# **VoiceRAG**: Habla con tus datos multimodales en tiempo real

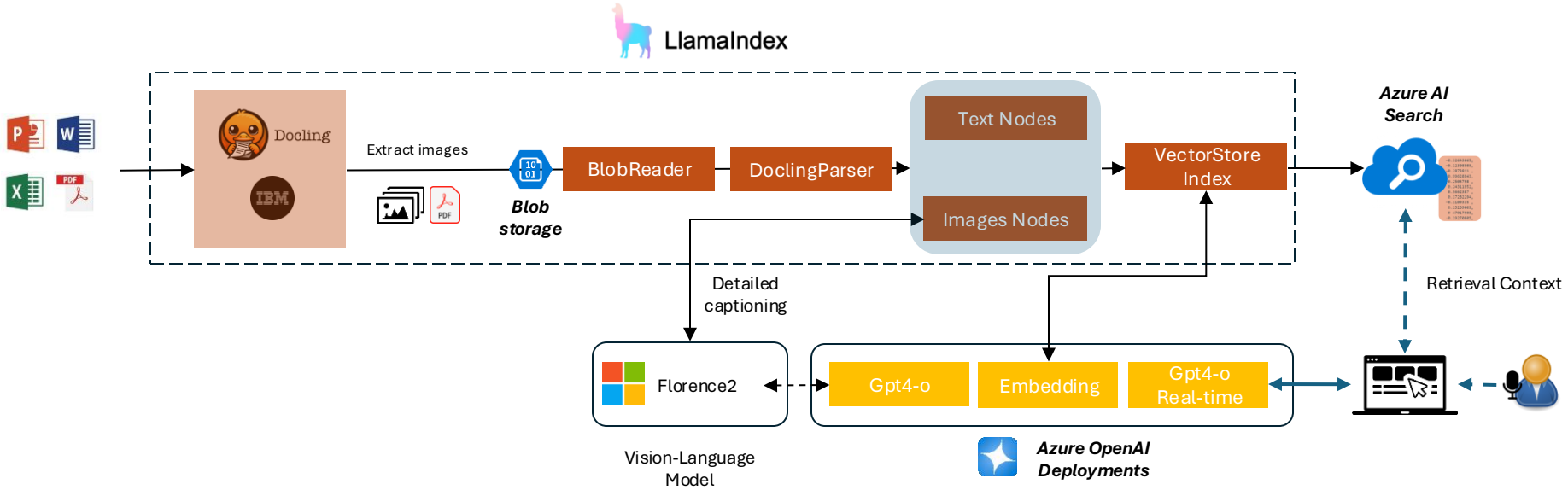
Rodrigo Cabello

Principal AI Research Engineer@Plain Concepts

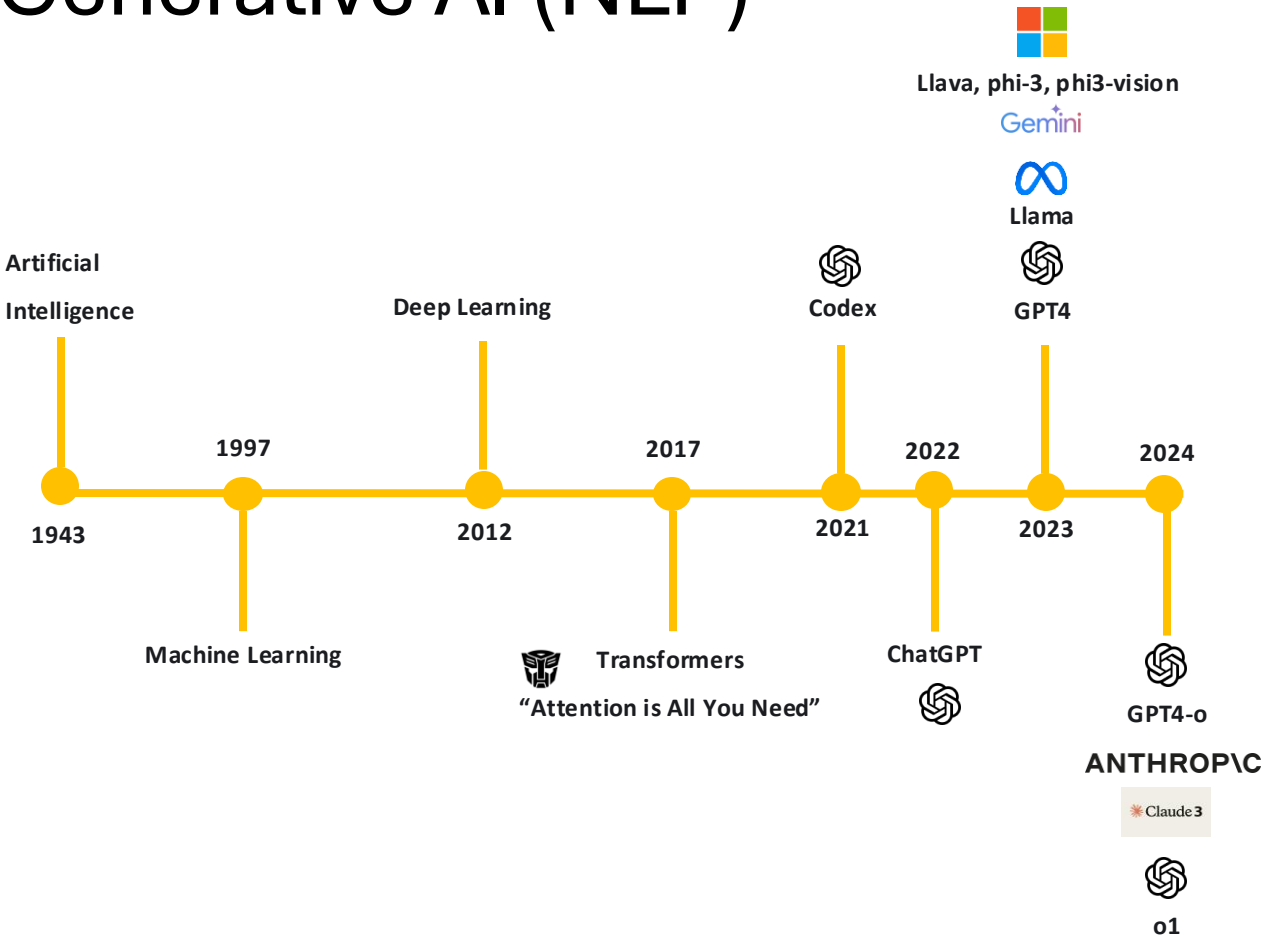
Microsoft® Most Valuable Professional

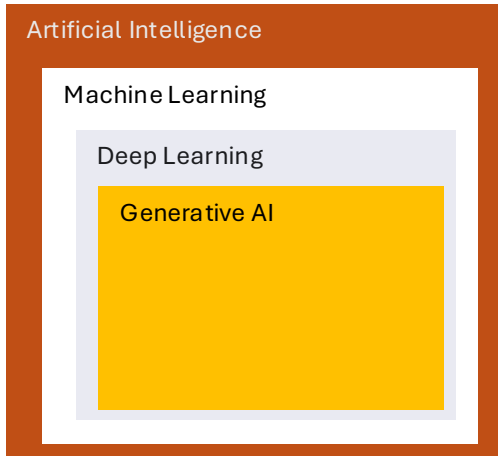
rodrigocabello

# Architecture



# Generative AI (NLP)



# Generative AI (NLP)

DOT NET MLG

## Classical NLP

- Sentiment Analysis
- Named Entity Recognition
- Entity Extraction
- Classification
- Summary
- ...

## Generative AI NLP

Sentiment Analysis
Entity Recognition
Entity Extraction
Classification

Q&A
Style transfer
Summary
Code Generation
...


Verisk

# How LLM's works?


Probability Distribution over next word/token

0.1 Cafe

0.05 Hospit

0.15

0.1

in out

We need to stop

We nee

nizing

thropomorphizing Chat

d to stop anthropomorphizing ChatG

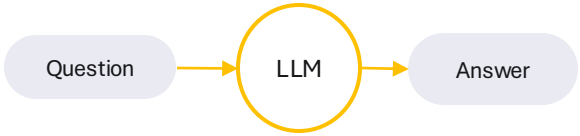
We need to stop anthropomorphizing ChatGPT

We need to stop anthropomorphizing ChatGPT.

LLM's are not search engines!!!

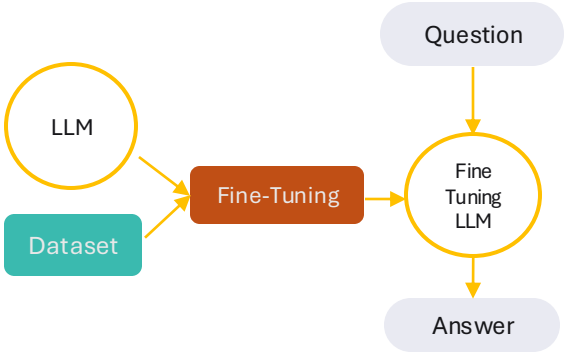
# Customizing LLM's

## Prompt Engineering



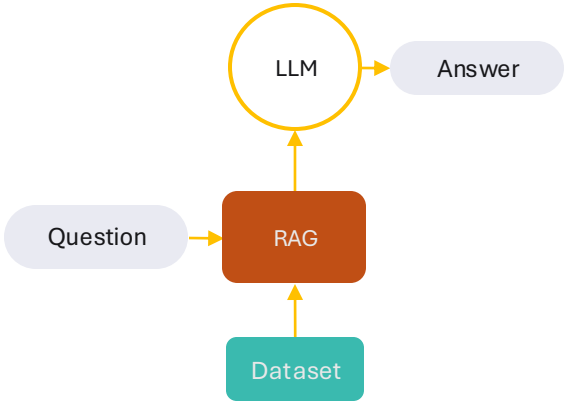


## Fine-tuning




*"Learning new behavior"*

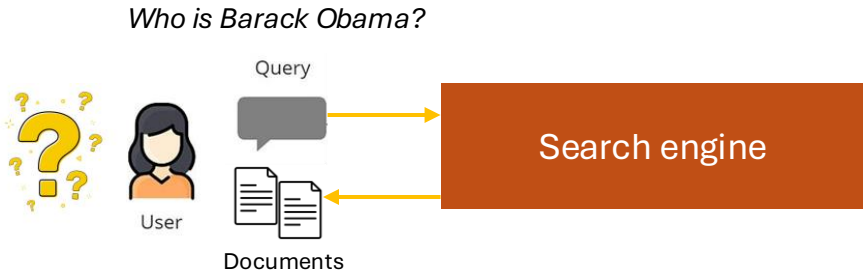
## Retrieval Augmented Generation




*"External Knowledge"*

# RAG
# (Retrieval Augmented Generation)

# Classical search



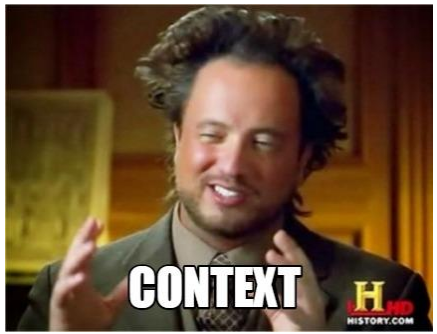

Keyword

Semantic search

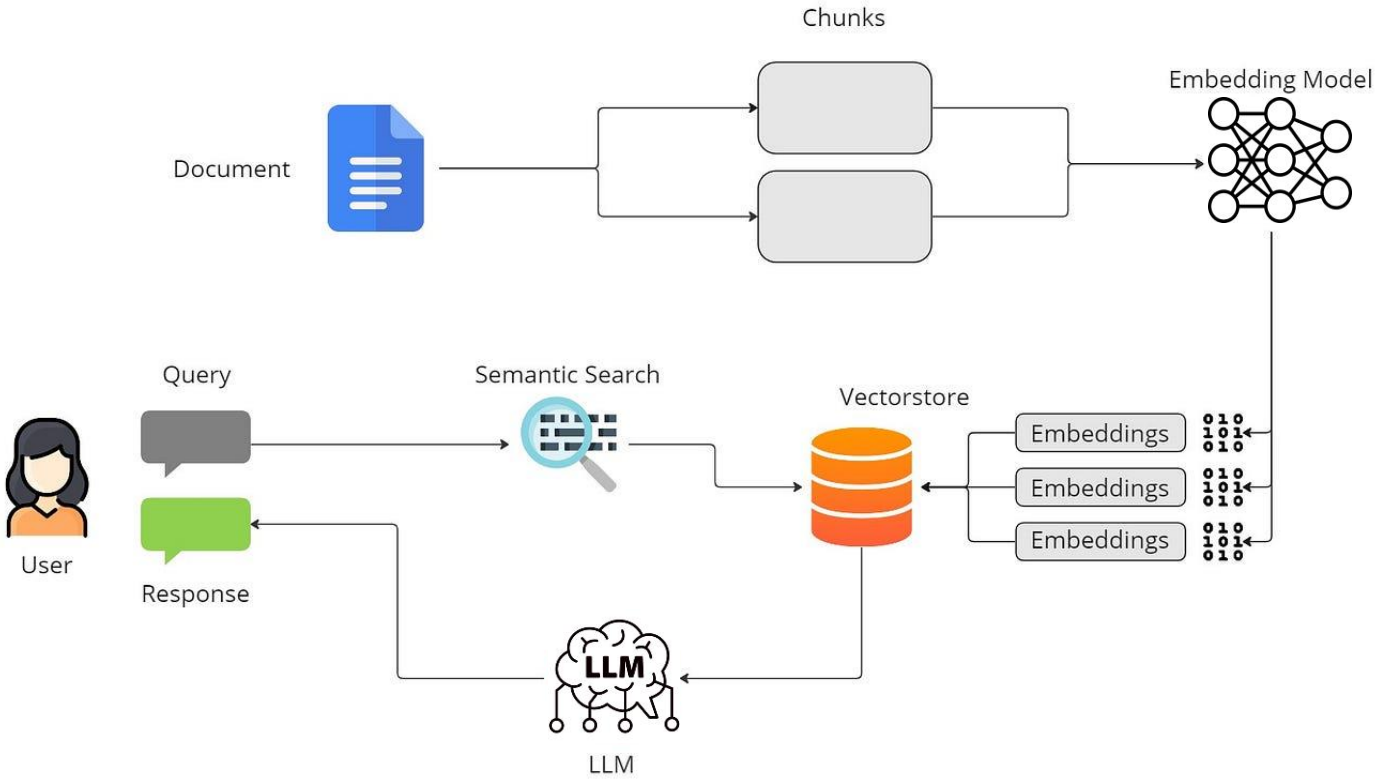
Hybrid search

Azure AI Search

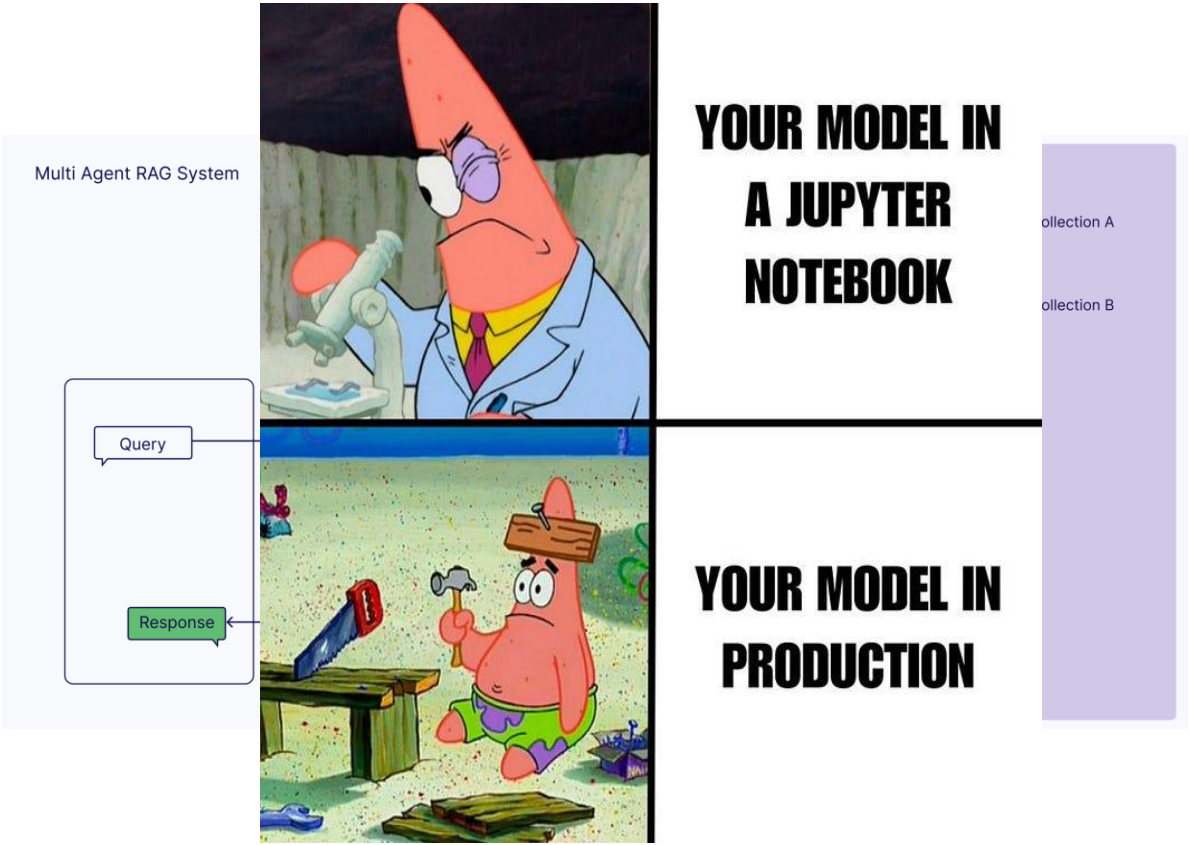
Elastic search



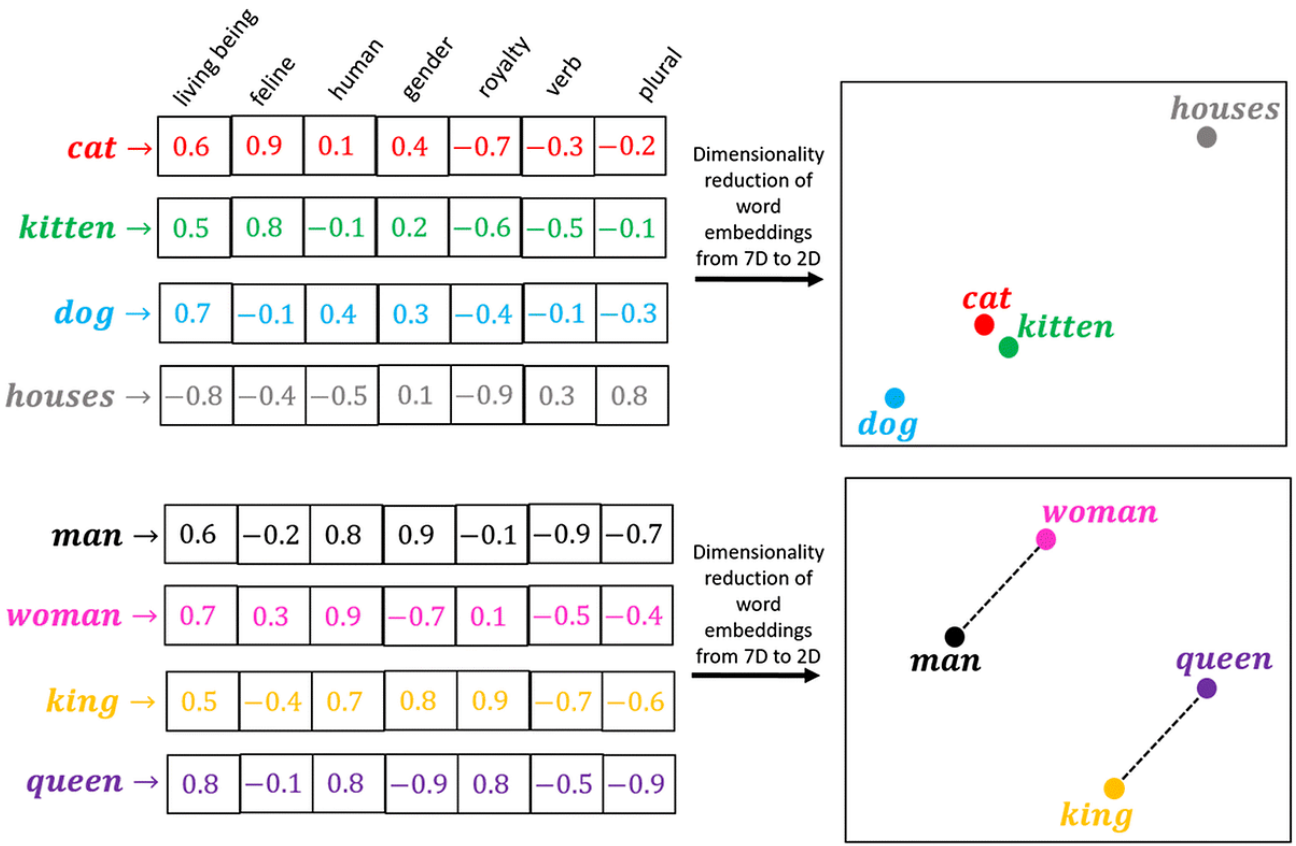
# RAG - Text


Chunks
Embedding Model
Document
Query
Semantic Search
Vectorstore
Embeddings
Embeddings
Embeddings
User
Response
LLM
LLM

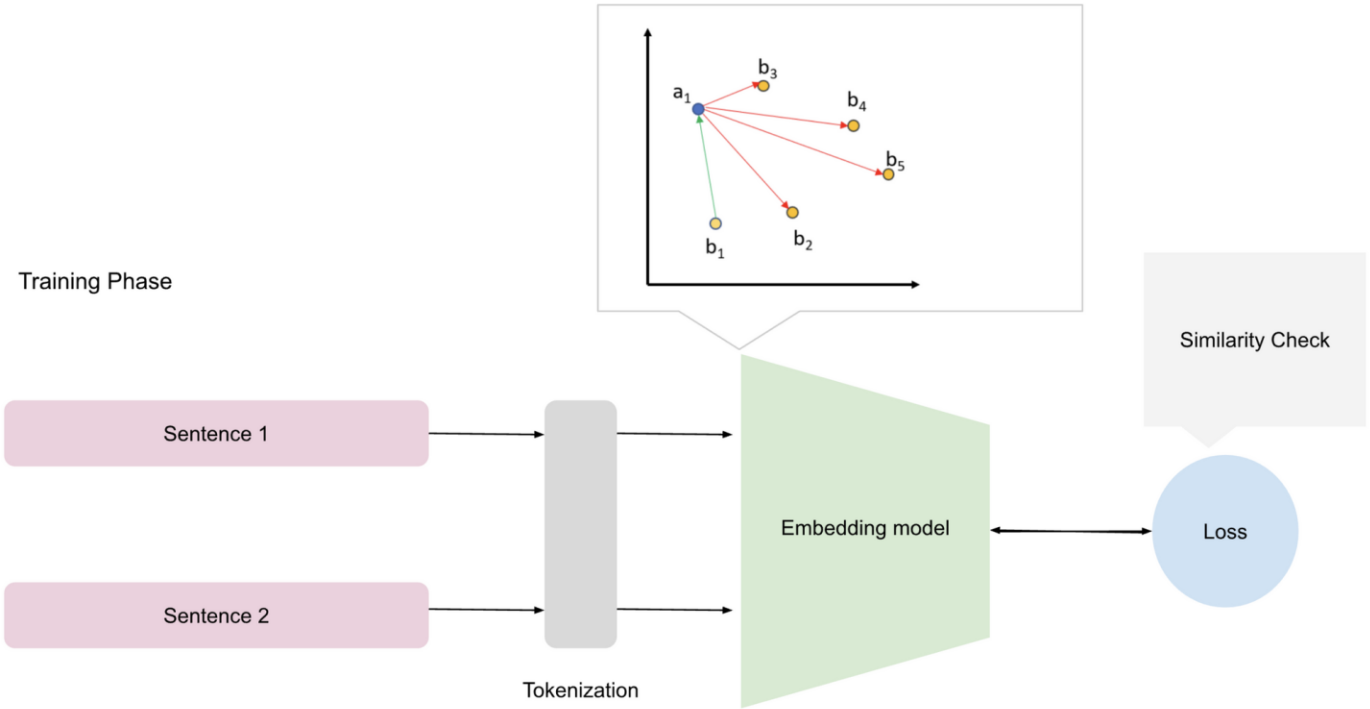
# Agentic RAG


Multi Agent RAG System
Query
Response
YOUR MODEL IN A JUPYTER NOTEBOOK
YOUR MODEL IN PRODUCTION
ollection A
ollection B

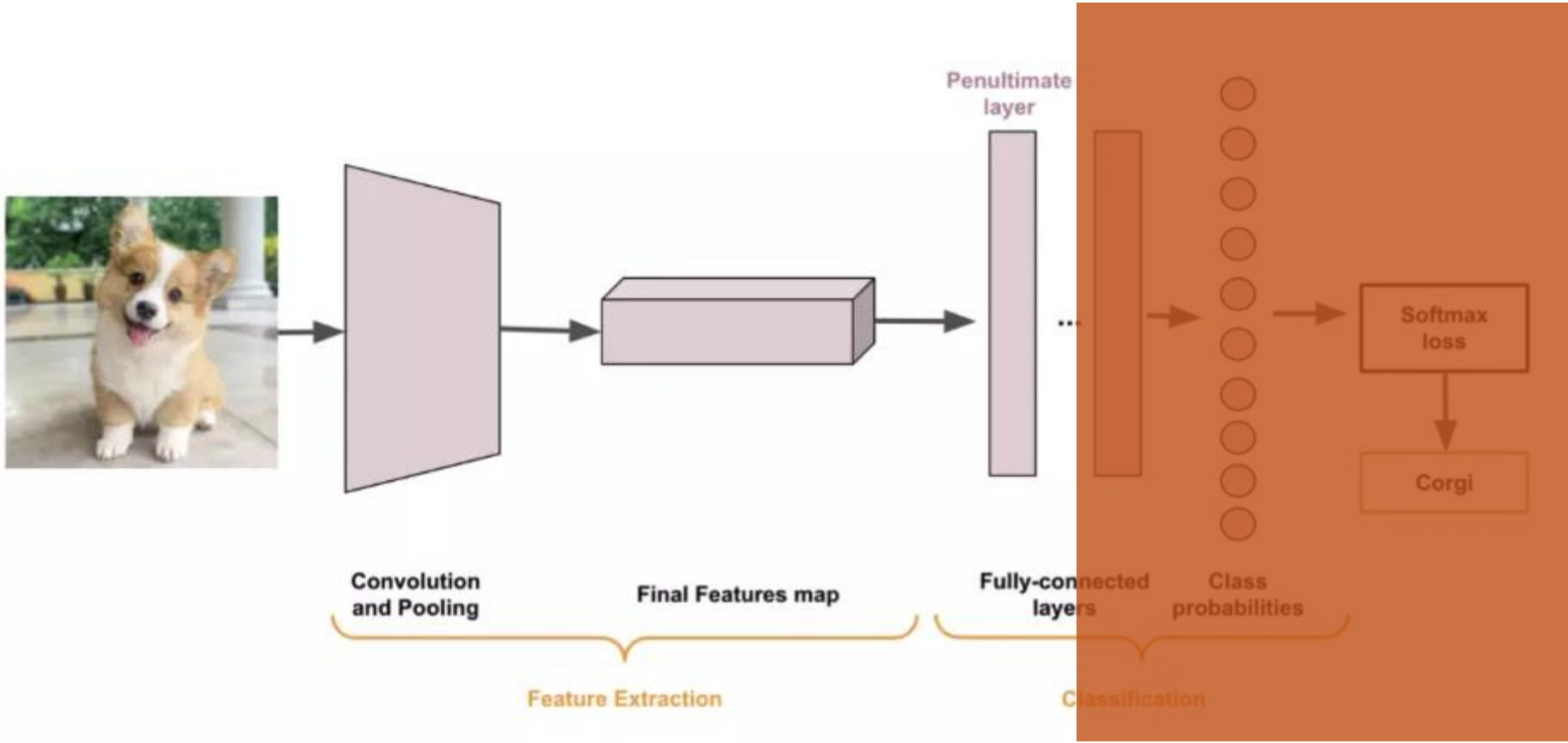
# Word Embedding



# Sentence embeddings



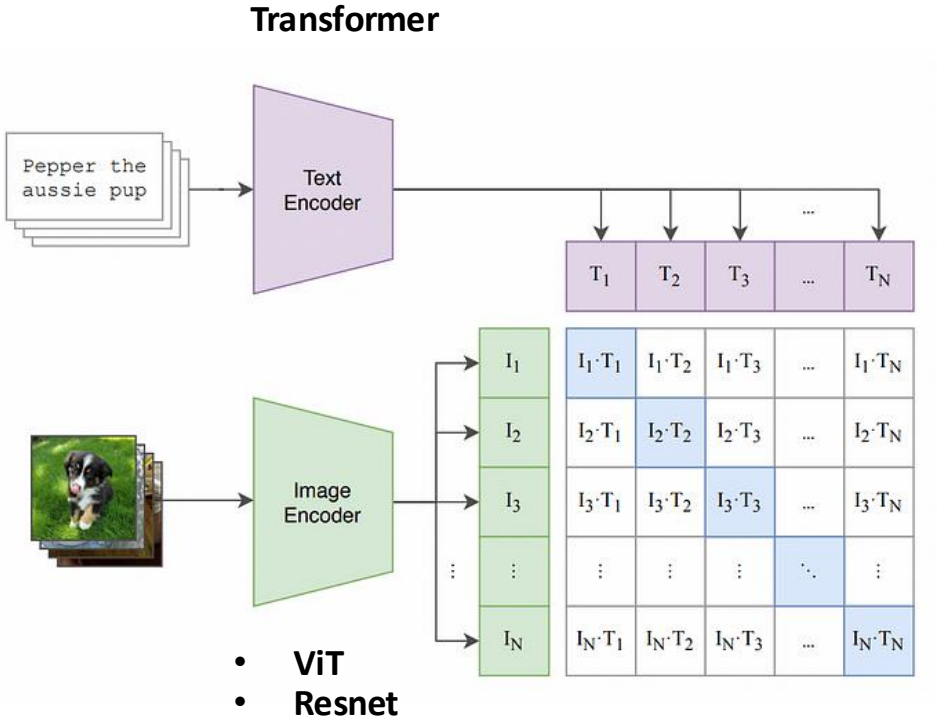
# Embeddings (Image)



# Embeddings

# Embeddings: Image + Text

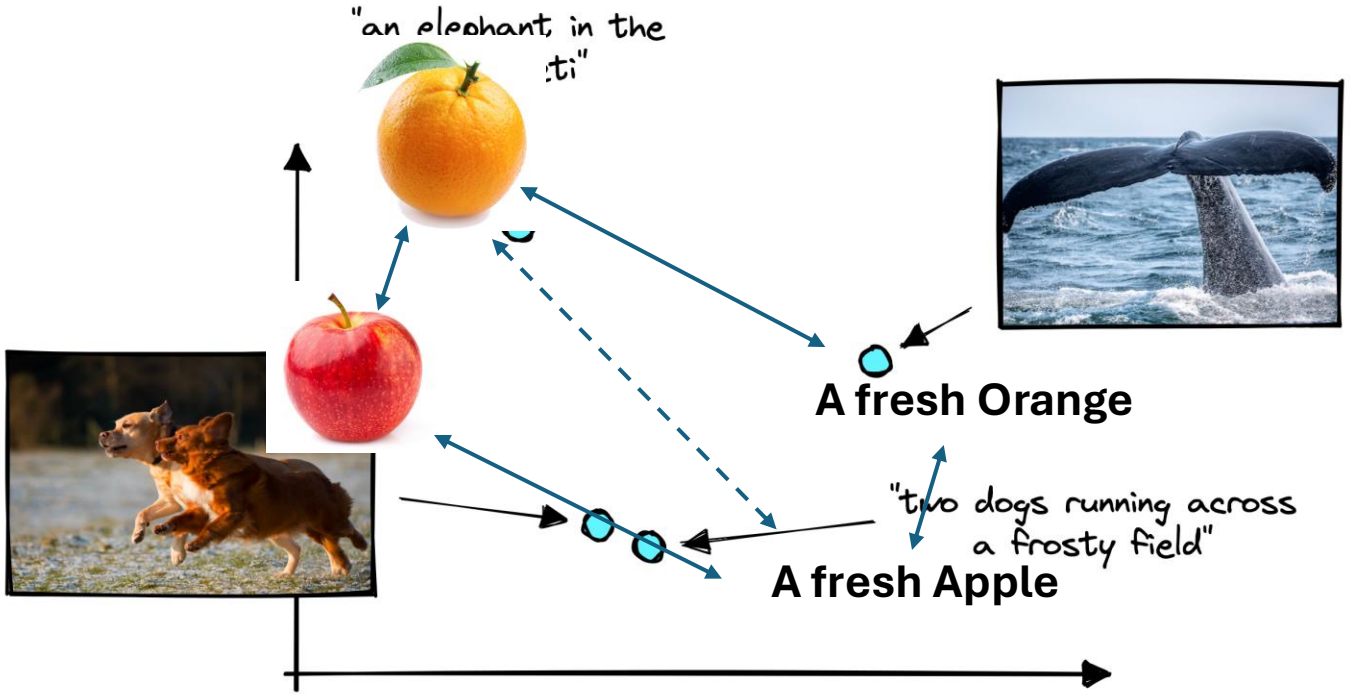
CLIP (Constrative Language-Image Pretraining)



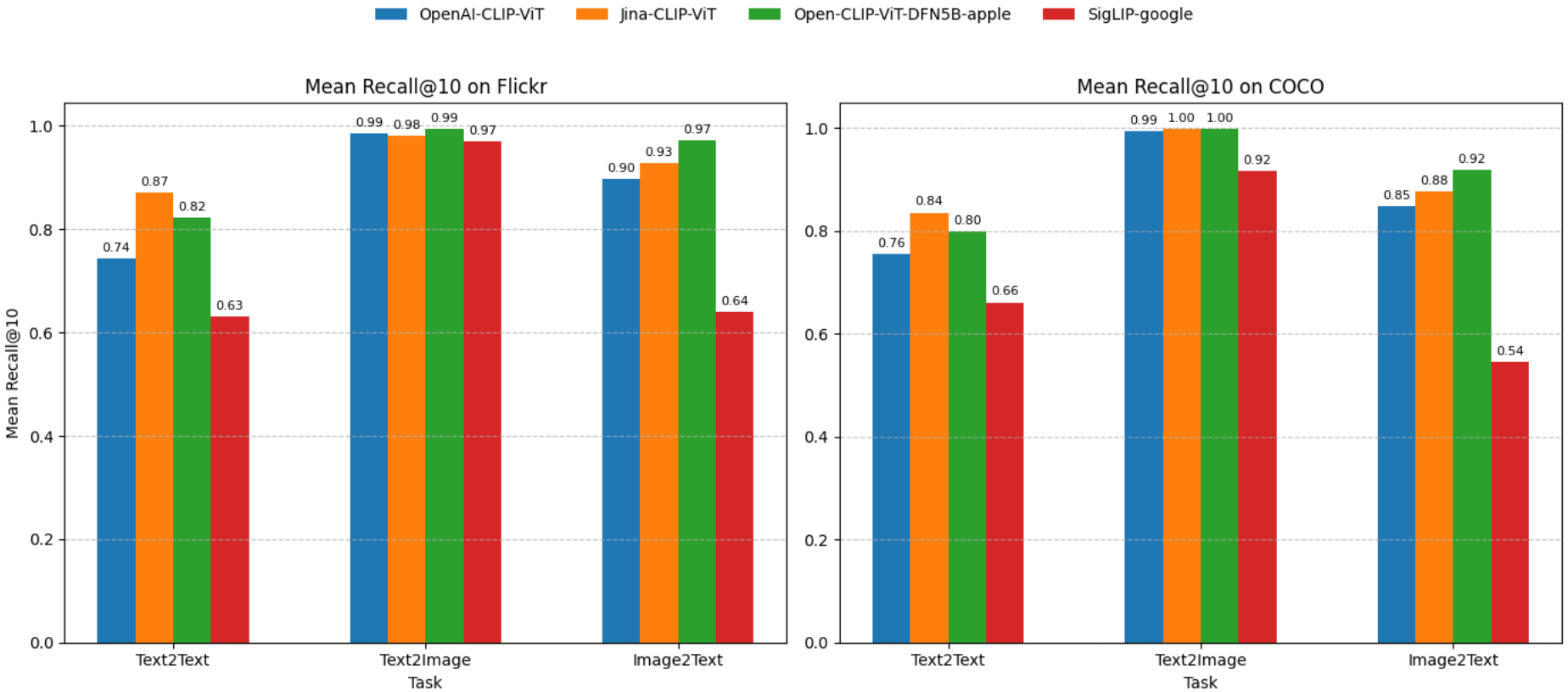

- **ViT**
- **Resnet**

# Embeddings: Image + Text

Multimodality GAP



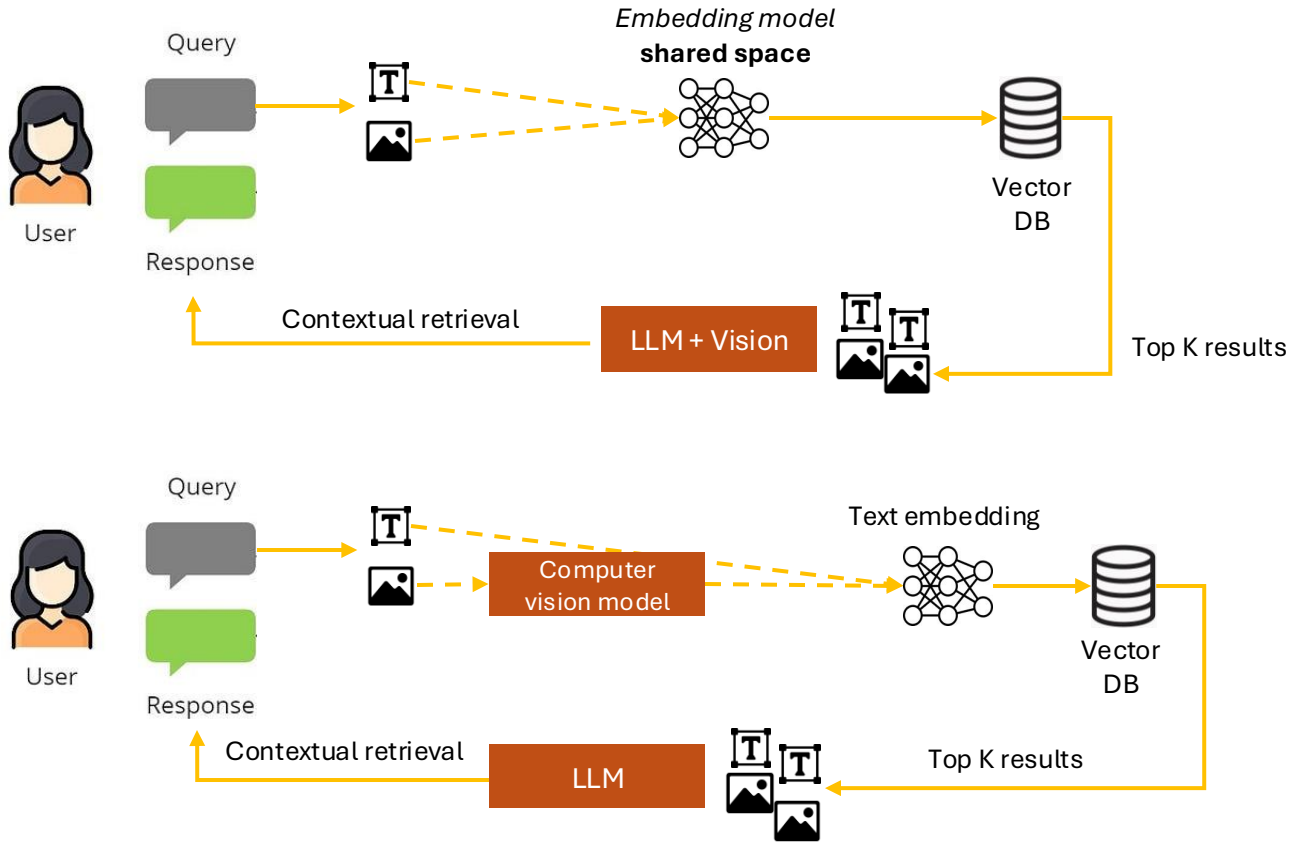
# Embeddings: Benchmark



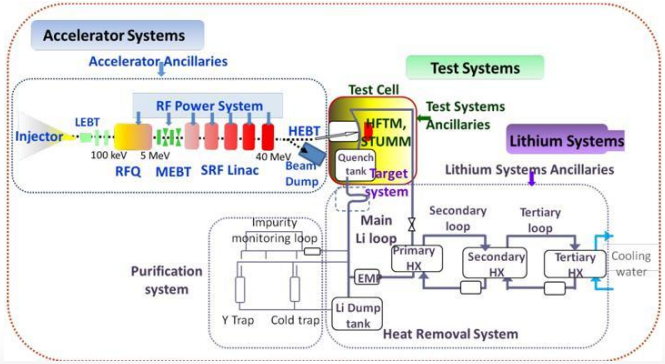
# Multimodal Embeddings

# Multimodal-RAG

# Multimodal RAG



# Image captioning




Florence2

{'<MORE_DETAILED_CAPTION>': 'The image is a flowchart that shows the process of a test system. It shows the different types of test systems that are used in the process.\n\nThe flowchart is divided into two sections. The top section is labeled "Accelerator Systems" and shows the accelerator system, which is a type of accelerator that is used to accelerate the acceleration of an animal\'s body. The accelerator system consists of an injector, a RF power system, and a HEBT (HEBT) system. The HEBT system is responsible for the transfer of energy from the body to the body, while the HEBT and HEBT systems are responsible for storing and transferring energy from one device to another. \n\nIn the center of the flowchart, there is a test cell, which allows the user to test the body\'s energy levels. The test cell is a device that helps to regulate the flow of energy through the body and the body. It also helps to monitor and monitor the energy levels in the body for better performance.\n- Injection system, the RF Power System is an injectable system that converts RF energy into RF energy, which can be used to power the body from the left side of the image to the right side. This system helps to reduce the amount of energy generated by the body in the test cell and to improve the overall performance of the animal.\n– Heat Removal System, which consists of a primary loop, secondary loop, tertiary loop, and cooling water, which helps to keep the body cool and prevent the animal from overheating.'}

The image depicts a schema
fusion materials testing or p
organized into three primar
Lithium Systems. Here's a d

1. **Accelerator Systems:**
   - **Injector** initiates th
   - **LEBT (Low Energy**
   - **RFQ (Radio Frequ**
   - **MEBT (Medium En**
     SRF Linac.
   - **SRF Linac (Superc**
     Propels particles to
   - **HEBT (High Energ**
     toward the beam d
   - The **Beam Dump** i
2. **Test Systems:**
   - The test systems a
     - **HFTM (High-**
       **Material Mod**
       radiation flux.

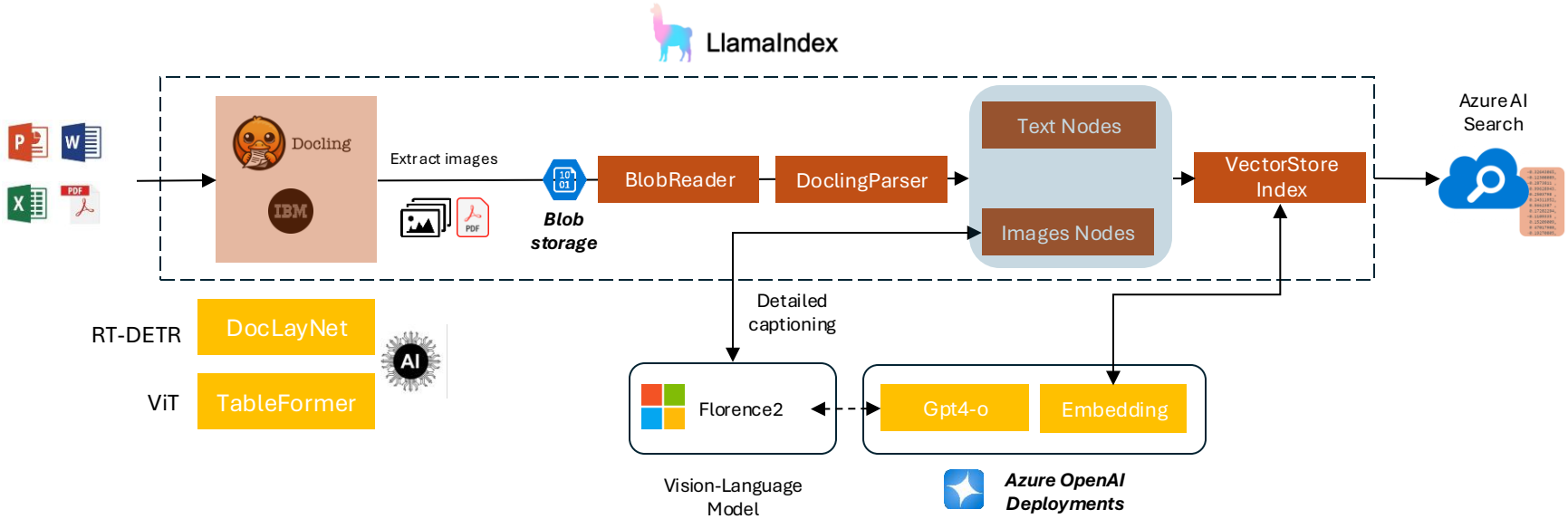
The image depicts a schematic layout of an experimental facility, likely related to fusion materials testing or particle accelerator-based experiments. The diagram is organized into three primary sections: Accelerator Systems, Test Systems, and Lithium Systems. Here's a detailed breakdown of the elements in the diagram:

1. **Accelerator Systems:**
   - **Injector** initiates the particle acceleration process.
   - **LEBT (Low Energy Beam Transport):** Transports the low-energy beam.
   - **RFQ (Radio Frequency Quadrupole):** Accelerates the beam to 100 keV.
   - **MEBT (Medium Energy Beam Transport):** Further beam transport to the SRF Linac.
   - **SRF Linac (Superconducting Radiofrequency Linear Accelerator):** Propels particles to 40 MeV.
   - **HEBT (High Energy Beam Transport):** Directs the high-energy beam toward the beam dump or test systems.
   - The **Beam Dump** is a mechanism to safely dissipate unused beam energy.
2. **Test Systems:**
   - The test systems are housed in a **Test Cell**, which includes:
     - **HFTM (High-Flux Test Module)** and **STUMM (Specific Test Unit for Material Modules):** Designed for testing materials under high-radiation flux.
     - **Target System:** Receives the high-energy beam and interacts with it.
   - A **Quench Tank** is included, possibly for cooling or emergency shutdown.
3. **Lithium Systems:**
   - The lithium systems manage the flow and heat removal of liquid lithium used in the target or other processes.
   - **Main Lithium Loop:** Includes a primary heat exchanger (HX) for removing heat generated in the target system.
   - **Secondary and Tertiary Loops:** Further remove heat, with cooling water as the final heat sink.
   - **Lithium Dump Tank:** A safety mechanism to hold lithium during shutdowns.
   - **Purification System:** Removes impurities from the lithium loop, utilizing a Y trap and cold trap for monitoring and impurity management.
4. **Ancillary Systems:**
   - Support subsystems for both the accelerator and lithium systems, including RF power systems and impurity monitoring.
5. **Heat Removal System:**
   - A series of heat exchangers (primary, secondary, tertiary) ensure the efficient dissipation of heat to maintain operational stability.

Overall, the diagram represents an integrated and highly specialized experimental setup, likely for testing materials under extreme conditions such as high radiation flux and thermal loads, as part of fusion or advanced material research.

# Ingestion Pipeline

# Ingestion Pipeline

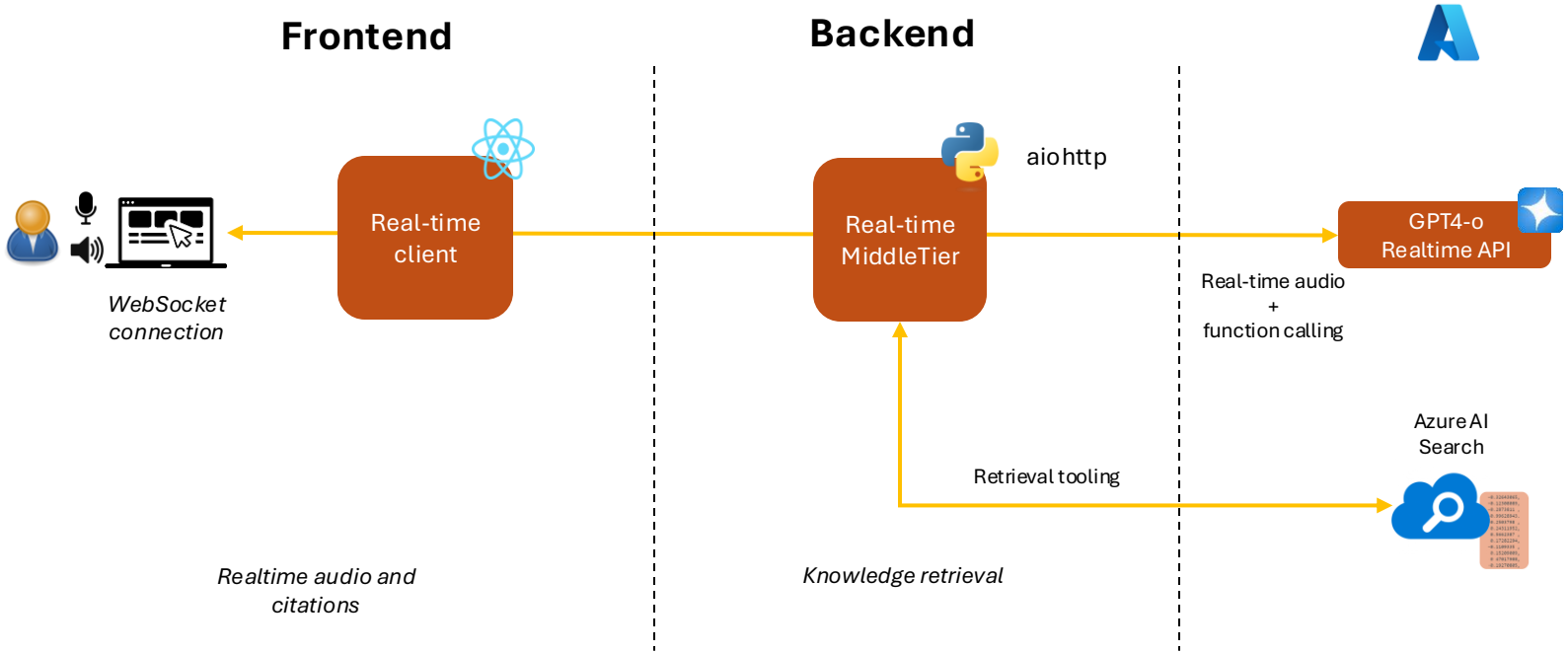


# Llama pipeline

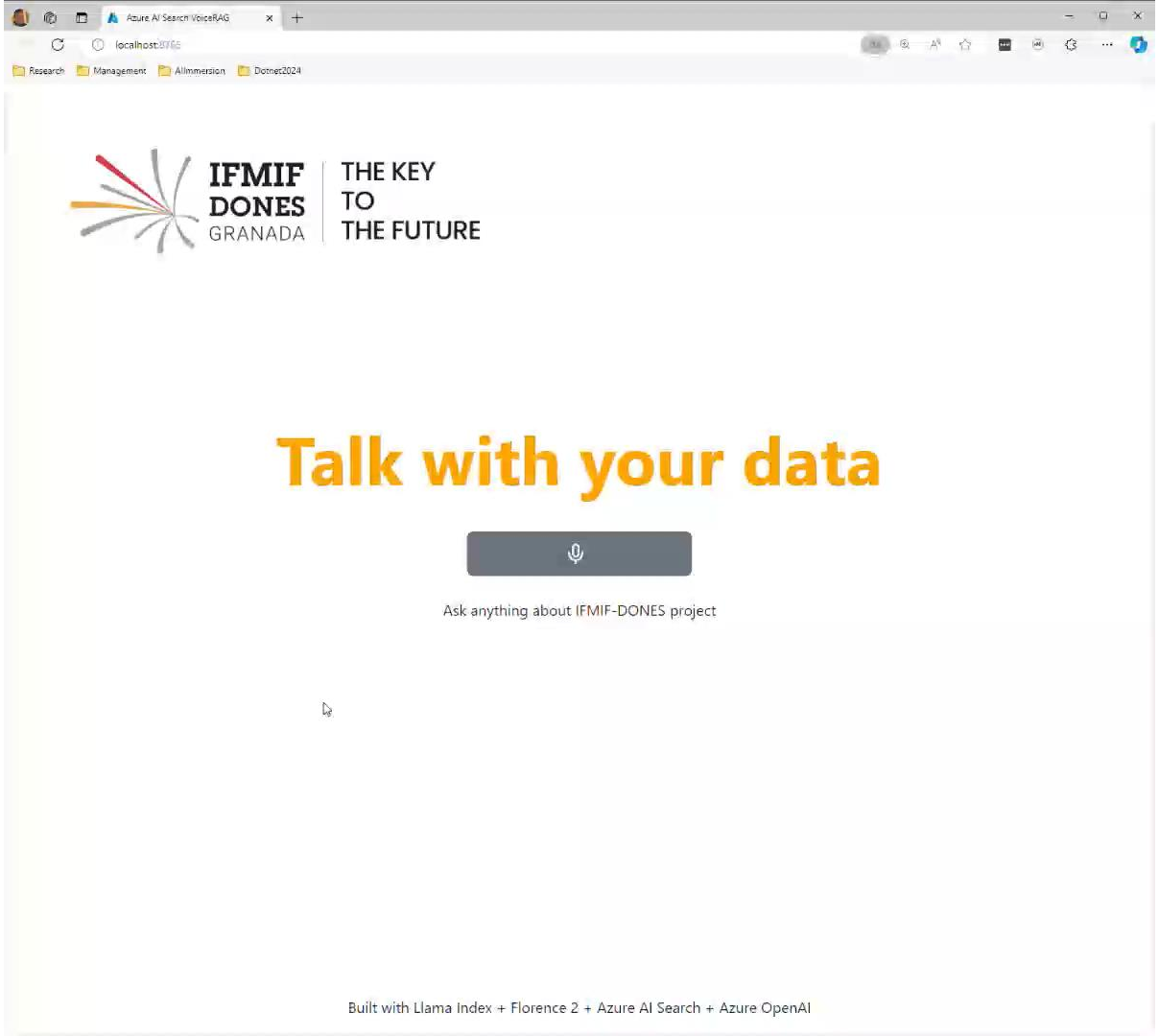
# VoiceRAG

# Real-time conversation

# Real-time conversation architecture



# VoiceRAG


Azure AI Search VoiceRAG
localhost:8765
Research
Management
AIImmersion
Dotnet2024
IFMIF
DONES
GRANADA
THE KEY
TO
THE FUTURE
Talk with your data
Ask anything about IFMIF-DONES project
Built with Llama Index + Florence 2 + Azure AI Search + Azure OpenAI


EL EVENTO SOBRE

# TECNOLOGÍAS CLOUD, WEB Y DATA

Verisk

SECTOR ALARM

prodware

LEMONCODE

METADEV

plain concepts