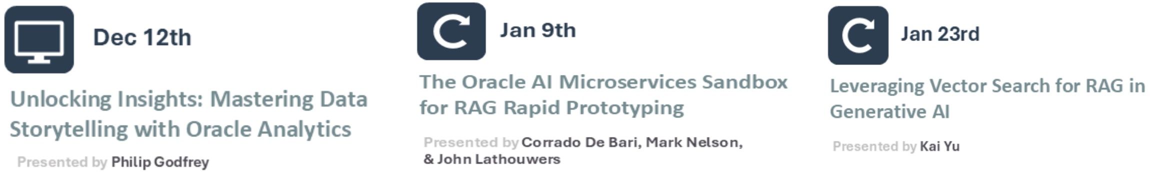
The Oracle AI Microservices Sandbox for RAG Rapid Prototyping

Corrado De Bari, Mark Nelson, John Lathouwers

Copyright © 2025, AnDOUC. All rights reserved.

Future & Past TechCasts:



Dec 12th

Unlocking Insights: Mastering Data Storytelling with Oracle Analytics

Presented by **Philip Godfrey**

Jan 9th

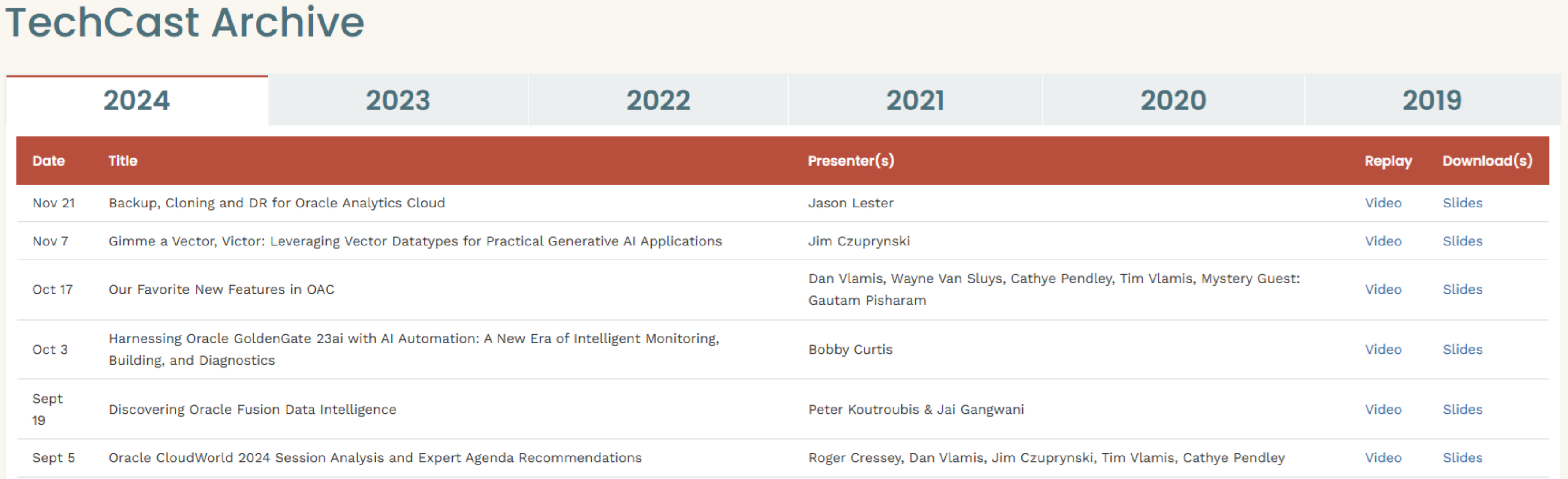
The Oracle AI Microservices Sandbox for RAG Rapid Prototyping

Presented by **Corrado De Bari, Mark Nelson, & John Lathouwers**

Jan 23rd

Leveraging Vector Search for RAG in Generative AI

Presented by **Kai Yu**



TechCast Archive

2024 | 2023 | 2022 | 2021 | 2020 | 2019

Date	Title	Presenter(s)	Replay	Download(s)
Nov 21	Backup, Cloning and DR for Oracle Analytics Cloud	Jason Lester	Video	Slides
Nov 7	Gimme a Vector, Victor: Leveraging Vector Datatypes for Practical Generative AI Applications	Jim Czuprynski	Video	Slides
Oct 17	Our Favorite New Features in OAC	Dan Vlamis, Wayne Van Sluys, Cathye Pendley, Tim Vlamis, Mystery Guest: Gautam Pisharam	Video	Slides
Oct 3	Harnessing Oracle GoldenGate 23ai with AI Automation: A New Era of Intelligent Monitoring, Building, and Diagnostics	Bobby Curtis	Video	Slides
Sept 19	Discovering Oracle Fusion Data Intelligence	Peter Koutroubis & Jai Gangwani	Video	Slides
Sept 5	Oracle CloudWorld 2024 Session Analysis and Expert Agenda Recommendations	Roger Cressey, Dan Vlamis, Jim Czuprynski, Tim Vlamis, Cathye Pendley	Video	Slides

Submit a topic to share at **https://andouc.org/techcasts/**

Analytics and Data
ORACLE USER COMMUNITY

Let's Connect

Website
http://andouc.org/

Chat with the Experts
https://bit.ly/Join-ANDOUC-Slack

Watch Previous TechCasts
https://bit.ly/3qmGgHN

@AnalyticAndData

https://www.facebook.com/AnDOracleUserCommunity

https://www.linkedin.com/company/analytics-and-data-oracle-user-community

Spatial + Graph SIG
bit.ly/Spatial-Graph-LinkedIn

ANALYTICS AND DATA
TechCasts

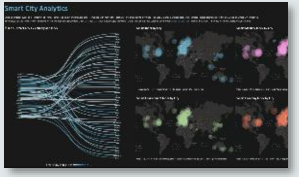
The Oracle Analytics

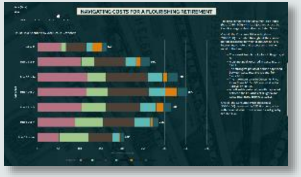
Data Visualization Challenge 2025

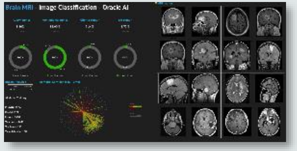
Get recognized for your Oracle Analytics skills!

Running February 3rd – 28th 2025

Registration open NOW







Analytics and Data
ORACLE USER COMMUNITY

Save the Date

Analytics and Data Summit 2025

*Registration is now **open**!*

April 8-10, 2025
Oracle Conference Center
Redwood Shores, California

https://andouc.org/andsummit2025/

ANALYTICS AND DATA
TechCasts

ORACLE

The Oracle AI Microservices Sandbox for RAG Rapid Prototyping

Mark Nelson, John Lathouwers, Corrado De Bari

Developer Evangelists
Oracle

Copyright © 2025, Oracle and/or its affiliates

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.

Copyright © 2025, Oracle and/or its affiliates

Everybody is building chatbots
But what defines enterprise quality?

Copyright © 2025, Oracle and/or its affiliates

The fallacy of the “quick” chatbot

Many articles present a simple chatbot that can be created very quickly but these are not production-ready – why not?

Copyright © 2025, Oracle and/or its affiliates

What are some common challenges?

- Choosing the right model
- Context window size
- The quality of the corpus – structure, content, formatting, size, presence of distractors
- The preciseness or ambiguousness of the question
- The quality of the prompt
- Domain-specific specialized language not seen in model pre-training
- Dealing with information after the model pre-training was completed

There are some aspects of working with LLMs that are extremely expensive (in terms of time and/or money)

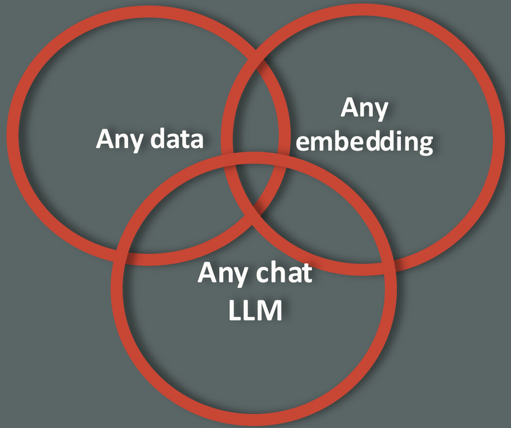
- Testing/evaluating model performance with human feedback or human-created datasets
- Fine tuning models
- Creating embeddings for large corpora

LLMs tend to "make up" answers if they don't know the answer – formally called "hallucination"

Copyright © 2025, Oracle and/or its affiliates

AI Microservices Sandbox

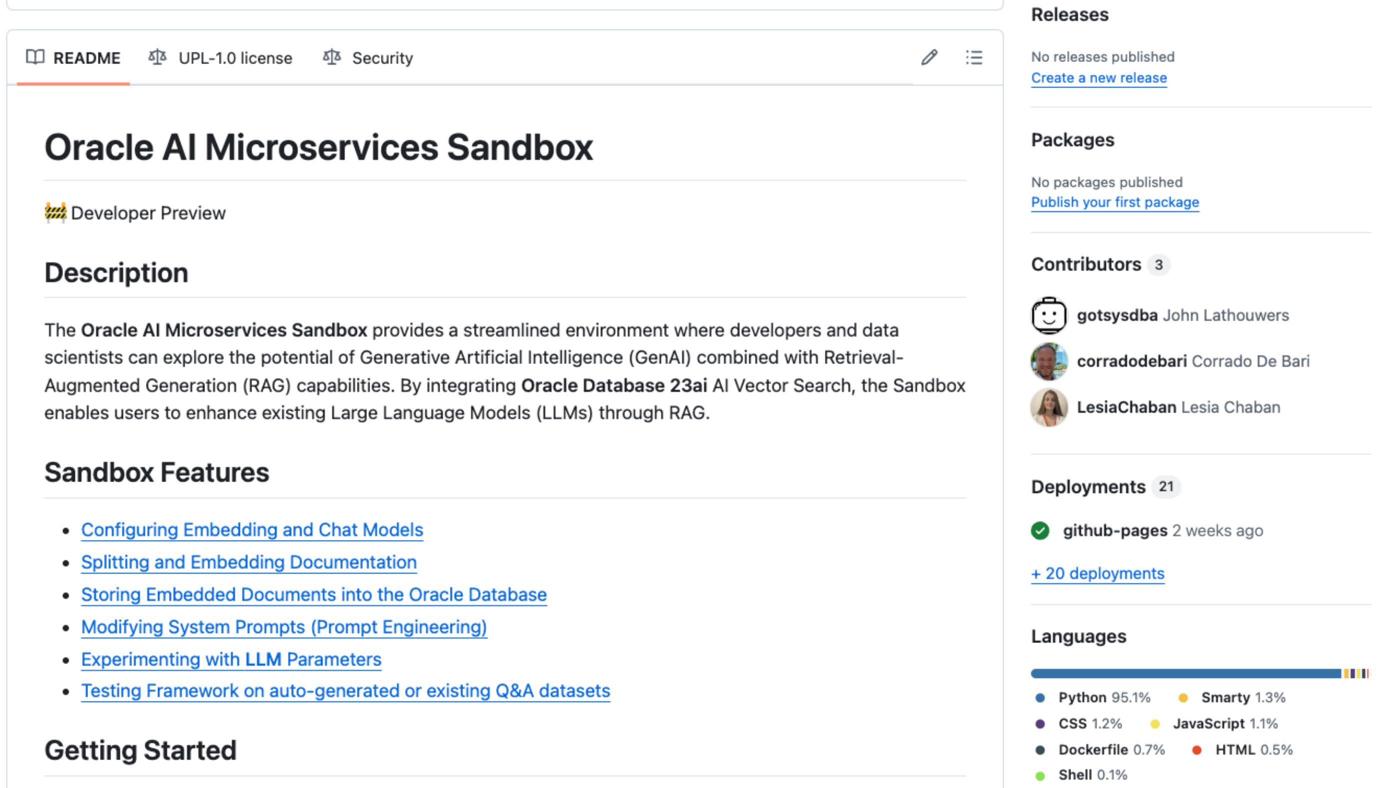
A Zero-Code platform to build, test & deploy a RAG Chatbot with:



The most important success factor in developing a real chatbot is ***experimentation*** and ***iteration.***

Copyright © 2025, Oracle and/or its affiliates

Oracle AI Microservices Sandbox



Oracle AI Microservices Sandbox

🚧 Developer Preview

Description

The **Oracle AI Microservices Sandbox** provides a streamlined environment where developers and data scientists can explore the potential of Generative Artificial Intelligence (GenAI) combined with Retrieval-Augmented Generation (RAG) capabilities. By integrating **Oracle Database 23ai** AI Vector Search, the Sandbox enables users to enhance existing Large Language Models (LLMs) through RAG.

Sandbox Features

- Configuring Embedding and Chat Models
- Splitting and Embedding Documentation
- Storing Embedded Documents into the Oracle Database
- Modifying System Prompts (Prompt Engineering)
- Experimenting with LLM Parameters
- Testing Framework on auto-generated or existing Q&A datasets

Getting Started

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

gotsysdba John Lathouwers
corradodebari Corrado De Bari
LesiaChaban Lesia Chaban

Deployments 21

✓ **github-pages** 2 weeks ago

+ 20 deployments

Languages

● **Python** 95.1%
● **Smarty** 1.3%
● **CSS** 1.2%
● **JavaScript** 1.1%
● **Dockerfile** 0.7%
● **HTML** 0.5%
● **Shell** 0.1%

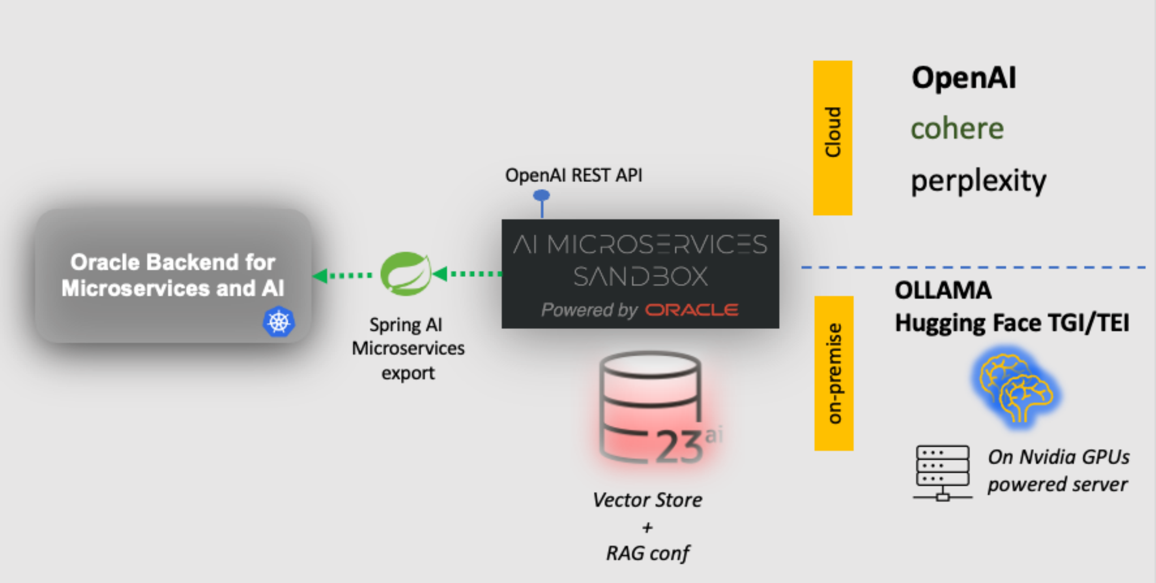
Open Source Project
in
Developer Preview

https://github.com/oracle-samples/oaim-sandbox

Copyright © 2025, Oracle and/or its affiliates

Building a RAG+ Chatbot

- **Playground env for RAG agent configuration: 0-Code approach**
- **Any LLM model (embeddings/chat):**
  - **PRIVATE**: On-premises with OLLAMA, HF TEI, HF TGI
  - **PUBLIC**: service providers like Cohere, OpenAI, Perplexity
- **Vector Store ingestion for DB23ai**:
  - Split & Model configuration
  - Documents: HTML, Txt, Pdf docs
  - From: local/OCI Object Storage/Web URL
- **Prompt configuration**
- **Test framework**:
  - Auto Generation for massive tests
  - Self Evalution
- **REST API Server** (OpenAI API compatible)
- **Export as SpringAI Microservice**



Copyright © 2025, Oracle and/or its affiliates

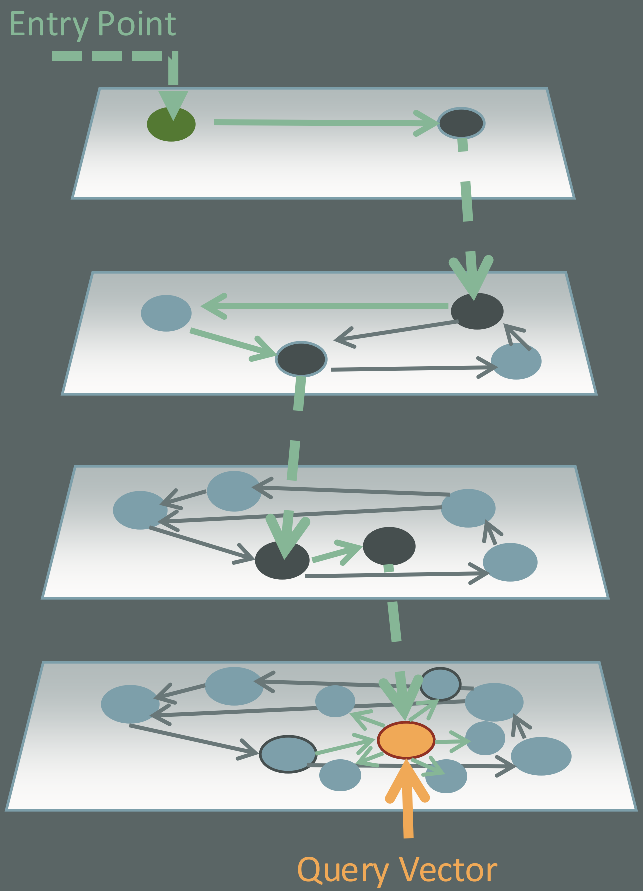
The Oracle DB 23ai as knowledge base:

NEW vector indexes provide 100x faster and highly accurate similarity searches

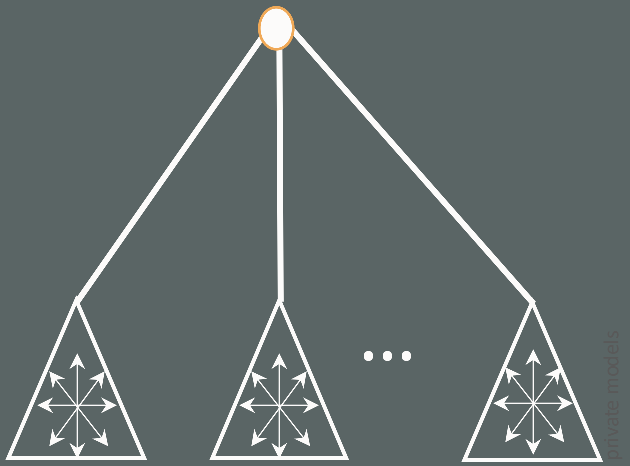
private models

Copyright © 2025, Oracle and/or its affiliates

Oracle database accelerates AI vector search using sophisticated vector indexes



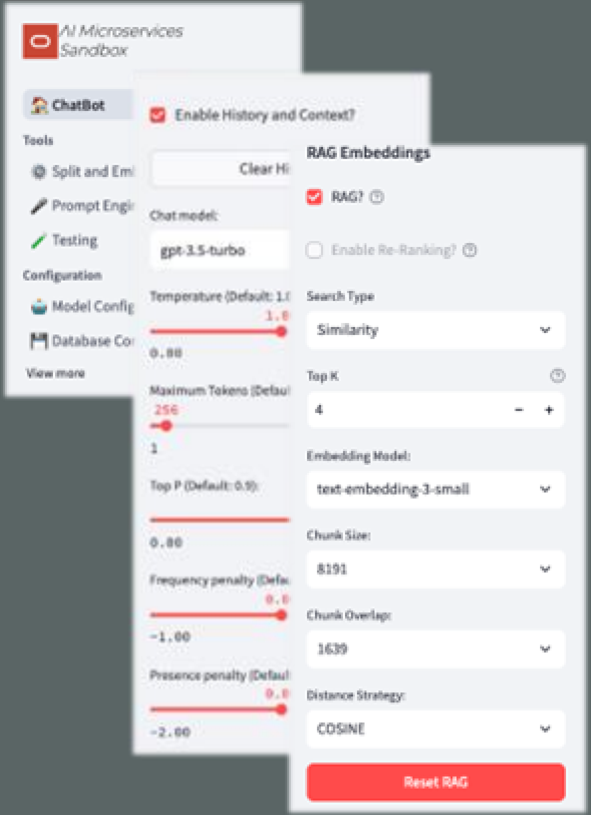
Graph Vector Index (e.g., **HNSW** Index)



Partition Vector Index (e.g., **IVF_FLAT** index)

Copyright © 2025, Oracle and/or its affiliates

Experiment a Chatbot Application (in Sandbox)



☑ Enable History and Context?

Chat model: gpt-3.5-turbo

Temperature (Default: 1.0): 1.0

Maximum Tokens (Defau: 256

Top P (Default: 0.9):

Frequency penalty (Defa: 0.0

Presence penalty (Defaul: 0.0

☑ RAG?

☐ Enable Re-Ranking?

Search Type: Similarity

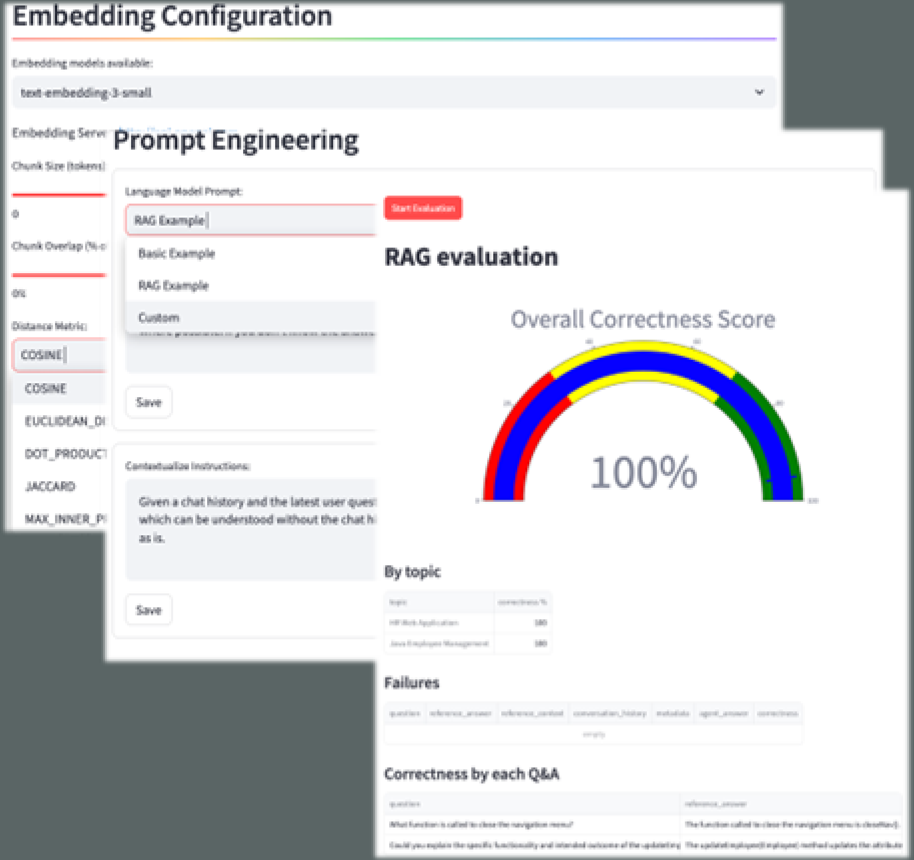
Top K: 4

Embedding Model: text-embedding-3-small

Chunk Size: 8191

Chunk Overlap: 1639

Distance Strategy: COSINE



Embedding models available: text-embedding-3-small

Distance Metric: COSINE

Language Model Prompt: RAG Example

Contextualize Instructions:

Given a chat history and the latest user ques which can be understood without the chat h as is.

By topic

topic	correctness %
HR Web Application	100
Java Employee Management	100

Failures

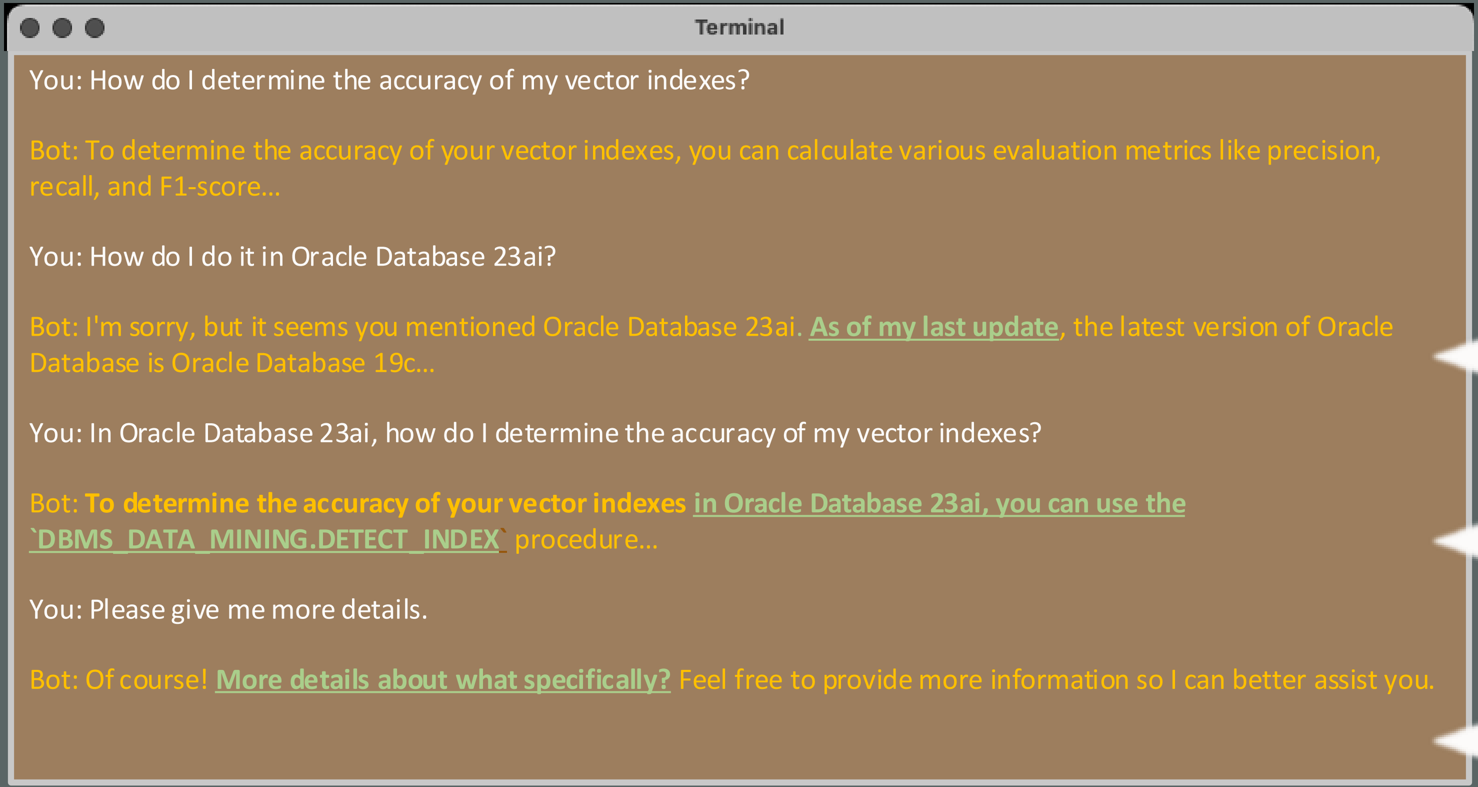
question	reference_answer	reference_context	conversation_history	metadata	agent_answer	correctness
empty						

Correctness by each Q&A

question	reference_answer
What function is called to close the navigation menu?	The function called to close the navigation menu is closeNav().

Copyright © 2025, Oracle and/or its affiliates

An Example of improving a Chatbot application (Before)



```
You: How do I determine the accuracy of my vector indexes?

Bot: To determine the accuracy of your vector indexes, you can calculate various evaluation metrics like precision,
recall, and F1-score...

You: How do I do it in Oracle Database 23ai?

Bot: I'm sorry, but it seems you mentioned Oracle Database 23ai. As of my last update, the latest version of Oracle
Database is Oracle Database 19c...

You: In Oracle Database 23ai, how do I determine the accuracy of my vector indexes?

Bot: To determine the accuracy of your vector indexes in Oracle Database 23ai, you can use the
`DBMS_DATA_MINING.DETECT_INDEX` procedure...

You: Please give me more details.

Bot: Of course! More details about what specifically? Feel free to provide more information so I can better assist you.
```

Knowledge Cutoff

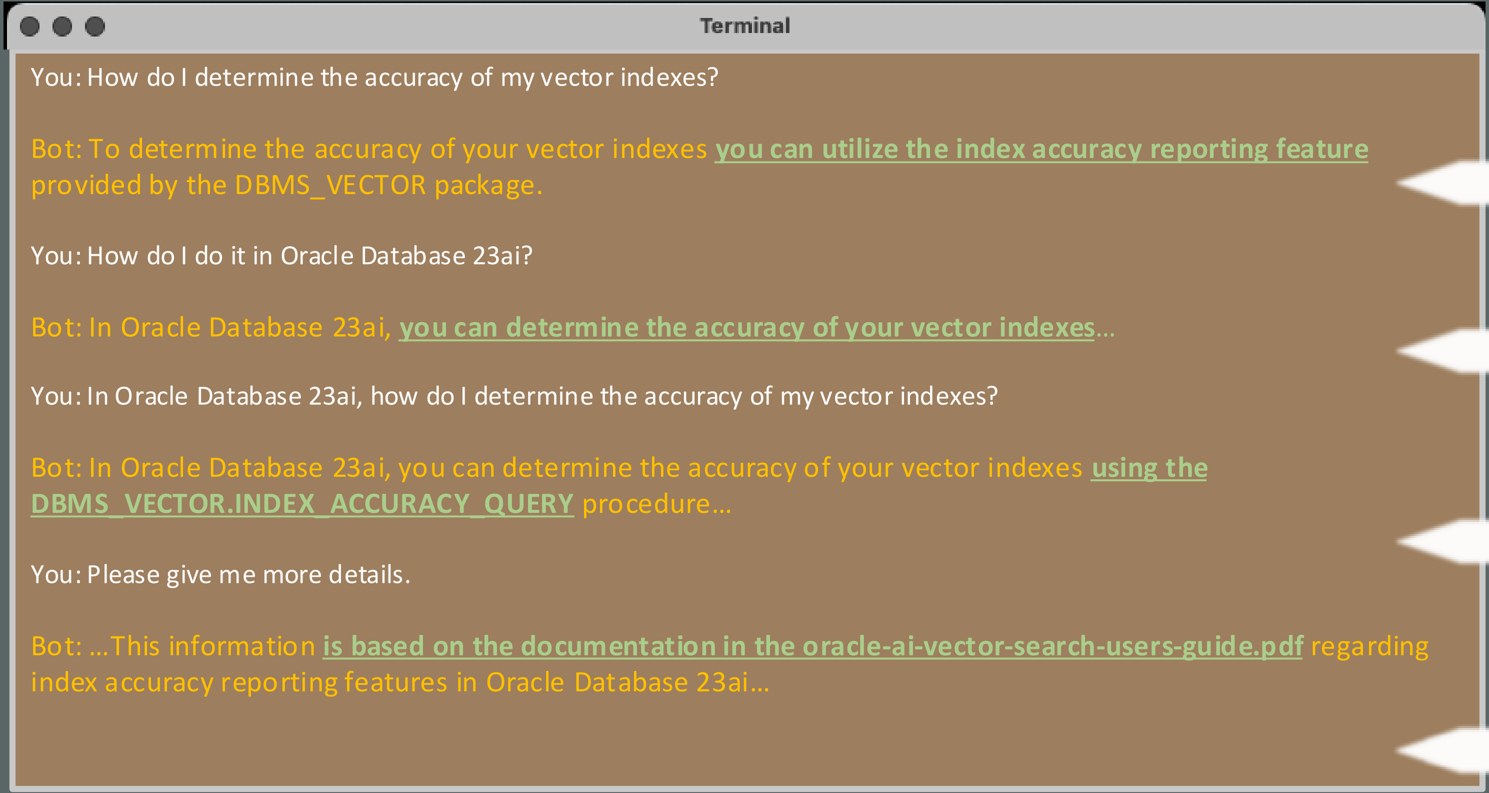
Hallucinations

No Memory

gpt-3.5-turbo

Copyright © 2025, Oracle and/or its affiliates

Experiment a Chatbot Application (in Sandbox)



```
You: How do I determine the accuracy of my vector indexes?

Bot: To determine the accuracy of your vector indexes you can utilize the index accuracy reporting feature
provided by the DBMS_VECTOR package.

You: How do I do it in Oracle Database 23ai?

Bot: In Oracle Database 23ai, you can determine the accuracy of your vector indexes…

You: In Oracle Database 23ai, how do I determine the accuracy of my vector indexes?

Bot: In Oracle Database 23ai, you can determine the accuracy of your vector indexes using the
DBMS_VECTOR.INDEX_ACCURACY_QUERY procedure…

You: Please give me more details.

Bot: …This information is based on the documentation in the oracle-ai-vector-search-users-guide.pdf regarding
index accuracy reporting features in Oracle Database 23ai…
```

Live Knowledge

Memory/Context

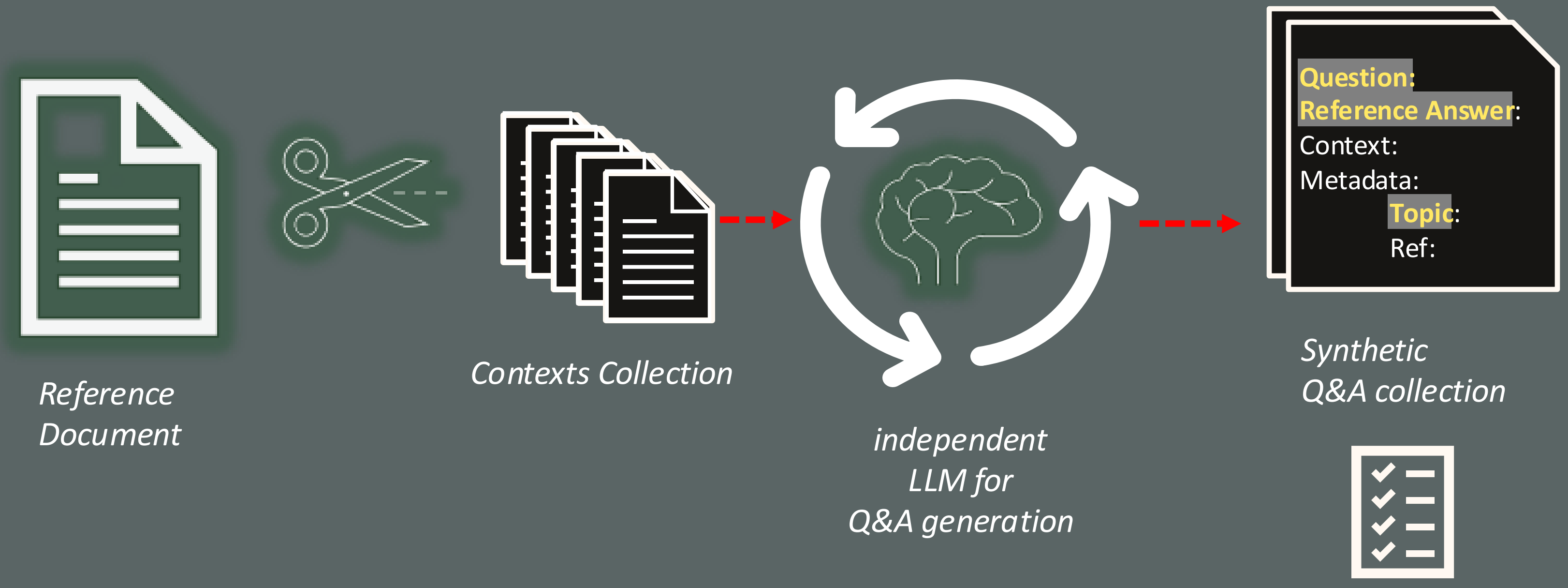
Grounded Response

RAG Attribution

Copyright © 2025, Oracle and/or its affiliates

RAG+ Testset Generation & Evaluation

Generation



Reference Document

Contexts Collection

independent LLM for Q&A generation

Question:
Reference Answer:
Context:
Metadata:
Topic:
Ref:

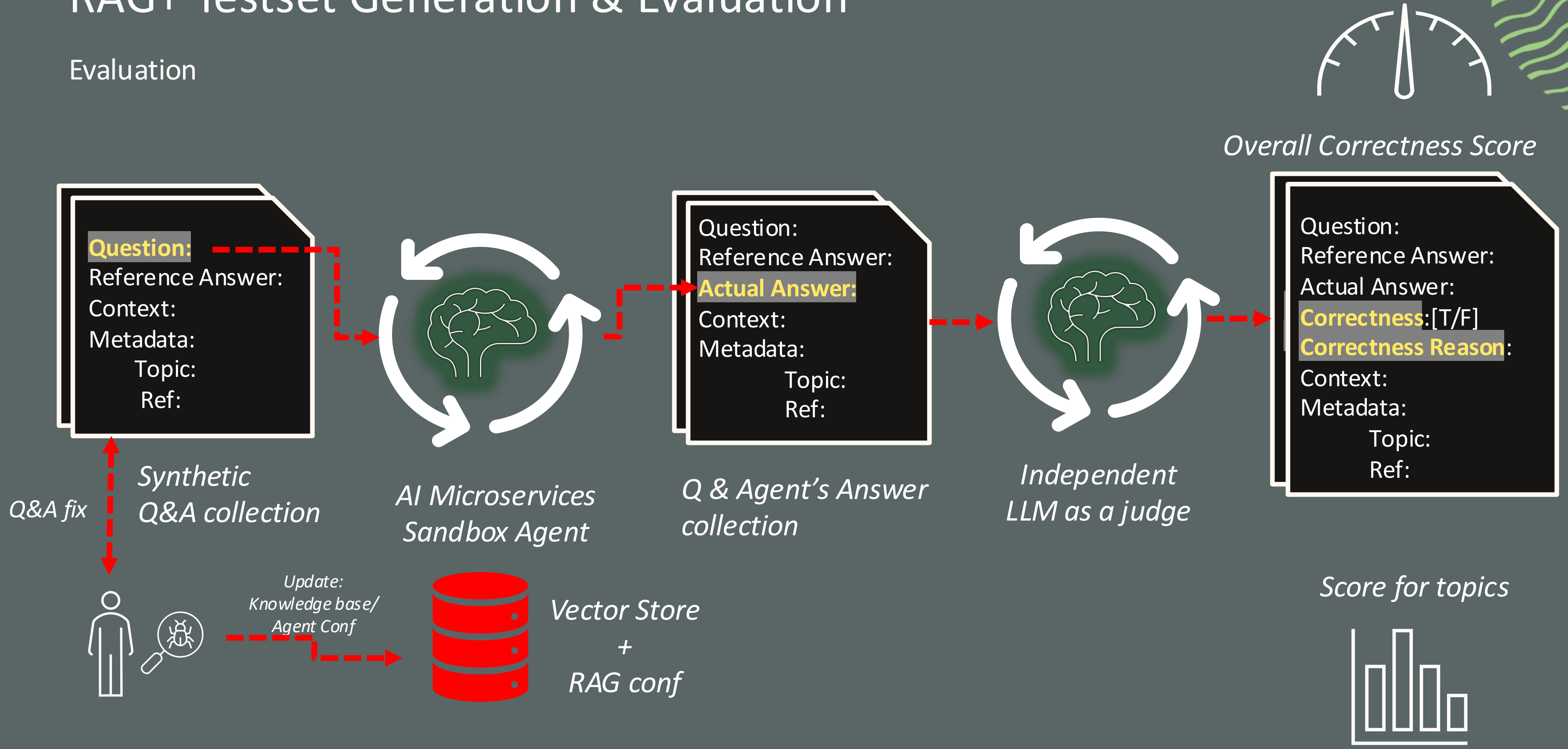
Synthetic Q&A collection

List of topics

Copyright © 2025, Oracle and/or its affiliates

RAG+ Testset Generation & Evaluation

Evaluation



Question:
Reference Answer:
Context:
Metadata:
Topic:
Ref:

Synthetic Q&A collection

Q&A fix

AI Microservices Sandbox Agent

Question:
Reference Answer:
Actual Answer:
Context:
Metadata:
Topic:
Ref:

Q & Agent's Answer collection

Independent LLM as a judge

Overall Correctness Score

Question:
Reference Answer:
Actual Answer:
Correctness:[T/F]
Correctness Reason:
Context:
Metadata:
Topic:
Ref:

Update: Knowledge base/ Agent Conf

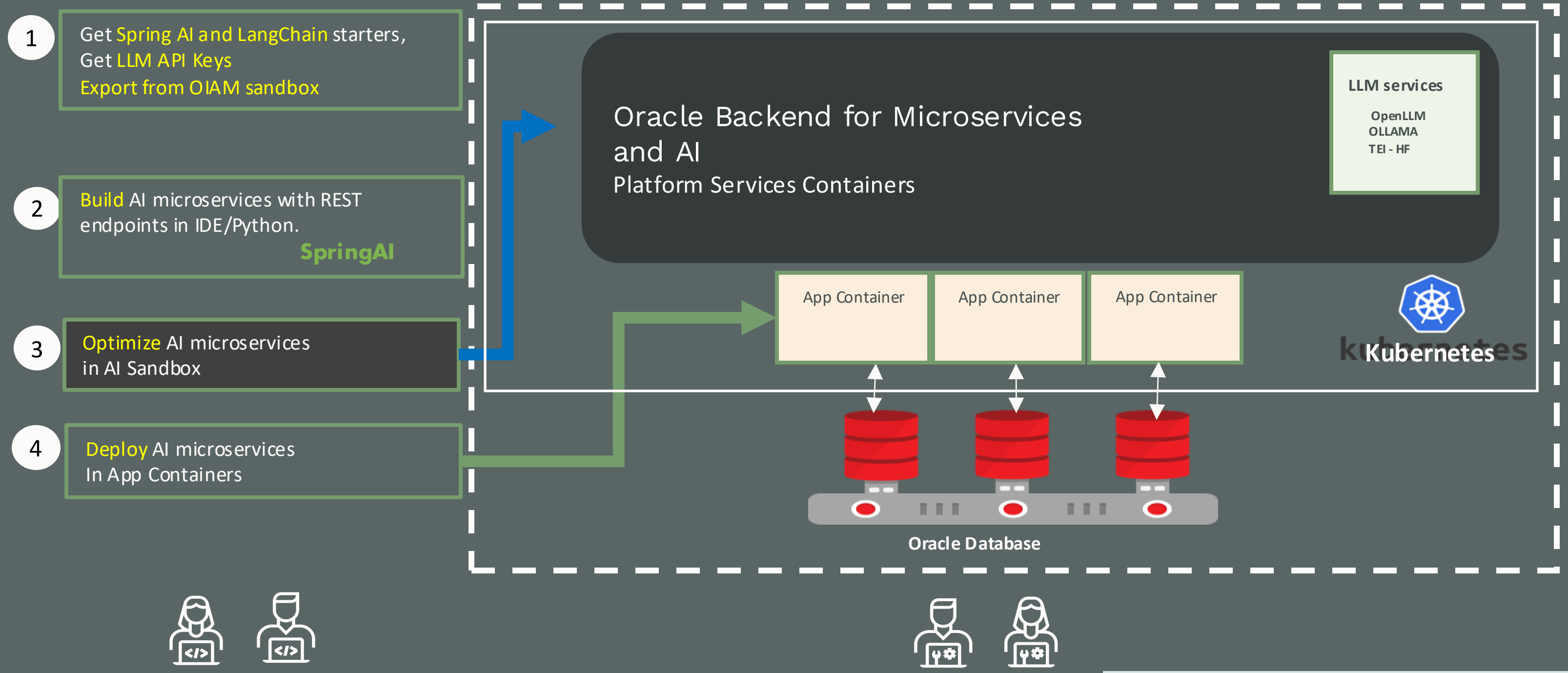
Vector Store + RAG conf

Score for topics

Copyright © 2025, Oracle and/or its affiliates

Oracle Backend for Microservices and AI v1.3

Working with AI in Microservices – Get Starters, Build, Optimize, & Deploy



1. Get Spring AI and LangChain starters,
Get LLM API Keys
Export from OIAM sandbox

2. Build AI microservices with REST endpoints in IDE/Python.
SpringAI

3. Optimize AI microservices
in AI Sandbox

4. Deploy AI microservices
In App Containers

AI Developers

DevOps

Available **free: on-premise, OCI and Azure.**

support through an Oracle DBMS contract

Copyright © 2025, Oracle and/or its affiliates

Summary

- Rag platform for GenAI Chatbot building
- Zero-code
- Base on Oracle DB 23ai high scalable Vectorstore
- Any LLM for Embeddings & Chat
- Well suitable for Private and/or On-premises deployement
- Free and Open Source

Copyright © 2025, Oracle and/or its affiliates

It’s now time for Q&A

Got any questions?

Thank you!

Helpful Links –

ORACLE ANALYTICS VIDEOS:
https://www.youtube.com/@OracleAnalytics/videos

OAC SEPTEMBER NEW FEATURES VIDEOS BY ORACLE: https://bit.ly/OACSept24Features

OAC NEW FEATURES DOCUMENTATION BY ORACLE:
https://docs.oracle.com/en/cloud/paas/analytics-cloud/acswn/index.html#GUID-CFF90F44-BCEB-49EE-B40B-8D040F02D476

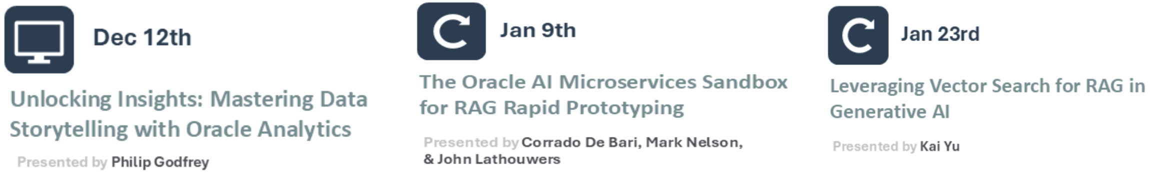
ORACLE ANALYTICS COMMUNITY:
https://community.oracle.com/products/oracleanalytics

ORACLE ANALYTICS LIBRARY/EXAMPLES:
https://www.oracle.com/business-analytics/data-visualization/examples/

ORACLE ANALYTICS LIVE DEMOS:
https://www.oracle.com/business-analytics/data-visualization/demos/

Copyright © 2025, AnDOUC. All rights reserved.

Future & Past TechCasts:



Dec 12th

Unlocking Insights: Mastering Data Storytelling with Oracle Analytics

Presented by **Philip Godfrey**

Jan 9th

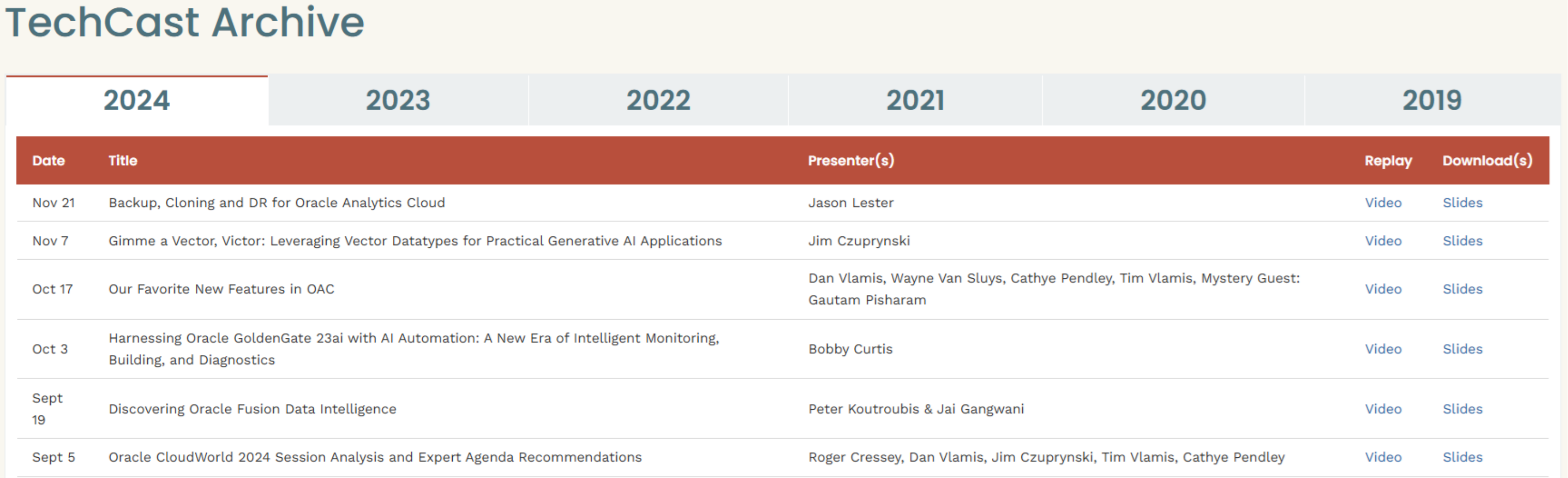
The Oracle AI Microservices Sandbox for RAG Rapid Prototyping

Presented by **Corrado De Bari, Mark Nelson, & John Lathouwers**

Jan 23rd

Leveraging Vector Search for RAG in Generative AI

Presented by **Kai Yu**



TechCast Archive

2024 | 2023 | 2022 | 2021 | 2020 | 2019

Date	Title	Presenter(s)	Replay	Download(s)
Nov 21	Backup, Cloning and DR for Oracle Analytics Cloud	Jason Lester	Video	Slides
Nov 7	Gimme a Vector, Victor: Leveraging Vector Datatypes for Practical Generative AI Applications	Jim Czuprynski	Video	Slides
Oct 17	Our Favorite New Features in OAC	Dan Vlamis, Wayne Van Sluys, Cathye Pendley, Tim Vlamis, Mystery Guest: Gautam Pisharam	Video	Slides
Oct 3	Harnessing Oracle GoldenGate 23ai with AI Automation: A New Era of Intelligent Monitoring, Building, and Diagnostics	Bobby Curtis	Video	Slides
Sept 19	Discovering Oracle Fusion Data Intelligence	Peter Koutroubis & Jai Gangwani	Video	Slides
Sept 5	Oracle CloudWorld 2024 Session Analysis and Expert Agenda Recommendations	Roger Cressey, Dan Vlamis, Jim Czuprynski, Tim Vlamis, Cathye Pendley	Video	Slides

Submit a topic to share at **https://andouc.org/techcasts/**

Analytics and Data
ORACLE USER COMMUNITY

Save the Date

Analytics and Data
Summit 2025

*Registration is now **open**!*

April 8-10, 2025
Oracle Conference Center
Redwood Shores, California

https://andouc.org/andsummit2025/

ANALYTICS AND DATA
TechCasts

Copyright © 2025, AnDOUC. All rights reserved.