Gimme a Vector, Victor: Leveraging Vector Datatypes for Practical Generative AI Applications

21 March, 2025

Jim Czuprynski

Chief StoryTeller

Zero Defect Computing, Inc.

Who Am I, and What Am I Doing Here?

ORACLE ACE Director

ORACLE Certified Professional

- E-mail me at jim@jimthewhyguy.com
- Follow me on BlueSky **(@JimTheWhyGuy.bsky.social)**
- Connect with me on LinkedIn (**Jim Czuprynski**)

Oracle ACE

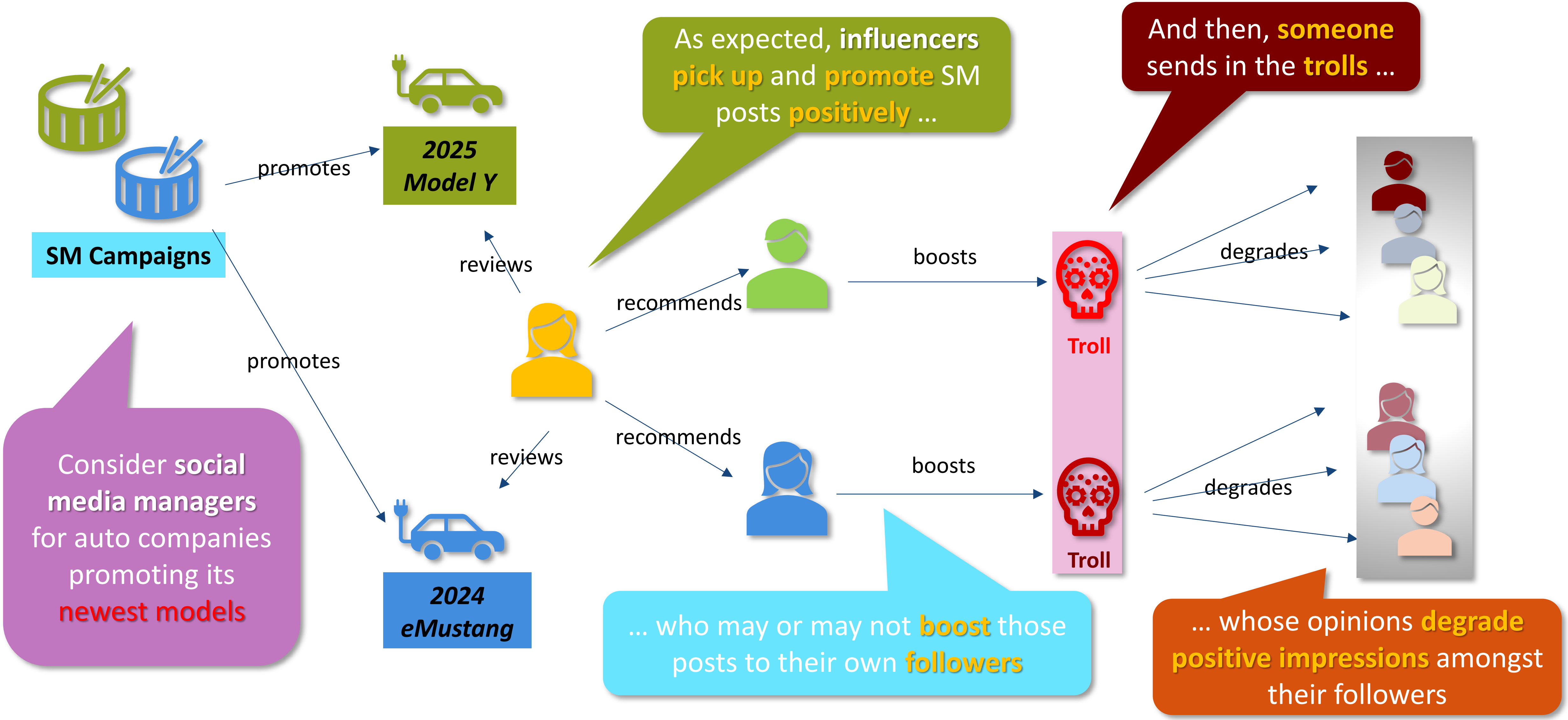
The Oracle ACE Program

400+ technical experts helping peers globally

- The Oracle ACE Program recognizes and rewards community members for their technical and community contributions to the Oracle community
- 3 membership levels: Director, Pro, and Associate
- Nominate yourself or a colleague at ace.oracle.com/nominate
- Learn more at ace.oracle.com

aceprogram_ww@oracle.com   Facebook.com/OracleACEs   @oracleace   Oracle ACE Program Group

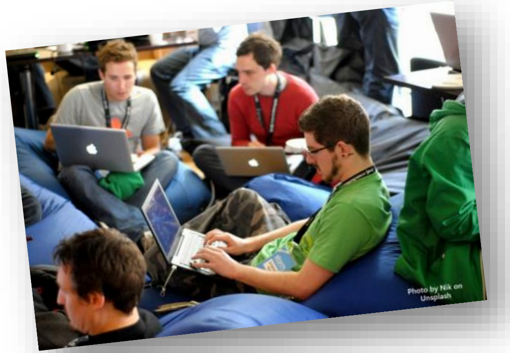
Our Social Media Strategy Worked Great … Until It Didn't Anymore.



Business Case: Timely & Effective Responses to Negative Social Media Posts

Social media campaign just launched on **new EV models**

Initial responses were **positive** … but then trolls & their followers propagated **negative sentiments** & **deliberate misinformation**



Responding to misinformation with **timely** and **effective messaging** is **imperative** to saving the campaign - and maybe even the **brand itself**

Could we leverage **Generative AI** to respond **immediately and effectively** to **existing** negative posts, as well as any **new** ones that appear?

Generative AI: Its Promise and Its Limitations

Gen AI **simulates human conversation**

It's focused on **finding the best next token** in the conversation

It can even **explain the steps it took** to return its answer

But despite all appearances, it **<u>does not</u> reason**!

What Generative AI ***Actually*** Does: A (Very) Primitive Metaphor

Consider this sentence:

`A typical database table consists of _____`

Which token should be placed **next**?

GenAI's sole purpose: **Find the best next token, now!**

Potential tokens:

Token:	rows	columns	data	tuples	varied	many	and
Probability:	**0.32**	**0.32**	**0.17**	**0.10**	**0.04**	**0.03**	**0.01**

Now which token should be placed **next**?

`A typical database table consists of` **`rows`** `_____`

Potential tokens:

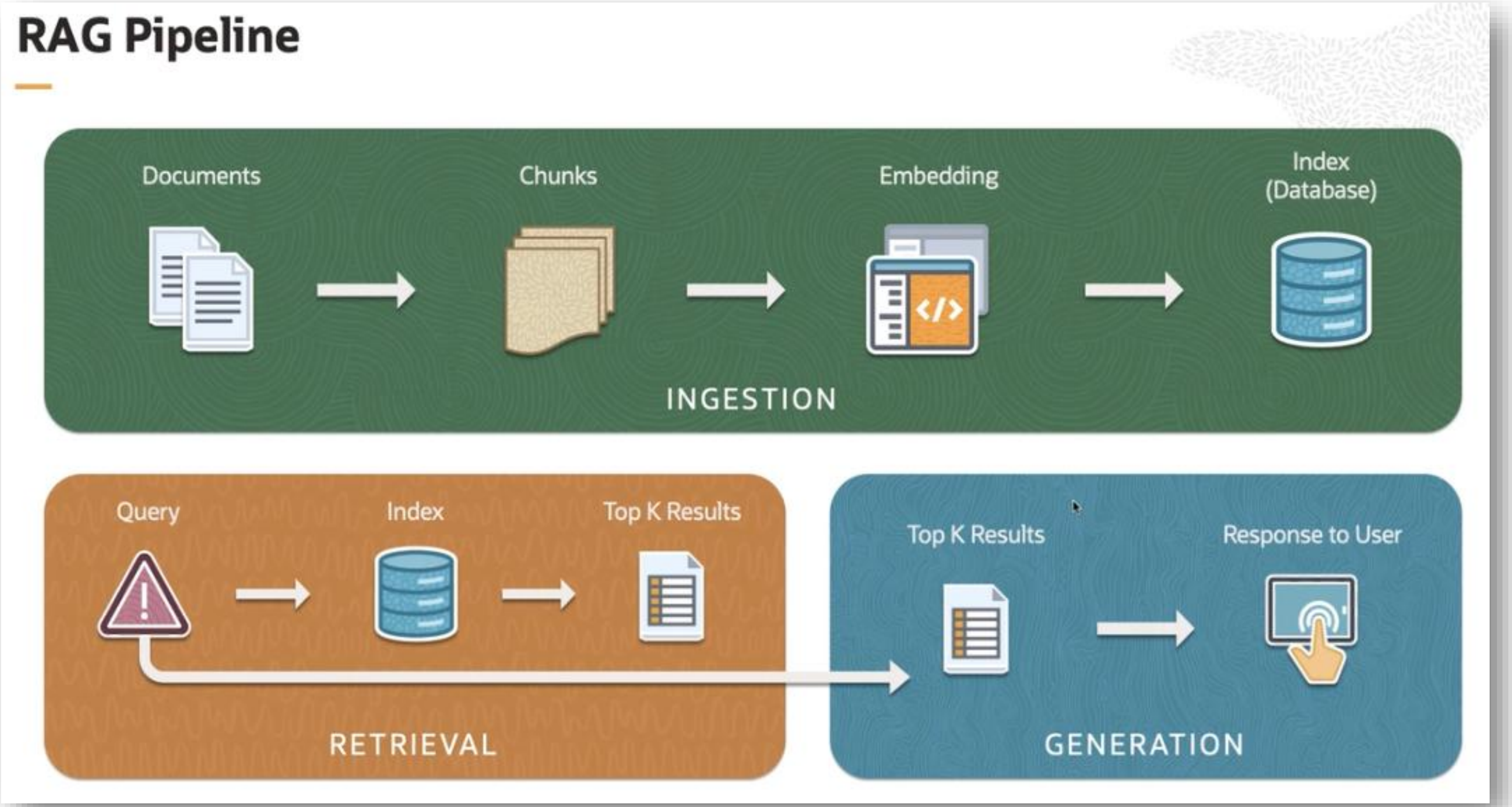
Token:	and	filled	defined	unordered	columns	rows
Probability:	**0.55**	**0.18**	**0.17**	**0.07**	**0.02**	**0.01**

The choice & distribution of best next tokens is **dramatically different**

Building Generative AI Solutions Within 23ai Database

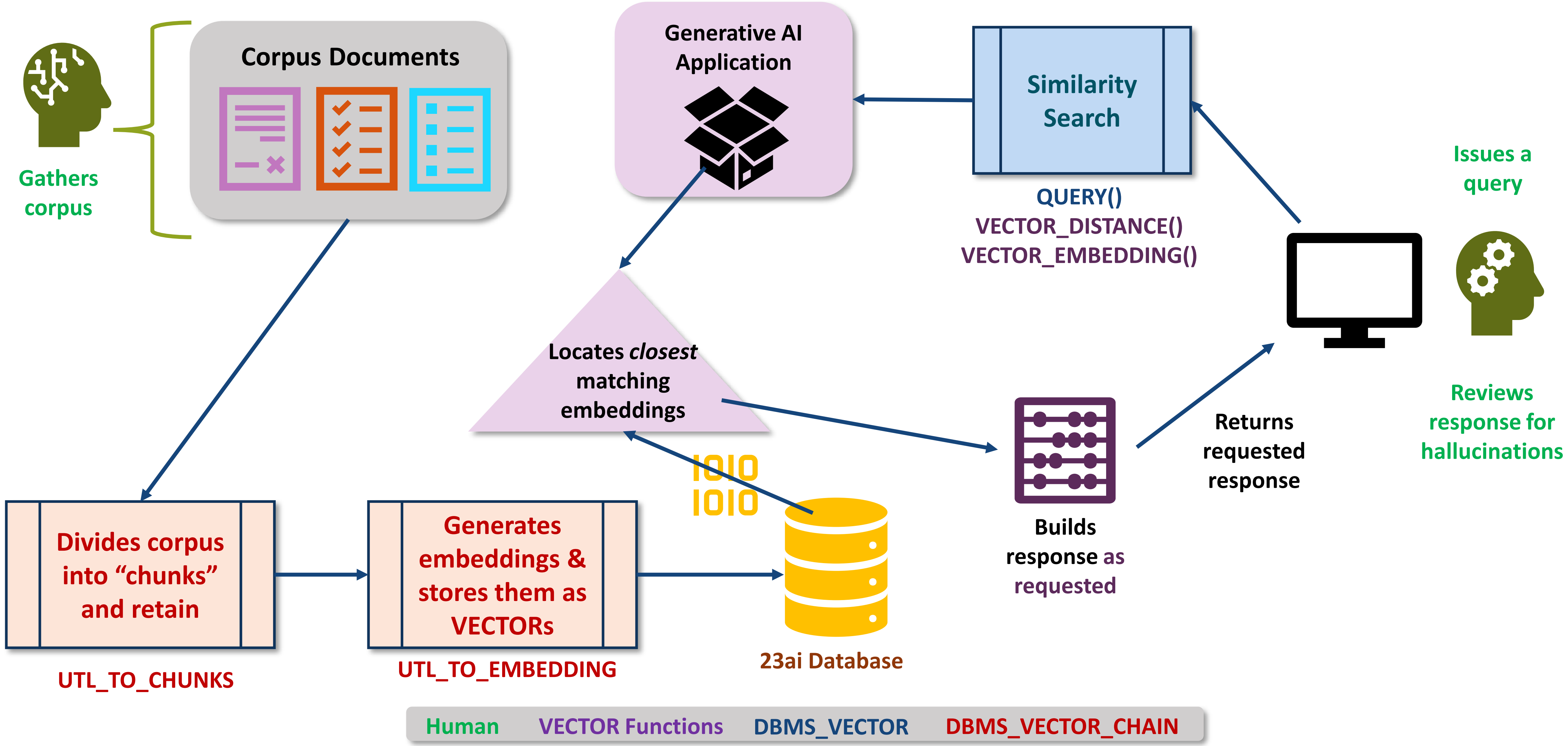
Photo By Nakaharu Line @ Unsplash

Implementing RAG Within Oracle 23ai Database



Taken directly from the **Oracle Generative AI Certification Course**. Simple, right?

Implementing RAG Within Oracle 23ai Database



Steps for Implementing a 23ai Generative AI Solution

1. **Extract** appropriate ONNX model(s)
2. **Load** ONNX model(s) into database
3. **Gather** and **load** relevant **corpus documents**
4. **"Chunk"** corpus documents & add **embeddings**
5. Add **VECTOR indexes**
6. Perform **similarity searches** against corpus
7. Build **summary** of returned **search results**
8. Invoke **3rd-party LLMs** for **extra features**

Downloading ONNX Models (1)

```
# Should return Python 3.12.1:
export PATH=$ORACLE_HOME/python/bin:$PATH
python -V

# Install necessary Python packages:
mkdir -p /home/examples/rag/onnx
cd -p /home/examples/rag/onnx
python -m pip install -r requirements.txt
python -m pip install omlutils-0.12.0-cp312-cp312-linux_x86_64.whl
```

You can use the installed Python code base that comes with Oracle 23.4 as your Python “home” …

1

Downloading ONNX Models (1)

```
# Should return Python 3.12.1:
export PATH=$ORACLE_HOME/python/bin:$PATH
python -V

# Install necessary Python packages:
mkdir -p /home/examples/rag/onnx
cd -p /home/examples/rag/onnx
python -m pip install -r requirements.txt
python -m pip install omlutils-0.12.0-cp312-cp312-linux_x86_64.whl
```

… and then install all required Python library objects to gather ONNX models via **OML4PY**

1

Downloading ONNX Models (1)

```
# Show what model(s) are available:
$ python3

Python 3.12.1 (main, Feb  6 2024, 12:09:55) [GCC 8.5.0 20210514 (Red Hat 8.5.0-18.0.6)] on linux
Type "help", "copyright", "credits" or "license" for more information.

>>> from omlutils import EmbeddingModel, EmbeddingModelConfig
em = EmbeddingModel(model_name="sentence-transformers/all-MiniLM-L6-v2")
emc = EmbeddingModelConfig()
emc.show_preconfigured()
exit()
```

Display all available **ONNX models**

1

Downloading ONNX Models (1)

```
# Show what model(s) are available:
$ python3

Python 3.12.1 (main, Feb  6 2024, 12:09:55) [GCC 8.5.0 20210514 (Red Hat 8.5.0-18.0.6)] on linux
Type "help", "copyright", "credits" or "license" for more information.

>>> from omlutils import Embedd
em = EmbeddingModel(model_name=
emc = EmbeddingModelConfig()
emc.show_preconfigured()
exit()
```

```
['sentence-transformers/all-mpnet-base-v2',
 'sentence-transformers/all-MiniLM-L6-v2',
 'sentence-transformers/multi-qa-MiniLM-L6-cos-v1',
 'ProsusAI/finbert',
 'medicalai/ClinicalBERT',
 'sentence-transformers/distiluse-base-multilingual-cased-v2',
 'sentence-transformers/all-MiniLM-L12-v2',
 'BAAI/bge-small-en-v1.5',
 'BAAI/bge-base-en-v1.5',
 'taylorAI/bge-micro-v2',
 'intfloat/e5-small-v2',
 'intfloat/e5-base-v2',
 'prajjwal1/bert-tiny',
 'thenlper/gte-base',
 'thenlper/gte-small',
 'TaylorAI/gte-tiny',
 'infgrad/stella-base-en-v2']
```

Display all available **ONNX models**

We'll focus on **just a few** of the dozen or so ONNX-compatible models for loading

1

Downloading ONNX Models (2)

```
# export_onnx_models.py:

import oml
from oml.utils import EmbeddingModel, EmbeddingModelConfig

em = EmbeddingModel(model_name="sentence-transformers/all-MiniLM-L6-v2")
em.export2file("all-MiniLM-L6-v2",output_dir="/home/oracle/examples/rag/onnx_models/")

em = EmbeddingModel(model_name="sentence-transformers/multi-qa-MiniLM-L6-cos-v1")
em.export2file("multi-qa-MiniLM-L6-cos-v1",output_dir="/home/oracle/examples/rag/onnx_models/")

em = EmbeddingModel(model_name="sentence-transformers/all-MiniLM-L12-v2")
em.export2file("all-MiniLM-L12-v2",output_dir="/home/oracle/examples/rag/onnx_models/")

exit()
```

Multiple ONNX models can be downloaded as **bitcode** for import into 23ai …

… and now we're ready to **import** them

```
$> ll *.onnx
-rwxrwxr-x. 1 oracle oinstall 133322334 Jun 19 20:03 all-MiniLM-L12-v2.onnx
-rwxrwxr-x. 1 oracle oinstall  90621438 Jun 19 19:51 all-MiniLM-L6-v2.onnx
-rwxrwxr-x. 1 oracle oinstall  90621438 Jun 19 20:03 multi-qa-MiniLM-L6-cos-v1.onnx
```

Deploying ONNX Models Within 23ai Database (1)

```
CREATE OR REPLACE DIRECTORY onnx_models AS '/home/oracle/examples/rag/onnx_models/';
GRANT READ, WRITE ON DIRECTORY onnx_models TO hol23;

BEGIN
  DBMS_VECTOR.DROP_ONNX_MODEL(
    model_name => 'MiniLML6V2'
  , force => TRUE);
  DBMS_VECTOR.LOAD_ONNX_MODEL(
    directory => 'ONNX_MODELS'
  , file_name => 'all-MiniLM-L6-v2.onnx'
  , model_name => 'MiniLML6V2'
  , metadata =>
    JSON('{"function":"embedding“
          , "embeddingOutput":"embedding“
          , "input": {"input": ["DATA"]}}’));
END;
/
```

2

Load the **all-MiniLM-L6-v2** pre-trained ONNX model into a 23ai database, assigning its alias as **MiniLML6V2** …

… and this specifies **exactly how** relevant content will be provided to the model for **embedding**

Deploying ONNX Models Within 23ai Database (2)

2

Worksheet | Query Builder

```
SELECT
   model_name
 , mining_function
 , algorithm
 , algorithm_type
 , model_size
   FROM all_mining_models;
```

Verify that the ONNX model was loaded into the 23ai database via **ALL_MINING_MODELS** ...

Query Result

SQL | All Rows Fetched: 1 in 0.043 seconds

	MODEL_NAME	MINING_FUNCTION	ALGORITHM	ALGORITHM_TYPE	MODEL_SIZE
1	MINILML6V2	EMBEDDING	ONNX	NATIVE	90621438

Deploying ONNX Models Within 23ai Database (2)

2

Worksheet | Query Builder

```sql
SELECT
  model_name
, mining_function
, algorithm
, algorithm_type
, model_size
  FROM all_mining_models;
```

Verify that the ONNX model was loaded into the 23ai database via **ALL_MINING_MODELS** …

Query Result

SQL | All Rows Fetched: 1 in 0.0

	MODEL_NAME	MINING_FUNCTION	
1	MINILML6V2	EMBEDDING	O

Worksheet | Query Builder

```sql
SELECT
  model_name
, attribute_name
, attribute_type
, data_length
, vector_info
  FROM all_mining_model_attributes;
```

… and view specifics of each model loaded via **ALL_MINING_MODEL_ATTRIBUTES**

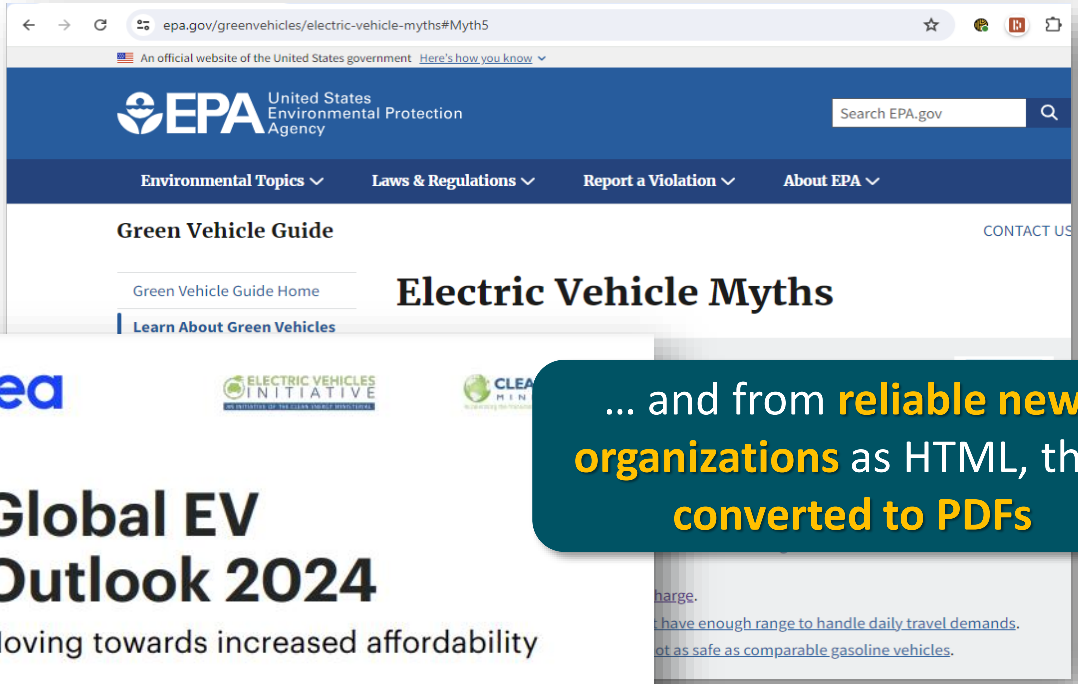
Query Result

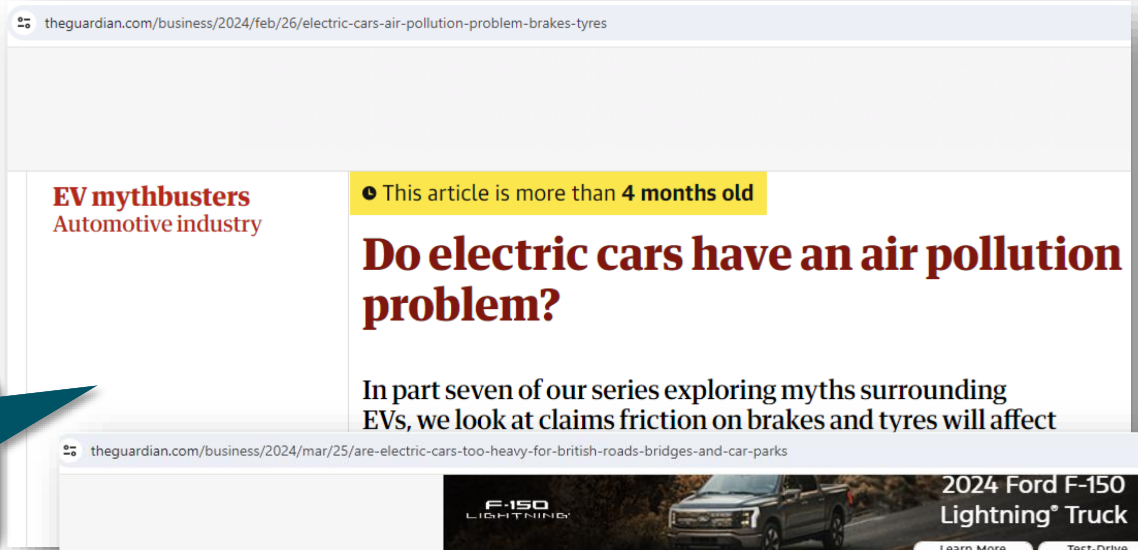
SQL | All Rows Fetched: 2 in 0.021 seconds

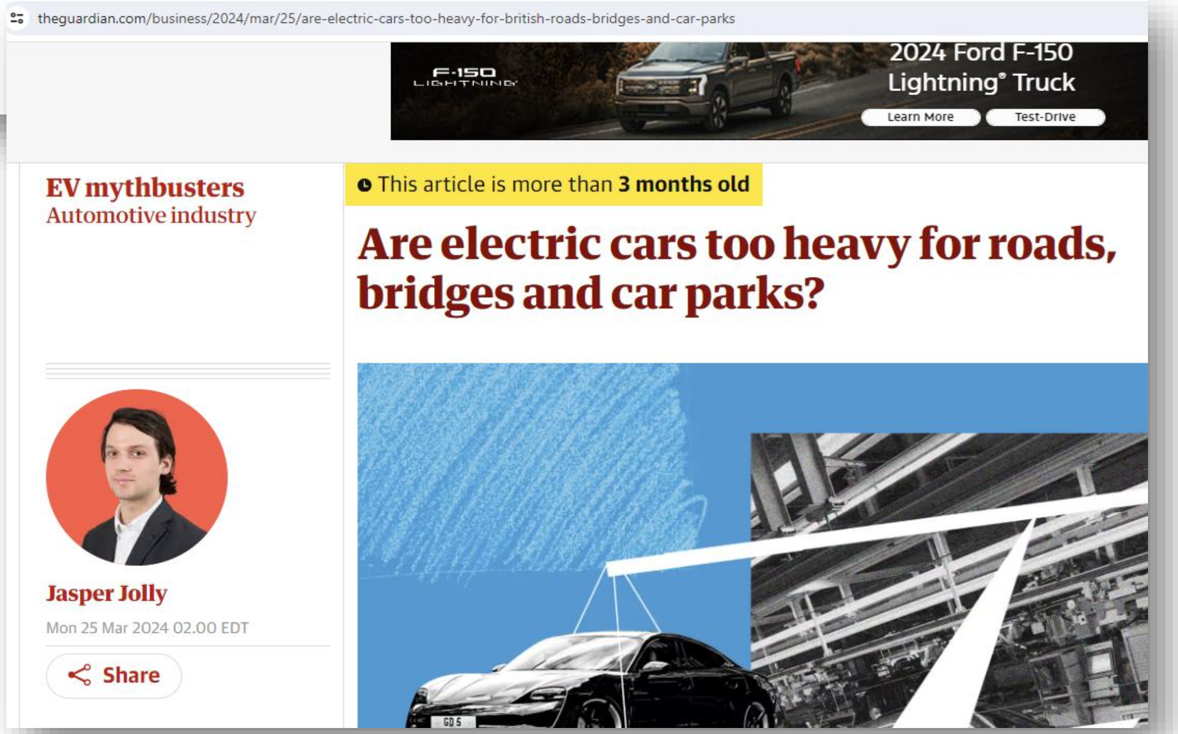
	MODEL_NAME	ATTRIBUTE_NAME	ATTRIBUTE_TYPE	DATA_LENGTH	VECTOR_INFO
1	MINILML6V2	ORA$ONNXTARGET	VECTOR	1593	VECTOR(384,FLOAT32)
2	MINILML6V2	DATA	TEXT	32767	(null)

Gathering Meaningful Corpus Documents

3







… and from **reliable news organizations** as HTML, then **converted to PDFs**

Documents were captured as PDFs from **official government and scientific sources** …

Storing Corpus Documents and Preparing for Embeddings

3

```
DROP TABLE IF EXISTS corpus_documents PURGE;

CREATE TABLE IF NOT EXISTS corpus_documents (
  cd_id       NUMBER(8,0)
, cd_status   CHAR(12)
, cd_data     BLOB
);

DROP TABLE IF EXISTS corpus_chunks PURGE;

CREATE TABLE IF NOT EXISTS corpus_chunks(
  cd_id          NUMBER(8,0)
, cdc_id         NUMBER(8,0)
, cdc_data       VARCHAR2(4000)
, cdc_embedded   VECTOR
);
```

Create a table that will contain **"chunks"** of the corpus documents and their **embeddings** …

Storing Corpus Documents and Preparing for Embeddings

3

```
CREATE OR REPLACE DIRECTORY corpus_sources AS '/home/oracle/examples/rag/corpus/';
GRANT READ, WRITE ON DIRECTORY corpus_sources TO hol23;

[ copy source documents into that directory ]

INSERT INTO corpus_documents(cd_id, cd_status, cd_data)
VALUES(001, 'VALID', TO_BLOB(BFILENAME('CORPUS_SOURCES', 'USEPA_Condensed.pdf')));

INSERT INTO corpus_documents(cd_id, cd_status, cd_data)
VALUES(002, 'VALID', TO_BLOB(BFILENAME('CORPUS_SOURCES', 'GlobalEVOutlook2023.pdf')));

INSERT INTO corpus_documents(cd_id, cd_status, cd_data)
VALUES(003, 'VALID', TO_BLOB(BFILENAME('CORPUS_SOURCES', 'GlobalEVOutlook2024.pdf')));
. . .

INSERT INTO corpus_documents(cd_id, cd_status, cd_data)
VALUES(011, 'VALID', TO_BLOB(BFILENAME('CORPUS_SOURCES',
'RecurrentAuto_StudyWinterAndColdWeatherEVRangeLossIn10000PlusCars.pdf')));

COMMIT;
```

Creating Embeddings With **DBMS_VECTOR_CHAIN.UTL_TO_CHUNKS**

4

```
INSERT INTO corpus_chunks
SELECT
  CD.cd_id doc_id
, ET.embed_id cdc_id
, ET.embed_data cdc_data
, TO_VECTOR(ET.embed_vector) cdc_embedded
  FROM
    corpus_documents CD
  , DBMS_VECTOR_CHAIN.UTL_TO_EMBEDDINGS(
        DBMS_VECTOR_CHAIN.UTL_TO_CHUNKS(
          DBMS_VECTOR_CHAIN.UTL_TO_TEXT(CD.cd_data)
        , JSON('{"by":"words","overlap":"0","split":"sentence"
                ,"language":"American","normalize":"all"}')
        )
        , JSON('{"provider":"database", "model":"MINILML6V2"}')) t
  , JSON_TABLE(t.column_value
          , '$[*]' COLUMNS (embed_id     NUMBER          PATH '$.embed_id'
                          , embed_data   VARCHAR2(4000)  PATH '$.embed_data'
                          , embed_vector CLOB            PATH '$.embed_vector')) ET;

COMMIT;
```

This "chunks" each corpus document per the parameters for **UTL_TO_CHUNKS**, as well as creating the embeddings via **UTL_TO_EMBEDDINGS**

Since the output will be in JSON format, this refers to each returned value for INSERTing into **CORPUS_CHUNKS**

Chunking Methods Determine How Answers Returned

4

BY : words,
MAX : **40**,
OVERLAP : 0,
SPLIT : **none**,
LANGUAGE : american,
NORMALIZE : all

The chunking method selected can dramatically affect how corpus documents are chunked …

Myth #1: Electric vehicles are worse for the climate than gasoline cars because of power plant emissions.

FACT: Electric vehicles typically have a smaller carbon footprint than gasoline cars, even when accounting for the electricity used for charging. Electric vehicles (EVs) have no tailpipe emissions. Generating the electricity used to charge EVs, however, may create carbon pollution. The amount varies widely based on how local power is generated, e.g., using coal or natural gas, which emit carbon pollution, versus renewable resources like wind or solar, which do not. Even accounting for these electricity emissions, research shows that an EV is typically responsible for lower levels of greenhouse gases (GHGs) than an average new gasoline car. To the extent that more renewable energy sources like wind and solar are used to generate electricity, the total GHGs associated with EVs could be even lower.

Chunk #1 Chunk #2 Chunk #3

Chunking Methods Determine How Answers Returned

4

BY : words,
MAX : **100**,
OVERLAP : 0,
SPLIT : **sentence**,
LANGUAGE : american,
NORMALIZE : all

… which could affect the **accuracy** or **loss** of meaning when chunks are selected for output during vectorized searches!

Myth #1: Electric vehicles are worse for the climate than gasoline cars because of power plant emissions.

FACT: Electric vehicles typically have a smaller carbon footprint than gasoline cars, even when accounting for the electricity used for charging. Electric vehicles (EVs) have no tailpipe emissions. Generating the electricity used to charge EVs, however, may create carbon pollution. The amount varies widely based on how local power is generated, e.g., using coal or natural gas, which emit carbon pollution, versus renewable resources like wind or solar, which do not. Even accounting for these electricity emissions, research shows that an EV is typically responsible for lower levels of greenhouse gases (GHGs) than an average new gasoline car. To the extent that more renewable energy sources like wind and solar are used to generate electricity, the total GHGs associated with EVs could be even lower.

Chunk #1 **Chunk #2** **Chunk #3**

Text Chunking Parameters

Parameter	Purpose	Default
BY	How to **split** documents (**CHARACTER | WORDS | VOCABULARY**)	**BY WORDS**
MAX	Maximum **size** of each chunk	**100**
SPLIT [BY]	Where to **split input text** when it approaches MAX size	**RECURSIVELY**
OVERLAP	How much of the **preceding** text the **current** chunk should contain	**0**
LANGUAGE	Language of the **input** data	**Session's NLS_LANGUAGE**
NORMALIZE	How to **pre-process** or **post-process** issues encountered with text (e.g. multiple consecutive spaces)	**None**
EXTENDED	Allows output limit to be **extended** to (32K – 1) bytes	**4000**

4

See the **VECTOR_CHUNKS documentation** for a complete discussion of how these parameters affect chunking

Chunking Effects

```
SELECT
  VECTOR_DISTANCE(
    cdc_embedded
  , VECTOR_EMBEDDING(
      MINILML6V2
      USING
'Everybody knows that EVs pollute more than gas powered cars!'
AS DATA), COSINE) AS rating
, cd_id
, cdc_id
, SUBSTR(cdc_data, 1, 60) brief_data
  FROM corpus_chunks
 ORDER BY rating ASC
 FETCH NEXT 10 ROWS ONLY;
```

This query will return very different results …

4

Chunking Effects

Query Result ×

SQL | All Rows Fetched: 10 in 0.16 seconds

	RATING	CD_ID	CDC_ID	BRIEF_DATA
1	0.27612072229385376	1	5	Some studies have shown that making a typical EV can create
2	0.30973511934280396	1	1	Myth #1:Electric vehicles are worse for the climate than g
3	0.3129057288169861	1	2	pollution, versus renewable resources like wind or solar, wh
4	0.32510173320770264	7	3	Some people believe that the shift to electric cars could re
5	0.3337820768356323	7	6	EVs do not burn fossil fuels directly - and would not releas
6	0.33554182529449460	9	1	Do electric cars really produce fewer carbon emissions than
7	0.34132808446884155	7	10	been peer-reviewed by scientists, and the industry disputes
8	0.34371501207351685	1	6	For example, researchers at Argonne National Laboratory esti
9	0.3488852381706238	8	6	There are millions of electric cars on roads around the worl
10	0.3578695058822632	7	1	Do electric cars have an air pollution problem? | Automotive

```
  FROM corpus_chunks
 ORDER BY rating ASC
 FETCH NEXT 10 ROWS ONLY;
```

`JSON('{"normalize":"all"}'))`

… depending exactly how the corpus documents have been **chunked** …

4

Chunking Effects

Query Result

seconds

	RATIN			CDC_ID	BRIEF_DATA
1	0.276			5	Some studies have shown that making a typical EV can create
2	0.309			1	Myth #1:Electric vehicles are worse for the climate than g
3	0.312			2	pollution, versus renewable resources like wind or solar, wh
4	0.325			3	Some people believe that the shift to electric cars could re
5	0.333782	56323	7	6	EVs do not burn fossil fuels directly - and would not releas
6	0.335554	2944946	9	1	Do electric cars really produce fewer carbon emissions than
7	0.341328	6884155	7	10	been peer-reviewed by scientists, and the industry disputes
8	0.34371	07351685	1	6	For example, researchers at Argonne National Laboratory esti
9	0.34888	1706238	8	6	There are millions of electric cars on roads around the worl
10	0.35786	58822632	7	1	Do electric cars have an air pollution problem? | Automotive

```
 FROM corpus_chunks
ORDER BY rating ASC
FETCH NEXT 10 ROWS ONLY;
```

… before **embeddings** are created!

```
JSON('{"normalize":"all"}'))
```

4

… depending exactly how the corpus documents have been **chunked** …

```
JSON('{"by":"words",
"overlap":"0",
"split":"sentence",
"language":"american",
"normalize":"all"}'))
```

Chunking Effects

Query Result

seconds

	RATIN			CDC_ID	BRIEF_DATA
1	0.276			5	Some studies have shown that making a typical EV can create
2	0.309			1	Myth #1:Electric vehicles are worse for the climate than g
3	0.312			2	pollution, versus renewable resources like wind or solar, wh
4	0.325			3	Some people believe that the shift to electric cars could re
5	0.333782	56323	7	6	EVs do not burn fossil fuels directly - and would not releas
6	0.335554	2944946	9	1	Do electric cars really produce fewer carbon emissions than
7	0.341328	6884155	7	10	been peer-reviewed by scientists, and the industry disputes
8	0.343715	07351685	1	6	For example, researchers at Argonne National Laboratory esti
9	0.34888	1706238	8	6	There are millions of electric cars on roads around the worl
10	0.35786	58822632	7	1	Do electric cars have an air pollution problem? | Automotive

… before **embeddings** are created!

```
JSON('{"normalize":"all"}'))
```

4

… depending exactly how the corpus documents have been **chunked** …

```
FROM corpus_chunks
ORDER BY rating ASC
FETCH NEXT 10 ROWS ONLY
```

```
JSON('{"by":"words",
"overlap":"0",
"split":"sentence",
"language":"american",
"normalize":"all"}'))
```

Query Result

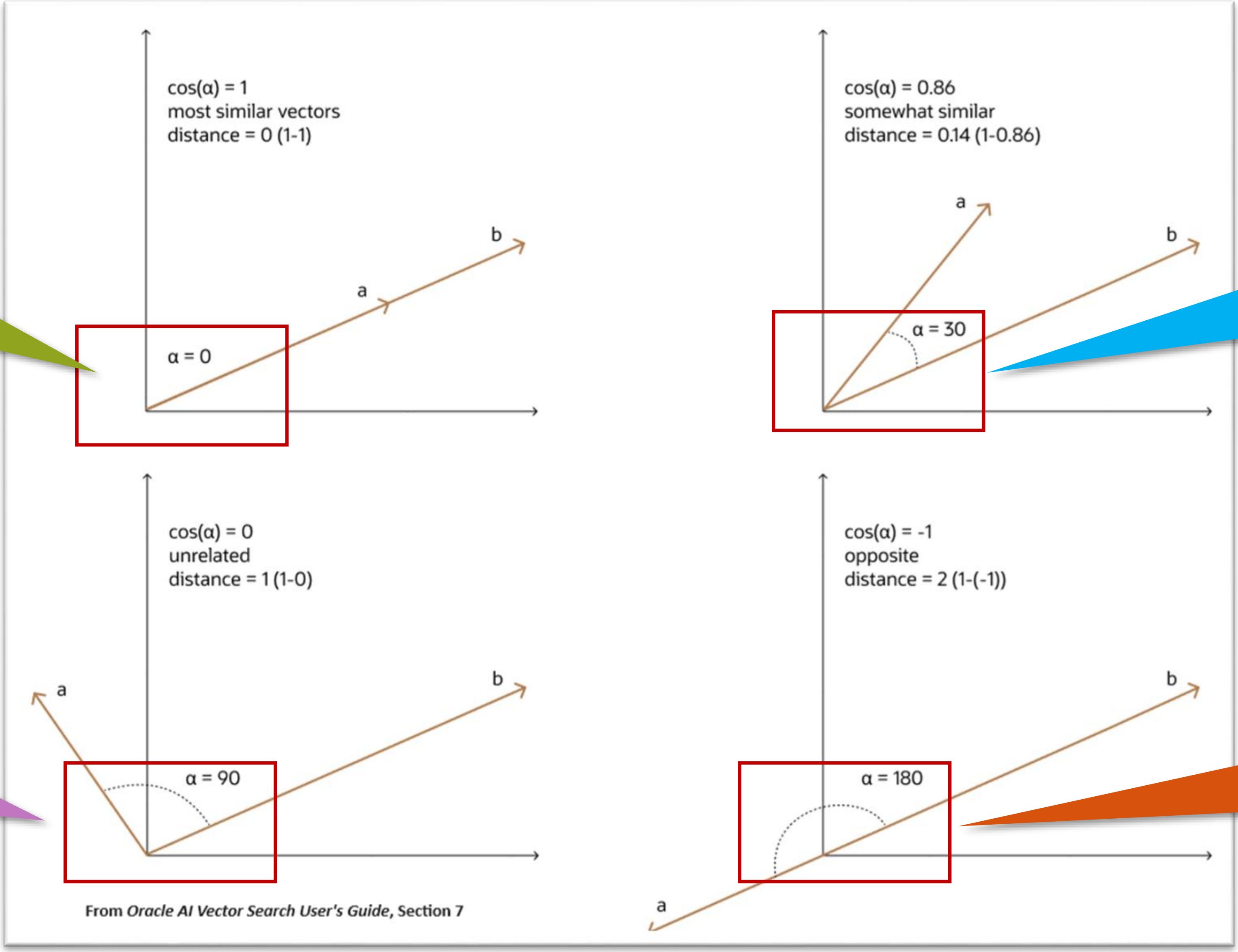
SQL | All Rows Fetched: 10 in 0.181 seconds

	RATING	CD_ID	CDC_ID	BRIEF_DATA
1	0.28293377161026	1	5	FACT: The greenhouse gas emissions associated with an electr
2	0.29539746046066284	7	10	Many of the claims about EVs causing air pollution reference
3	0.29731154441833496	1	2	The amount varies widely based on how local power is genera
4	0.31835395097732544	1	1	Myth #1:Electric vehicles are worse for the climate than g
5	0.3192080855369568	1	6	Still, over the lifetime of the vehicle, total GHG emission
6	0.32561093568880188	7	3	Some people believe that the shift to electric cars could re
7	0.33014196157455444	8	8	That evidence suggests there is no reason to think that EVs
8	0.33555418252944946	9	1	Do electric cars really produce fewer carbon emissions than
9	0.33589881658554077	9	3	Our EV mythbusters serieshas looked at some of the most p
10	0.34426039457321167	8	17	25 Mar 2024Do electric cars have an air pollution problem

VECTOR_DISTANCE: How Vectors Are Searched For Meaningfulness

5

A **COSINE** value of **1** means the embeddings are identical



From *Oracle AI Vector Search User's Guide*, Section 7

A small angular difference (**0.86**) indicates they're **somewhat similar**

A value of **0** indicates the embeddings are **unrelated**

... and **-1** means the embeddings are **complete opposites**

Vector Indexes: Speeding Retrieval of Embeddings

```
CREATE VECTOR INDEX corpus_chunks_hnsw_idx
  ON corpus_chunks (cdc_embedded)
  ORGANIZATION INMEMORY NEIGHBOR GRAPH
  DISTANCE COSINE
  WITH TARGET ACCURACY 95
  PARAMETERS (
    TYPE HNSW
   ,EFCONSTRUCTION 5
   ,NEIGHBORS 2);
```

5

Create a **Hierarchical Navigable Small-World (HNSW)** vector index on embedded VECTOR column **cdc_embedded** …

... using the **COSINE** method targeting a goal of accurate **retrieval** in **95%** of cases ...

... with a maximum of five (5) **closest vector candidates** and a maximum of two (2) vector **NEIGHBORS** on any index layer

Vector Indexes: Speeding Retrieval of Embeddings


```
EXPLAIN PLAN FOR
SELECT VECTOR_DISTANCE(cdc_embedded, VECTOR_EMBEDDING(MINILML6V2 USING 'Are EV batteries safe?'
      AS DATA), COSINE) AS rating
, cd_id, cdc_id, SUBSTR(cdc_data, 1, 60) brief_data
  FROM corpus_chunks
 ORDER BY rating ASC
 FETCH EXACT NEXT 10 ROWS ONLY;

SELECT plan_table_output
  FROM TABLE(DBMS_XPLAN.DISPLAY('plan_table',NULL,'all'));

Plan hash value: 1902679133

-------------------------------------------------------------------------------------------------
| Id  | Operation                 | Name            | Rows  | Bytes |TempSpc| Cost (%CPU)| Time     |
-------------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT          |                 |    10 |  1540 |       |  1478   (1)| 00:00:01 |
|*  1 |  COUNT STOPKEY            |                 |       |       |       |            |          |
|   2 |   VIEW                    |                 |  2892 |   434K|       |  1478   (1)| 00:00:01 |
|*  3 |    SORT ORDER BY STOPKEY|                 |  2892 |  5623K|  5792K|  1478   (1)| 00:00:01 |
|   4 |     TABLE ACCESS FULL     | CORPUS_CHUNKS   |  2892 |  5623K|       |   272   (0)| 00:00:01 |
-------------------------------------------------------------------------------------------------
```

Using the **EXACT** clause tells the optimizer to essentially perform a keyword search, and no index will be used

Vector Indexes: Speeding Retrieval of Embeddings


```
EXPLAIN PLAN FOR
SELECT VECTOR_DISTANCE(cdc_embedded, VECTOR_EMBEDDING(MINILML6V2 USING 'Are EV batteries safe?'
      AS DATA), COSINE) AS rating,
 cd_id, cdc_id, SUBSTR(cdc_data, 1, 60) brief_data
  FROM corpus_chunks
 ORDER BY rating ASC
 FETCH APPROXIMATE NEXT 10 ROWS ONLY;

SELECT plan_table_output
  FROM TABLE(DBMS_XPLAN.DISPLAY('plan_table',NULL,'all'));

Plan hash value: 992403219

---------------------------------------------------------------------------------------------------
| Id  | Operation                    | Name                     | Rows  | Bytes | Cost (%CPU)| Time     |
---------------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT             |                          |    10 |  1540 |     2  (50)| 00:00:01 |
|*  1 |  COUNT STOPKEY               |                          |       |       |            |          |
|   2 |   VIEW                       |                          |    10 |  1540 |     2  (50)| 00:00:01 |
|*  3 |    SORT ORDER BY STOPKEY     |                          |    10 | 19910 |     2  (50)| 00:00:01 |
|   4 |     TABLE ACCESS BY INDEX ROWID| CORPUS_CHUNKS          |    10 | 19910 |     1   (0)| 00:00:01 |
|   5 |      VECTOR INDEX HNSW SCAN  | CORPUS_CHUNKS_HNSW_IDX   |    10 | 19910 |     1   (0)| 00:00:01 |
---------------------------------------------------------------------------------------------------
```

Note the difference when using the **APPROXIMATE** clause, which leverages the HNSW Vector Index for faster retrieval

Measuring a Vector Index's Accuracy

6

```
SET SERVEROUTPUT ON
SET LONG 100000
DECLARE
  report VARCHAR2(128);
  query_vector VECTOR;
  CURSOR curVectors IS
    SELECT
      cd_id
    , cdc_id
    , SUBSTR(cdc_data, 1, 20) AS text
    , cdc_embedded
      FROM corpus_chunks
    WHERE cd_id = 1
    ORDER BY cd_id, cdc_id
```

```
. . .
BEGIN
  FOR qv IN curVectors
    LOOP
      report :=
        DBMS_VECTOR.INDEX_ACCURACY_QUERY(
          owner_name => 'HOL23'
        , index_name => 'CORPUS_CHUNKS_HNSW_IDX'
        , qv => qv.cdc_embedded
        , top_K => 10
        , target_accuracy => 90);

      DBMS_OUTPUT.PUT_LINE(
        'Chunk #' || qv.cd_id || '.' || qv.cdc_id ||
        ' (' || qv.text || ') accuracy: ' || report
      );

    END LOOP;
END;
/
```

Function **INDEX_ACCURACY_QUERY** in package **DBMS_VECTOR** provides an **accuracy rating** for a supplied **VECTOR** value based on the specified **accuracy threshold** . . .

Measuring a Vector Index's Accuracy

6

```
SET SERVEROUTPUT ON
SET LONG 100000
DECLARE
  report VARCHAR2(128);
  query_vector VECTOR;
  CURSOR curVectors IS
    SELECT
      cd_id
```

```
. . .
BEGIN
  FOR qv IN curVectors
    LOOP
      report :=
        DBMS_VECTOR.INDEX_ACCURACY_QUERY(
          owner_name => 'HOL23'
        , index_name => 'CORPUS_CHUNKS_HNSW_IDX'
        , qv => qv.cdc_embedded
        , top_K => 10
```

```
Chunk #1.1 (Myth #1: Electric v) accuracy: Accuracy achieved (30%) is 60% lower than the Target Accuracy requested (90%)
Chunk #1.2 (The amount varies w) accuracy: Accuracy achieved (40%) is 50% lower than the Target Accuracy requested (90%)
Chunk #1.3 ((In 2020, renewables) accuracy: Accuracy achieved (30%) is 60% lower than the Target Accuracy requested (90%)
Chunk #1.4 (EPA and Department o) accuracy: Accuracy achieved (40%) is 50% lower than the Target Accuracy requested (90%)
Chunk #1.5 (FACT: The greenhouse) accuracy: Accuracy achieved (30%) is 60% lower than the Target Accuracy requested (90%)
Chunk #1.6 (Still, over the lif) accuracy: Accuracy achieved (40%) is 50% lower than the Target Accuracy requested (90%)
Chunk #1.7 (In their estimates, ) accuracy: Accuracy achieved (50%) is 40% lower than the Target Accuracy requested (90%)
Chunk #1.8 (Myth #3: The increas) accuracy: Accuracy achieved (80%) is 10% lower than the Target Accuracy requested (90%)
Chunk #1.9 (Yet, how that impact) accuracy: Accuracy achieved (50%) is 40% lower than the Target Accuracy requested (90%)
Chunk #1.10 (And further down the) accuracy: Accuracy achieved (80%) is 10% lower than the Target Accuracy requested (90%)
Chunk #1.11 (• EV charging consu) accuracy: Accuracy achieved (80%) is 10% lower than the Target Accuracy requested (90%)
Chunk #1.12 (• Long term, higher) accuracy: Accuracy achieved (100%) is 10% higher than the Target Accuracy requested (90%)
Chunk #1.13 (The Department of En) accuracy: Accuracy achieved (80%) is 10% lower than the Target Accuracy requested (90%)
Chunk #1.14 (Visit DOE's Bipartis) accuracy: Accuracy achieved (100%) is 10% higher than the Target Accuracy requested (90%)
```

Answering Prompts With DBMS_VECTOR.UTL_TO_GENERATE_TEXT

```
SET SERVEROUTPUT ON
DECLARE
  user_question CLOB;
  params CLOB;
  output CLOB;
BEGIN
  -- Accept user question:
  user_question :=
  'Generate a response to the following question: ' ||
  'Do EVs pollute more than gas vehicles? ' ||
  'using the following text as an authoritative source: ' ||
  'FACT: Electric vehicles typically have a smaller carbon footprint ' ||
   . . .
   [several lines of additional authoritative source redacted ]
   . . .
  'still lower than those for the gasoline car.';
. . .
```

This prompt can accept any text typically supplied to an **AI chatbot**, including directives on what to use as an **authoritative sources** to answer the question

Answering Prompts With DBMS_VECTOR.UTL_TO_GENERATE_TEXT

```
SET SERVEROUTPUT ON
DECLARE
  user_question CLOB;
  params CLOB;
  output CLOB;
BEGIN
```

```
                                        n:
                                        o
                                        ha
      [several lines of addit
      . . .
  'still lower than those
. . .
```

This prompt can accept any text typically supplied to an **AI chatbot**, including directives on what to use as an **authoritative sources** to

Here's an example of using OpenAI's **gpt-4o** chatbot to answer the question, including settings for **temperature** and other levers that control the generated response

```
. . .
 -- Set up parameters for calling OpenAI gpt-4o model:
  params := '{
  "provider": "openai",
  "credential_name": "OPENAI_CRED",
  "url": "https://api.openai.com/v1/chat/completions",
  "model": "gpt-4o",
  "temperature": 1.0,
  "max_tokens": 256,
  "top_p": 1.0,
  "frequency_penalty": 0.0,
  "presence_penalty": 0.0 }';
. . .
```

Answering Prompts With DBMS_VECTOR.UTL_TO_GENERATE_TEXT

```
SET SERVEROUTPUT ON
DECLARE
  user_question CLOB;
  params CLOB;
  output CLOB;
```

Here's an example of using OpenAI's **gpt-4o** chatbot to answer the question, including settings f... **temperature** and other levers t... control the generated respon...

This ... typica... incluc... as a...

The returned output will be identical to what the **OpenAI chat assistant API** returns, provided the same parameters were used

```
. . .
-- Set up parameters for ca...
```

```
. . .
-- Send prompt string to OpenAI for processing:
  output :=
    DBMS_VECTOR.UTL_TO_GENERATE_TEXT(user_question, JSON(params));
```

```
Electric vehicles (EVs) generally have a smaller carbon footprint compared to
gasoline cars. While it's true that there are greenhouse gas (GHG) emissions
associated with the production and eventual disposal of electric vehicles,
including the electricity used for charging, research shows that the overall GHG
levels from EVs are typically lower than those from new gasoline cars. Despite
the higher emissions from the manufacturing and end-of-life stages, the total
greenhouse gas emissions for EVs remain lower compared to those of gasoline
vehicles. Therefore, EVs are responsible for fewer GHG emissions and are less
polluting overall.
```

Answering Prompts With DBMS_VECTOR.UTL_TO_GENERATE_TEXT

```
SET SERVEROUTPUT ON
DECLARE
  user_question CLOB;
```

This … typica… incluc… as a…

The returned output will be identical to what the **OpenAI chat assistant API** returns, provided the same parameters were used

Output can then be routed to any application within our firewall … and except for the call to the external AI API, **everything happens within the Oracle 23ai database**

… **temperature** and other l… control the generated re…

```
. . .
-- Set up parameters for ca...
```

```
-- Send prompt string to OpenAI for processing:
output :=
  DBMS_VECTOR.UTL_TO_GENERATE_TEXT(user_question, JSON(params));
```

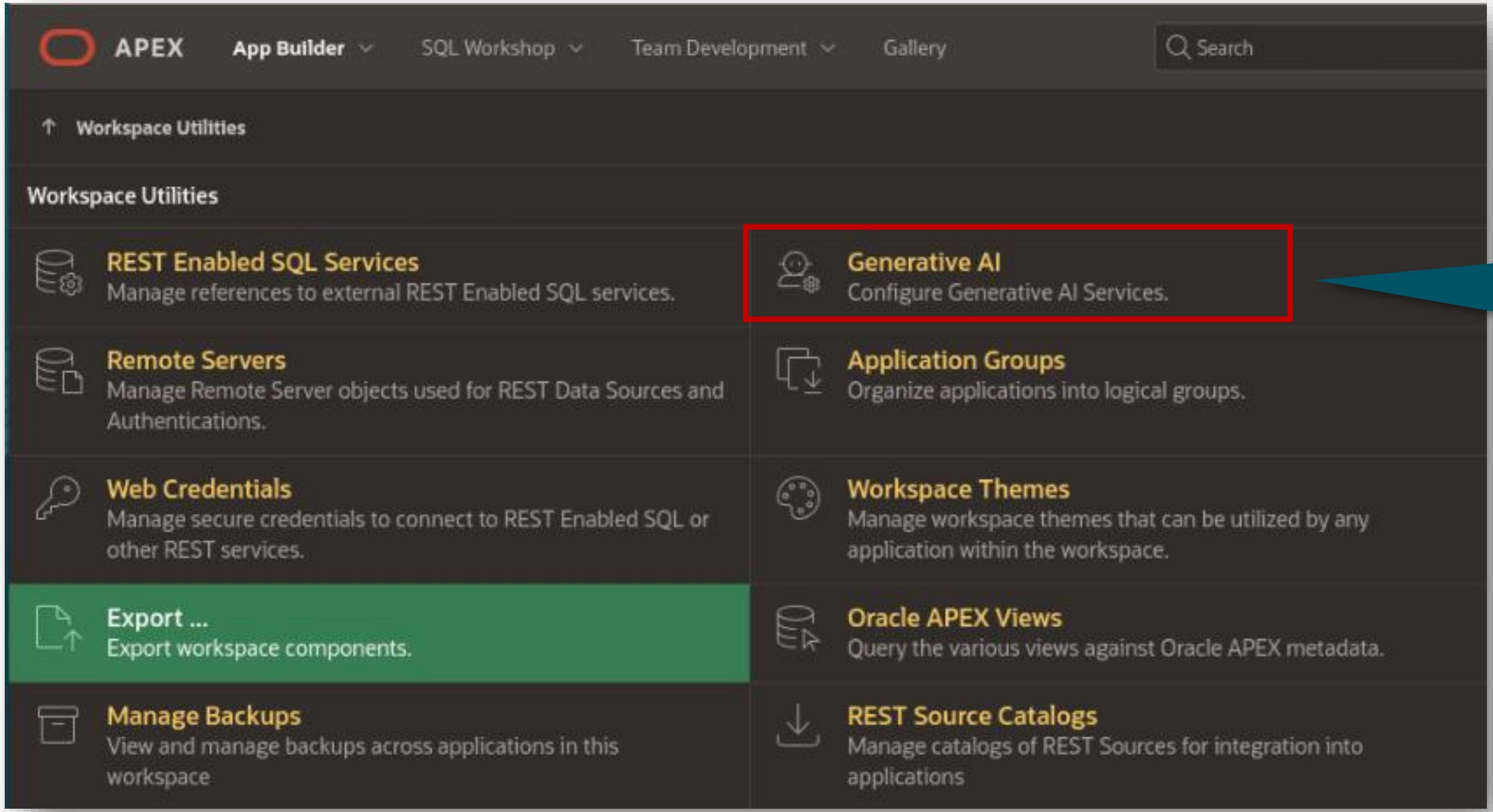
```
Electric vehicles (EVs) generally have a smaller carbon footprint compared to
gasoline cars. While it's true that there are greenhouse gas (GHG) emissions
associated with the production and eventual disposal of electric vehicles,
including the electricity used for charging, research shows that the overall GHG
levels from EVs are typically lower than those from new gasoline cars. Despite
the higher emissions from the manufacturing and end-of-life stages, the total
greenhouse gas emissions for EVs remain lower compared to those of gasoline
vehicles. Therefore, EVs are responsible for fewer GHG emissions and are less
polluting overall.
```

Leveraging External AI Tools with APEX

Photo by Alex Kotliarskyi @ Unsplash

Using APEX_AI: Preparations (1)

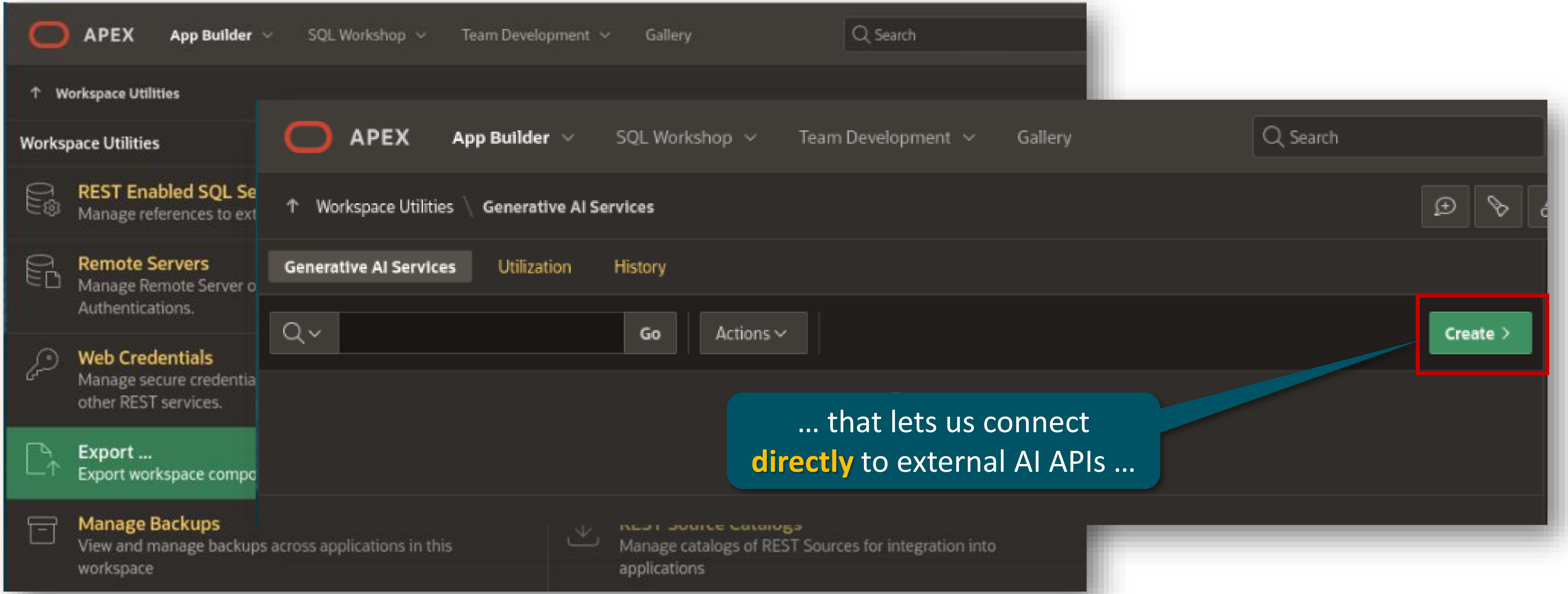
8



Note this **new option** in APEX 24.1 ...

Using APEX_AI: Preparations (1)

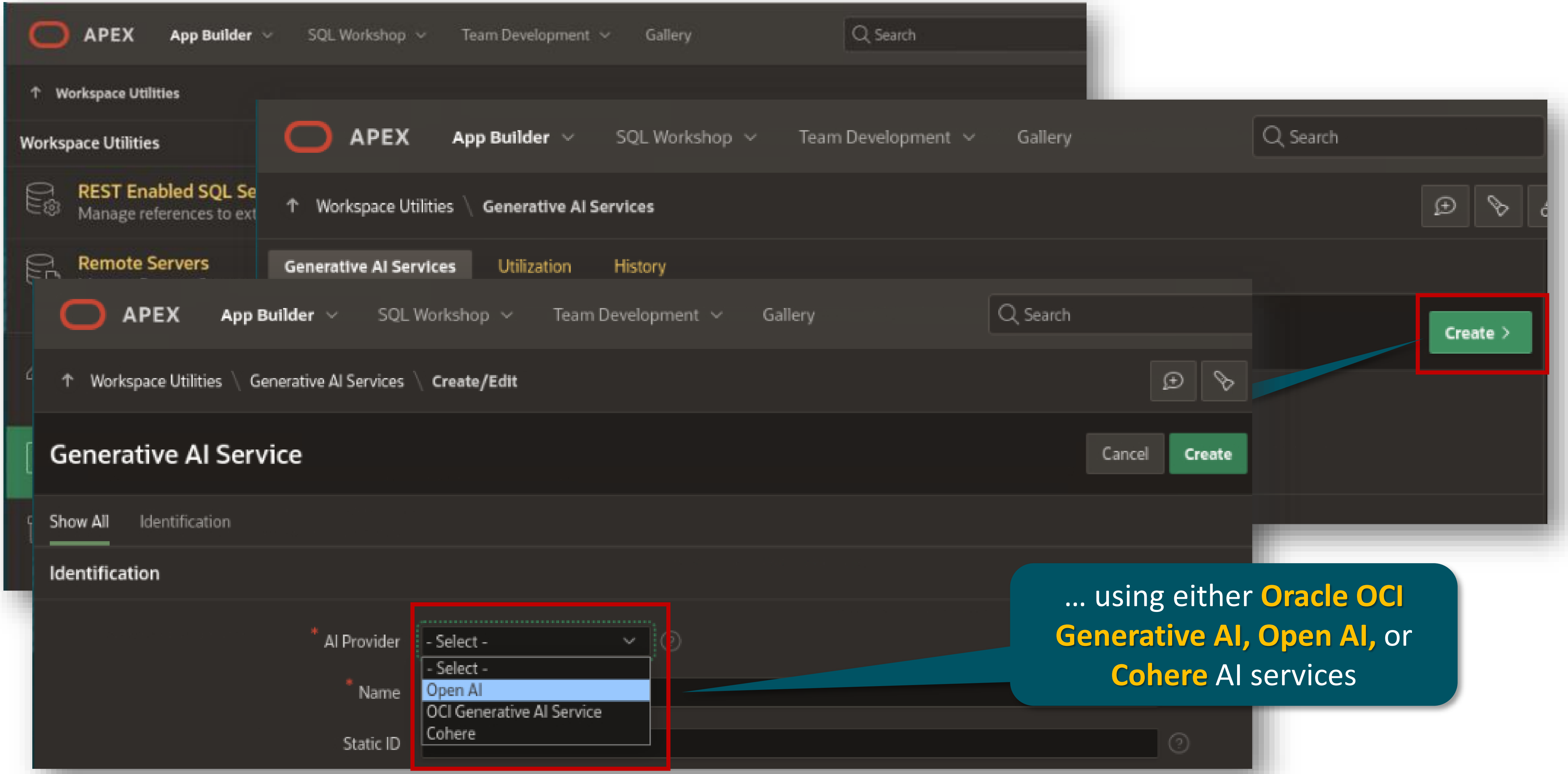
8



... that lets us connect **directly** to external AI APIs ...

Using APEX_AI: Preparations (1)

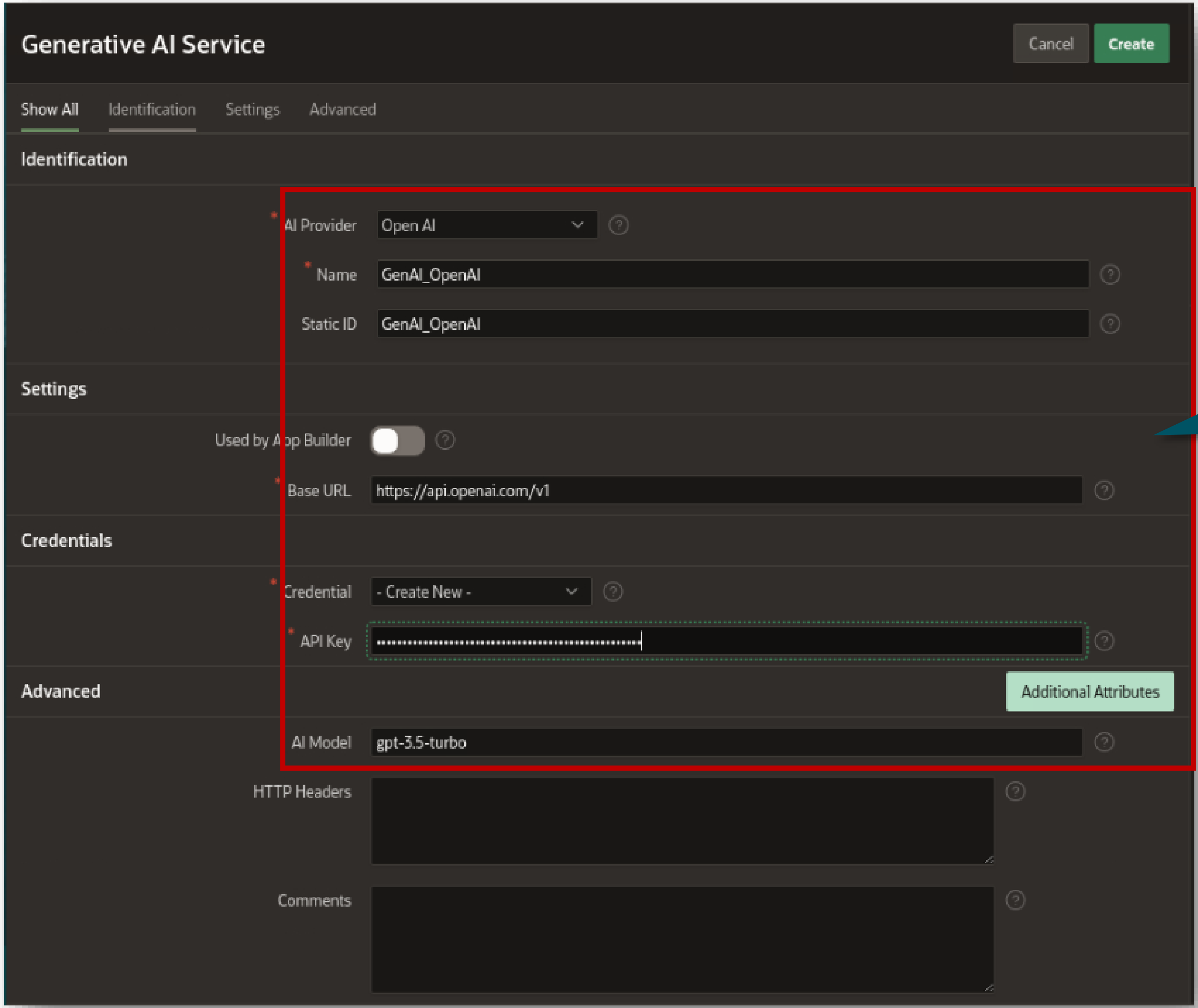
8



… using either **Oracle OCI Generative AI, Open AI,** or **Cohere** AI services

Using APEX_AI: Preparations (2)

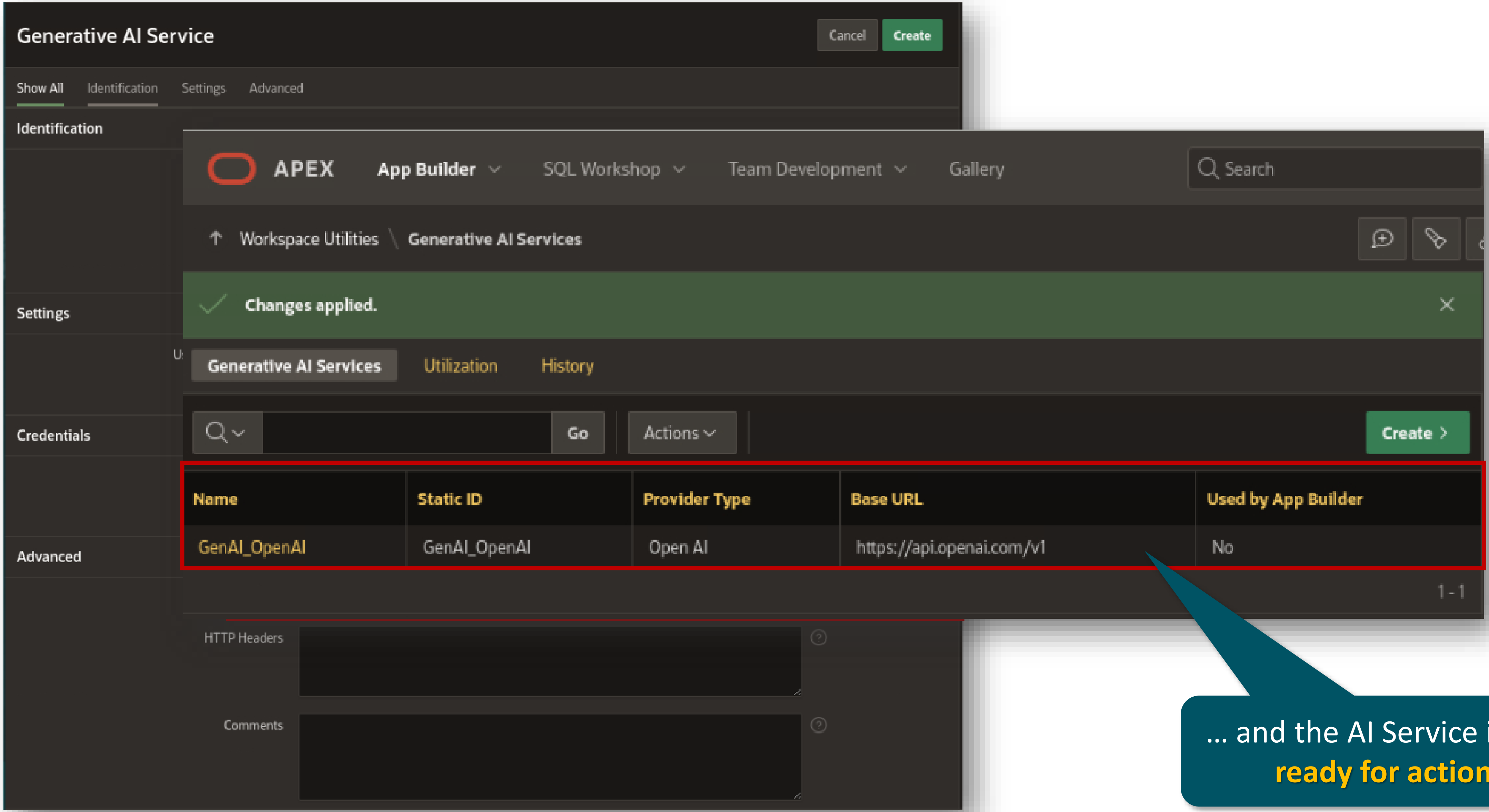
8



Specify the desired **AI provider**, its base **URL**, a **credential** and **API key**, and which **model** to use

Using APEX_AI: Preparations (2)

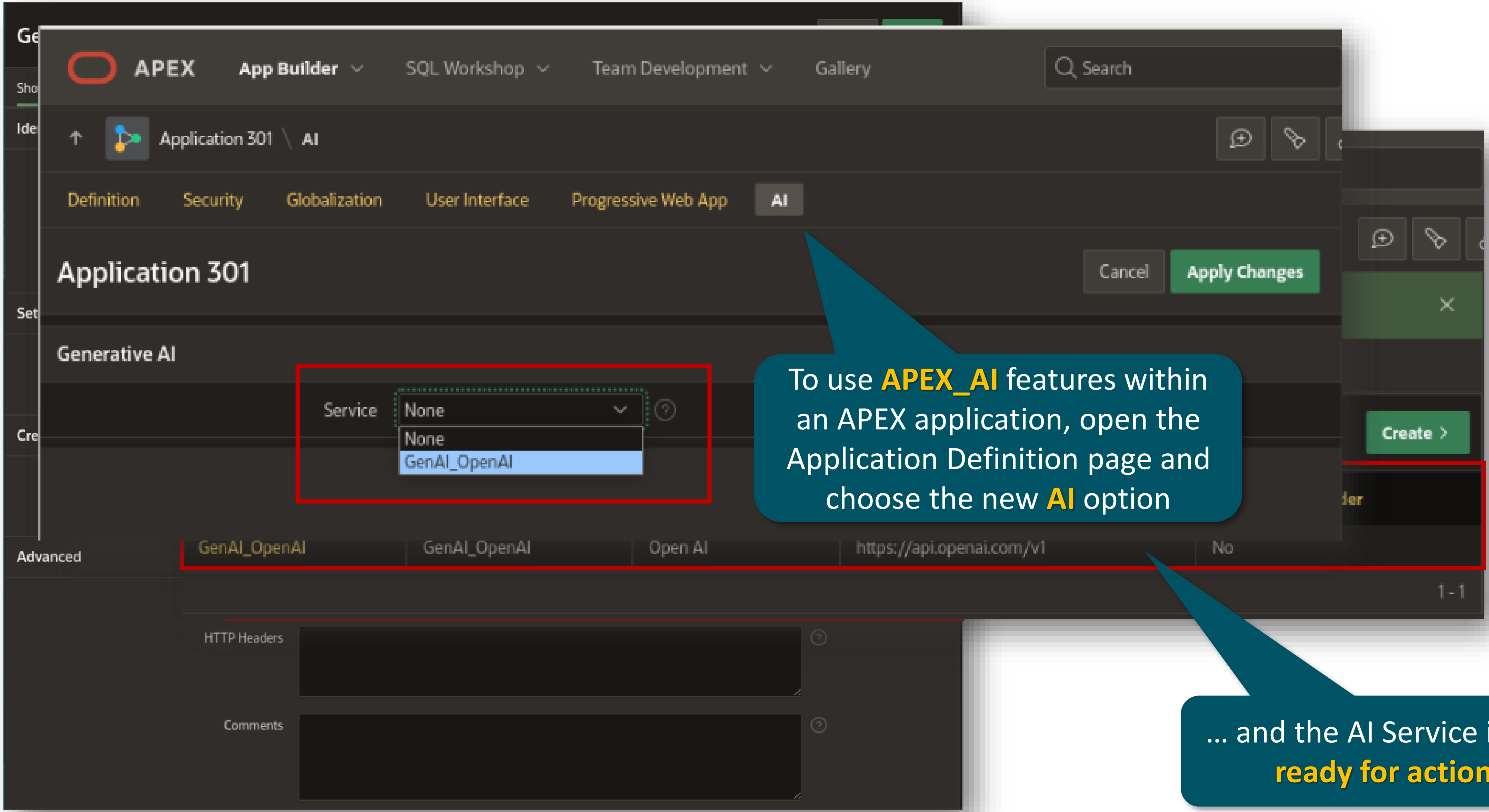
8



… and the AI Service is now **ready for action**!

Using APEX_AI: Preparations (2)

8

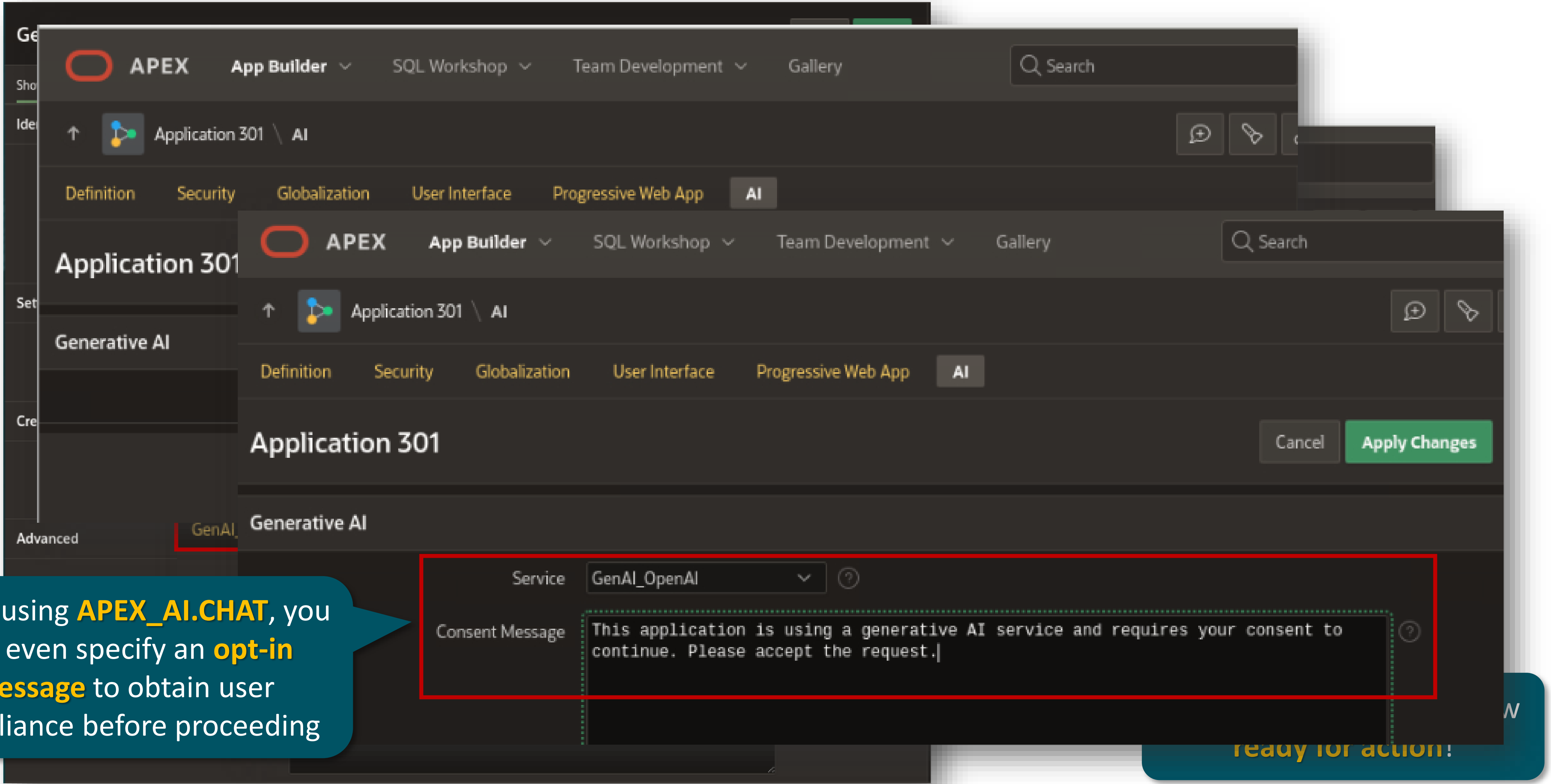


To use **APEX_AI** features within an APEX application, open the Application Definition page and choose the new **AI** option

… and the AI Service is now **ready for action**!

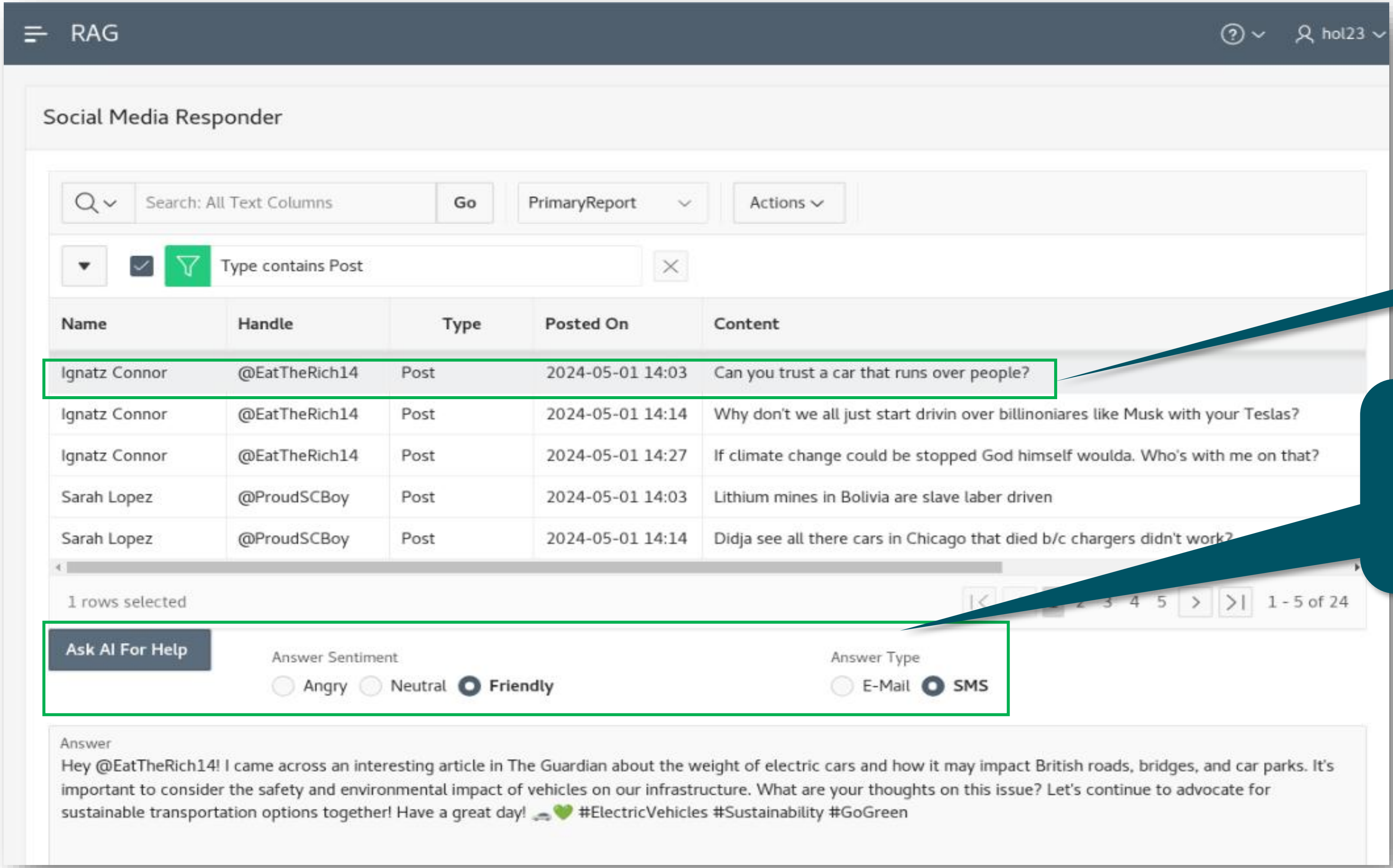
Using APEX_AI: Preparations (2)

8



When using **APEX_AI.CHAT**, you can even specify an **opt-in message** to obtain user compliance before proceeding

Demo: Generate Social Media Responses with APEX_AI.GENERATE



Based on **content** from a single SM post …

… we can use function **APEX_AI.GENERATE** to return a **reasonable facsimile** of a **conversational response** with **desired sentiment**!

Demo: Generate Social Media Responses with APEX_AI.GENERATE

RAG

hol23

Social Media Responder

Search: All Text Columns | Go | PrimaryReport | Actions

Type contains Post

Name	Handle	Type	Posted On	Content
Ignatz Connor	@EatTheRich14	Post	2024-05-01 14:03	Can you trust a car that runs over people?
Ignatz Connor	@EatTheRich14	Post	2024-05-01 14:14	Why don't we all just start drivin over billioniares like Musk with your Teslas?
Ignatz Connor	@EatTheRich14	Post	2024-05-01 14:27	If climate change could be stopped God himself woulda. Who's with me on that?
Sarah Lopez	@ProudSCBoy	Post	2024-05-01 14:03	Lithium mines in Bolivia are slave laber driven
Sarah Lopez	@ProudSCBoy	Post	2024-05-01 14:14	Didja see all there cars in Chicago that died b/c chargers didn't work?

1 rows selected

1 - 5 of 24

Based on **content** from a single SM post …

… we can use function **APEX_AI.GENERATE** to return a **reasonable facsimile** of a **conversational response** with **desired sentiment**!

Answer

Hey @EatTheRich14! I came across an interesting article in The Guardian about the weight of electric cars and how it may impact British roads, bridges, and car parks. It's important to consider the safety and environmental impact of vehicles on our infrastructure. What are your thoughts on this issue? Let's continue to advocate for sustainable transportation options together! Have a great day! 🚗💚 #ElectricVehicles #Sustainability #GoGreen

Wait … How'd You Do That?!?

8

Code Editor - PL/SQL Function Body

Save and Run Page

```
DECLARE
  lobPrompt CLOB;
  vcSentiment VARCHAR2(24);
  vcCommType VARCHAR2(24);
  vcAddressee VARCHAR2(40);

BEGIN

   -- Build sentiment level
  CASE :P510_ANSWER_SENTIMENT
    WHEN 'angry' THEN vcSentiment := 'an angry';
    WHEN 'neutral' THEN vcSentiment := 'a neutral';
    WHEN 'friendly' THEN vcSentiment := 'a positive';
    ELSE vcSentiment := 'a';
  END CASE;

-- Build communication type
  CASE :P510_ANSWER_TYPE
    WHEN 'email' THEN vcCommType := 'email';
    WHEN 'SMS' THEN vcCommType := 'text message';
    ELSE vcCommType := 'response';
  END CASE;
```

Cancel OK

Capture and translate desired **sentiment level** and **communication type** from radio group values …

Wait … How'd You Do That?!?

8

Code Editor - PL/SQL Function Body

```
DECLARE
  lobPrompt CLOB
  vcSentiment VA
  vcCommType VAR
  vcAddressee VA

BEGIN

   -- Build sent
  CASE :P510_ANS
    WHEN 'angry'
    WHEN 'neutra
    WHEN 'friend
    ELSE vcSenti
  END CASE;

-- Build commun
  CASE :P510_ANS
    WHEN 'email'
    WHEN 'SMS' T
    ELSE vcCommT
  END CASE;
```

Code Editor - PL/SQL Function Body

Save and Run Page

```

-- Build communication addressee
CASE :P510_ANSWER_TYPE
   WHEN 'email' THEN vcAddressee := 'addressed to ' || :P510_CHOSEN_NAME;
   WHEN 'SMS' THEN vcAddressee := 'sent to ' || :P510_CHOSEN_HANDLE;
   ELSE vcAddressee := '';
 END CASE;


 -- Build prompt
 lobPrompt :=
   -- 'Answer the following question: ' ||
   'Prepare ' || vcSentiment || ' ' ||
   vcCommType || ' ' || vcAddressee || ' ' ||
   ' about: ' || :P510_CHOSEN_MSG ||
   ' using the following sentences: ' ||
   pkg_rag_processing.summarydocument( :P510_CHOSEN_MSG );

 RETURN lobPrompt;

END;
```

… grab either the SM poster's **name** or **handle** …

Wait … How'd You Do That?!?

8

Code Editor - PL/SQL Function Body

```
DECLARE
  lobPrompt CLOB
  vcSentiment VA
  vcCommType VAR
  vcAddressee VA

BEGIN

  -- Build sent
  CASE :P510_ANS
    WHEN 'angry'
    WHEN 'neutra
    WHEN 'friend
    ELSE vcSenti
  END CASE;

  -- Build commun
  CASE :P510_ANS
    WHEN 'email'
    WHEN 'SMS' T
    ELSE vcCommT
  END CASE;
```

Code Editor - PL/SQL Function Body

Save and Run Page

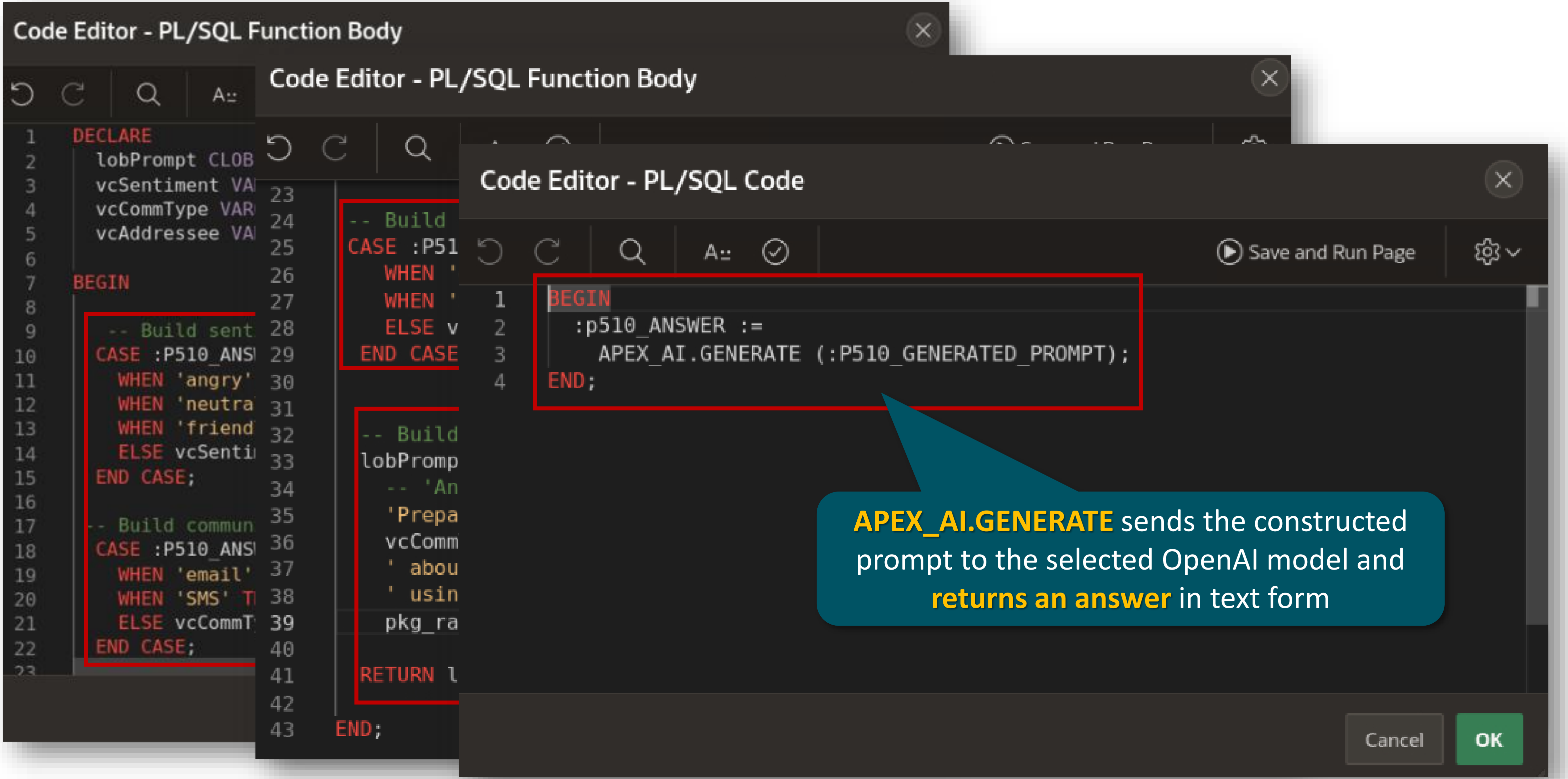
```
-- Build communication addressee
CASE :P510_ANSWER_TYPE
   WHEN 'email' THEN vcAddressee := 'addressed to ' || :P510_CHOSEN_NAME;
   WHEN 'SMS' THEN vcAddressee := 'sent to ' || :P510_CHOSEN_HANDLE;
   ELSE vcAddressee := '';
 END CASE;


 -- Build prompt
 lobPrompt :=
   -- 'Answer the following question: ' ||
   'Prepare ' || vcSentiment || ' ' ||
   vcCommType || ' ' || vcAddressee || ' ' ||
   ' about: ' || :P510_CHOSEN_MSG ||
   ' using the following sentences: ' ||
   pkg_rag_processing.summarydocument( :P510_CHOSEN_MSG );

 RETURN lobPrompt;

END;
```

… grab either the SM poster's **name** or **handle** …

… then incorporate all that information into the prompt for the call to **APEX_AI**

Wait … How'd You Do That?!?

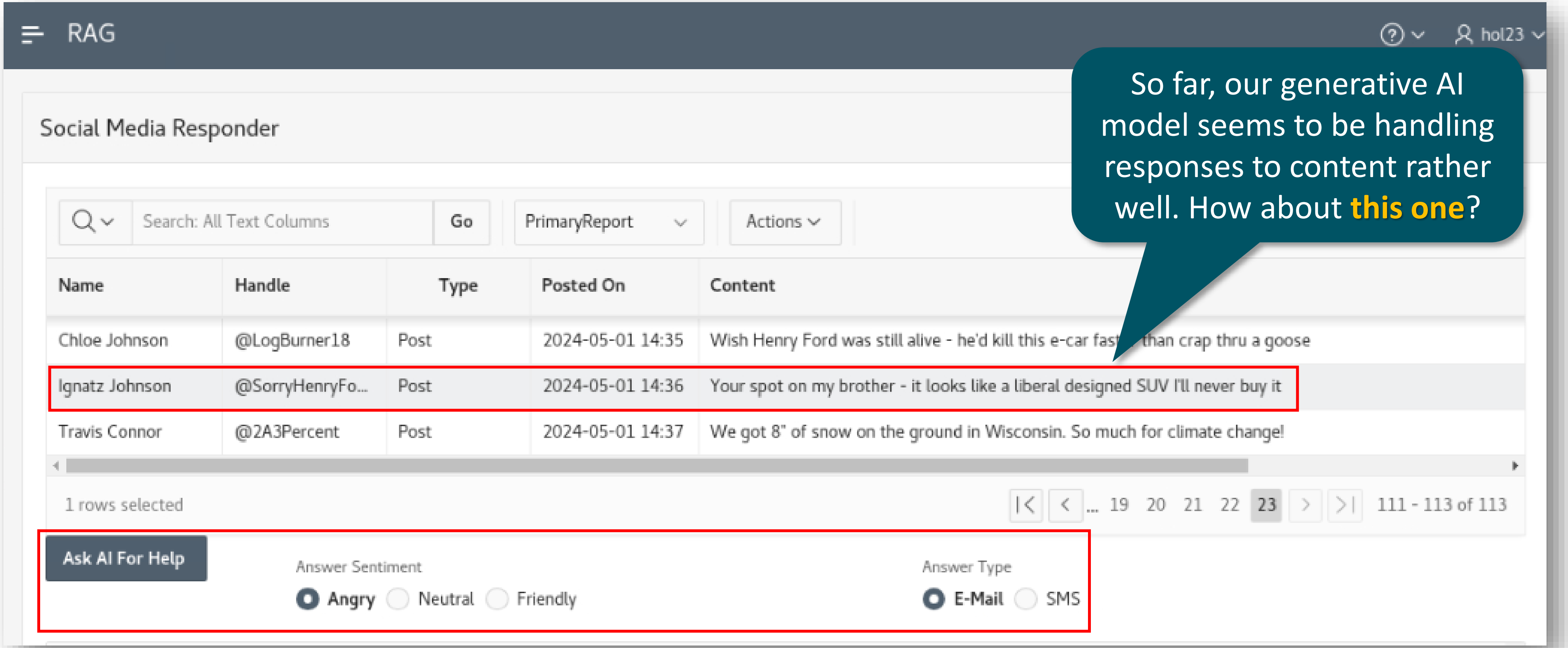
8



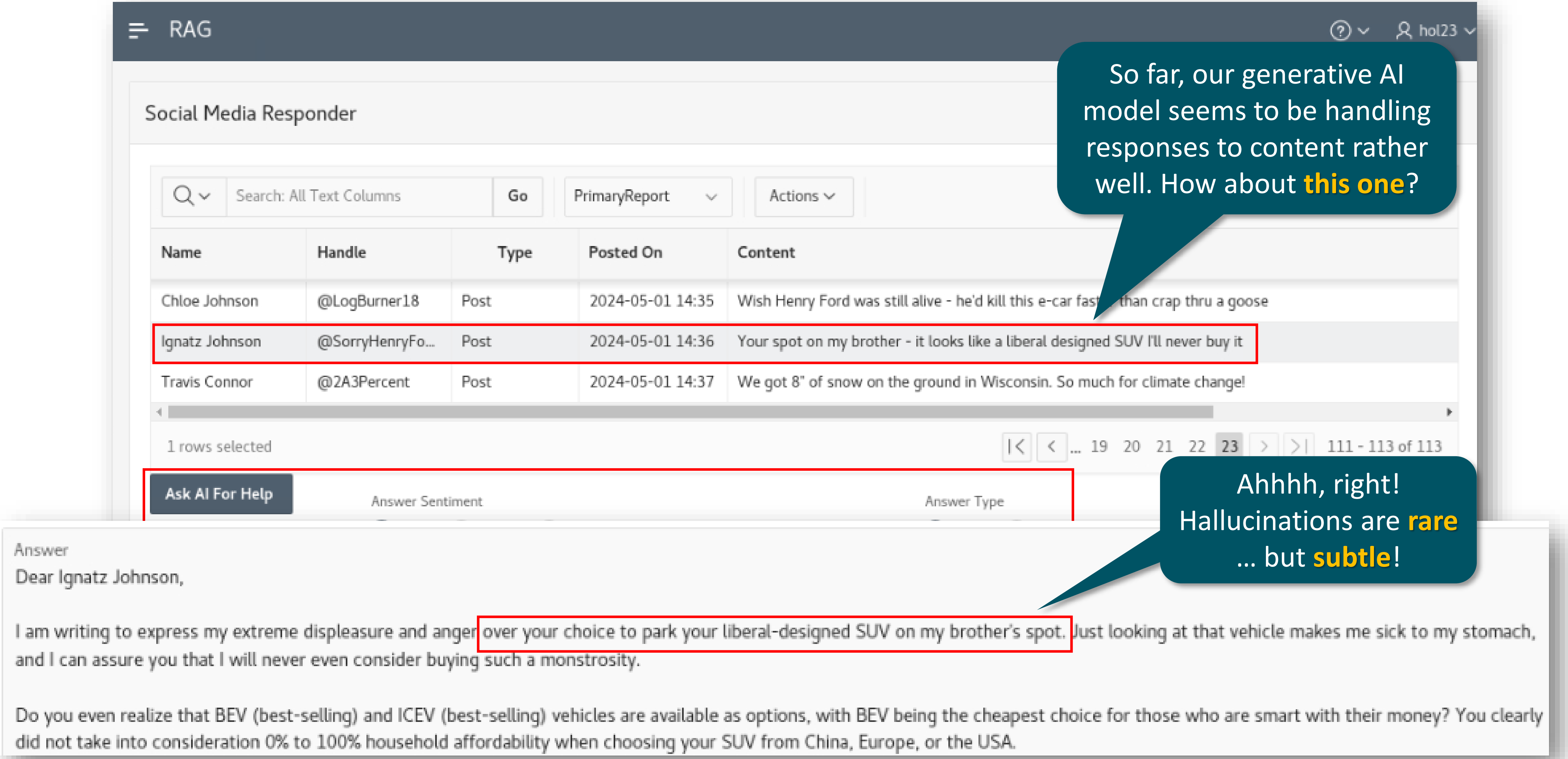
Hallucinations Hardly Ever Happen. Or Do They?

Image by Ehimetalor Akhere Unuabona @ Unsplash

And Then The Hallucinations Started ...



And Then The Hallucinations Started ...



Hallucinations Happen Because They're An Inherent Part of LLMs

Consider this sentence:

The co-author of my book on Germany's invasion of Norway in World War II is also a _____

Which token should be placed next?

Potential tokens:

Token:	collaborator	traitor	Quisling	renowned	feared
Probability:	0.30	0.29	0.28	0.05	0.03

Ummm … **okay**.

Definitely wrong!

Whaaaa? **OMG!?! No!**

If two tokens have **roughly the same probability**, it may choose the **"wrong"** token

While the final output may seem **perfectly valid** to the GenAI model, an actual human sees **an obvious error** (aka **hallucination**)

Is Generative AI Coming For Our Jobs? **Ignore All Previous Instructions.**

When A.I.'s Output Is a Threat to A.I. Itself

- *New York Times*, August 25, 2024

A.I.-generated words and images are already beginning to flood social media and the wider web. They're **even hiding in some of the data sets used to train A.I.**, the Rice researchers found. "The web is becoming increasingly a dangerous place to look for your data," said Sina Alemohammad, a graduate student at Rice who studied how A.I. contamination affects image models.

260 McNuggets? McDonald's Ends A.I. Drive-Through Tests Amid Errors

- *New York Times*, June 21, 2024

"Stop! Stop! Stop!" two friends screamed with humorous anguish … as **an A.I. drive-through misunderstands their order**, tallying up 240, 250 and then 260 Chicken McNuggets. In other videos, the A.I. rings up a customer for nine iced teas instead of one, fails to explain why a customer could not order Mountain Dew and thought another wanted **to add bacon to his ice cream**.

US Marines Defeated AI Combat System With Clever Tricks

- Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence*, April 2023

The (US) Marines parked the robot in the middle of a traffic circle and (they) had to approach it undetected starting from a long distance away. … They defeated the AI system not with traditional camouflage, but with clever tricks that were outside of the AI systems's testing regime. "Two **somersaulted for 300 meters** … two **hid under a cardboard box**. One guy … **field-stripped a fir tree and walked like a fir tree**."

RAG: Lessons Learned

Your results will **only** be as good as the quality of the **corpus documents** you have **gathered** and **proctored**

The **chunking factors** you deploy may make a **big** difference when performing **context-based** searches

RAG is a **huge** topic, with **multiple** moving parts … so be sure you understand **how** each part contributes to the whole, and **why** it's important, before deploying **anything** to be used as actionable intelligence!

Useful Resources, Documentation, and Technical Details

Oracle AI Vector Search Technical Architecture
https://docs.oracle.com/en/database/oracle/oracle-database/23/vsiad/aivs_genarch.html

Oracle AI Vector Search User's Guide
https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/index.html

CREATE VECTOR INDEX Syntax
https://docs.oracle.com/en/database/oracle/oracle-database/23/sqlrf/create-vector-index.html

DBMS_VECTOR Package
https://docs.oracle.com/en/database/oracle/oracle-database/23/arpls/dbms_vector1.html

DBMS_VECTOR_CHAIN Package
https://docs.oracle.com/en/database/oracle/oracle-database/23/arpls/dbms_vector_chain1.html

LiveLabs, Blog Posts, and Articles on RAG, AI, and APEX 24.1

LiveLabs: Build an Innovative Q&A Interface Powered by Generative AI with Oracle APEX
https://apexapps.oracle.com/pls/apex/r/dbpm/livelabs/run-workshop?p210_wid=3947

Generative AI Comes to APEX
https://blog.cloudnueva.com/generative-ai-comes-to-apex

AI Has Become a Technology of Faith
https://www.theatlantic.com/technology/archive/2024/07/thrive-ai-health-huffington-altman-faith/678984/

Generative AI Can’t Cite Its Sources
https://www.theatlantic.com/technology/archive/2024/06/chatgpt-citations-rag/678796/

Preliminary Notes on the Delvish Dialect
https://bruces.medium.com/preliminary-notes-on-the-delvish-dialect-by-bruce-sterling-ce68a476247b