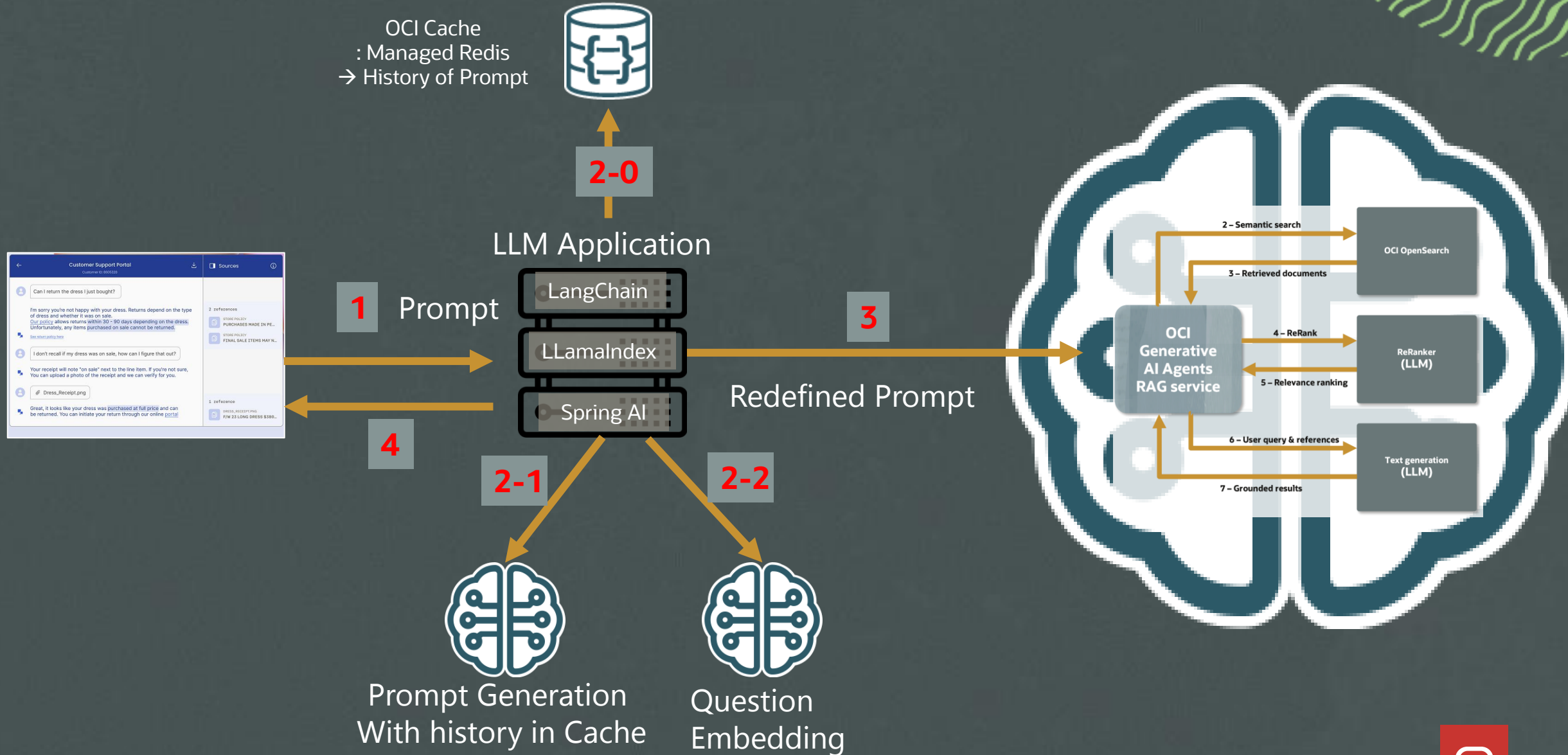


Oracle RAG(Retrieval Augmented Generation)

OCI Generative AI Agents (Beta)



RAG(Retrieval Augmented Generation) 아키텍처



OCI Generative AI Agent



RAG with OpenSearch in Beta Jan 2024

정보 수집

OCI Gen AI Agent 현재 베타 상태이며 데이터 저장소로 OCI OpenSearch 지원.

에이전트 작동 방식

OCI Gen AI Agent는 관련 정보를 조회하고 프롬프트의 컨텍스트에 설정하는 일련의 작업을 추상화 및 자동화.

고성능 생성형 모델과 통합

OCI Gen AI Agent는 OCI Generative AI 서비스 모델(Cohere 및 Llama-2)을 기본 지원.

Reasoning & Planning

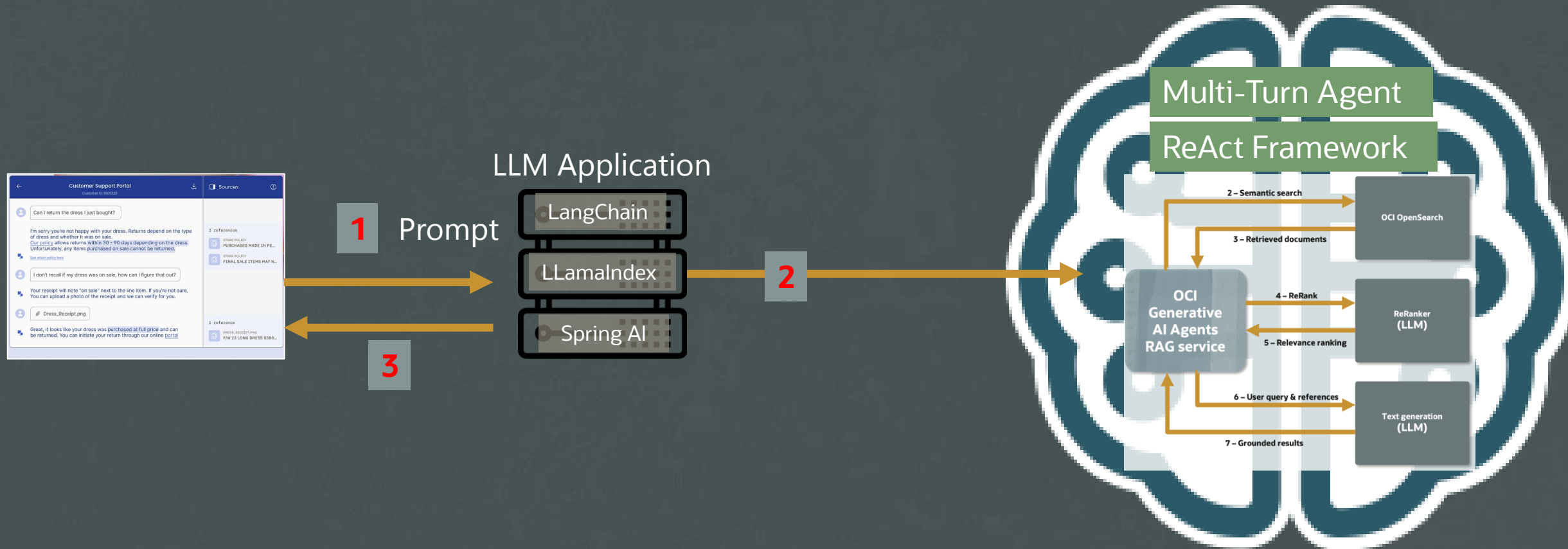
OCI Gen AI Agent는 ReAct framework을 통해서 일련의 생각, 행동, 관찰을 바탕으로 추론, 계획 및 동작

멀티-턴 에이전트

과거 요청 기록을 유지, 모델 컨텍스트와 응답을 재정의하여 명확성 향상

The screenshot shows the Oracle Cloud console for Generative AI Agents. The top navigation bar includes the Oracle Cloud logo, a 'Cloud Classic >' link, and a search bar. The main content area is titled 'Generative AI Agents' and has a sidebar with links for 'Overview', 'Agent', 'Data source', and 'Private endpoints'. The 'Overview' section contains a description of the service, a note about the beta release supporting RAG, and a 'Get started' section with two steps: 'Step 1: Create Agent' and 'Step 2: Create Data source (optional)'. A 'Get Started' button is visible at the bottom of the 'Get started' section.

RAG(Retrieval Augmented Generation) 아키텍처



OCI Generative AI Agent

RAG with OpenSearch in Beta Jan 2024

정보 수집

OCI Gen AI Agent 현재 베타 상태이며 데이터 저장소로 OCI OpenSearch 지원.

2024 상반기 Oracle Database 23c AI Vector Search와 MySQL HeatWave Vector Store 지원 예정

에이전트 작동 방식

OCI Gen AI Agent는 관련 정보를 조회하고 프롬프트의 컨텍스트에 설정하는 일련의 작업을 추상화 및 자동화.

고성능 생성형 모델과 통합

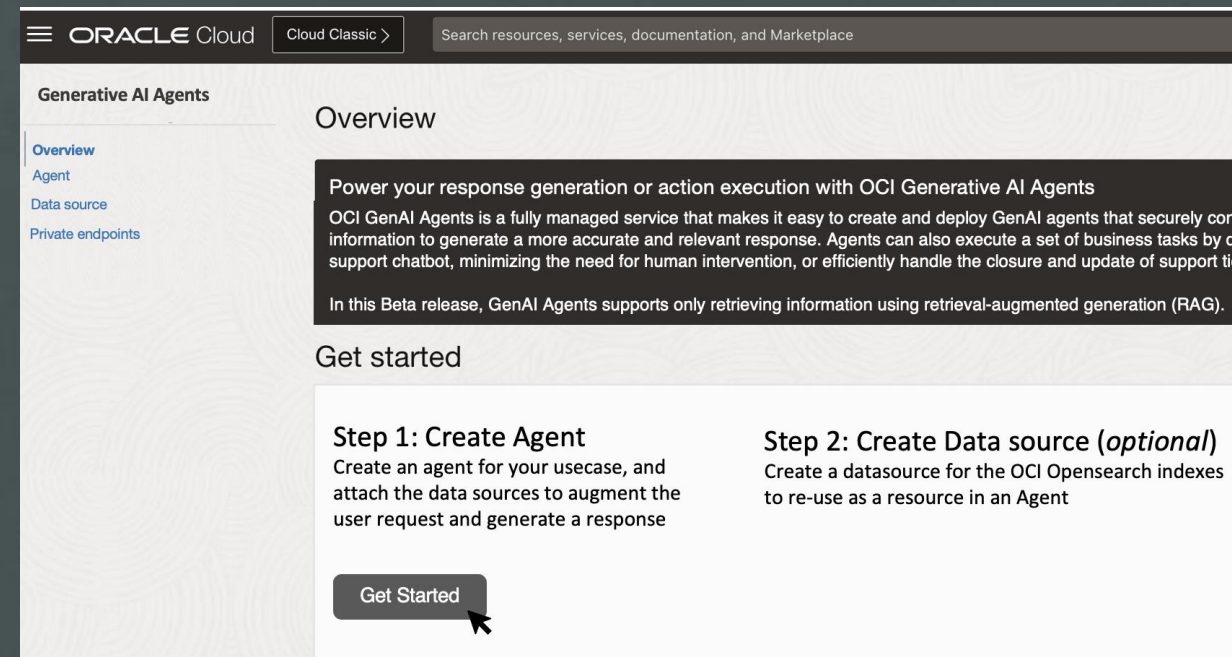
OCI Gen AI Agent는 OCI Generative AI 서비스 모델(Cohere 및 Llama-2)을 기본 지원.

Reasoning & Planning

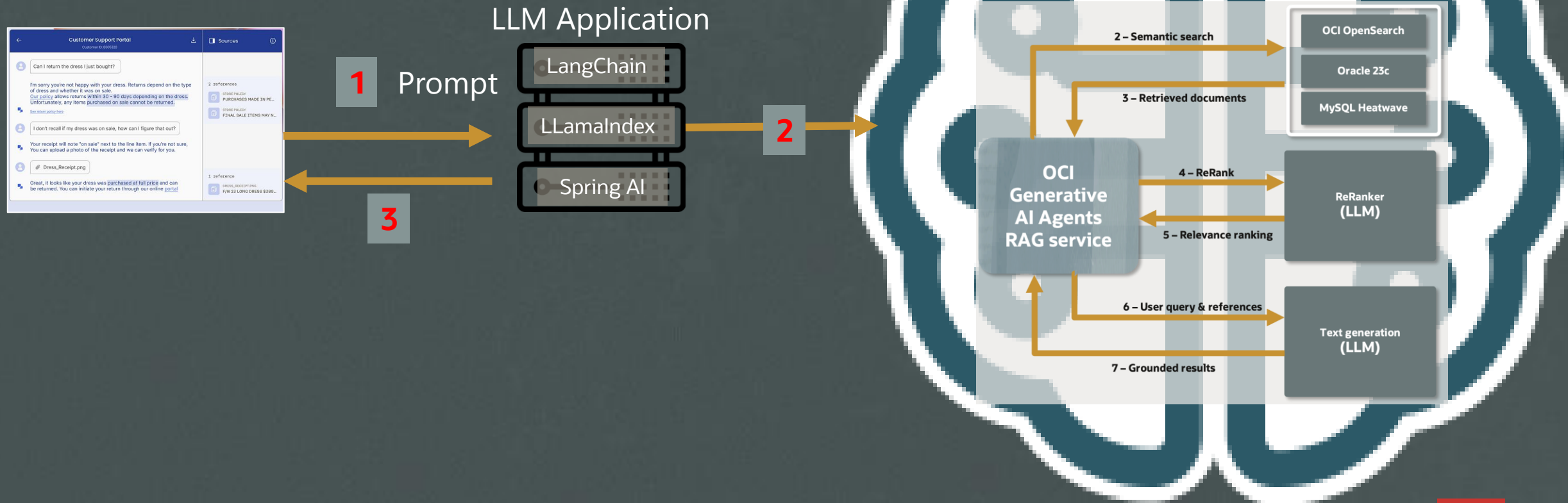
OCI Gen AI Agent는 ReAct framework을 통해서 일련의 생각, 행동, 관찰을 바탕으로 추론, 계획 및 동작

멀티-턴 에이전트

과거 요청 기록을 유지, 모델 컨텍스트와 응답을 재정의하여 명확성 향상



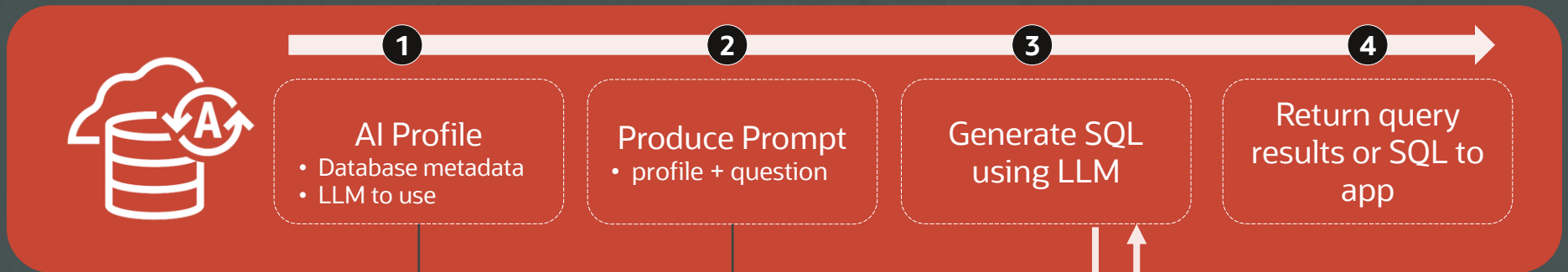
RAG(Retrieval Augmented Generation) 아키텍처



LLM을 이용한 SQL 생성 지원(23c)

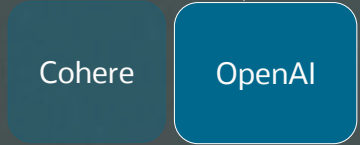


SELECT AI <ACTION>
Give me the average
salary of employees in
each department;



dbms_cloud_ai.create_profile

dbms_cloud_ai.set_profile



ACTION

runsql	return sql result set
narrate	return a conversational result
showsql	return the generated query
chat	general AI-chat conversation



LLM을 이용한 SQL 생성 지원(23c)

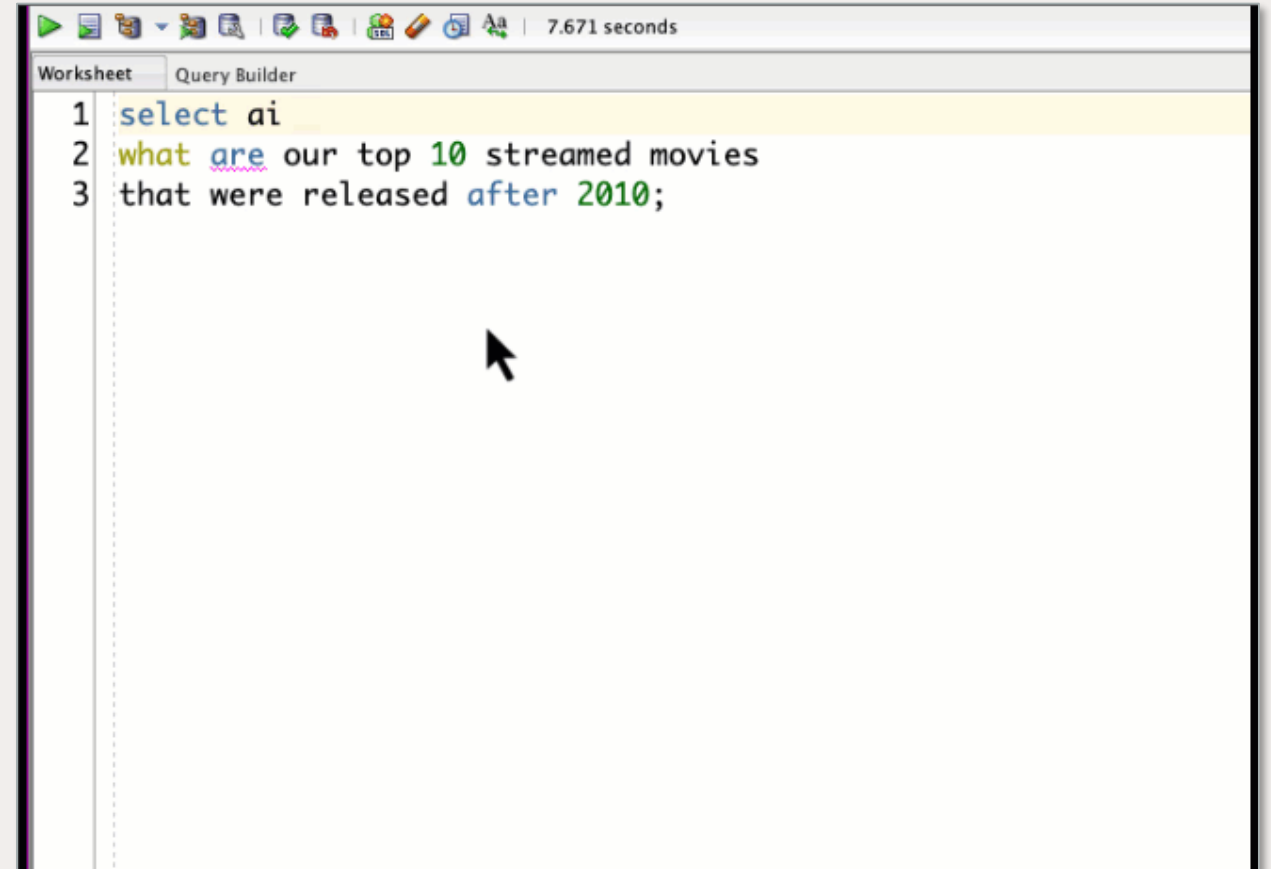
새로운 DBMS_CLOUD_AI PL/SQL 패키지를 사용하여 사용 가능

키워드 AI 및 질문과 함께 표준 SELECT 문을 사용합니다.

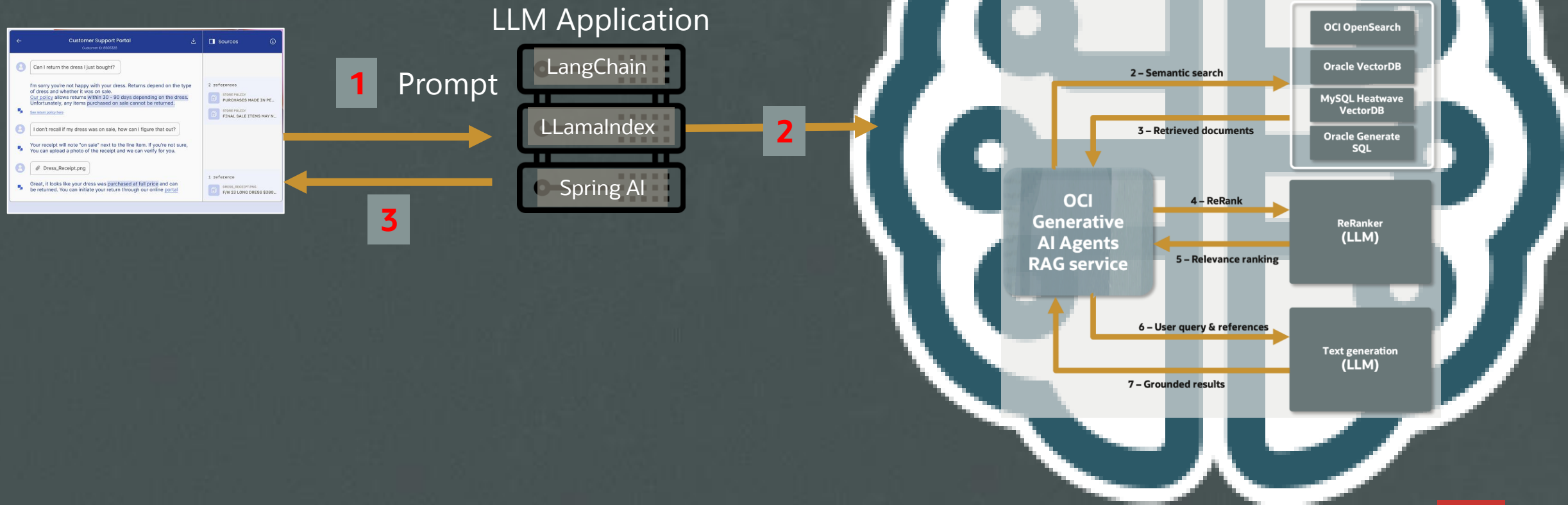
다른 SQL 결과 집합과 마찬가지로 결과를 처리합니다.

Actions

- **runsql** - SQL 결과 집합 반환(기본값)
- **showsql** – 생성된 쿼리문 반환
- **narrate** – 대화 결과 반환
- **Chat** – 일반 AI 챗



RAG(Retrieval Augmented Generation) 아키텍처



Summary: Oracle Generative AI & RAG 아키텍처



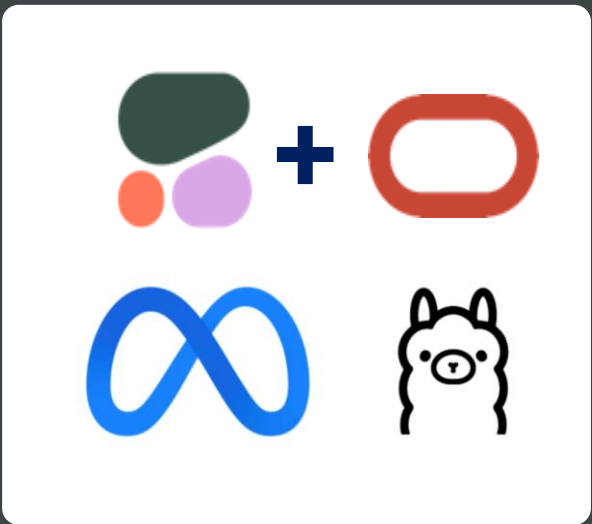
Model



DATA



INFRASTRUCTURE



Oracle
Database 23C



OpenSearch



Heatwave



ADW



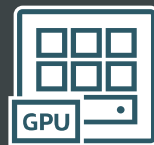
ATP



EXACI



EXACC



H100/A100/A10 GPU



Cloud Native
OKE



Super Cluster



Chatbot

Oracle Enterprise LLM/Gen AI Platform



감사합니다.



김태완 상무, 한국오라클
taewan.kim@oracle.com