Artificial Intelligence in DAQ Systems (WP9.1)

AVOLIO G. FOR THE WP9.1 TEAM

Outline

Hiring and hardware procurement

Organization matters

LLM and Anomaly Detection studies (highlights)

Conclusions and outlook

Hiring and Hardware Procurement

ORIGIN hired

- Contract started on Feb 2025

Hardware for AI

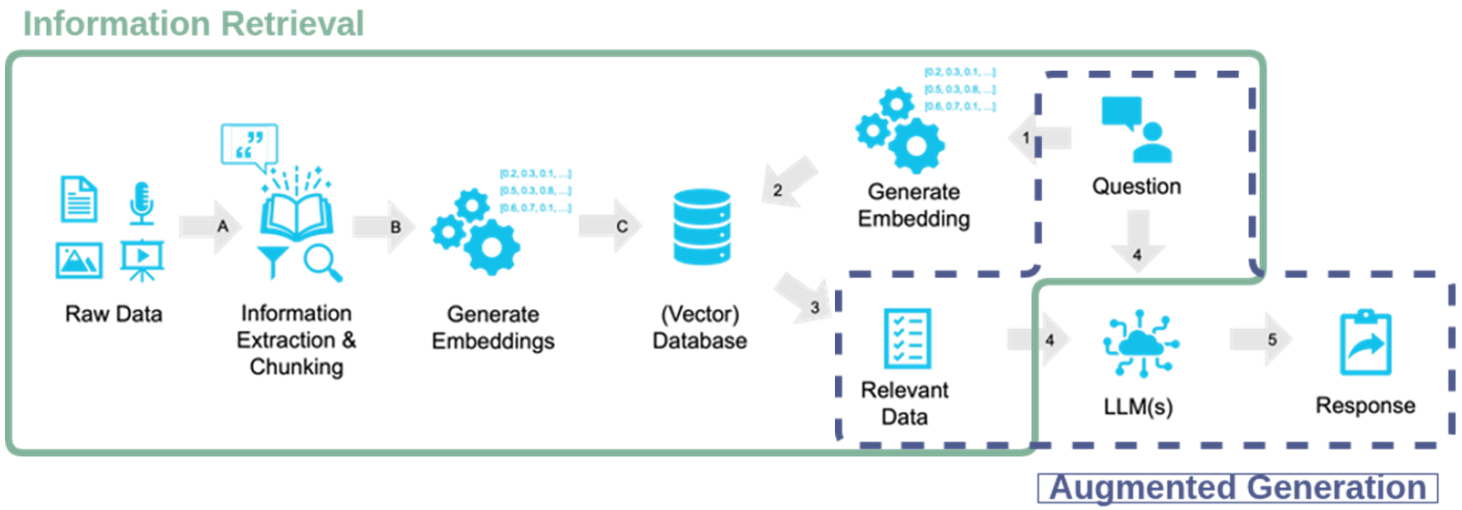
- **GPU** ordered and already **delivered**
  - NVIDIA RTX PRO 6000 Blackwell 96 GB
- **Server** to host the GPU ordered and ~~**waiting for delivery**~~ **delivered yesterday**

Organization Matters

Mattermost team: https://mattermost.web.cern.ch/rd-wp91

E-group: ep-dep-rnd-daq-wp91@cern.ch

Exploring LLMs and RAG Systems



A **Retrieval-Augmented Generation (RAG)** system is an AI model that combines two key steps:

- **Retrieval**: It searches a large collection of documents or data to find the most relevant information based on a user's query
- **Generation**: It uses a language model (like GPT) to generate a natural-language response, using the retrieved information as context

LLM - Information Retrieval

The Pipeline

LangChain

(Open-source Framework)

- Document processing, embedding creation and retrieval
- Allows integration with different models and databases

Hugging Face

(Open-source Platform)

- Provides access to thousands of pre-trained models
- Embedding creation, re-ranking and answer generation

ChromaDB

(Vector Database)

- Stores document embeddings for efficient retrieval
- Provides fast similarity search based on vector representations

LLM - Information Retrieval

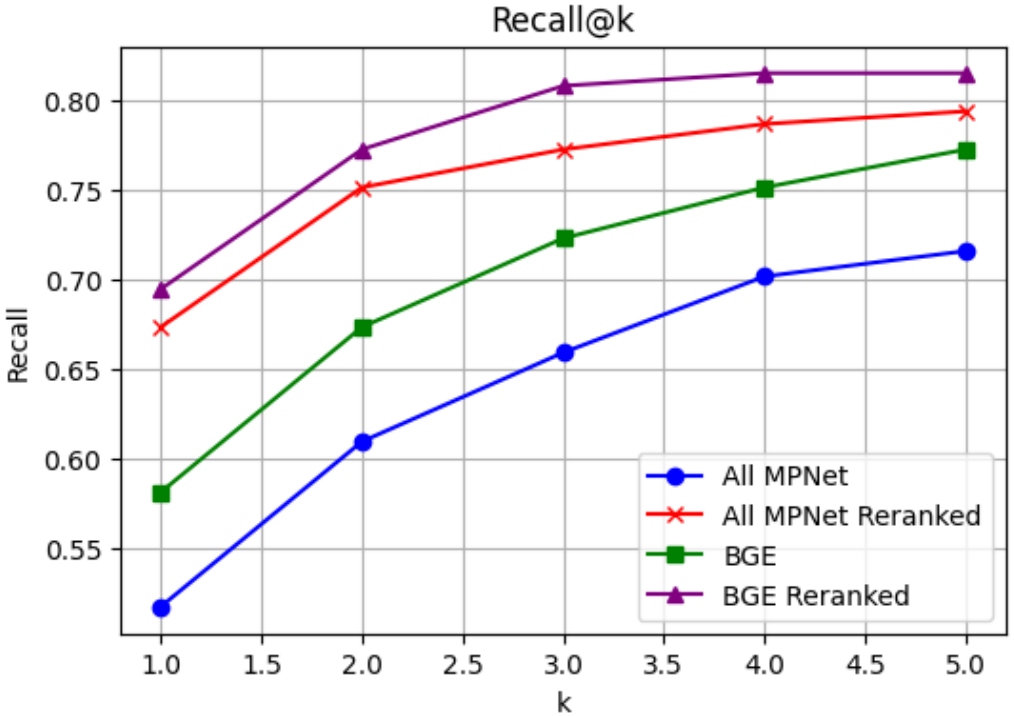
The Document Re-ranking

Why re-ranking

- Documents are retrieved using a vector search
- The resulting content may be semantically close but less relevant
- A system that is good at retrieval is not necessarily good at ranking

How

- After the vector search, documents are re-scored and re-ordered based on **semantic relevance** to the query



$$recall@K = \frac{TP}{TP + FN} = \frac{Number\ of\ relevant\ items\ in\ K}{Total\ number\ relevant\ items}$$

K is the number of selected top documents

LLM - Information Retrieval

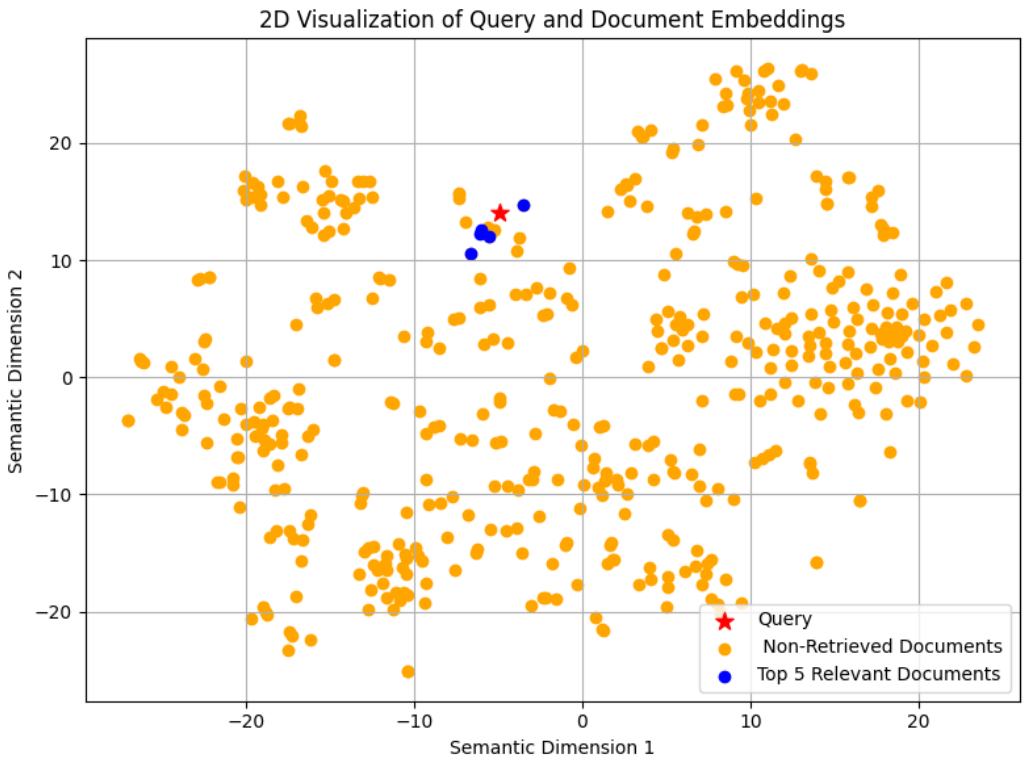
Matching Query and Documents

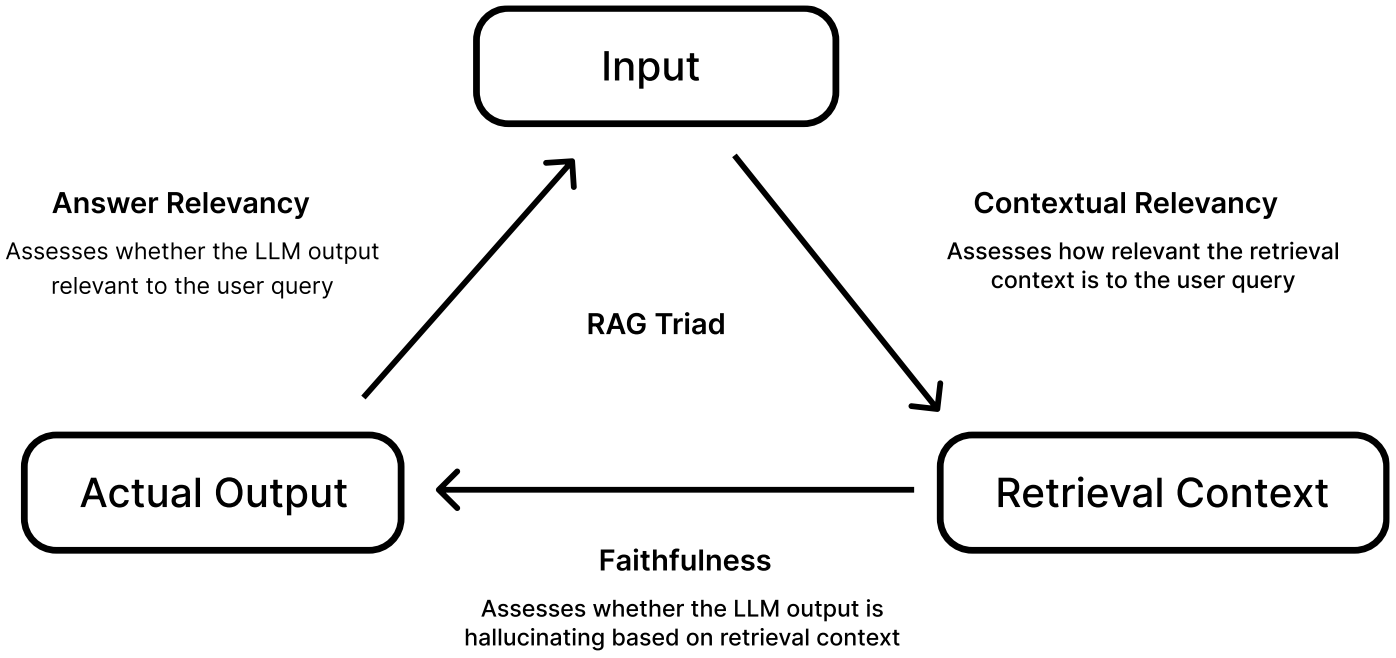
Simplified projection of a much higher-dimensional document representation

- Make it easier to visualize their semantic relationships

How are documents positioned based on their similarity to a given a query?

- The closer they are, the more relevant they are likely to be





LLM - Generation

Evaluating the Quality of the Answers

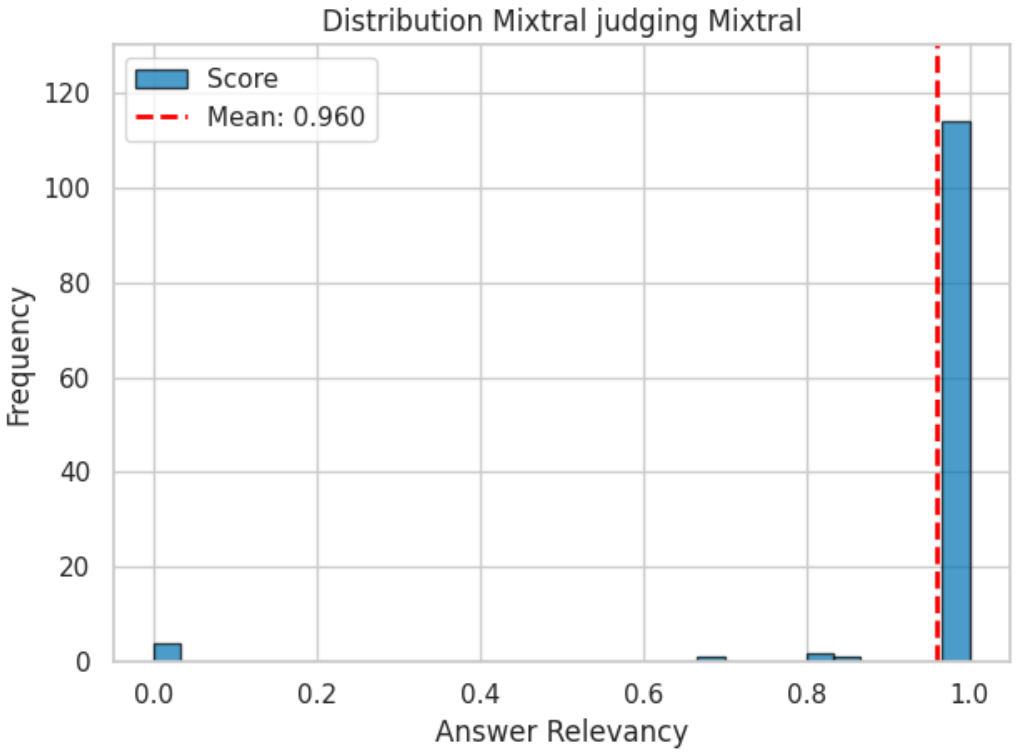
Use LLM as a judge

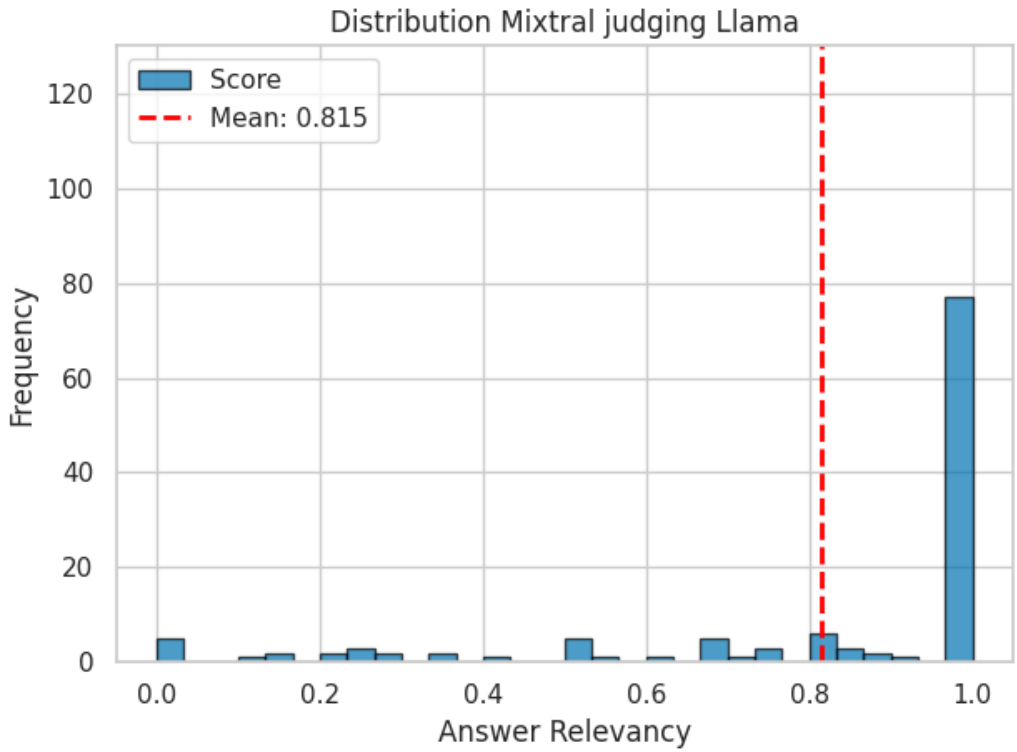
- A LLM **evaluates** or **scores** outputs from other models or systems
- Instead of generating answers, the LLM acts like a reviewer or referee by
  - **Reading** the input (e.g., a question, a reference answer, and a model's response)
  - **Assessing** the quality, accuracy, relevance, or coherence of the response
  - **Providing** a score, explanation, or ranking based on predefined criteria

Answer Relevancy	Faithfulness	Contextual Relevancy
• Ask to the Judge to extract a list of **Statements** (main sentences) from the generated **Answer** • Ask to the Judge to determine whether each statement is relevant to **Answer** the given **Question**	• Ask to the Judge to extract a list of **Claims** (main sentences) from the generated **Answer** • Ask to the Judge to extract a list of **Truths** (main sentences) from the retrieved **Context** • Ask to the Judge to determine whether each Claim contradicts any fact in the retrieval **context**	• Ask to the Judge to determine whether each **Statement** in the context is relevant to answer the **Question**

LLM - Generation

Evaluating the Quality of the Answers





LLM - Example

Query

Which website address hosts the official ALICE FLP documentation?

Ground Truth

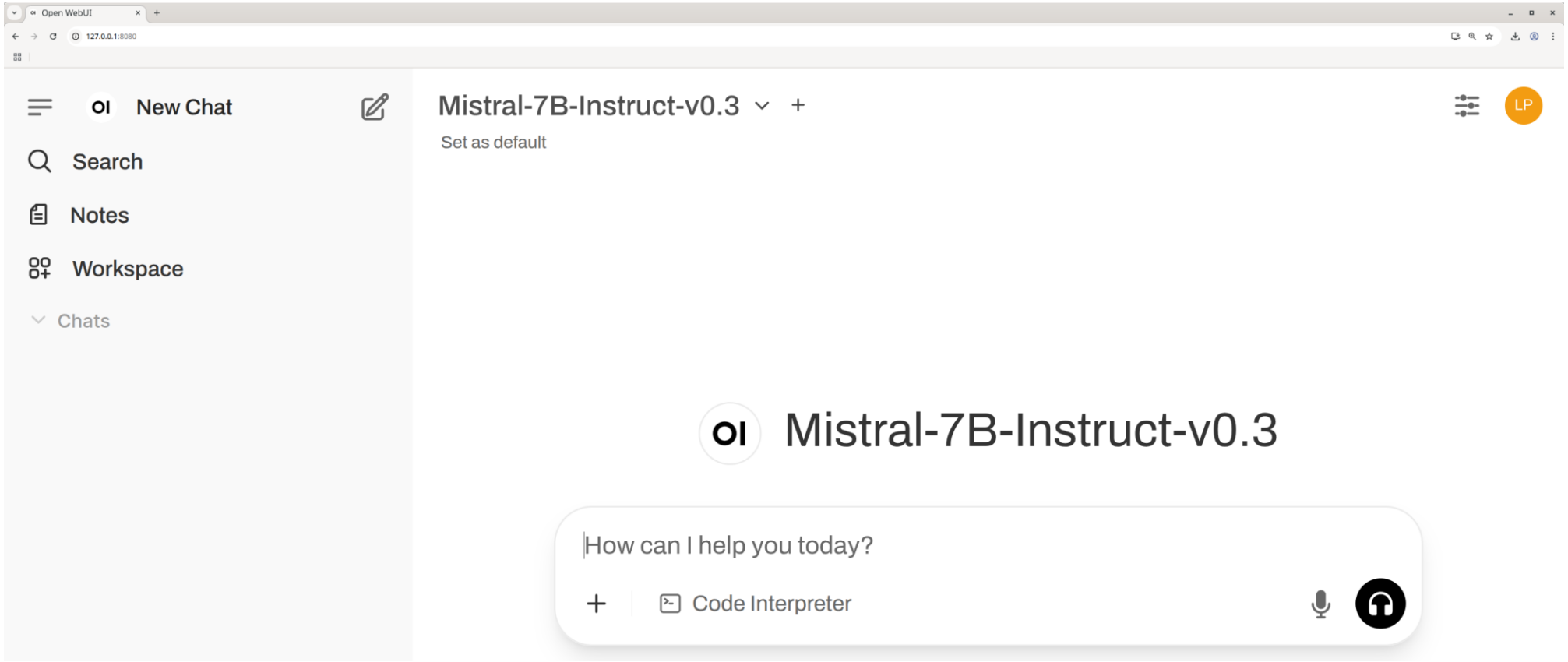
The official ALICE FLP documentation can be accessed at https://alice-flp.docs.cern.ch/

RAG Answer

The official ALICE FLP documentation is accessible at https://alice-flp.docs.cern.ch/

***LLM**: Mixtral-7B-Instruct-v0.2 (locally installed)*

Open WebUI Interface



Anomaly Detection

Objective

- **Real-time** analysis of monitoring data, with timely responses to changing conditions and anomalies

Status

- Ongoing projects based on ATLAS TDAQ monitoring data
- Early detection of **anomalies** in the **HLT farm** with a hybrid approach (DeepHydra)
  - Clustering T-DBSCAN: detects short dropout of a node
  - Deep neural network: identifies long term outliers
- **Time series prediction** of trigger rates with an LSTM-based autoencoder architecture

Plans

- Add **continuous** and **reinforcement learning** to adapt to changing conditions
- **Interface** with our **RAG system** to provide instructions to shifters, collect feedbacks and eventually take actions

Connections Inside CERN

Started discussions with CERN IT representative	Future planning	
	Available resources	
In contact with AccGPT people to join the [AI Chatbot Collaboration]	Not an official service	For sure, an initiative providing a valuable platform for collaboration
	Gathered more than 50 use cases at CERN	
	Secure access to any large language model	Hosted in the cloud or on-premises
	Support for a variety of RAG pipelines	
	OpenAI subscription available	Strictly for testing purposes

Summary & Outlook

What are we exlporing

A **Retrieval-Augmented Generation** (RAG) system using LLMs

Anomaly detection systems using **deep neural networks** and **autoencoder-based** architectures

Toward agentic systems

RAG systems are a foundational step toward building **agentic systems** that go beyond answering questions

Capable of tool use, code execution, and autonomous actions

From answering "*what should I do?*" to "*let me do it for you*"

Ultimate goal is to achieve **full automation** of detector operation