

# 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

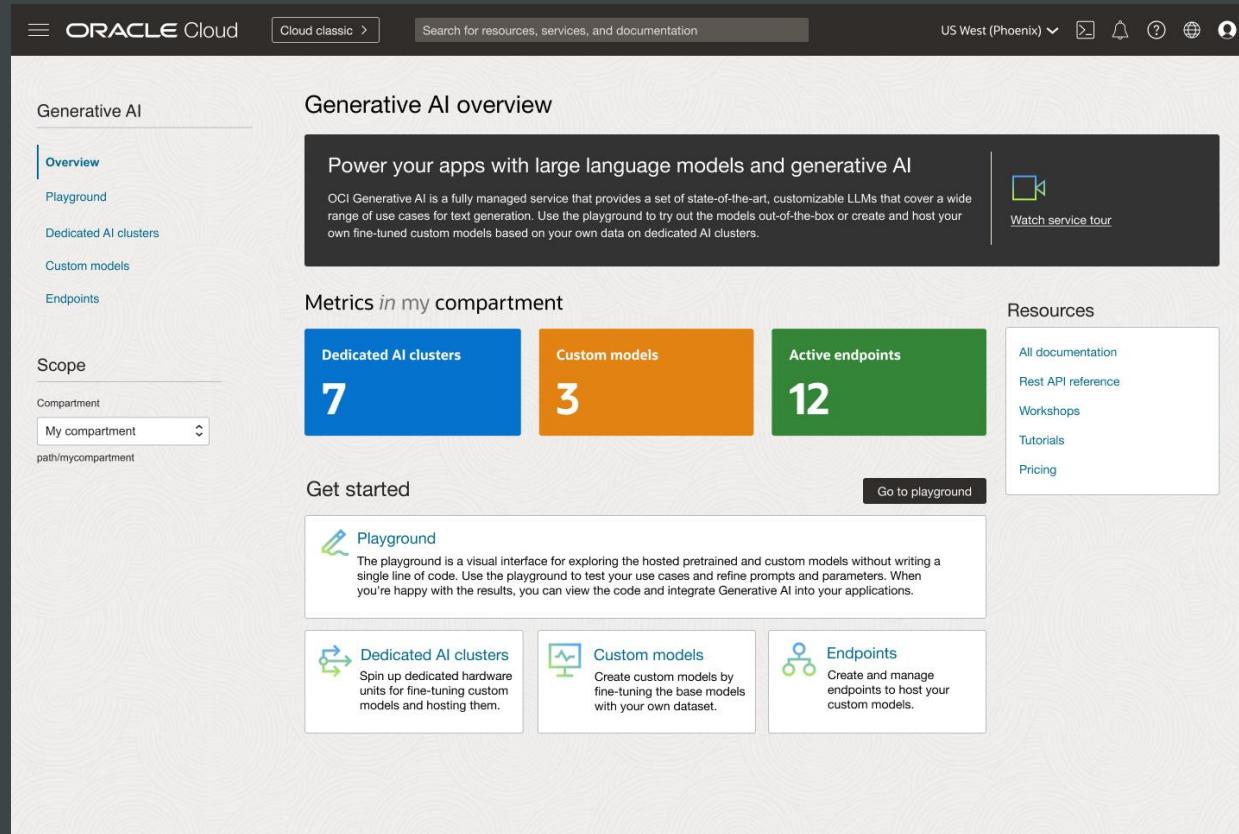
# Safe harbor statement

---

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



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.

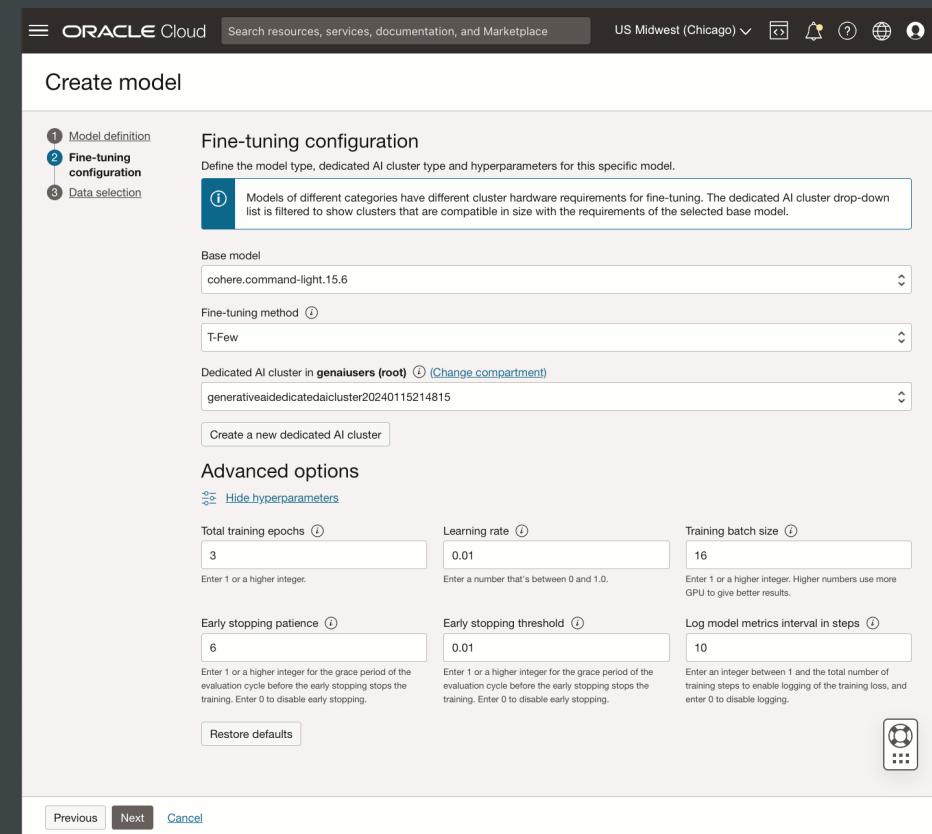
Watch service tour

Metrics in my compartment

Dedicated AI clusters	Custom models	Active endpoints
7	3	12

Get started

- Playground
- Dedicated AI clusters
- Custom models
- Endpoints



Create model

Fine-tuning configuration

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

Model definition: 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): generativeaidedicatedaicleuster20240115214815

Create a new dedicated AI cluster

Advanced options

Hide hyperparameters

Total training epochs: 3

Learning rate: 0.01

Training batch size: 16

Early stopping patience: 6

Early stopping threshold: 0.01

Log model metrics interval in steps: 10

Restore defaults

GA of OCI Generative AI  
2024.01.23

# OCI Generative AI 서비스

## 엔터프라이즈 기업을 위한 LLM 서비스 지원 모델

GA of OCI Generative AI  
2024.01.23

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.

Watch service tour

Metrics in my compartment

Dedicated AI clusters	Custom models	Active endpoints
7	3	12

Resources

- All documentation
- Rest API reference
- Workshops
- Tutorials
- Pricing

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.

Dedicated AI clusters

Custom models

Endpoints



# 오라클의 AI 서비스 스택

## Applications

Fusion Applications

NetSuite

Fusion Analytics

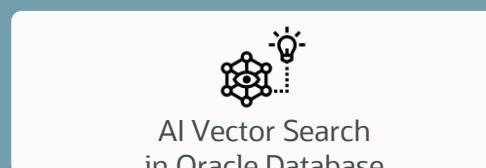
Industry Applications

3rd Party Applications

## AI Services

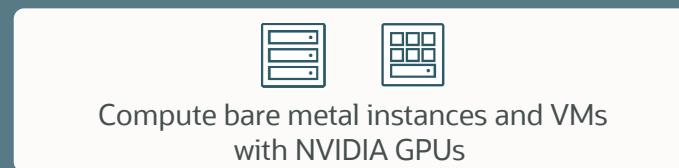


## Data Platforms



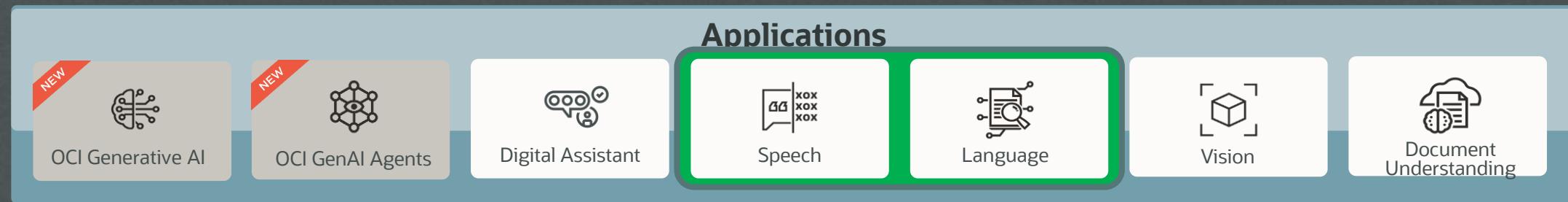
## Data

## AI Infrastructure



Oracle AI Partners

# 오라클의 AI 서비스 스택



### ORACLE AI Demos

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

Results

banking-1.wav

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 **Renault**? 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, **Cosser** with me.

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

00:00:45 - 00:00:56 I'm 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 crisis began in 2009 with the collapse of Iceland's banking system, then spread primarily to Portugal, Italy, Ireland, Greece, and Spain in 2010, leading to the popularization of an offensive monitor (PHIGS). 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).

42015000 characters remaining

Analyze

Language Detection

English

Text Classification

Finance/Banking

Named Entity Recognition

The European sovereign debt crisis (bank) 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 crisis began in 2009 with the collapse of Iceland's banking system, then spread primarily to Portugal, Ireland, Greece, and Spain (bank). 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 (economy) and financial contagion, and by the International Monetary Fund (monetary) (IMF) (monetary).

Key phrase extraction

collapse (bank), european sovereign debt crisis (bank), several european countries (bank), collapse of financial institutions (bank), high government debt (bank), rising bond yield spreads (bank), government securities (bank), debt crisis (bank), Iceland's banking system (bank), offensive monitor (bank), loss of confidence (bank), european businesses and economies (bank), financial guarantees (bank), european countries (bank), euro and financial contagion (bank), international monetary fund (bank), portugal (bank), italy (bank), greece (bank), popularisation (bank), page (bank), orials (bank), inf (bank), ireland (bank), spain (bank)

Sentiment

Document sentiment: Negative

No sentiment aspects returned

Sentence level sentiment

### ORACLE AI Demos

Translation

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

Source language: English

Target language: Korean

OCI has successfully implemented Renault to run its production high-performance computing (HPC) workloads on OCI. Specifically, workloads for Crash Simulation using RAMP-CRAACH, they have already outperformed previous on-premise performance by 19%. This is a significant achievement, as the improved performance allows Renault to perform hundreds more crash and simulation runs per year, leading to better in-car and more fuel-efficient vehicles. Renault is the third largest automobile firm for OCI, extending OCI's reach to cover the largest three automotive markets: US, APAC, and EMEA.

# 오라클의 AI 서비스 스택

## AI Services



# Oracle Generative AI 전략 & 차별성



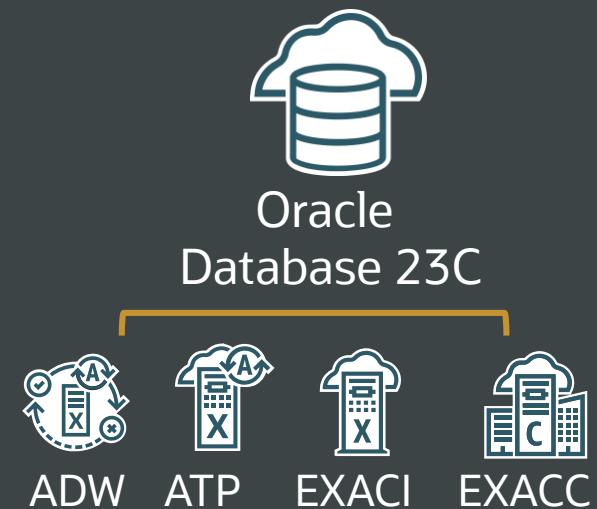
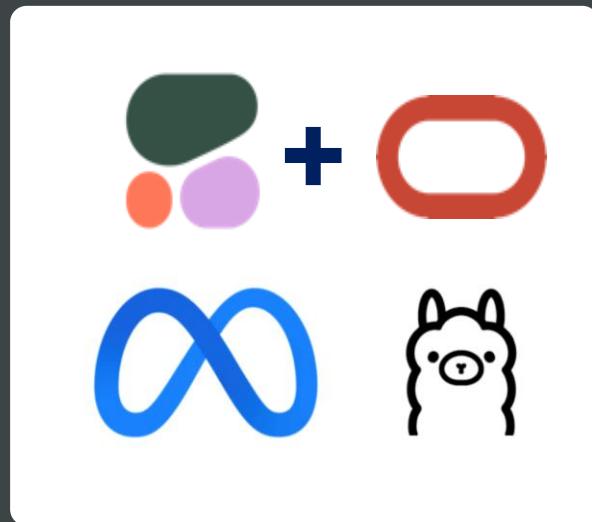
Model



DATA



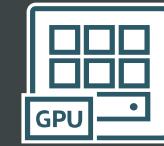
INFRASTRUCTURE



OpenSearch



Heatwave



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 요약을 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 파트너쉽을 통해 엔터프라이즈를 위한 Generative AI 서비스 개발

스탠포드에서 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) <small>not publicly available or viable to serve given size</small>	Base	85.5%
ANTHROPIC	Anthropic v4 (52B)	Command	84.2%
AI21labs	J1 Grande v2 (17B)	Command	80.6%
ALPHALPHA	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%
ALPHALPHA	Luminous Extended (30B)	Command	58.2%
Hugging Face	BLOOM (176B)	Base	52.9%

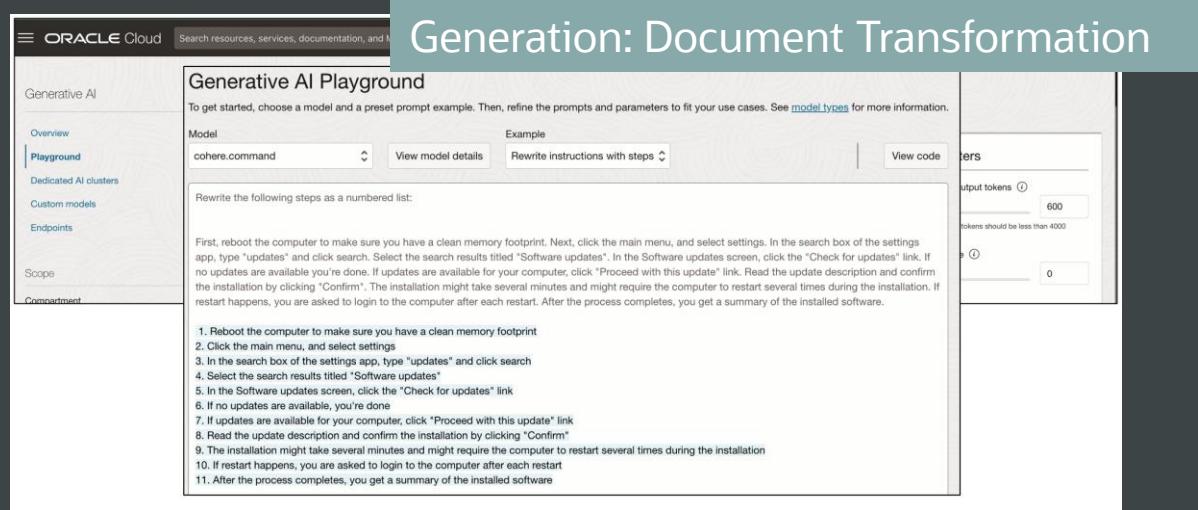
Cohere delivers top-tier LLM performance, outperforming peers in independent LLM benchmarks

(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



Generative AI Playground

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.

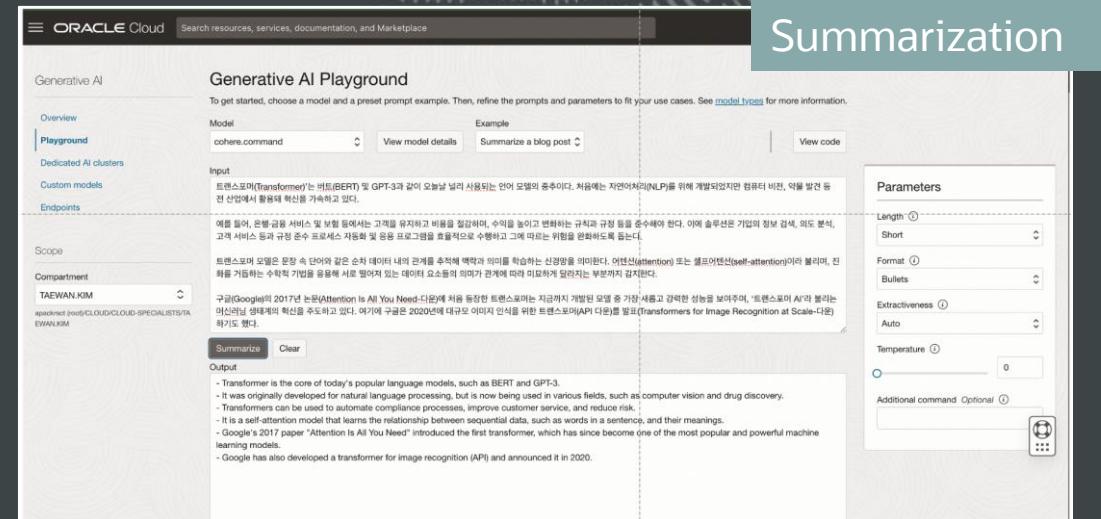
Model: cohene.command Example: Rewrite instructions with steps

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 that happens, you are asked to login to the computer after each restart. After the process completes, you get a summary of the installed software.

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



Generative AI Playground

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.

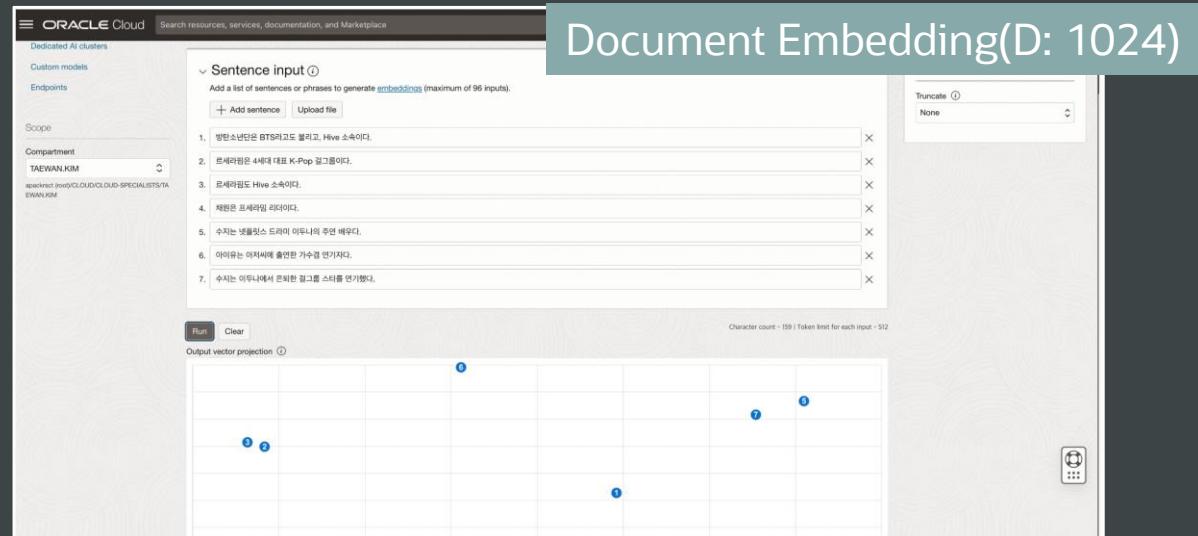
Model: cohene.command Example: Summarize a blog post

Input: 트랜스포머(Transformer)는 베트(BERT) 및 GPT-3과 같이 오늘날 널리 사용되는 언어 모델의 종주이다. 처음에는 자연어 처리(NLP)를 위해 개발되었지만 컴퓨터 비전, 음악 발견 등 전 산업에서 활용돼 혁신을 가져오고 있다.

Output: 예를 들어, 은행 금융 서비스 및 보험 등에서는 고객을 유치하고 비용을 절감하며, 수익을 높이고 번화하는 규칙과 규정 등을 추적해 맥락과 의미를 학습하는 신경망을 마련한다. 어텐션(Attention) 또는 셀프어텐션(self-attention)이라 불리며, 진화를 거듭하는 수학적 기법을 활용해 서로 엮여 있는 데이터 요소의 의미가 관계에 따라 미묘하게 부여까지 감지한다.

Additional command: Optional

## Document Embedding(D: 1024)



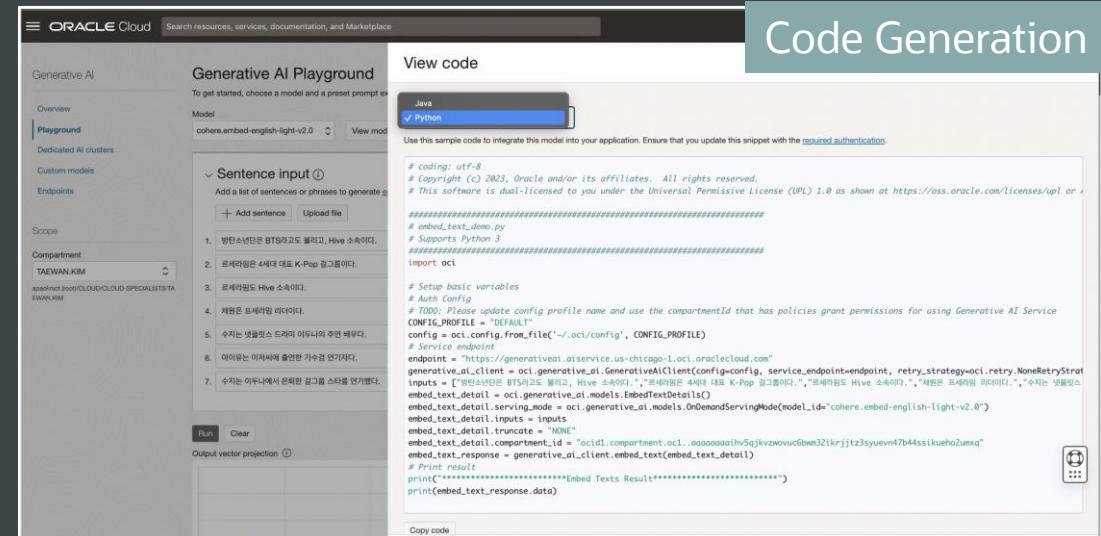
Generative AI Playground

Add a list of sentences or phrases to generate embeddings (maximum of 96 inputs).

Sentence input: 1. 방탄소년단은 BTS라고 불리고, Hive 소속이다.  
2. 트세리원은 4세대 대표 K-Pop 그룹이다.  
3. 르세리원은 Hive 소속이다.  
4. 차원은 트세리원 라이이다.  
5. 수지는 넷플릭스 드라마 아수나의 주인 배우다.  
6. 아이유는 아수나에 출연한 가수겸 연기자다.  
7. 수지는 아수나에서 은퇴한 걸그룹 스타일 연기했다.

Output vector projection: Character count - 199 | Taken limit for each input - 512

## Code Generation



Generative AI Playground

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.

Model: cohene.embed-english-light-v2.0 Example: View code

Sentence input: 1. 방탄소년단은 BTS라고 불리고, Hive 소속이다.  
2. 트세리원은 4세대 대표 K-Pop 그룹이다.  
3. 르세리원은 Hive 소속이다.  
4. 차원은 트세리원 라이이다.  
5. 수자는 넷플릭스 드라마 아수나의 주인 배우다.  
6. 아이유는 아수나에 출연한 가수겸 연기자다.  
7. 수자는 아수나에서 은퇴한 걸그룹 스타일 연기했다.

Output code: Java

```
# coding: utf-8
# Copyright (c) 2023, Oracle and/or its affiliates. All rights reserved.
# This software is dual-licensed to you under the Universal Permissive License (UPL) 1.0 as shown at https://oss.oracle.com/licenses/upl and the MIT License (MIT) as shown at https://oss.oracle.com/licenses/mit.
```

Output code: Python

```
# embed_text_demo.py
# Supports Python 3
# Import OCI
# Setup basic variables
# Auth Config
# T000: Please update config profile name and use the compartmentId that has policies grant permissions for using Generative AI Service
CONFIG_PROFILE = "DEFAULT"
config = oci.config.from_file(f'{os.path.expanduser("~")}/oci/config', CONFIG_PROFILE)
# Service Endpoint
endpoint = "https://generativeai.ociservice.us-ashburn1.oci.oraclecloud.com"
generative_ai_client = oci.generativai.Client(config=config, service_endpoint=endpoint, retry_strategy=oci.retry.NoneRetryStrategy())
# Inputs
inputs = ["방탄소년단은 BTS라고 불리고, Hive 소속이다.", "트세리원은 4세대 대표 K-Pop 그룹이다.", "르세리원은 Hive 소속이다.", "차원은 트세리원 라이이다.", "수자는 넷플릭스 드라마 아수나의 주인 배우다.", "아이유는 아수나에 출연한 가수겸 연기자.", "수자는 아수나에서 은퇴한 걸그룹 스타일 연기했다."]
# Embed Text
embed_text_detail_service = oci.generativai.Models.OnDemandService(model_id="cohene.embed-english-light-v2.0")
embed_text_detail_inputs = inputs
embed_text_detail_truncate = None
embed_text_detail_compartment_id = "ocid1.compartment.oc1..aaaaaaaahv5gjkwvovuc6bw32ikrjjt23yuevn47b4ssssikueh02umq"
embed_text_response = generative_ai_client.embed_text(embed_text_detail)
# Print result
print("*****Embed Texts Result*****")
print(embed_text_response.data)
```

# OCI Generative AI 서비스: 전용 모델

### Create dedicated AI cluster

**Info** 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.

**Compartiment**  
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

**General**

Namespace: axk4z7krhqfx  
Compartment: [genaiusers](#)  
Created: Thu, Aug 31, 2023, 16:08:56 UTC  
ETag: 093eab27-5c55-446e-80e1-0012d028819e  
OCID: ...xycuh76q [Show](#) [Copy](#)

**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](#) 

**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 

**Resources**

**Objects**

<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	

**Objects**

Upload More Actions  Search by prefix

<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.

**ⓘ** 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)** **ⓘ** [\(Change compartment\)](#)  
generativeaiadicatedaicluster20240115214815

Create a new dedicated AI cluster

### Advanced options

**ⓘ** [Hide hyperparameters](#)

Total training epochs <b>ⓘ</b> 3 Enter 1 or a higher integer.	Learning rate <b>ⓘ</b> 0.01 Enter a number that's between 0 and 1.0.	Training batch size <b>ⓘ</b> 16 Enter 1 or a higher integer. Higher numbers use more GPU to give better results.
Early stopping patience <b>ⓘ</b> 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 <b>ⓘ</b> 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 <b>ⓘ</b> 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 **geniausers (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.	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 Inappcs, 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, Inappcs 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 its employees with self-service capabilities and career development tools. ###
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 Asia-Pacific with aggressive 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 would like to share a success story from Madero Restaurants. This prospect is a rapidly growing 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. [Your Contact Information] END

Previous  Cancel

# OCI Generative AI 서비스: LLM 모델 Fine-Tuning

# OCI Generative AI 서비스: LLM 모델 관리

ORACLE Cloud Search resources, services, documentation, and Marketplace

## 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: ...jm3kucc7le920qeq47q Show Copy OCID: ...yuzzatbf7bwhwtoria Show Copy Description: State: Active Lifecycle details: Created Endpoint Dedicated AI Cluster: oci1:generative:dedicated:aicluster.oc1.us-chicago-1.aaaaaaaaacqy6o-koamhlnrnt2ut3lk2c4qalbxrbkm2a5/72vof64c3a	Model name: cohere.command-light Model version: 15.6 Created on: Mon, 15 Jan 2024 21:50:41 UTC Created by: tele.haghshenass@oracle.com Content moderation: Enabled

Resources

Endpoint metrics

Start time: Jan 15, 2024 22:33:43 UTC End time: Jan 15, 2024 23:33:43 UTC Quick selects: Last hour

Work requests

Total processing time

Number of calls

Service Errors Count

Client Errors Count

Total number of input

Total number of output

Terms of Use and Privacy Cookie Preferences Copyright © 2024, Oracle and/or its affiliates. All rights reserved.

# OCI Generative AI 서비스

## SOTA Open-Source & Commercial Model 지원 서비스

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.

Watch service tour

Metrics in my compartment

Dedicated AI clusters	Custom models	Active endpoints
7	3	12

Resources

- All documentation
- Rest API reference
- Workshops
- Tutorials
- Pricing

Get started

- Playground
- Dedicated AI clusters
- Custom models
- Endpoints



GA of OCI Generative AI  
2024.01.23



## 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.



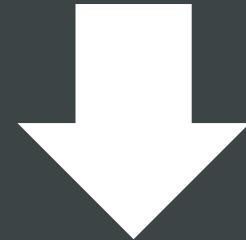
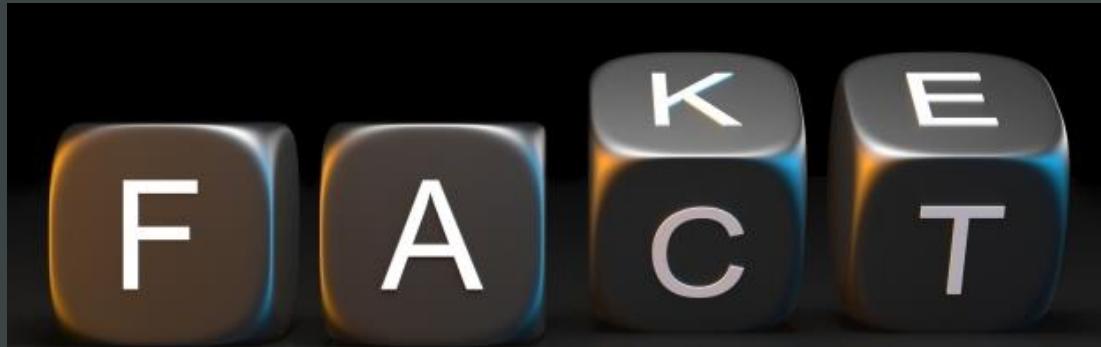
Security and  
privacy



Predictable performance  
and pricing

# Oracle Enterprise LLM Platform

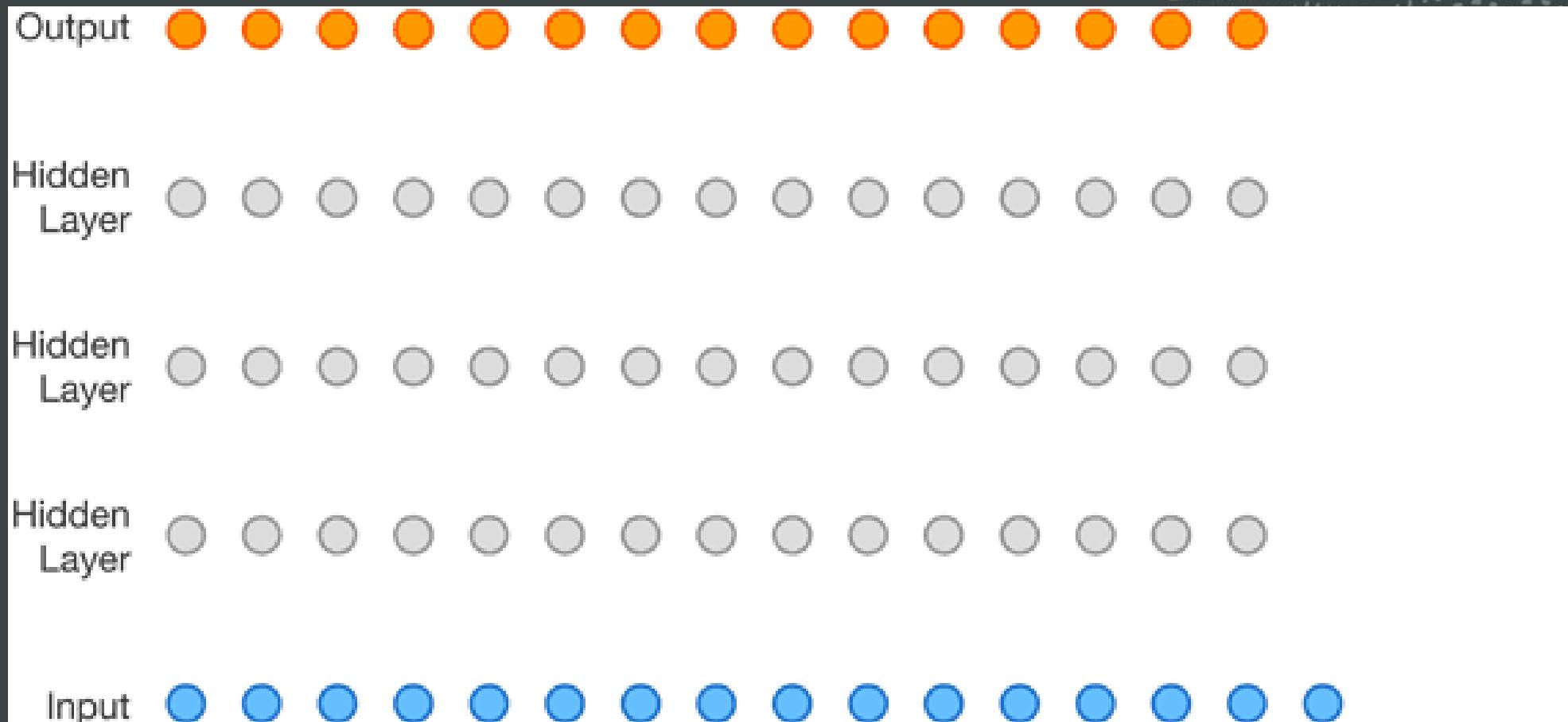
기업용 LLM의 도전 과제



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<sup>†‡</sup>, Ethan Perez\*,

Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Naman Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,

Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>

<sup>†</sup>Facebook AI Research; <sup>‡</sup>University College London; <sup>\*</sup>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.

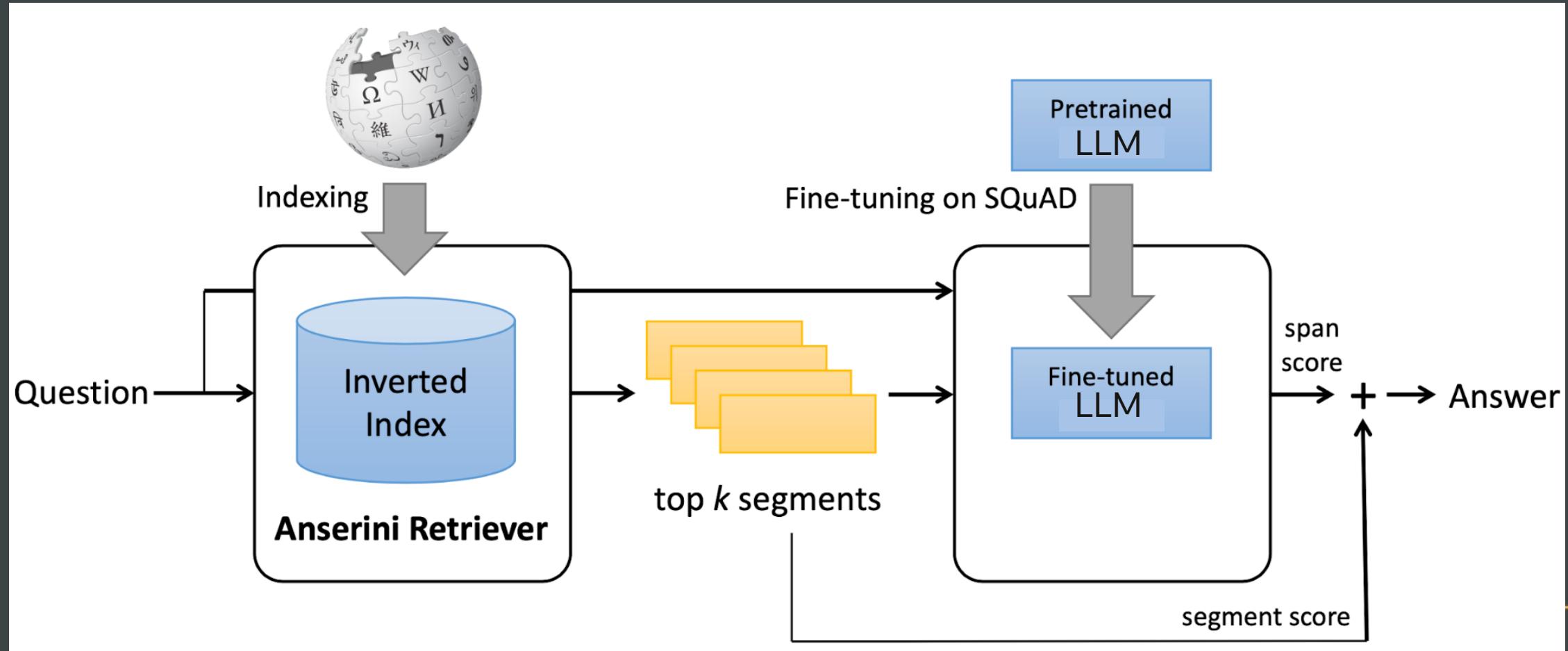
I work at the intersection of information retrieval techniques (IR) and large language models (LLMs), and have done a lot of work on *Retrieval-Augmented Language Models*.

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 updateable 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 ([Twitter](#)) [here](#) and toot ([@](#)) [here](#)

# RAG(Retrieval Augmented Generation)

Open Domain Question 해결 기법 (2021.04)



# 인간 커뮤니케이션

## W5H1, 육하원칙



Who 누가

When 언제

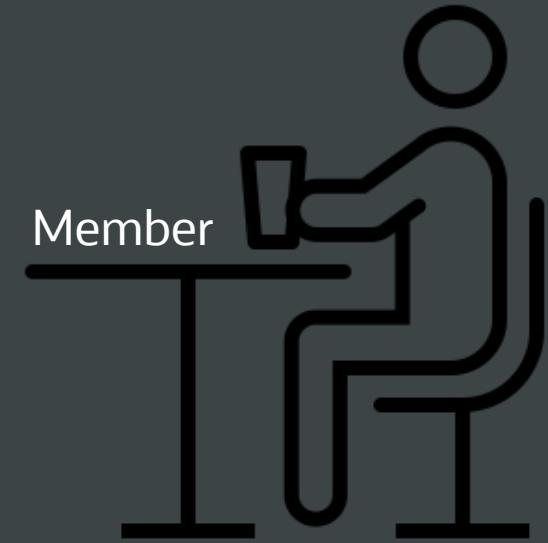
Where 어디서

**Prompt**

How 어떻게

What 무엇을

Why 왜

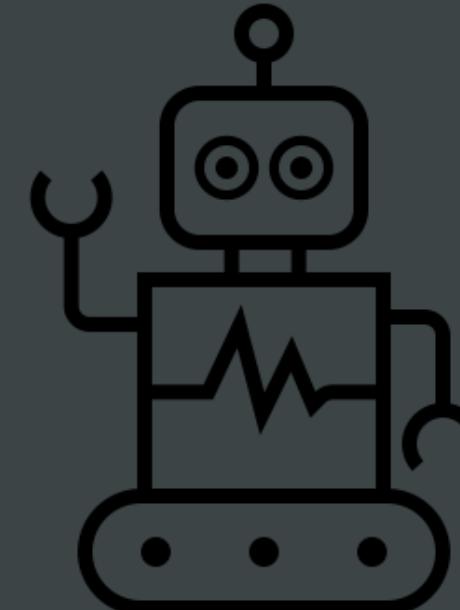


# LLM과 커뮤니케이션

## Prompt Engineering



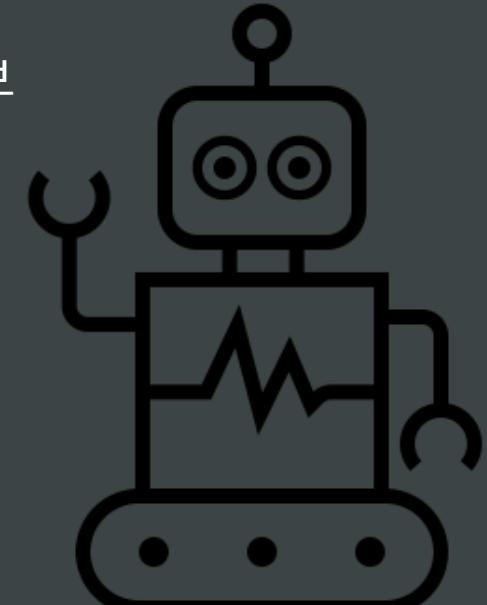
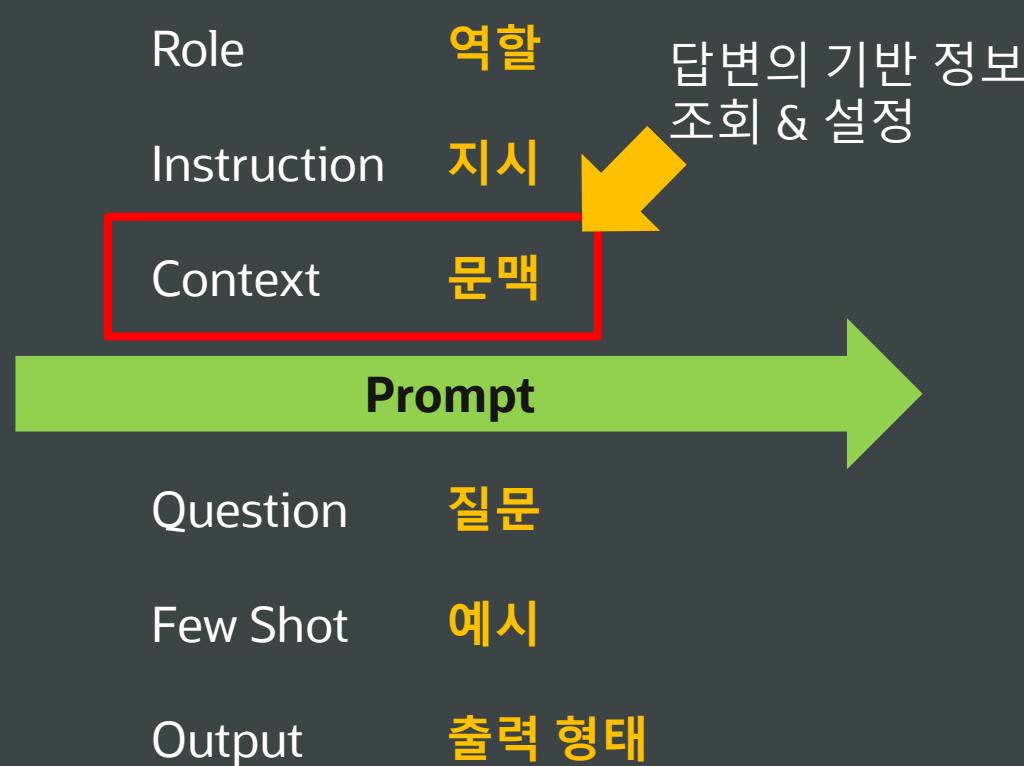
Role	역할
Instruction	지시
Context	문맥
<b>Prompt</b>	
Question	질문
Few Shot	예시
Output	출력 형태



Generative AI (LLM,  
Large Language Model)

# LLM과 커뮤니케이션

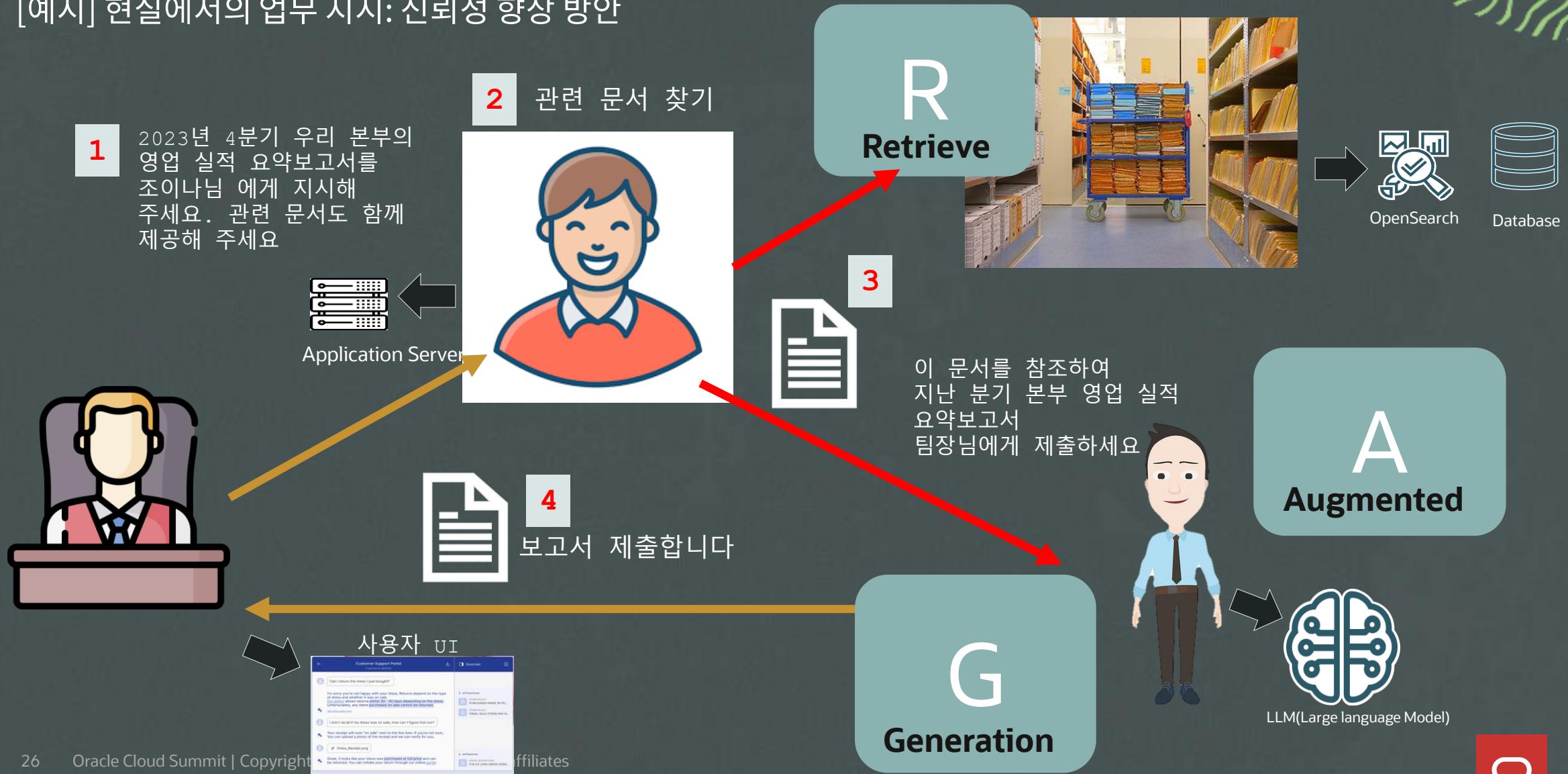
## Prompt Engineering



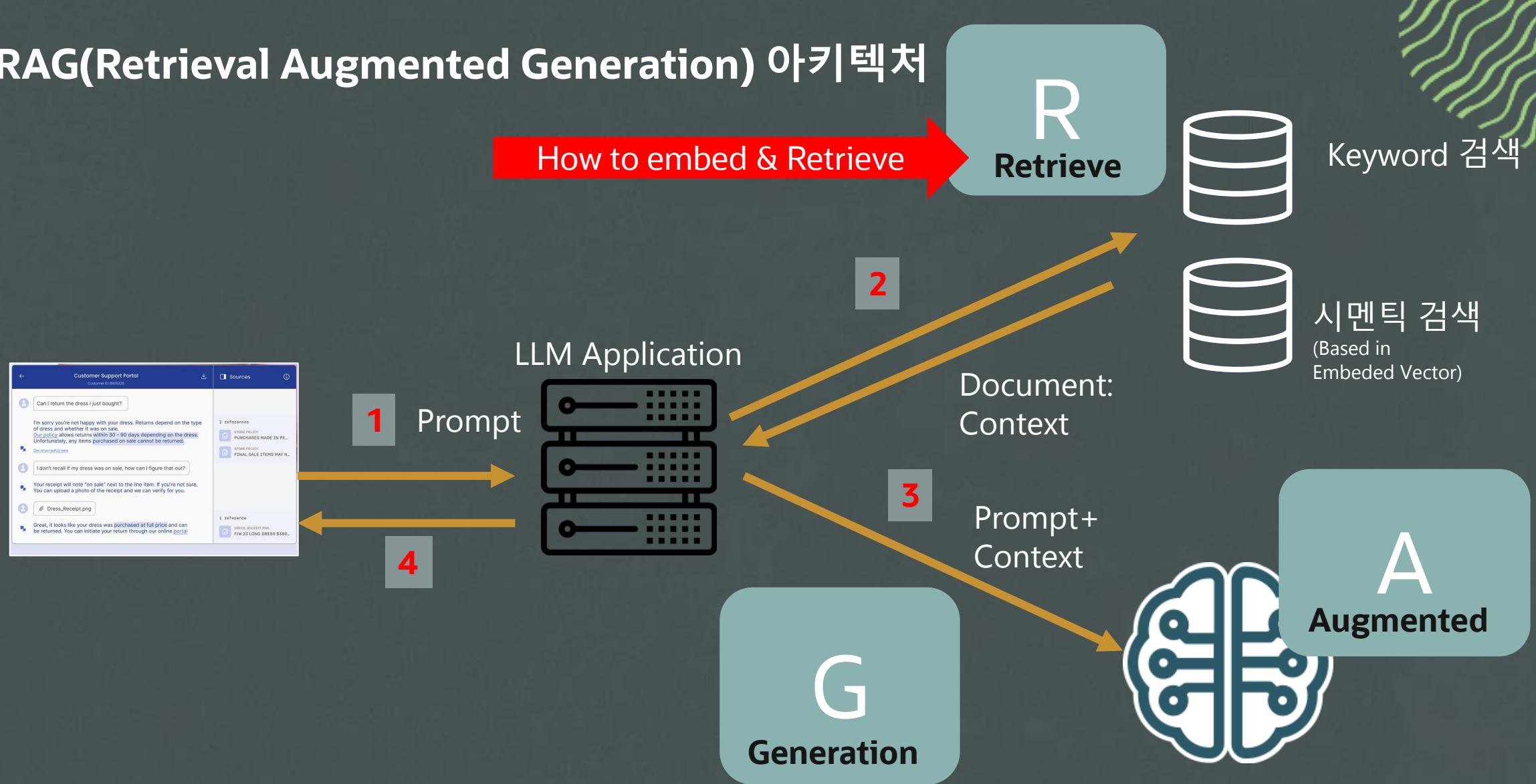
Generative AI (LLM,  
Large Language Model)

# RAG (Retrieval Augmented Generation)

[예시] 현실에서의 업무 지시: 신뢰성 향상 방안

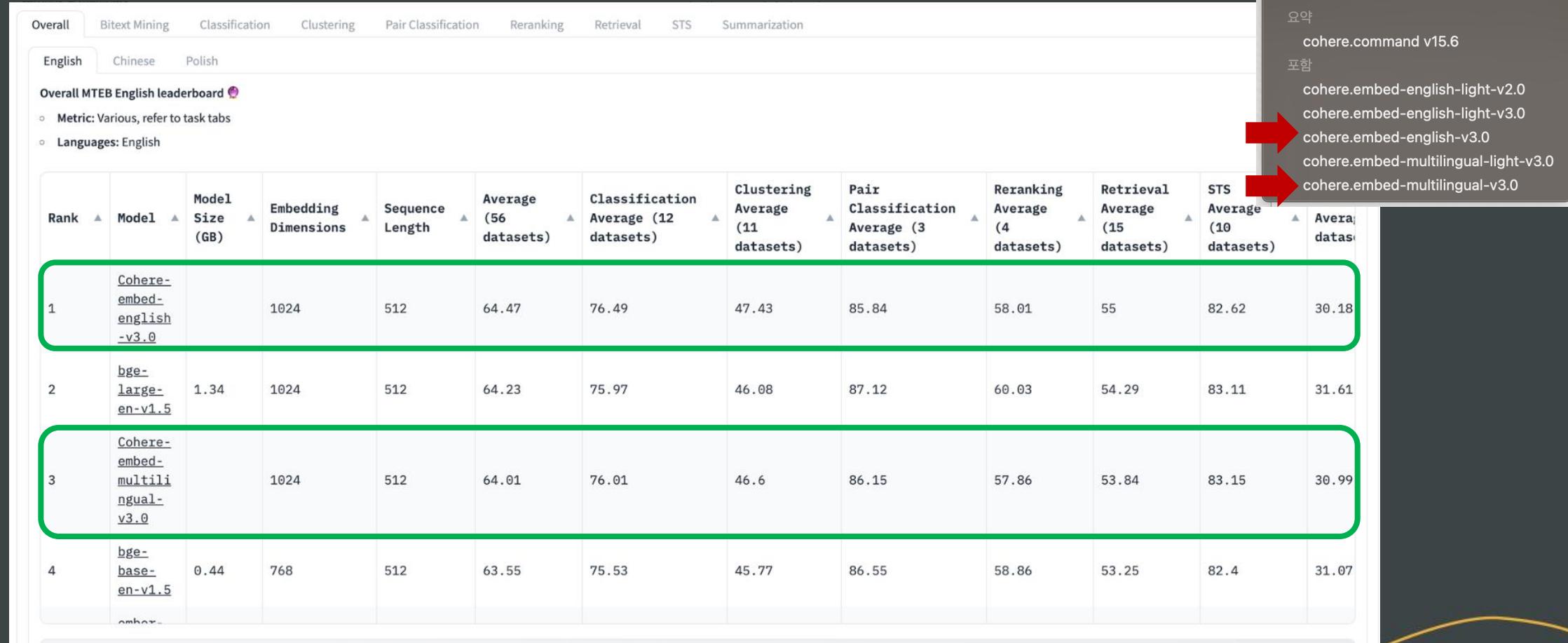


# RAG(Retrieval Augmented Generation) 아키텍처



# RAG(Retrieval Augmented Generation) 아키텍처

## Embedding model의 성능 지표 (Huggingface Leaderboard)



Overall MTEB English leaderboard												
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	<a href="#">Cohere-embed-english-v3.0</a>	1024	512	64.47	76.49	47.43	85.84	58.01	55	82.62	30.18	30.18
2	<a href="#">bge-large-en-v1.5</a>	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29	83.11	31.61
3	<a href="#">Cohere-embed-multilingual-v3.0</a>	1024	512	64.01	76.01	46.6	86.15	57.86	53.84	83.15	30.99	30.99
4	<a href="#">bge-base-en-v1.5</a>	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

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

# RAG(Retrieval Augmented Generation) 아키텍처

VectorDBs in OCI



OpenSearch



Oracle 23C



MySQL Heatwave



PostgreSQL

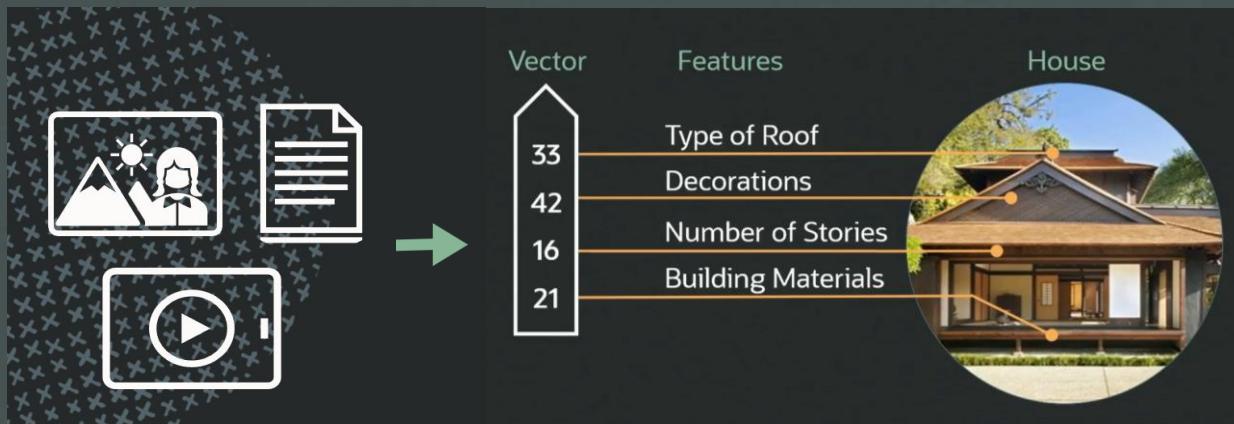


OCI Cache  
Managed Redis

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

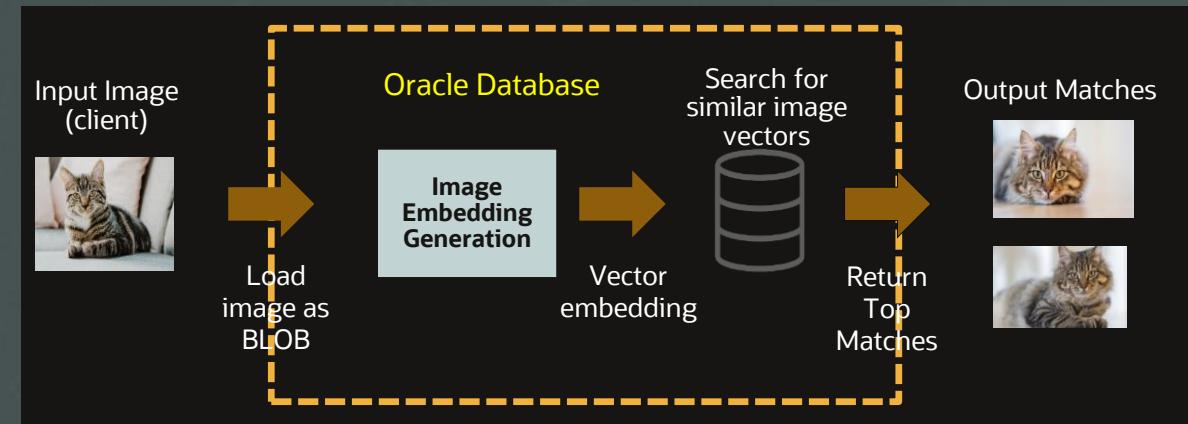
## Vector Search: 오라클 DB 내에서 Vector Type 제공

- 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 운영 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



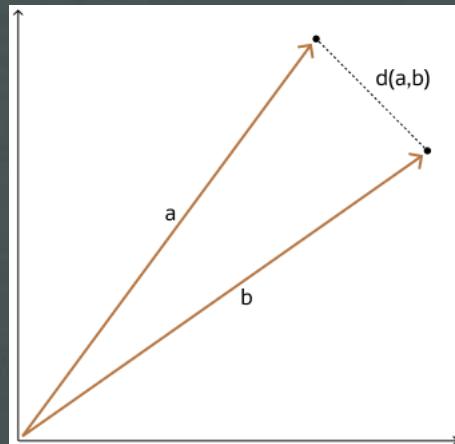
```
create table galaxies (
  id number,
  name varchar2(50),
  doc varchar2(500),
  embedding vector
);
```

```
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]');
```

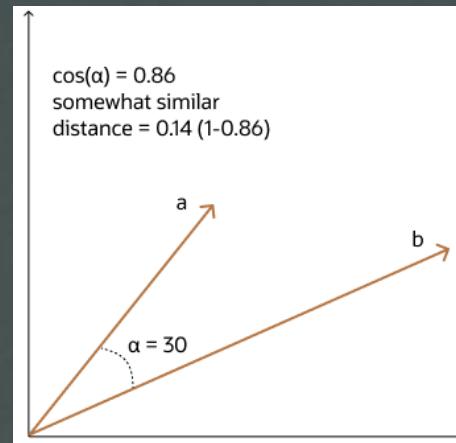
```
insert into galaxies values (2, 'M33',
  'Messier 33 is a spiral galaxy in the
  Triangulum constellation.', '[0,0,1,0,0]');
```

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

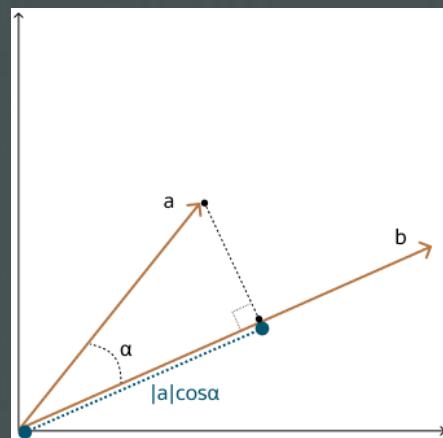
# RAG 지원을 위한 Oracle DB의 전략 : Vector 유사도 연산



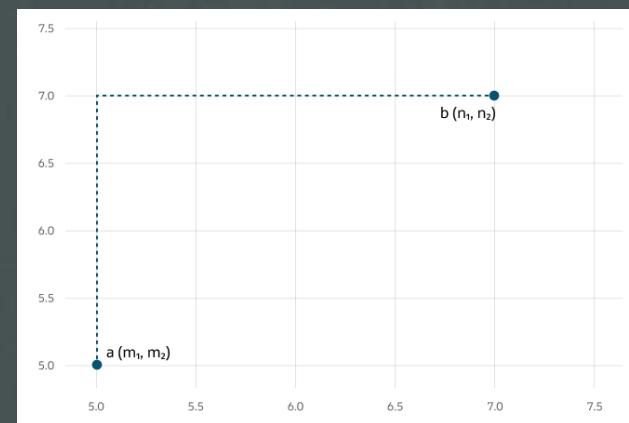
Euclidean Distance



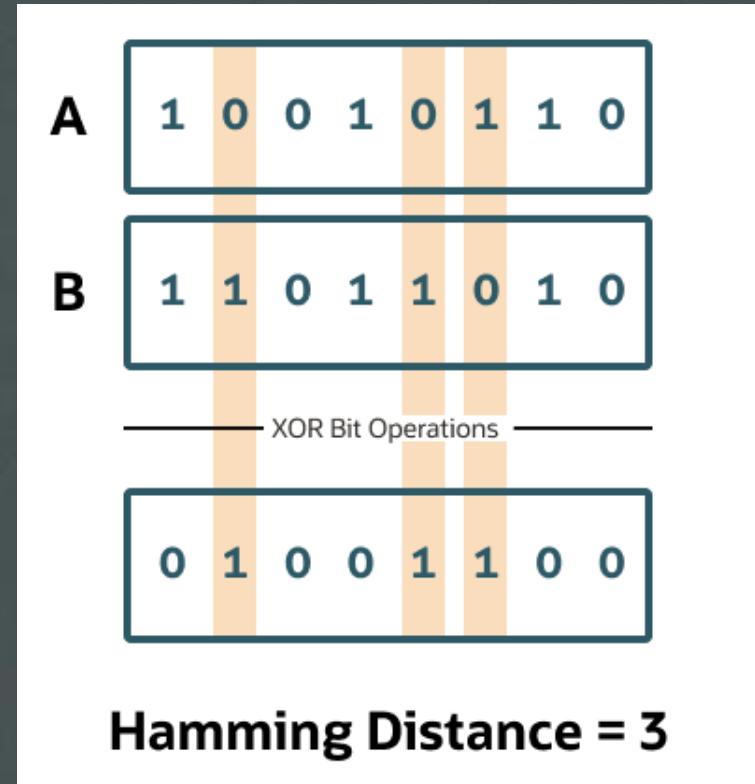
Cosine Similarity



Dot Product Similarity



Manhattan Distance



Hamming Distance

# RAG 지원을 위한 Oracle DB의 전략 : Vector 유사도 쿼리



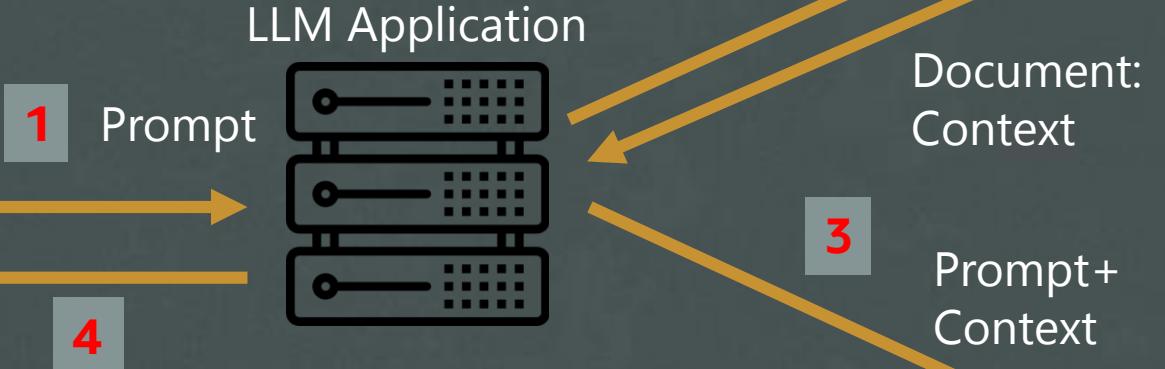
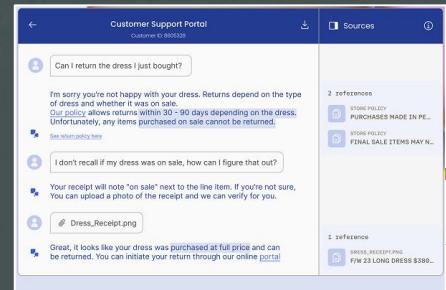
```
SELECT name,
       VECTOR_DISTANCE( embedding,
                          (SELECT embedding FROM galaxies WHERE name = 'NGC1073'),
                          COSINE ) as vector_similarity
  FROM galaxies
 WHERE name <> 'NGC1073'
 ORDER BY VECTOR_DISTANCE( embedding,
                            (SELECT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

## NAME      VECTOR\_SIMILARITY

---

M31	0.000000000000000220446049250313
M91	0.000000000000000220446049250313
M77	0.000000000000000220446049250313

# RAG(Retrieval Augmented Generation) 아키텍처



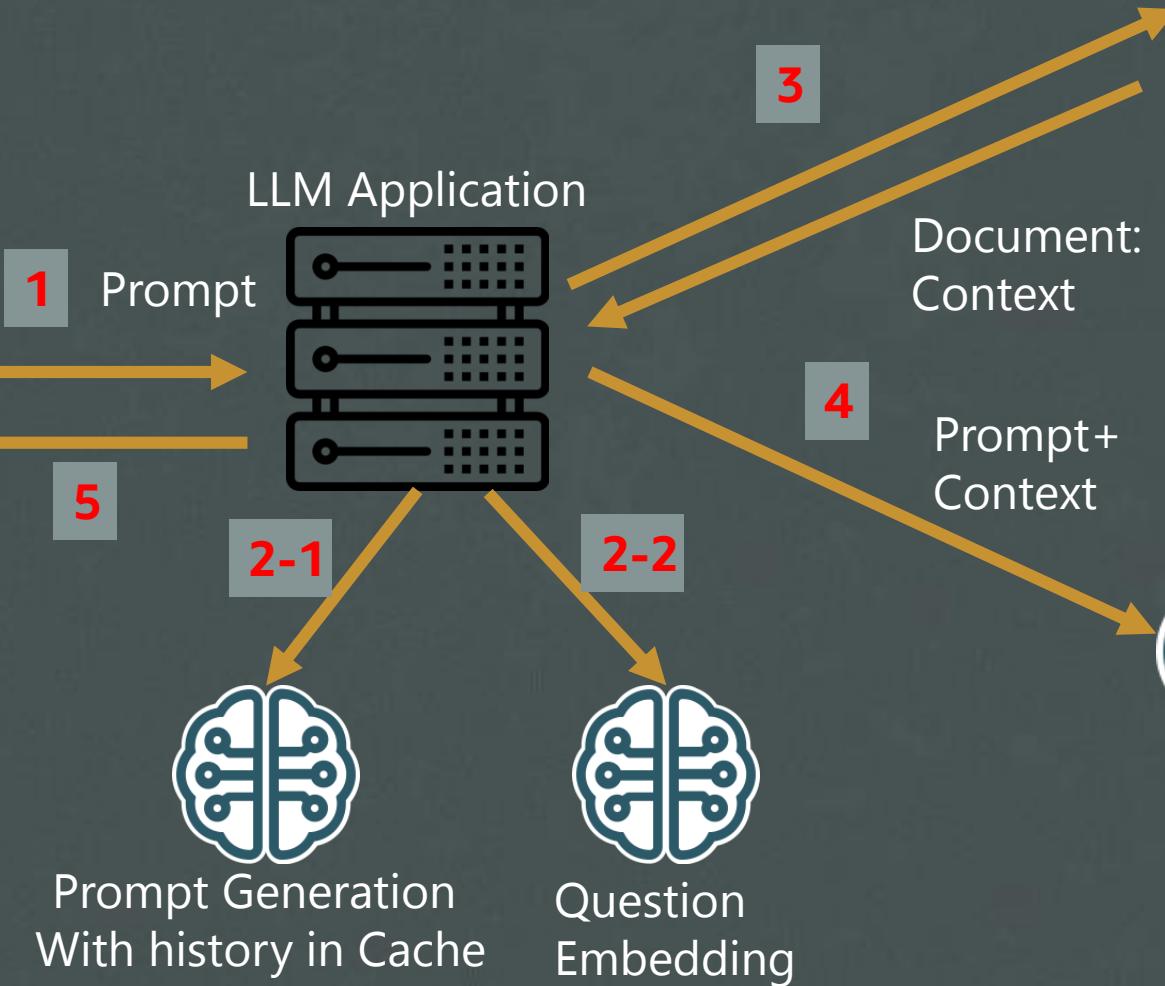
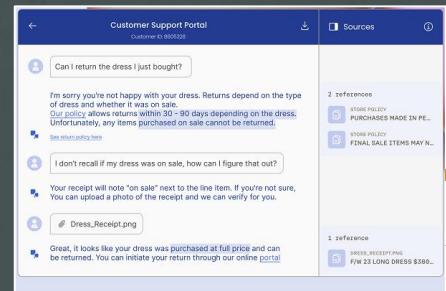
Keyword 검색



시멘틱 검색  
(Based on Embedded Vector)



# RAG(Retrieval Augmented Generation) 아키텍처



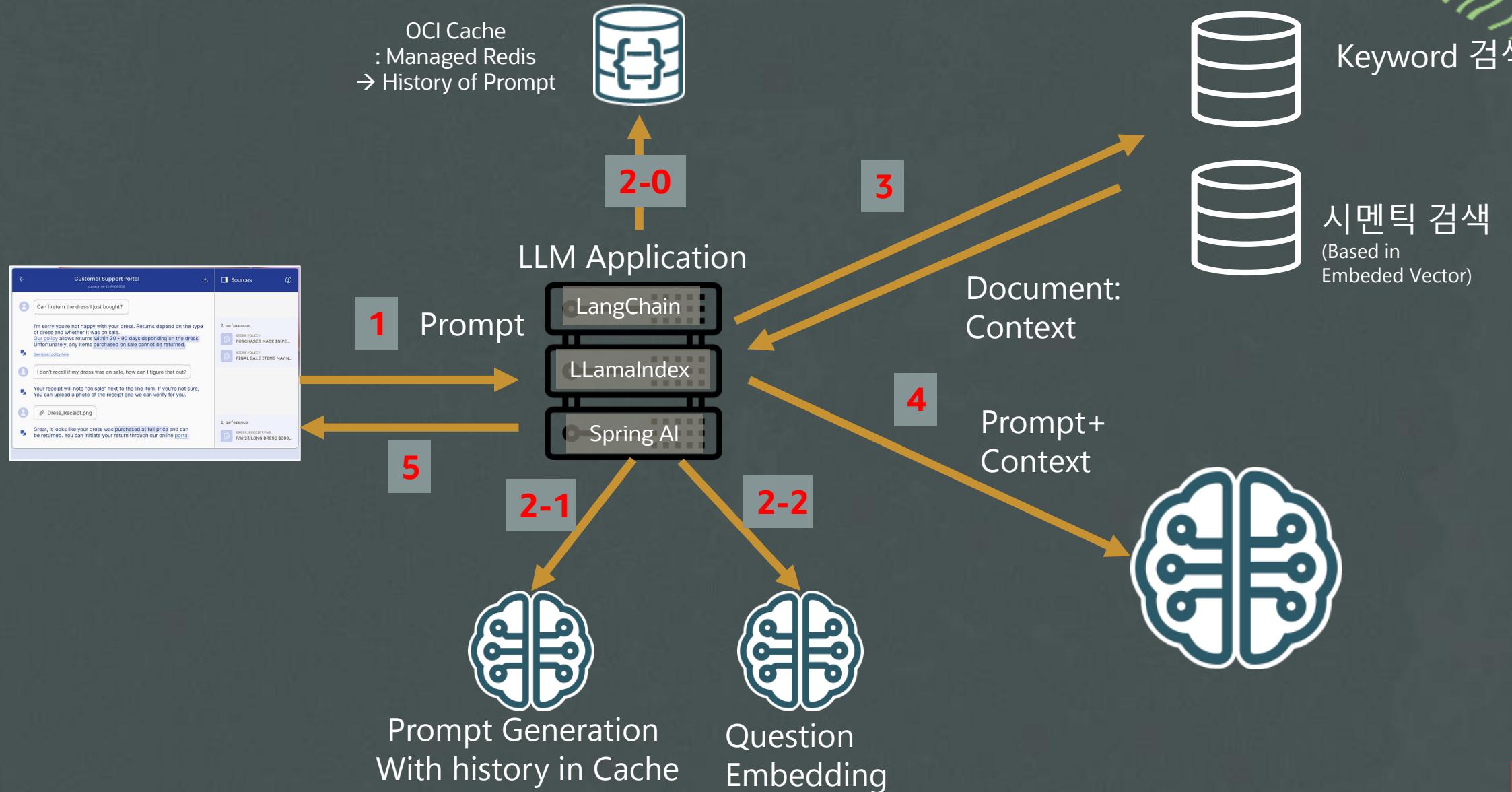
Keyword 검색



시멘틱 검색  
(Based on Embedded Vector)



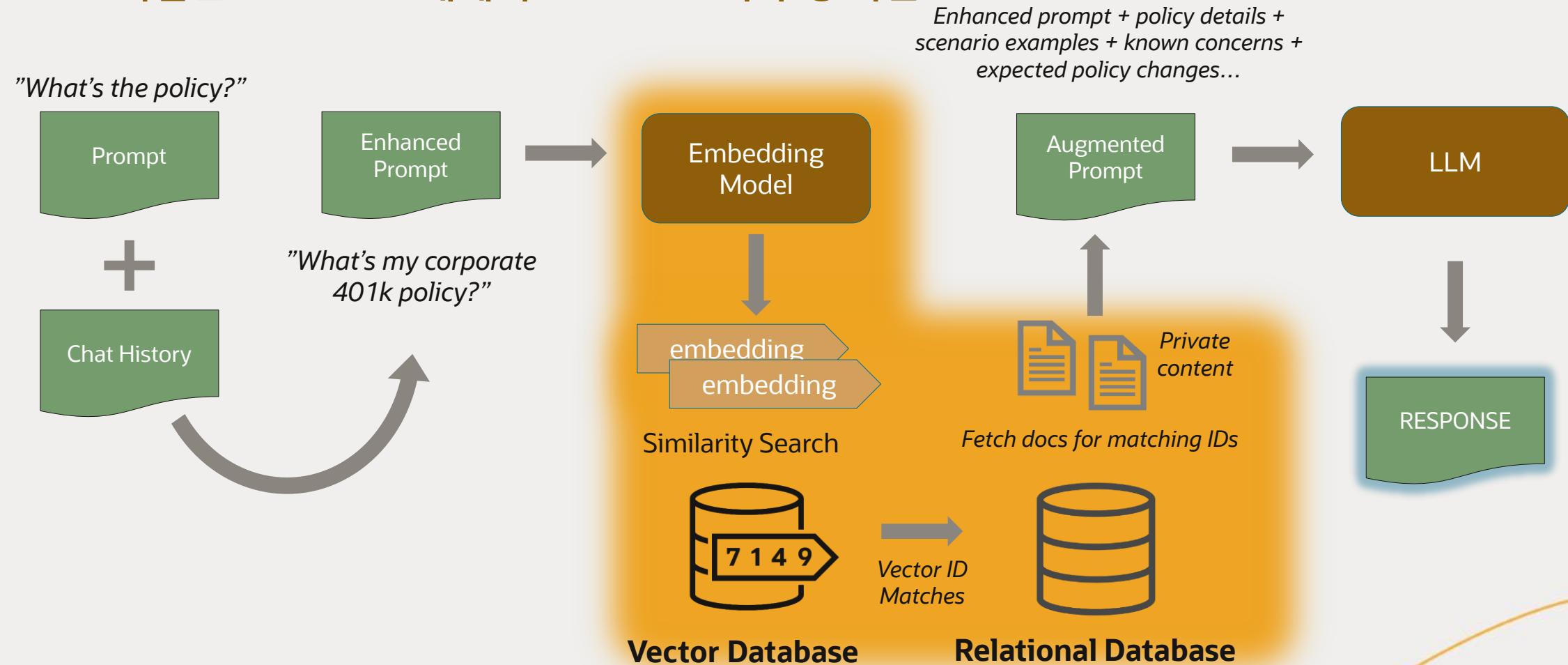
# RAG(Retrieval Augmented Generation) 아키텍처



# RAG(Retrieval Augmented Generation)

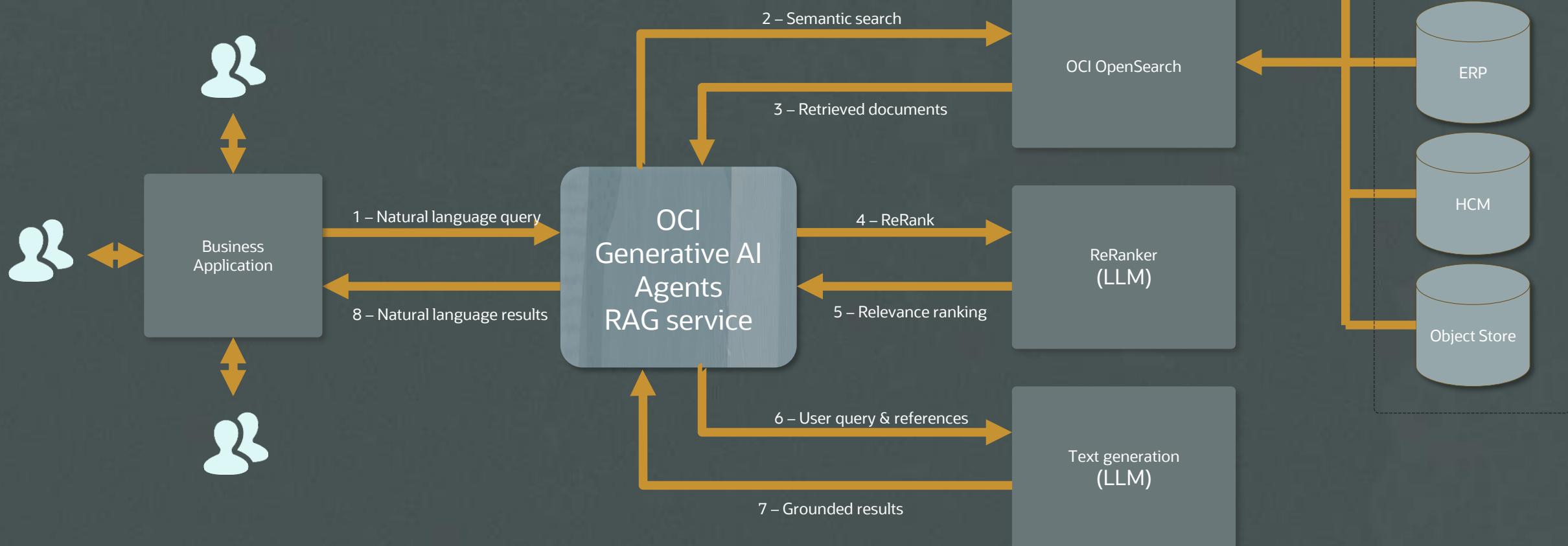
기업 지식/데이터를 활용하여 hallucination 없는 Q&A 지원

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

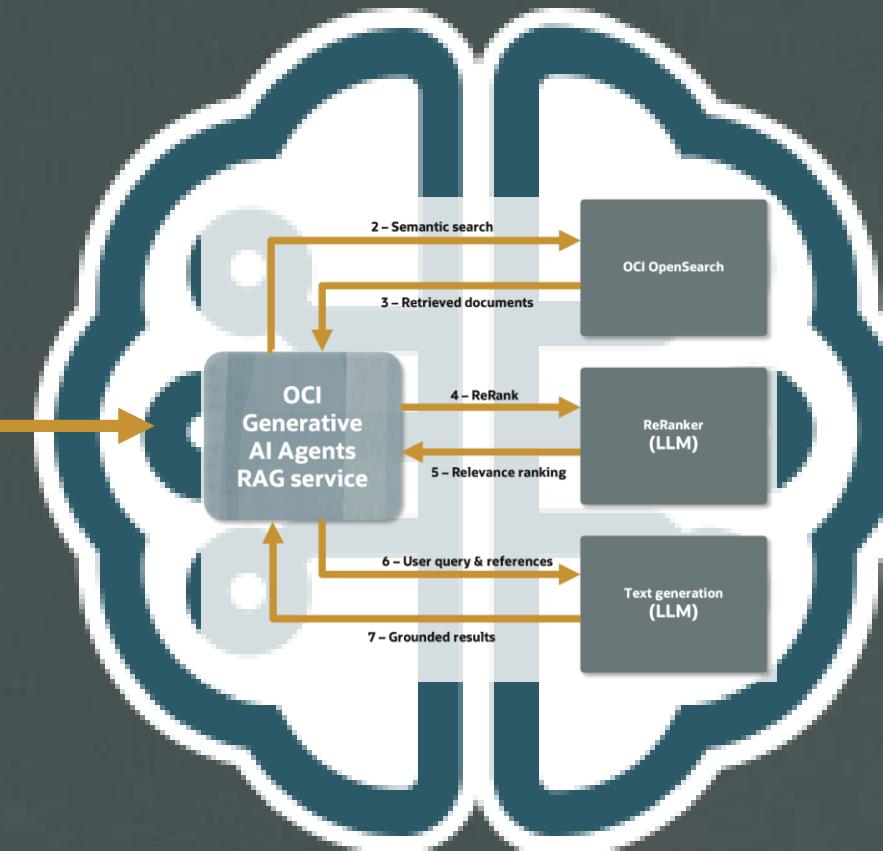
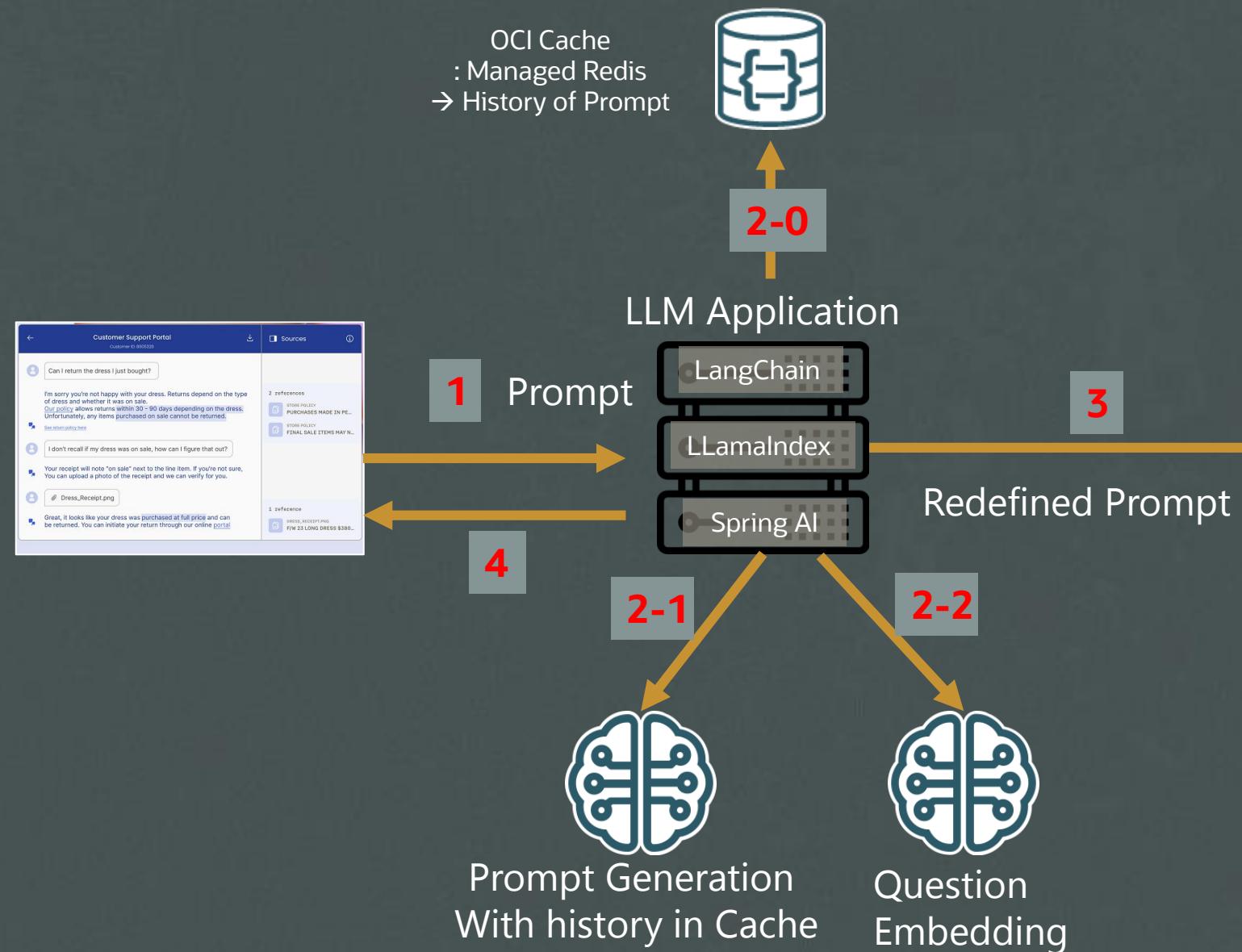


# Oracle RAG(Retrieval Augmented Generation)

## 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



ORACLE Cloud Cloud Classic > Search resources, services, documentation, and Marketplace

Generative AI Agents

- Overview
- Agent
- Data source
- Private endpoints

Overview

Power your response generation or action execution with OCI Generative AI Agents

OCI GenAI Agents is a fully managed service that makes it easy to create and deploy GenAI agents that securely connect to your data sources to generate a more accurate and relevant response. Agents can also execute a set of business tasks by connecting to external services like Oracle Database or Oracle Analytics Cloud. This allows you to build intelligent chatbots, support chatbot, minimize the need for human intervention, or efficiently handle the closure and update of support tickets.

In this Beta release, GenAI Agents supports only retrieving information using retrieval-augmented generation (RAG).

Get started

Step 1: Create Agent

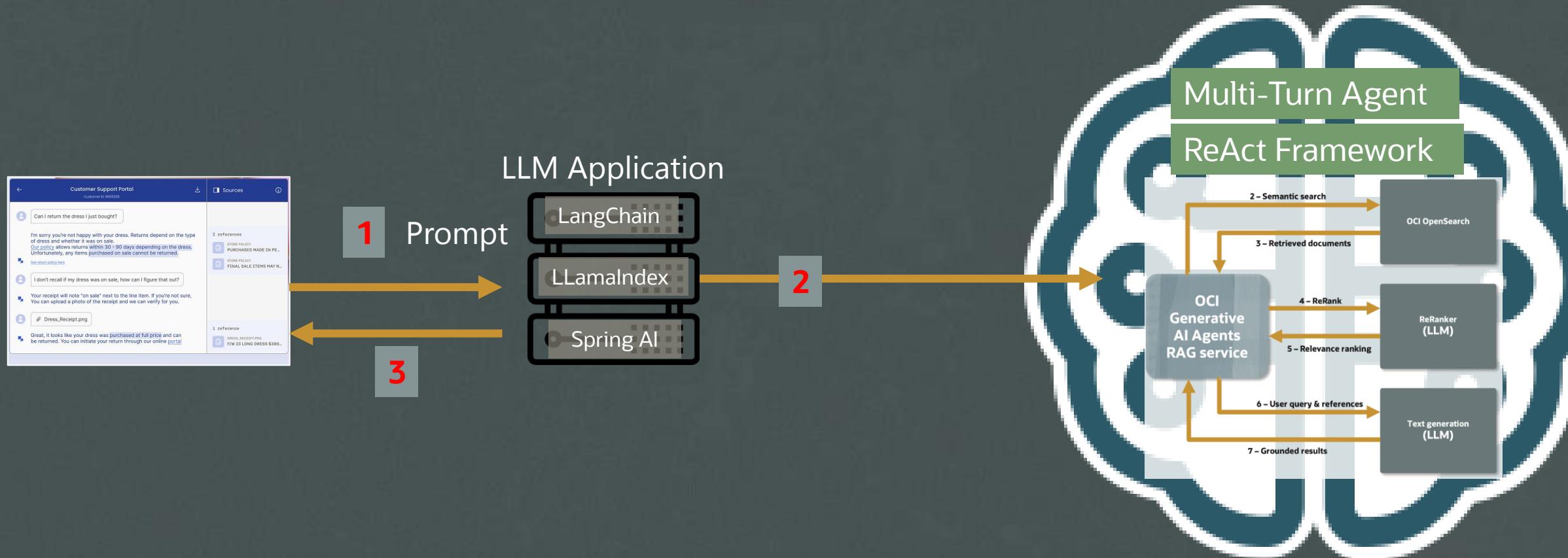
Create an agent for your usecase, and attach the data sources to augment the user request and generate a response

Step 2: Create Data source (optional)

Create a datasource for the OCI OpenSearch indexes to re-use as a resource in an Agent

Get Started

# 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

Oracle Cloud Cloud Classic Search resources, services, documentation, and Marketplace

Generative AI Agents

- Overview
- Agent
- Data source
- Private endpoints

Overview

Power your response generation or action execution with OCI Generative AI Agents. OCI GenAI Agents is a fully managed service that makes it easy to create and deploy GenAI agents that securely connect to your data sources to generate a more accurate and relevant response. Agents can also execute a set of business tasks by connecting to external services. Support for chatbot, minimizing the need for human intervention, or efficiently handle the closure and update of support tickets.

In this Beta release, GenAI Agents supports only retrieving information using retrieval-augmented generation (RAG).

Get started

Step 1: Create Agent

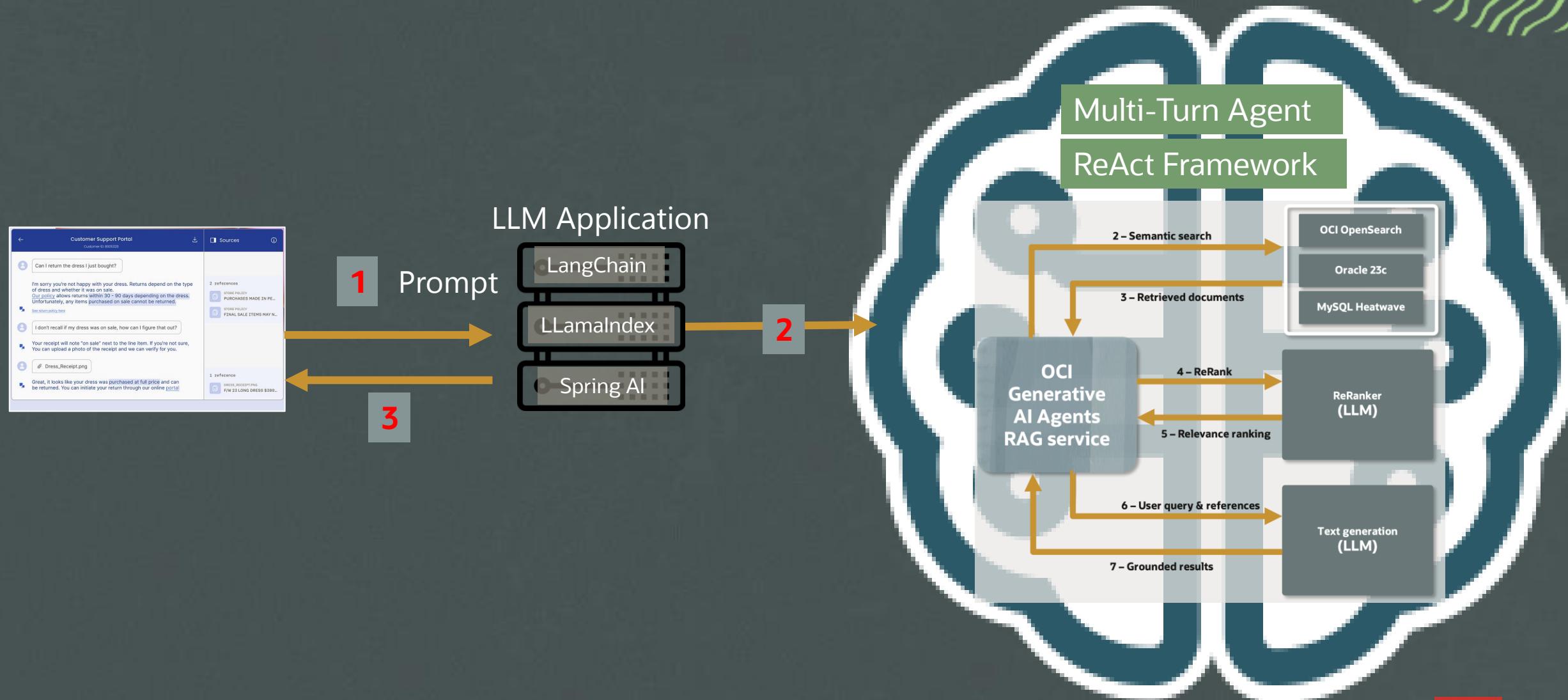
Create an agent for your usecase, and attach the data sources to augment the user request and generate a response

Step 2: Create Data source (optional)

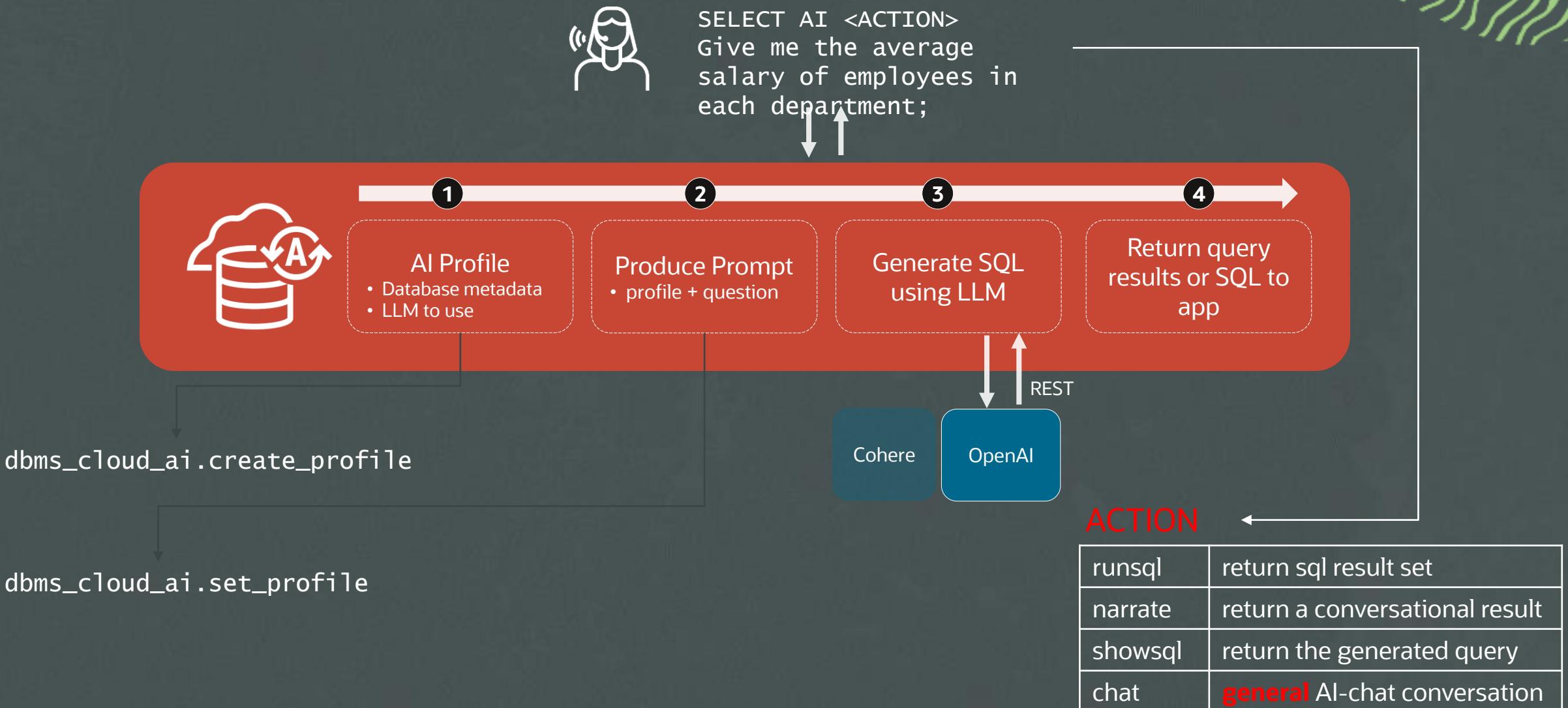
Create a datasource for the OCI OpenSearch indexes to re-use as a resource in an Agent

Get Started

# RAG(Retrieval Augmented Generation) 아키텍처



# LLM을 이용한 SQL 생성 지원(23c)



# LLM을 이용한 SQL 생성 지원(23c)

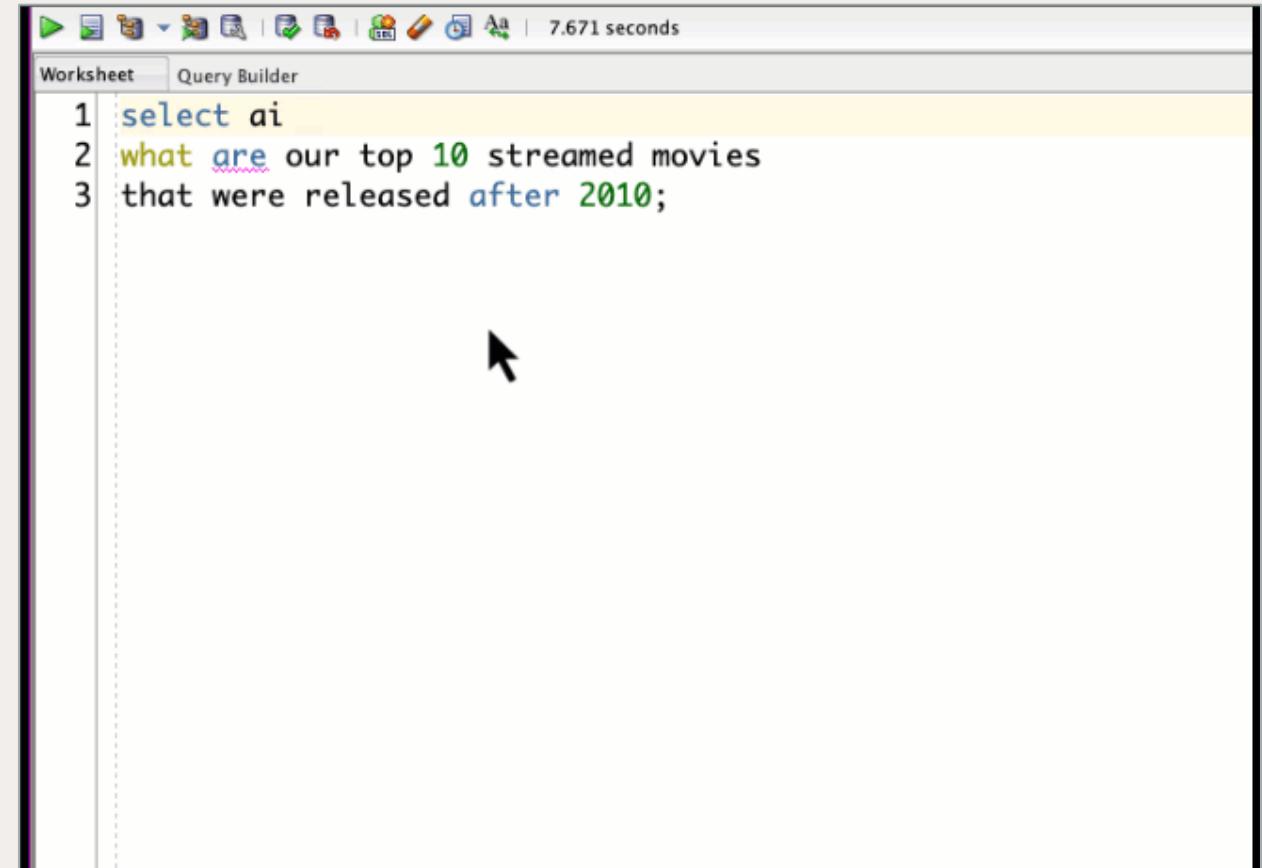
새로운 `DBMS_CLOUD_AI` PL/SQL 패키지를 사용하여 사용 가능

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

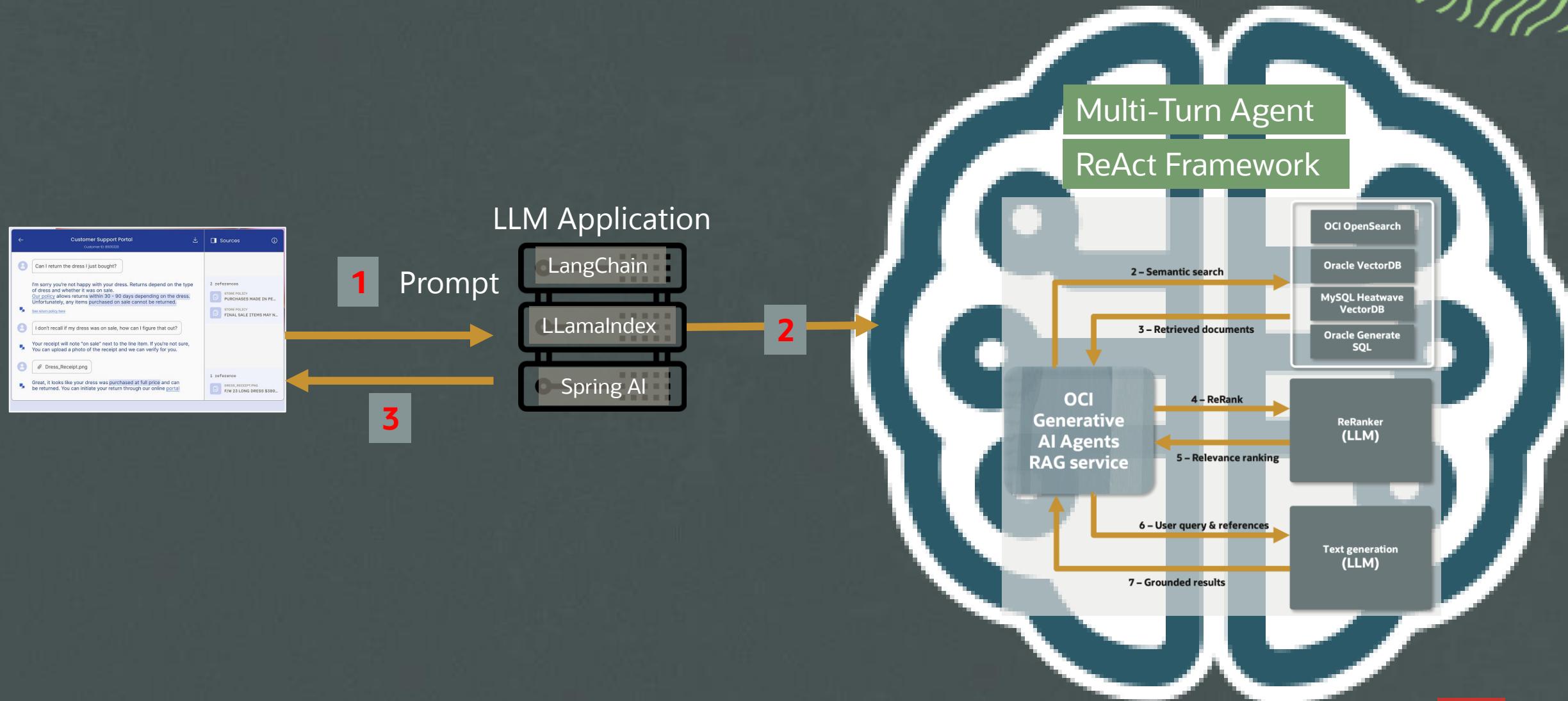


The screenshot shows a Oracle SQL Worksheet interface. The title bar indicates '7.671 seconds'. The 'Worksheet' tab is selected. The code area contains the following AI-generated SQL query:

```
1 select ai
2 what are our top 10 streamed movies
3 that were released after 2010;
```

A cursor arrow is visible in the bottom right corner of the code area.

# RAG(Retrieval Augmented Generation) 아키텍처



# Summary: Oracle Generative AI & RAG 아키텍처



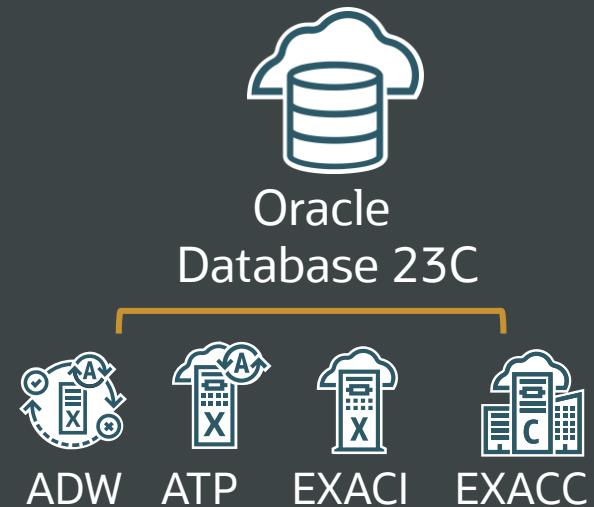
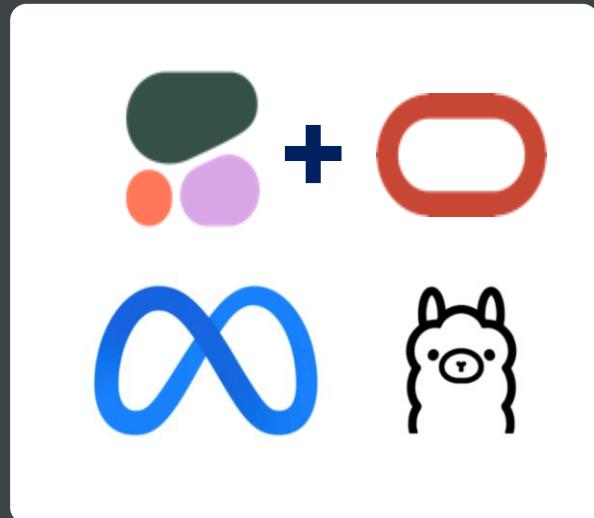
Model



DATA



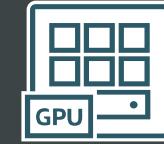
INFRASTRUCTURE



OpenSearch



Heatwave



H100/A100/A10 GPU



Cloud Native OKE



Super Cluster



Chatbot

## Oracle Enterprise LLM/Gen AI Platform



# 감사합니다.

---

김태완 상무, 한국오라클  
[taewan.kim@oracle.com](mailto:taewan.kim@oracle.com)