Build your own RAG solution

Heli Helskyaho

Pekka Kanerva

Miracle Finland Oy

Miracle Finland Oy

Heli

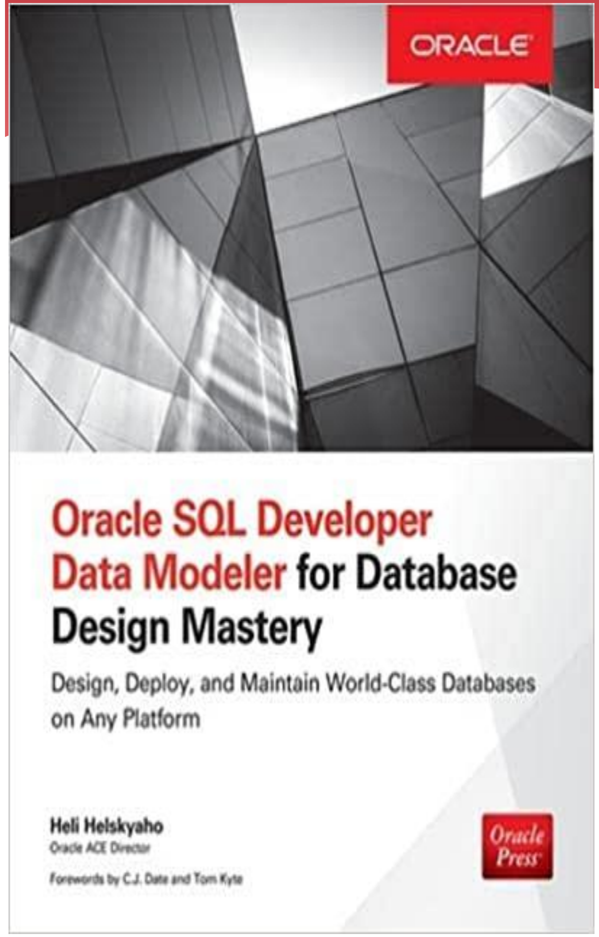
* Graduated from University of Helsinki (Master of Science, computer science), currently a doctoral student at University of Helsinki
* Worked with Oracle products since 1993, worked for IT since 1990
* Data and Database!
* CEO for Miracle Finland Oy
* Oracle ACE Director
* Public speaker and an author
* Author of the book Oracle SQL Developer Data Modeler for Database Design Mastery (Oracle Press, 2015), co-author for Real World SQL and PL/SQL: Advice from the Experts (Oracle Press, 2016), Machine Learning for Oracle Database Professionals: Deploying Model-Driven Applications and Automation Pipelines (Apress, 2021), and Extending Oracle Application Express with Oracle Cloud Features: A Guide to Enhancing APEX Web Applications with Cloud-Native and Machine Learning Technologies (Apress, 2022)

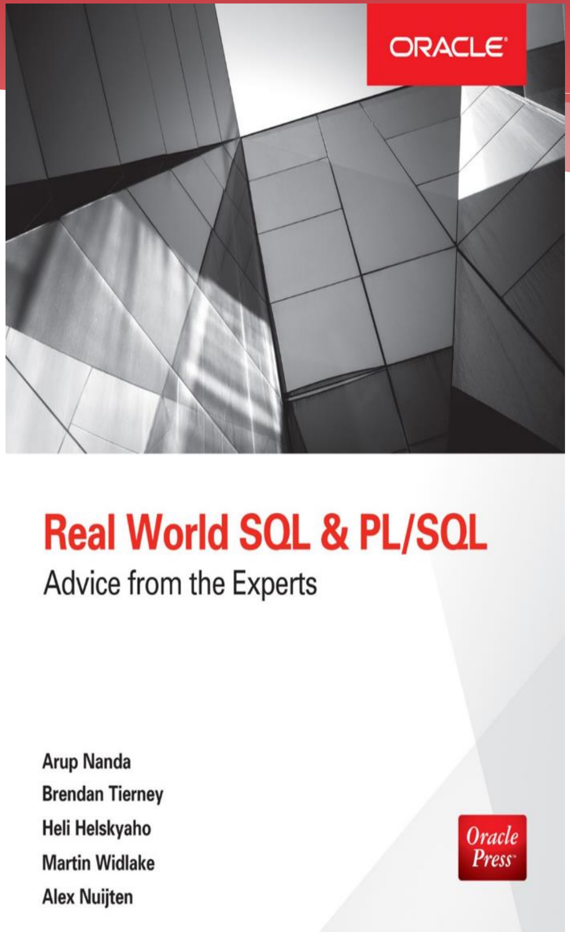
Oracle ACE Director

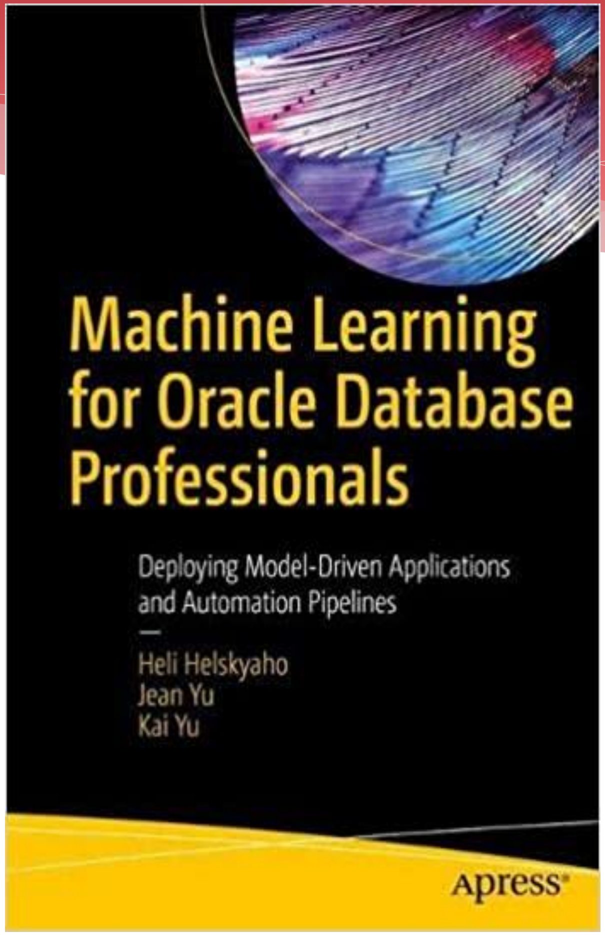
MIRACLE
Miracle Finland Oy

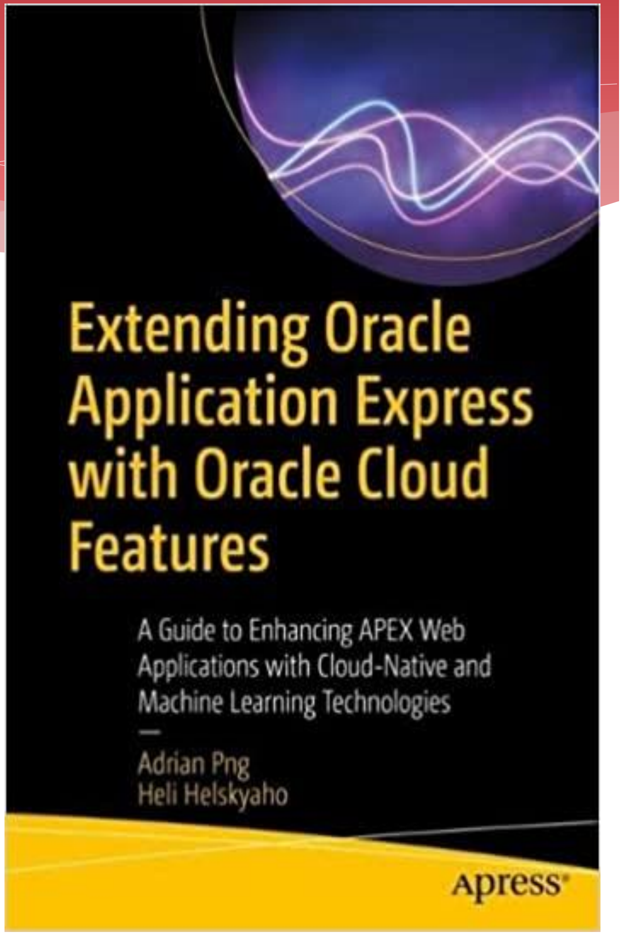
Copyright © Miracle Finland Oy

Books









Miracle Finland Oy

Copyright © Miracle Finland Oy

Pekka

* Pekka Kanerva is Oracle technology consultant working for Miracle Finland Oy
* Pekka holds a Master’s degree at the Helsinki University of Technology and he has been working with Oracle from the early 90'.
* He is specialized in Oracle technologies, application development with PL/SQL and APEX, and AI/ML Services in Oracle Cloud. Pekka has acquired several certificates on Oracle, Snowflake and Data Vault 2.0 technologies.

Copyright © Miracle Finland Oy

Generative AI

* Can help us
  * Automate tasks, create efficiency, better outcomes and money savings
  * Create possibilities that did not exist before
  * …

MIRACLE Miracle Finland Oy

Copyright © Miracle Finland Oy

Use cases (Tasks) are unlimited

* Content generation, augmentation
* Summarization
* Content personalization, customer segmentation
* Question answering, Conversations
* Virtual assistants
* Language style transferring, adjust the tone
* Creative writing, technical writing, articles, letters, emails,…
* Translations, localizations
* Feedback analysis, automated customer responses
* Code generation, error detection, debugging, code conversions to another language
* Code documentation, automated testing
* …

MiRACLE Miracle Finland Oy

Copyright © Miracle Finland Oy

Concerns about GenerativeAI

* Legal issues
  * Privacy
  * Security
  * Intellectual property rights (IPR), protection
  * Acts, laws,…
* Ethical issues
  * Bias
  * Misinformation
* …

Copyright © Miracle Finland Oy

Hallucination and old data

* When the LLM does not know (does not have the data for the answer), it invents the answer: hallucinates
* When was the foundation model trained? Long time ago? Years ago? Old data!
* The data used for model training might be erroneous (internet).
* How to check the answer is correct? Where did the model find the answer?
* Note: remember the ”job” of an LLM is to create content? So hallucination is actually exactly what we were expecting.

Miracle Finland Oy

Copyright © Miracle Finland Oy

What about my own data?

* Data in the database
* Data in our documentation
* Data in our knowledge base
* Data in our support system
* Data in our…

Copyright © Miracle Finland Oy

The solution?

* Retrieval-Augmented Generation, RAG?

Miracle Finland Oy

Copyright © Miracle Finland Oy

RAG

* Adds the **context**
* **Augments** the LLM with your **own data**
* Adding data **without** training
* **Grounds** the responses to your data
* Reduces **hallucination**

Copyright © Miracle Finland Oy

RAG

* Retrieval-Augmented Generation
* ”Retrieval”
  * the ability to retrive data from a data source; database, internet,…
* "Augmented Generation”
  * augmenting the generation process with retrieved data

Copyright © Miracle Finland Oy

Generative AI without the context (ChatGPT)

what is OML?

OML can refer to a few different things depending on the context:

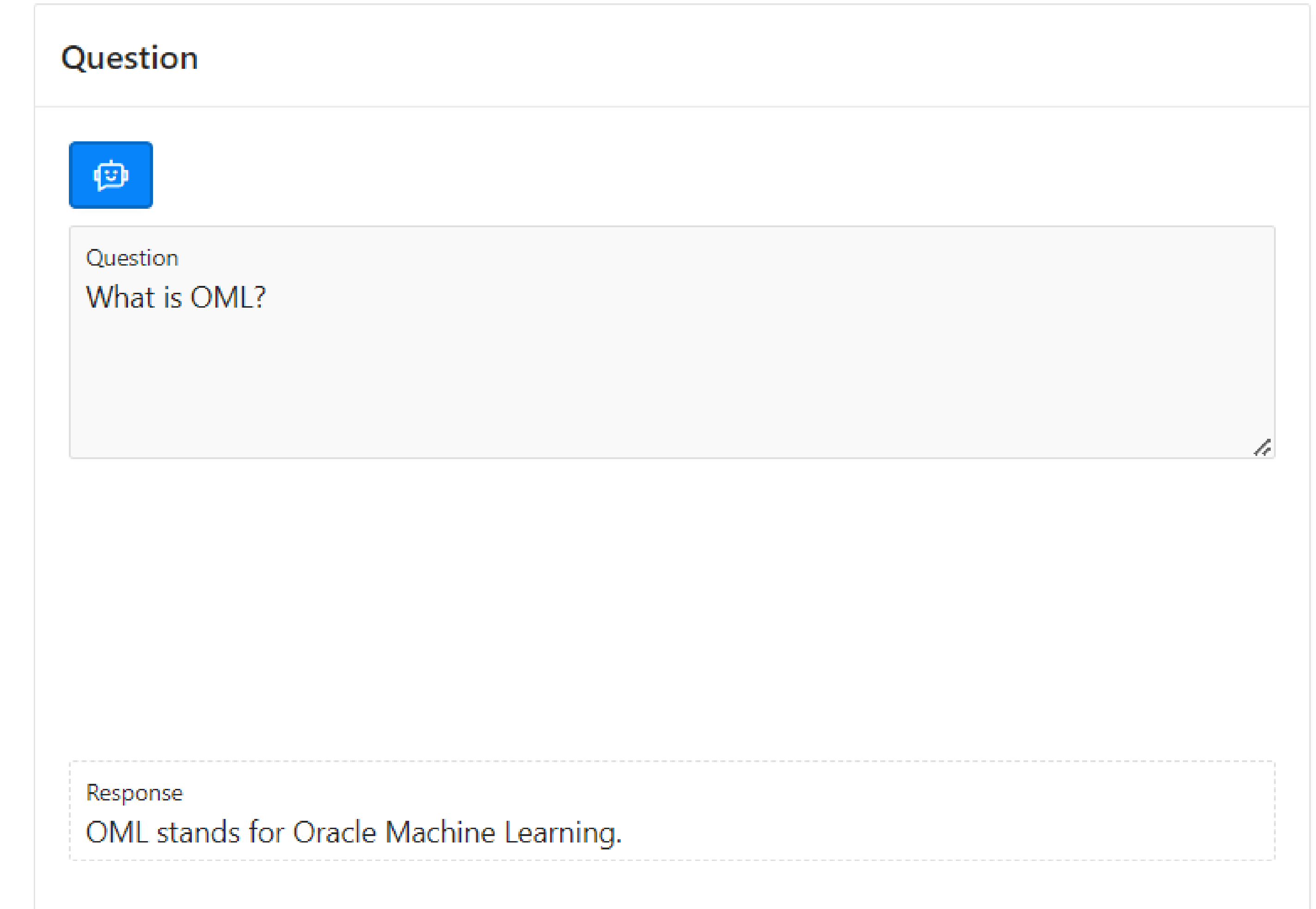
1. **Operational Machine Learning (OML)**: This refers to the practice of deploying, monitoring, and managing machine learning models in a production environment. It involves the entire lifecycle of machine learning, from model training and validation to deployment and ongoing maintenance. OML is important for ensuring that models perform well over time and adapt to new data.
2. **One More Level (OML)**: In gaming, "OML" is often used as an abbreviation for "One More Level," which refers to the addictive nature of some games that encourage players to keep playing "just one more level."
3. **Oracle Machine Learning (OML)**: This is a platform provided by Oracle that allows data scientists and developers to build, train, and deploy machine learning models within Oracle databases. It integrates machine learning capabilities directly into Oracle's database systems.
4. **Outright Monetary Transactions (OMT)**: In finance, "OML" might be confused with OMT, which refers to a program by the European Central Bank aimed at stabilizing bond markets in the Eurozone.
5. **Online Math Learning (OML)**: A platform or website that provides resources and tutorials for students to learn mathematics online.
6. **Oh My Lord (OML)**: A casual expression used in online conversations or social media to express surprise, excitement, or disbelief, similar to "Oh My God (OMG)."

The meaning of OML depends on the context in which it's used.

Copyright © Miracle Finland Oy

MiRACLE
Miracle Finland Oy

RAG solution



Miracle Finland Oy
Copyright © Miracle Finland Oy

Augmentation

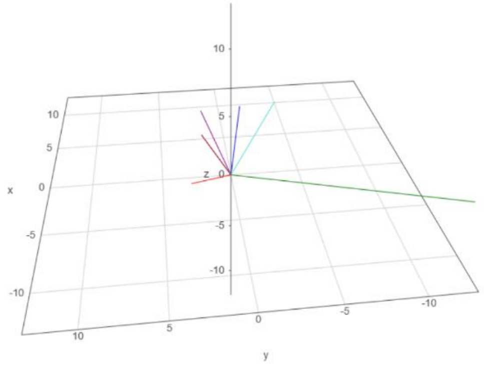
* The response can be augmented with data in the **prompt** (prompt engineering)
* Or in a **database**, **internet**, or any data source we can reach
* But if we want to augment it with, for example, by data from pdf documents or images, we need ***vectors***

Miracle Finland Oy

Copyright © Miracle Finland Oy

Vectors

* Unstructured data (text, image, voice, video,…) is transformed (encoded, embedded) into numeric representation and stored as vectors
* The data is stored as its *semantic content*, not the actual content, obtained by vectorizing



Copyright © Miracle Finland Oy

Distance and Similarity metrics

* Examples of these metrics
  * Euclidean (*EUCLIDEAN*) and Euclidean Squared Distances (*EUCLIDEAN_SQUARED or L2_SQUARED*)
  * Cosine Similarity (*COSINE*)
  * Dot Product Similarity (*DOT*)
  * Manhattan Distance (*MANHATTAN*)
  * Hamming Similarity (*HAMMING*)
* Used as a *vector distance operand* to the **VECTOR_DISTANCE** Function in the Oracle Database 23ai

Copyright © Miracle Finland Oy

Distance and Similarity metrics

* The metric to be chosen depends on the embedding model chosen!
* Use the distance/similarity metric that was used to train your embedding model.
* Documentation!

`embed-multilingual-v2.0`	multilingual classification and embedding support. See supported languages here.	768	256	Dot Product Similarity	Classify, Embed
`embed-english-v3.0`	A model that allows for text to be classified or turned into embeddings. English only.	1024	512	Cosine Similarity	Embed, Embed Jobs
`embed-english-light-v3.0`	A smaller, faster version of `embed-english-v3.0`. Almost as capable, but a lot faster. English only.	384	512	Cosine Similarity	Embed, Embed Jobs
`embed-multilingual-v3.0`	Provides multilingual classification and embedding support. See supported languages here.	1024	512	Cosine Similarity	Embed, Embed Jobs

Copyright © Miracle Finland Oy

Oracle Database Vector datatype

Define dimension and format.

Dimension: how many dimensions in a vector.
[1.1, 2.2, 3.3] has three dimensions.

Operations for using the new datatype.

```
CREATE TABLE t2 (
  v1  VECTOR,
  v2  VECTOR(384,   *),
  V3  VECTOR(768,   FLOAT32),
  V4  VECTOR(1024,  FLOAT64),
  V5  VECTOR(4096,  INT8),
  V6  VECTOR(*,     *)
);


 DESC t2;

 Name    Null?    Type
 ------------------------------
 V1               VECTOR(* , FLOAT32)
 V2               VECTOR(384 , *)
 V3               VECTOR(768 , FLOAT32)
 V4               VECTOR(1024 , FLOAT64)
 V5               VECTOR(4096 , INT8)
 V6               VECTOR(* , *)
```


Copyright © Miracle Finland Oy

A table with data and a vector

```
Create table MyText (
    TextID Number(16),
    TextClause (CLOB),
    Text_vector VECTOR);
```

Copyright © Miracle Finland Oy

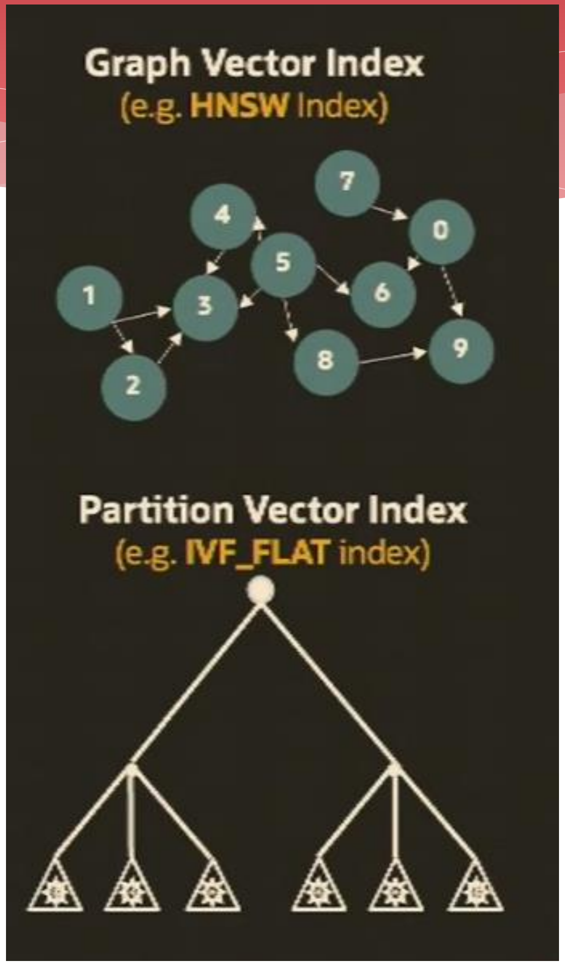
The RAG Process (simplified)

* **Generate vectors** for unstructured data using an embedding model
* **Save** vectors into the database in a column of VECTOR datatype
* Create Approximate Vector **Index** for the VECTOR column
* **Query** using AI Vector Search (SQL)

Copyright © Miracle Finland Oy

Approximate Vector Indexes

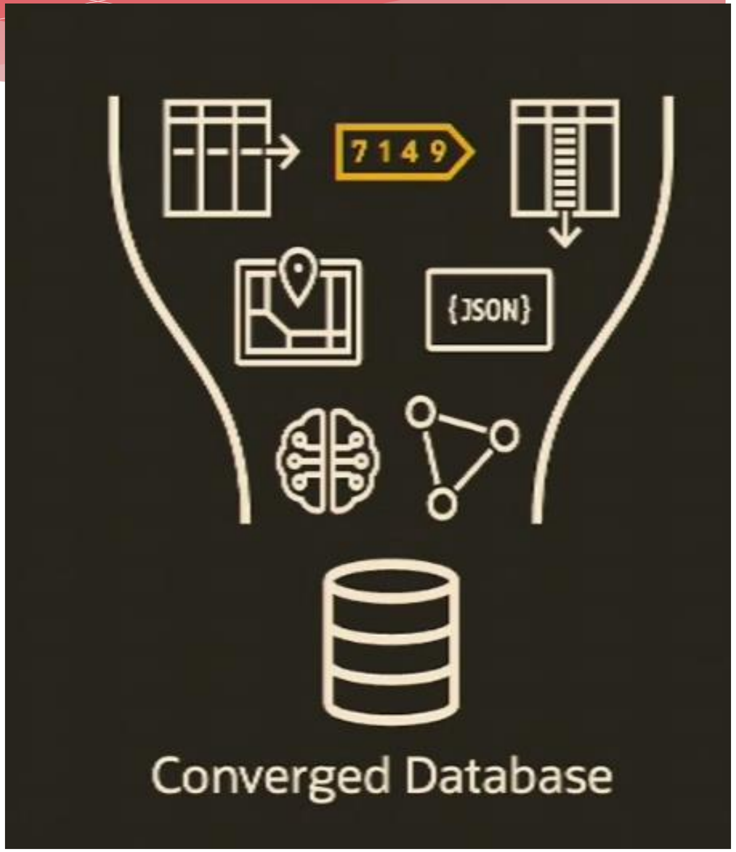
CREATE VECTOR INDEX text_idx ON Customer(text_vector)
ORGANIZATION [INMEMORY NEIGHBOR GRAPH | NEIGHBOR PARTITIONS]
DISTANCE EUCLIDEAN | COSINE_SIMILARITY | HAMMING ...



Copyright © Miracle Finland Oy

SQL joins (Authors, Books, Pages)

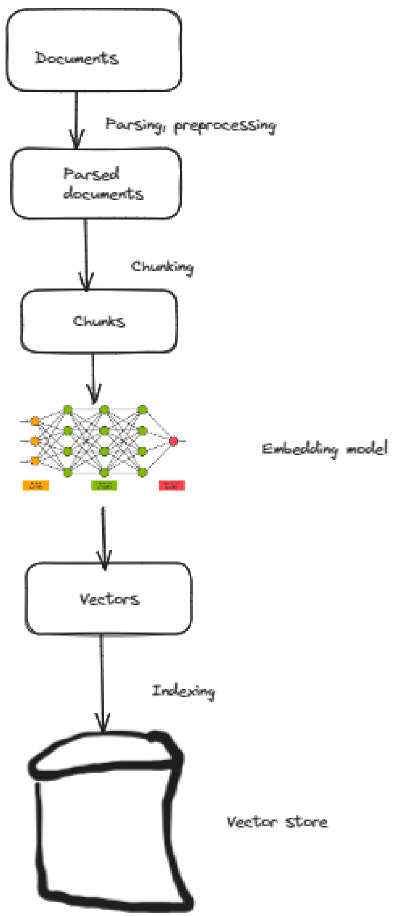
Select pageID
from Authors, Books, Pages
where Authors.authorID = Books.authorID and
Books.bookID = Pages.bookID and
Books.bookGenre = 'Fiction' and
Author.authorCountry = 'Finland'
order by vector_distance(pageVec, :queryVec)
fetch approx first 5 rows only;



MIRACLE
Miracle Finland Oy

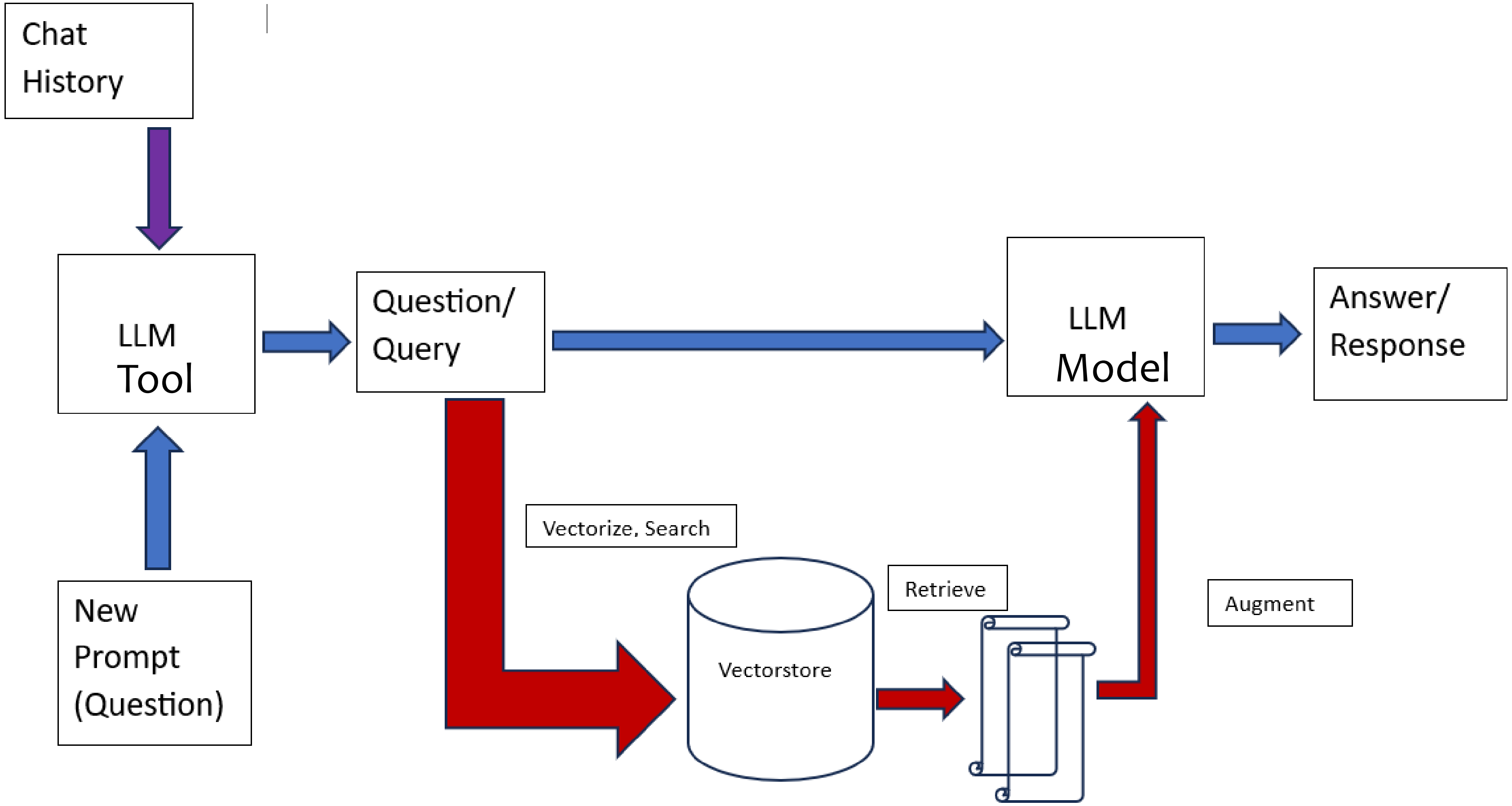
Copyright © Miracle Finland Oy

Storing Documents



Copyright © Miracle Finland Oy

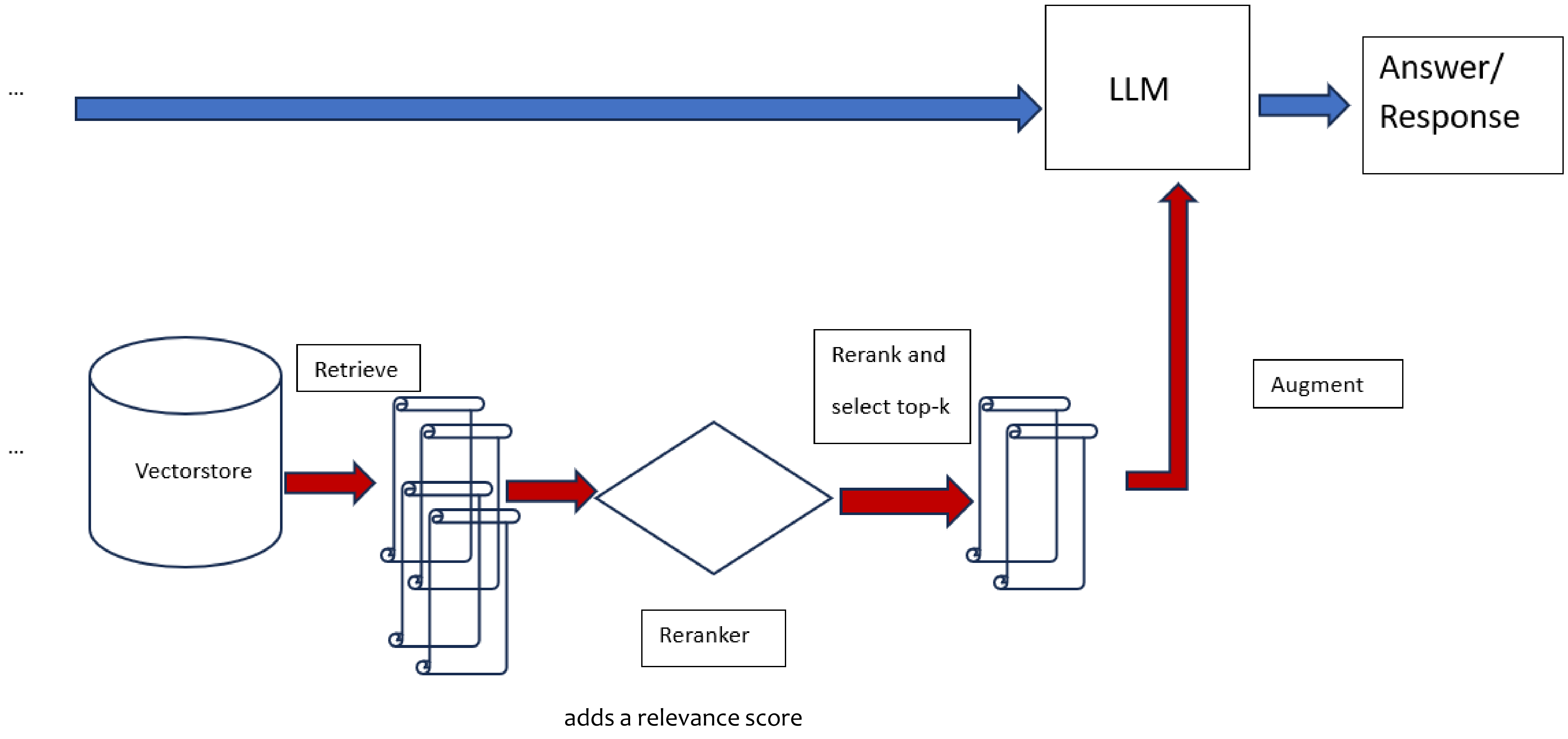
RAG



Miracle Finland Oy

Copyright © Miracle Finland Oy

Reranking

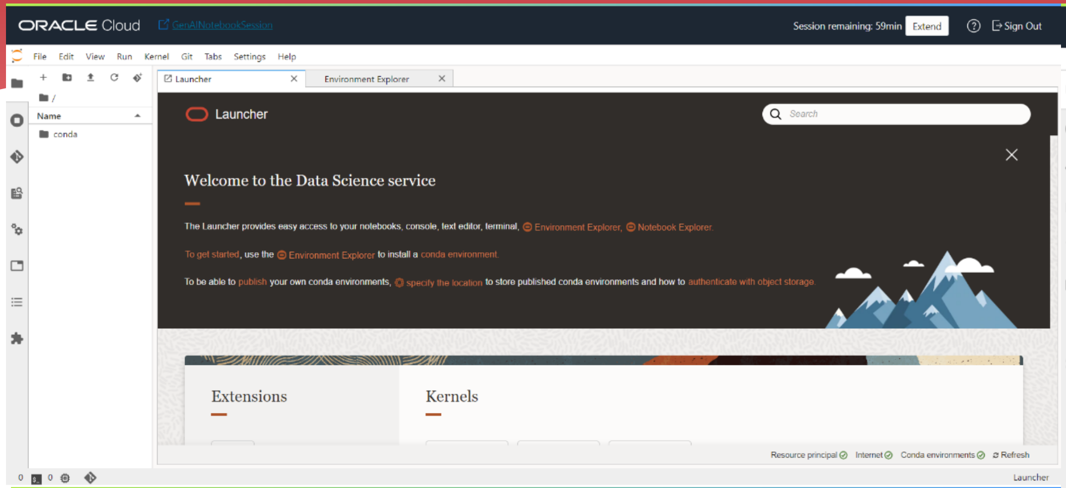


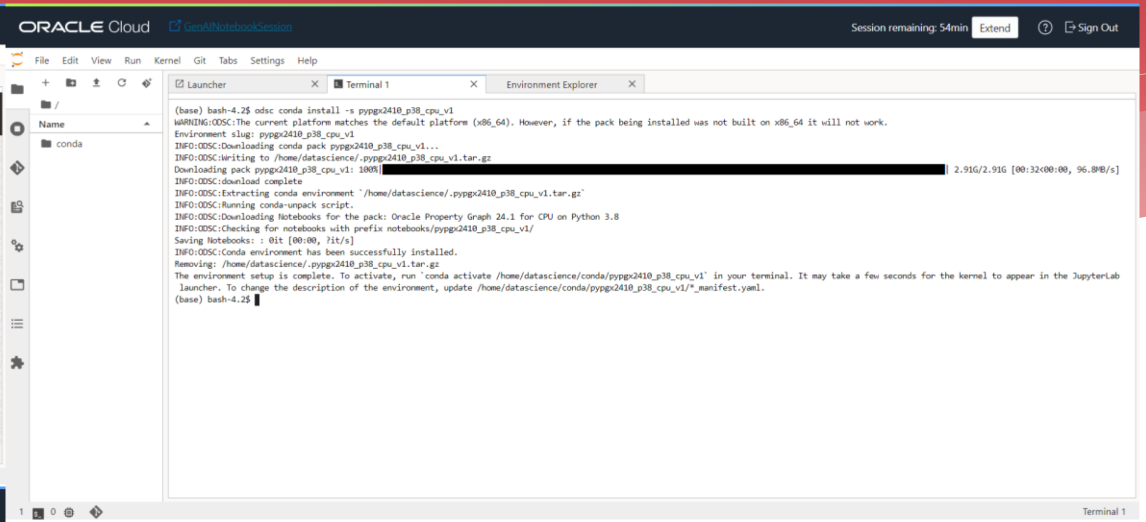
Copyright © Miracle Finland Oy

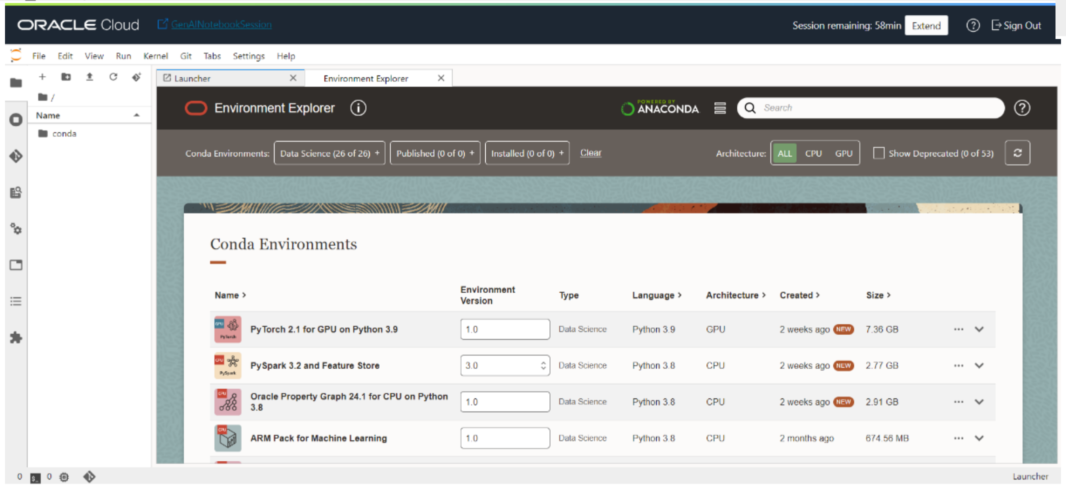
Oracle?

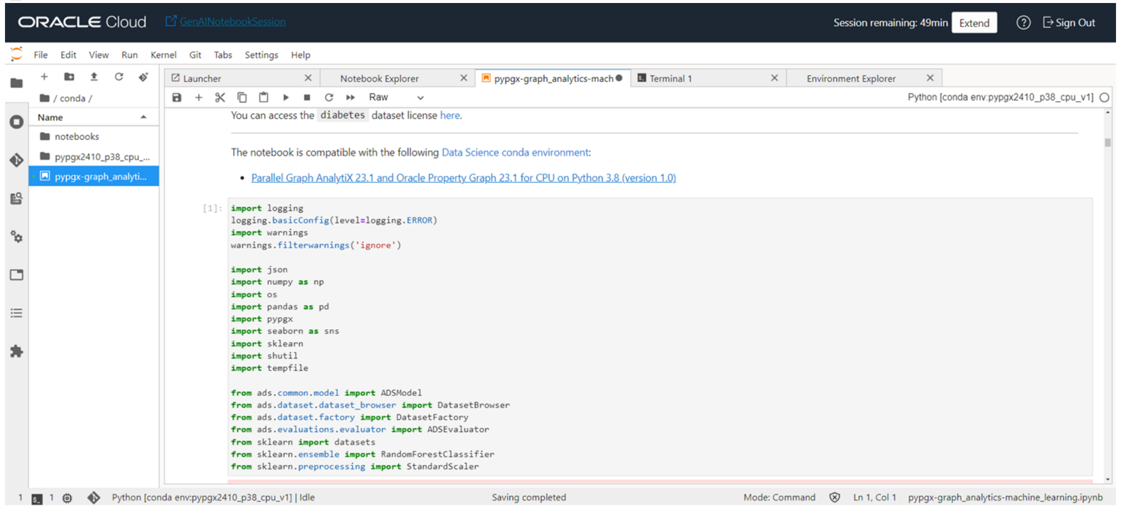
Copyright © Miracle Finland Oy

Data Science Service







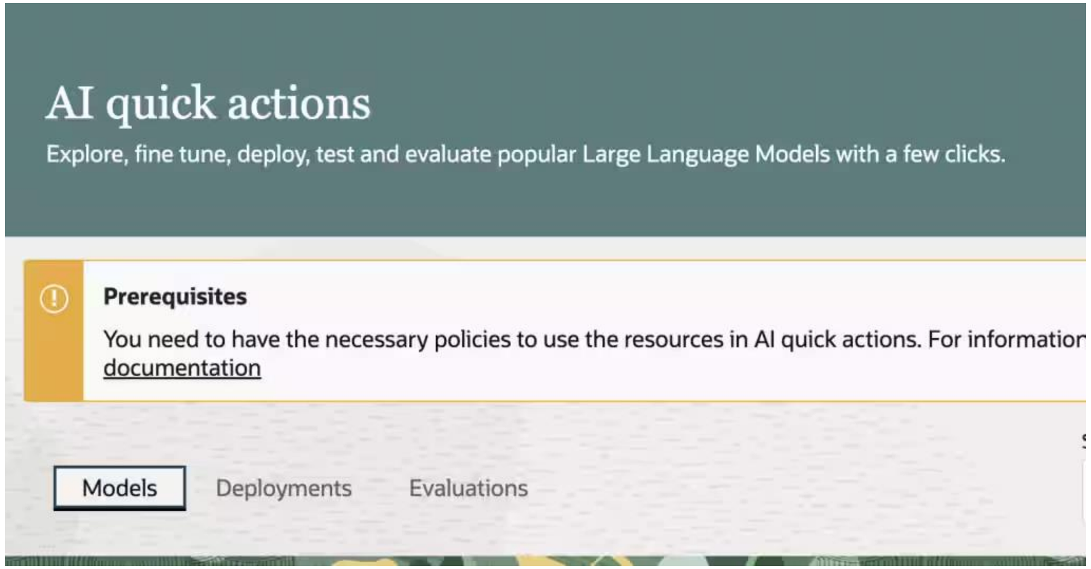


Copyright © Miracle Finland Oy

Miracle Finland Oy

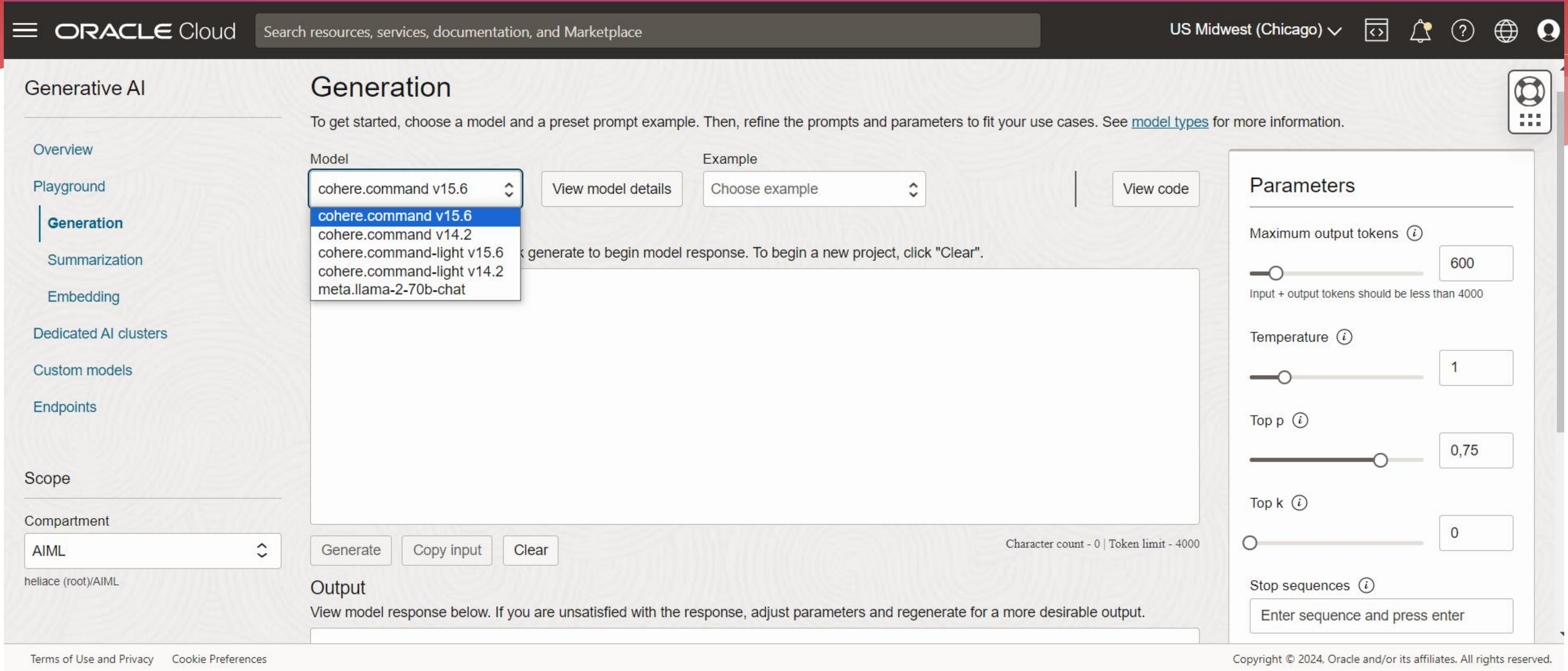
Data Science Service

* AI Quick Actions
  * *Deploying* a foundation model
  * *Fine-tuning* a foundation model
  * *Test* a foundation model
  * *Evaluating* a model
  * …



Copyright © Miracle Finland Oy

AI Service: Generative AI



https://blogs.oracle.com/ai-and-datascience/post/ga-oci-generative-ai

Copyright © Miracle Finland Oy

Oracle Database 23ai

* ONNX models and Oracle Machine Learning
* Vector datatype
* Approximate Vector Indexes
* PL/SQL Packages
* AI Vector Search
* Data dictionary views for DBMS_VECTOR
* Several data models (relational, Graph, Spatial, JSON, …)
* SQL query language to query all this data
* …

MiRACLE Miracle Finland Oy

Copyright © Miracle Finland Oy

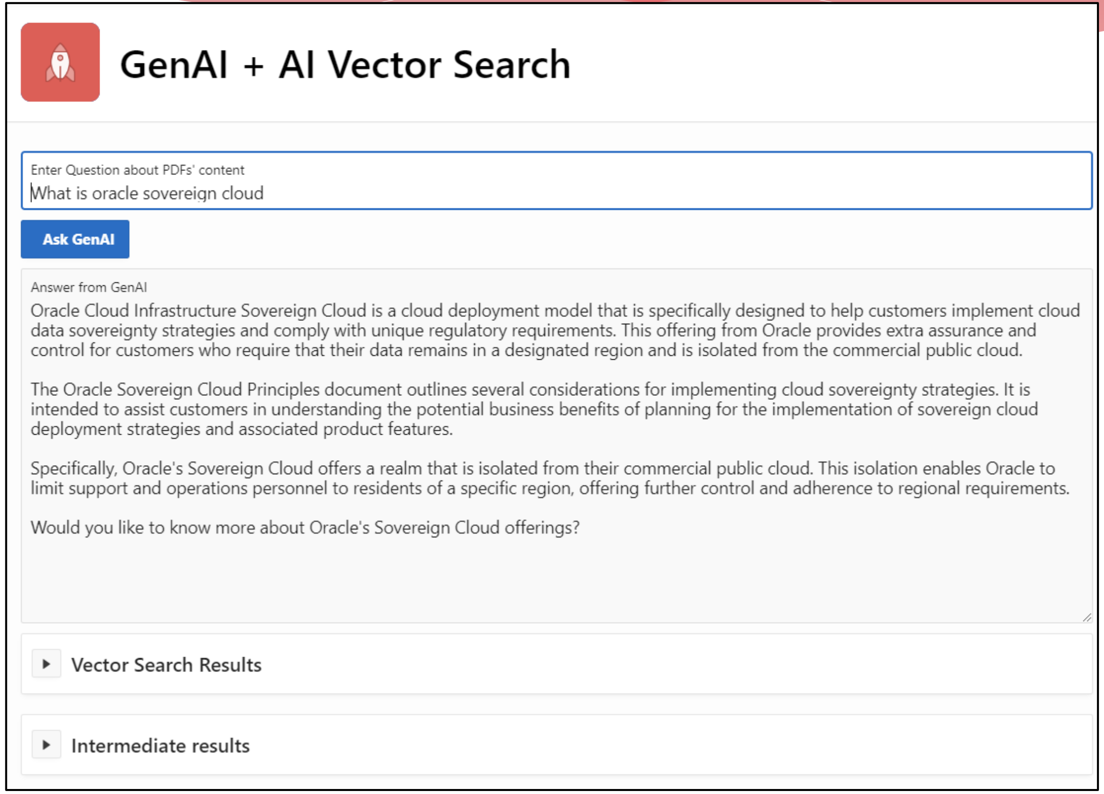
A RAG with APEX

Miracle Finland Oy

Copyright © Miracle Finland Oy

RAG Demo: GenAI + AI Vector Search
Search from your own PDFs

1. Chunk and embed PDFs (done inside the database) to vectors and store to DB
2. Ask a question
   a) Get closest matches using AI Vector Search
   b) Send question and matches to OCI GenAI and get answer



MIRACLE
Miracle Finland Oy

Thank you!

QUESTIONS?

Heli.helskyaho@miracleoy.fi
Pekka.kanerva@miracleoy.fi

@HeliFromFinland
Blog: Helifromfinland.com

MIRACLE
Miracle Finland Oy