João António Assis Reis
98474

joaoreis16@ua.pt

deti universidade de aveiro departamento de electrónica, telecomunicações e informática

IEETA

A conversational query builder on medical databases

Mestrado em Engenharia Informática

4th July 2024

Under the guidance of:

Advisor	João Rafael Almeida
Co-advisor	José Luís Oliveira
Collaborator	Tiago Almeida

Use case for a medical researcher

Task: Conduct a study about COVID-19.

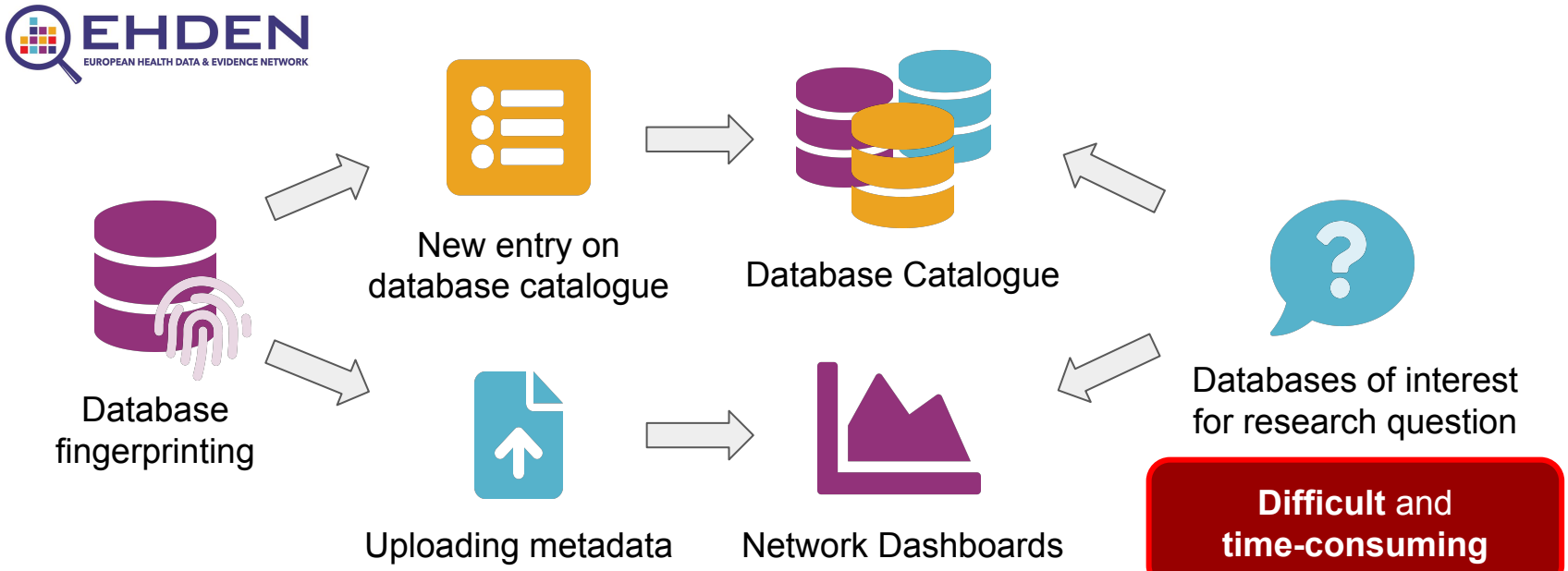
Procedure:

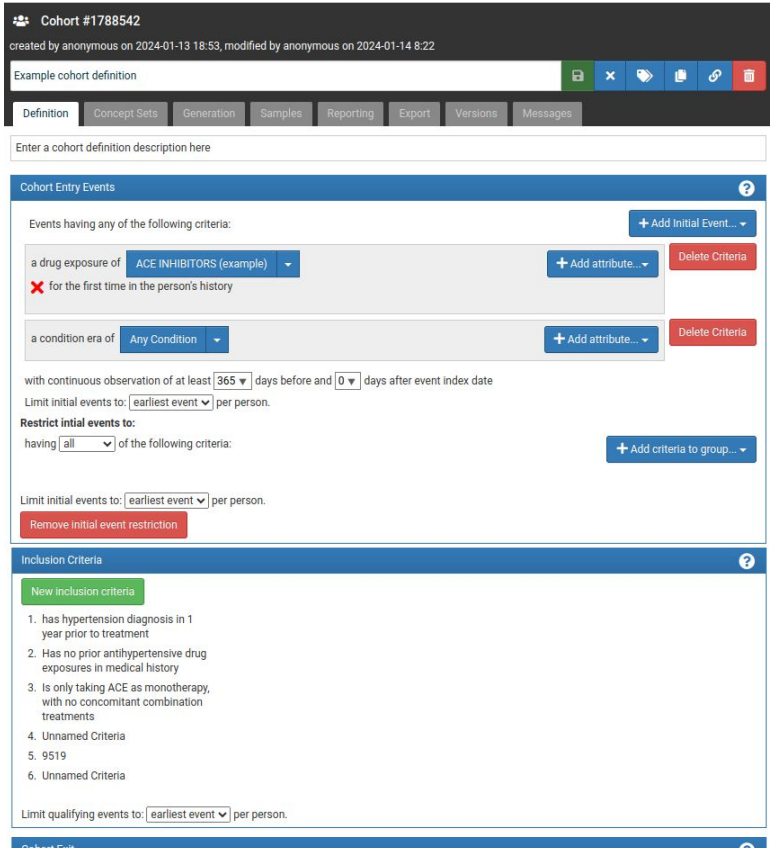
1) Definition of the study protocol.
2) Contact data owners interested in collaborating in the study.
3) Conducting the study.
4) Aggregate the results.
5) Publish the findings.

Complex and time-consuming process

Why?


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
OMOP CDM


EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK
New entry on database catalogue
Database Catalogue
Database fingerprinting
Uploading metadata
Network Dashboards
Databases of interest for research question
Difficult and time-consuming
Data Custodian
Researcher

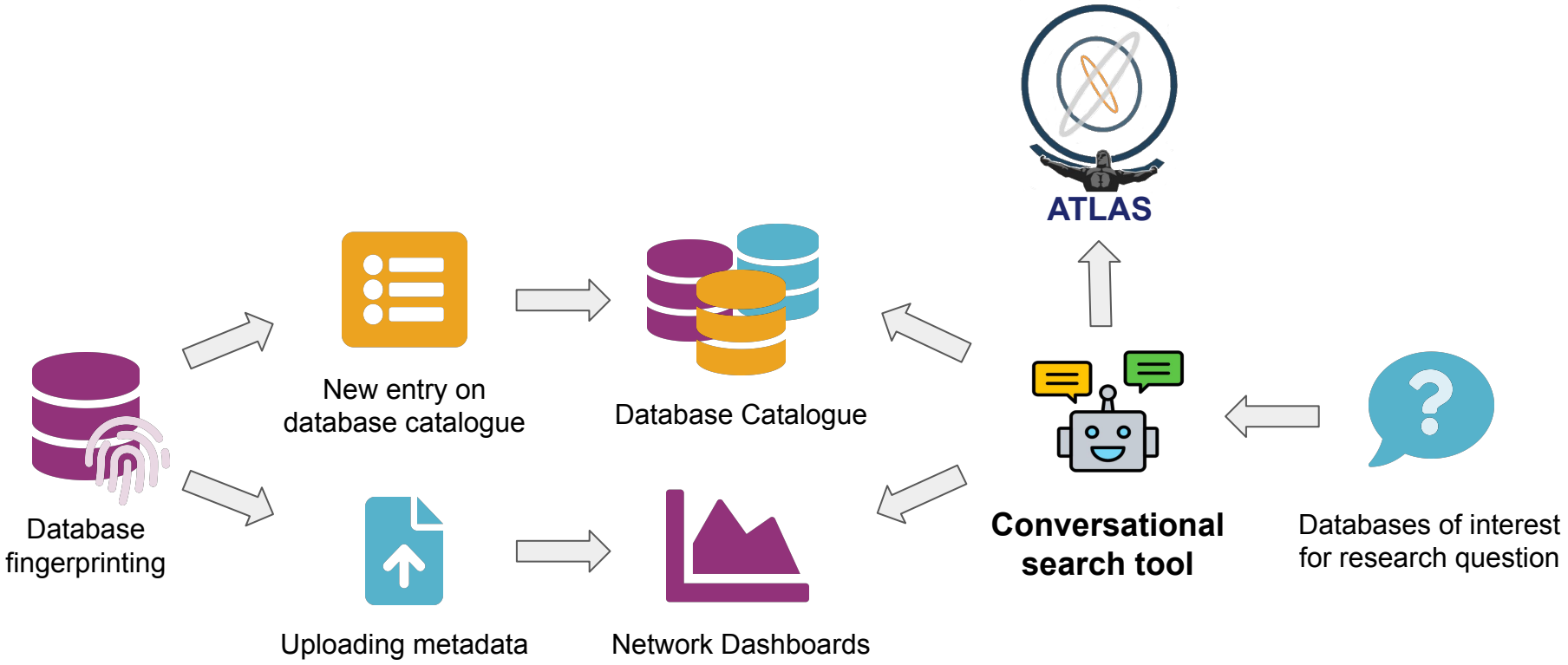


OHDSI
Observational Health Data Sciences and Informatics

ATLAS

Cohort Definition

- Query that **defines a set of persons** who meet certain criteria over a specified duration


ATLAS
New entry on database catalogue
Database Catalogue
Database fingerprinting
Uploading metadata
Network Dashboards
Conversational search tool
Databases of interest for research question

Information Retrieval

Retrieve relevant information from collections of unstructured data, such as documents

Techniques

- **Traditional methods**
  - Term Frequency - Inverse Document Frequency (TF-IDF)
  - Best Matching 25 (BM25)
- **Neural methods**
  - Interaction-based method

Question Answering

Focuses on providing a single and specific answer to a question posed in natural language

- Uses **Natural Language Processing (NLP)** and **Information Retrieval (IR)**

Definition

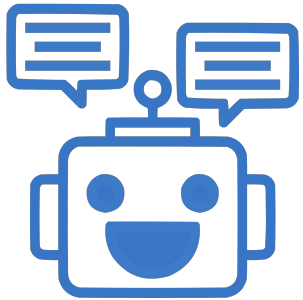
Designed to **comprehend and generate text** that is coherent and contextually relevant, engaging in human language interactions.

- These models process and **predict patterns** with accuracy.

Comparison

Model	Provider	Fine-tuneability	Open-source
~~GPT-4~~	~~OpenAI~~	~~No~~	~~No~~
LLaMa 2	Meta	Yes	Yes
~~PaLM 2~~	~~Google~~	~~No~~	~~No~~
Falcon	TII	Yes	Yes
Mistral	Mistral AI	Yes	Yes

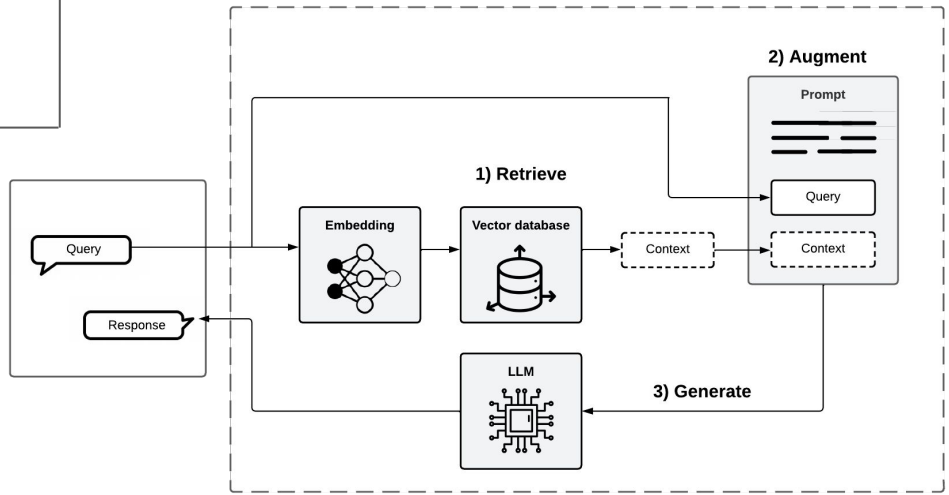
Due to the use of **sensitive and private data**



Types of chatbots based on their **response generation**:

- Rule-based chatbot
- Retrieval-based chatbot
- **Generative-based chatbot**

overcoming limitations of the LLM, such as hallucination → **Retrieval-Augmented Generation (RAG)**



improve and adapt the LLM → **Reinforcement Learning from Human Feedback (RLHF)**

Low-code tools to build customized LLM orchestration flow and AI agents

FloWiseAI

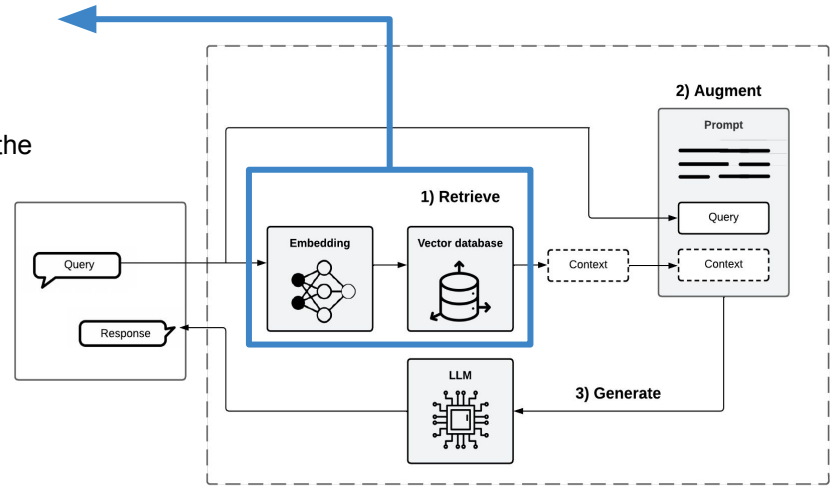
LangFlow

Information Retrieval

- **Model**: BM25
- Used to retrieve the most suitable databases
  - Based on the **databases concepts** extracted from the **Network Dashboards** tool

Large Language Model

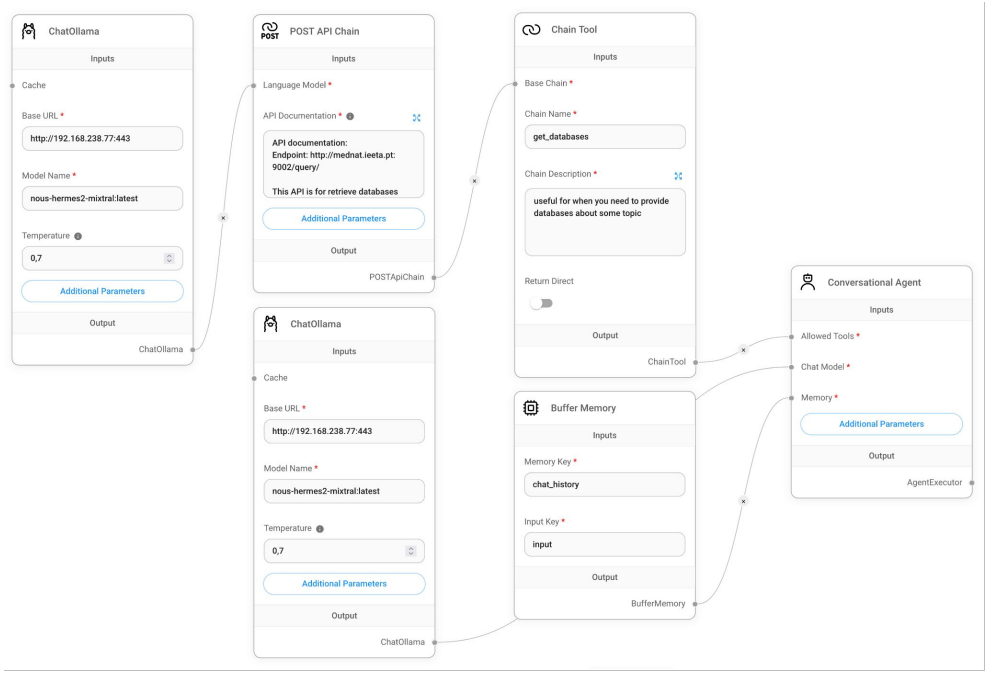
- **Model**: Nous Hermes 2 Mixtral 8x7B
- Used as a NLP tool
  - To extract information from user's messages
  - To generate responses to the user
- Also used to create a fluid conversation



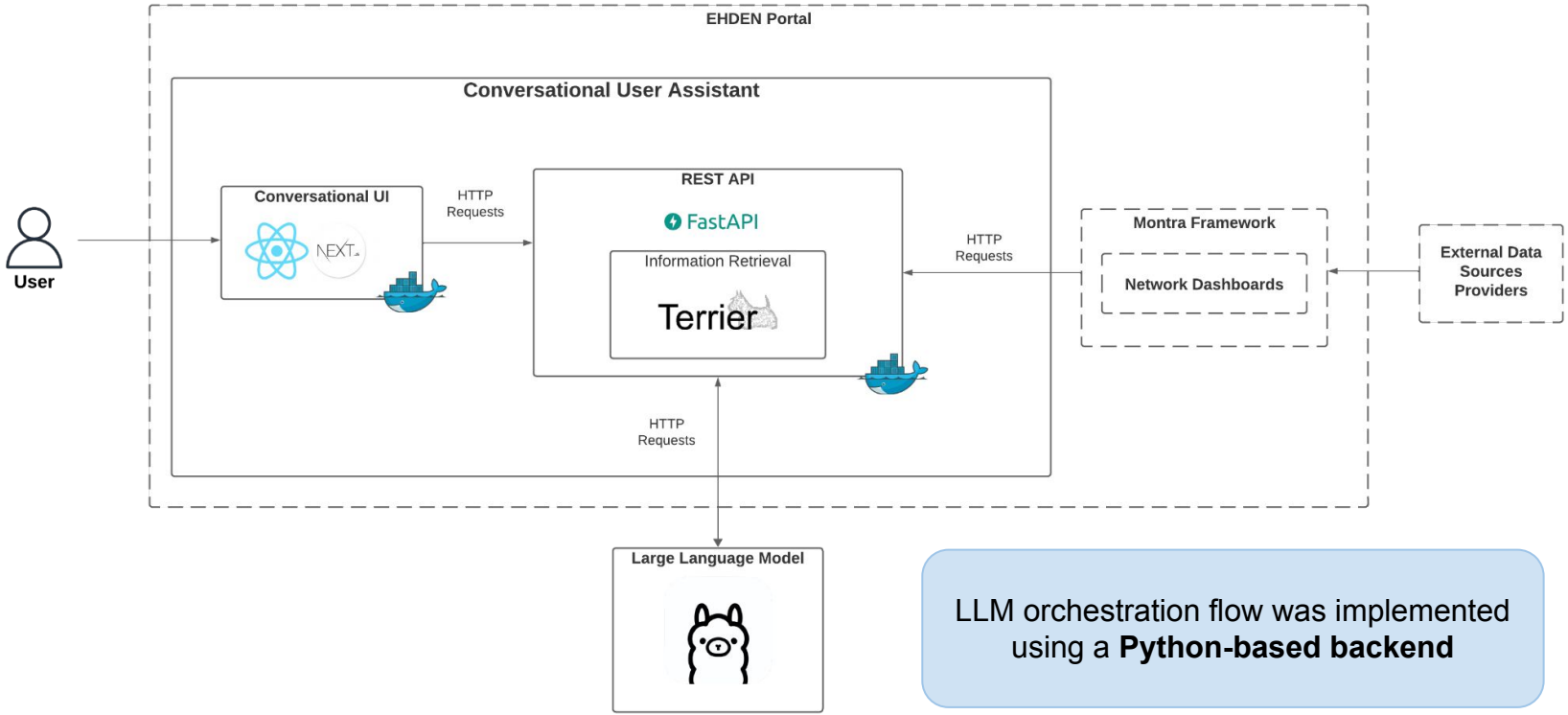
Retrieval-Augmented Generation (RAG)

Generate a response to retrieve the best databases to the user

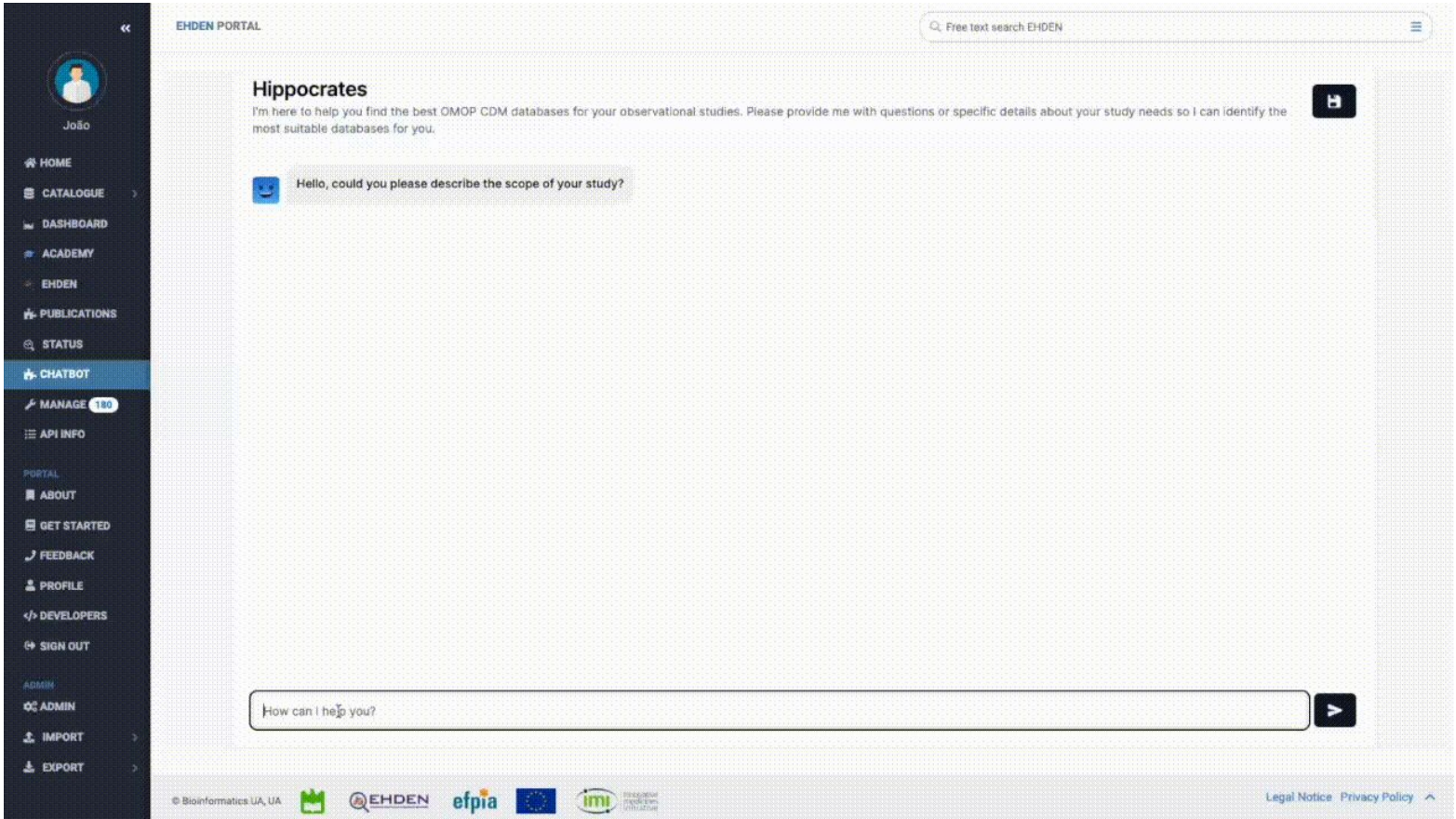
To **orchestrate the interactions** between the LLM, the user interface and the information retrieval component



- FlowiseAI proved to be more **promising** than Langflow
- Development of a chatbot to discover medical databases using FlowiseAI
- However, it not addresses the **complexity of the cohort definition requirements**

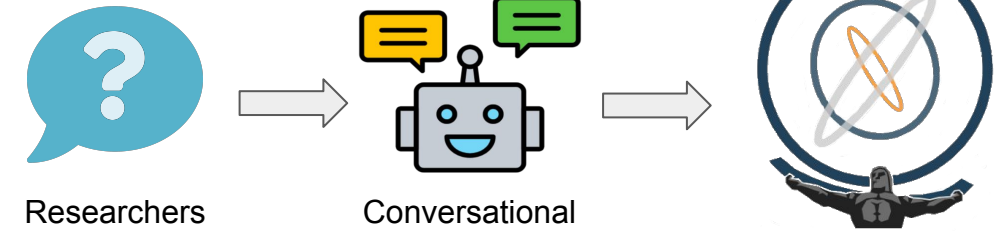


LLM orchestration flow was implemented using a **Python-based backend**


EHDEN PORTAL
Free text search EHDEN
João
HOME
CATALOGUE
DASHBOARD
ACADEMY
EHDEN
PUBLICATIONS
STATUS
CHATBOT
MANAGE 180
API INFO
PORTAL
ABOUT
GET STARTED
FEEDBACK
PROFILE
DEVELOPERS
SIGN OUT
ADMIN
ADMIN
IMPORT
EXPORT
Hippocrates
I'm here to help you find the best OMOP CDM databases for your observational studies. Please provide me with questions or specific details about your study needs so I can identify the most suitable databases for you.
Hello, could you please describe the scope of your study?
How can I help you?
© Bioinformatics UA, UA
EHDEN
efpia
imi
Legal Notice
Privacy Policy

Goal

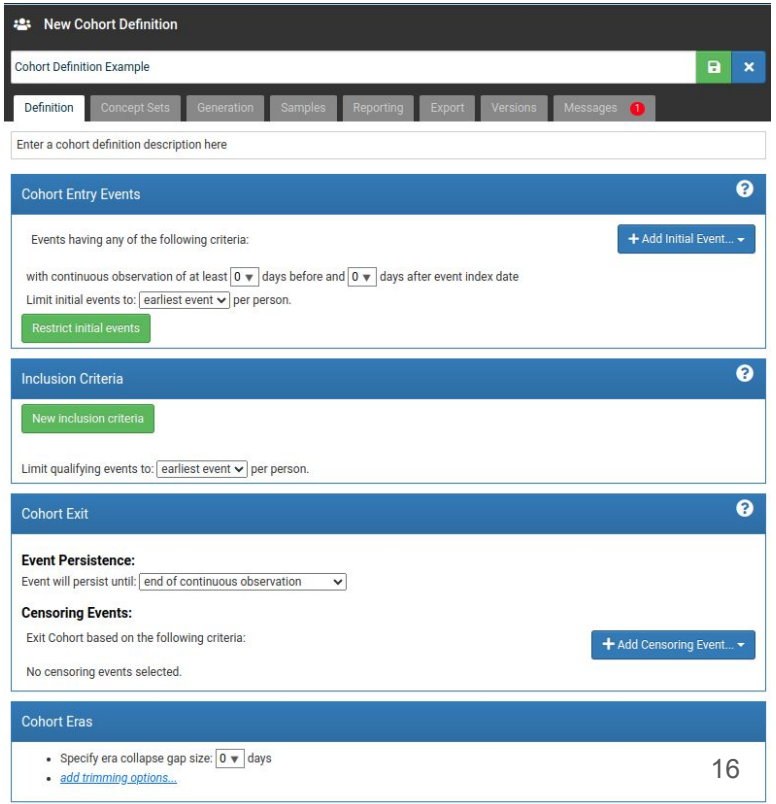
Enhance the conversational system to **support the cohort definition** for observational studies.



Researchers interested in conduct a study

Conversational system

ATLAS



Concept Set Collection of **medical concepts** used to define clinical elements, like diseases, drugs, procedures

Concept Set Name

Concept Test

ID	Concept	Select	Excluded	Descendants	Mapped
710158	COVID-19	☑	☑	☐	☐
756061	Asymptomatic COVID-19	☑	☐	☑	☐
35894915	COVID-19 vaccine	☑	☐	☐	☐
37310268	Suspected COVID-19	☑	☐	☐	☐
703431	COVID-19 excluded	☐			

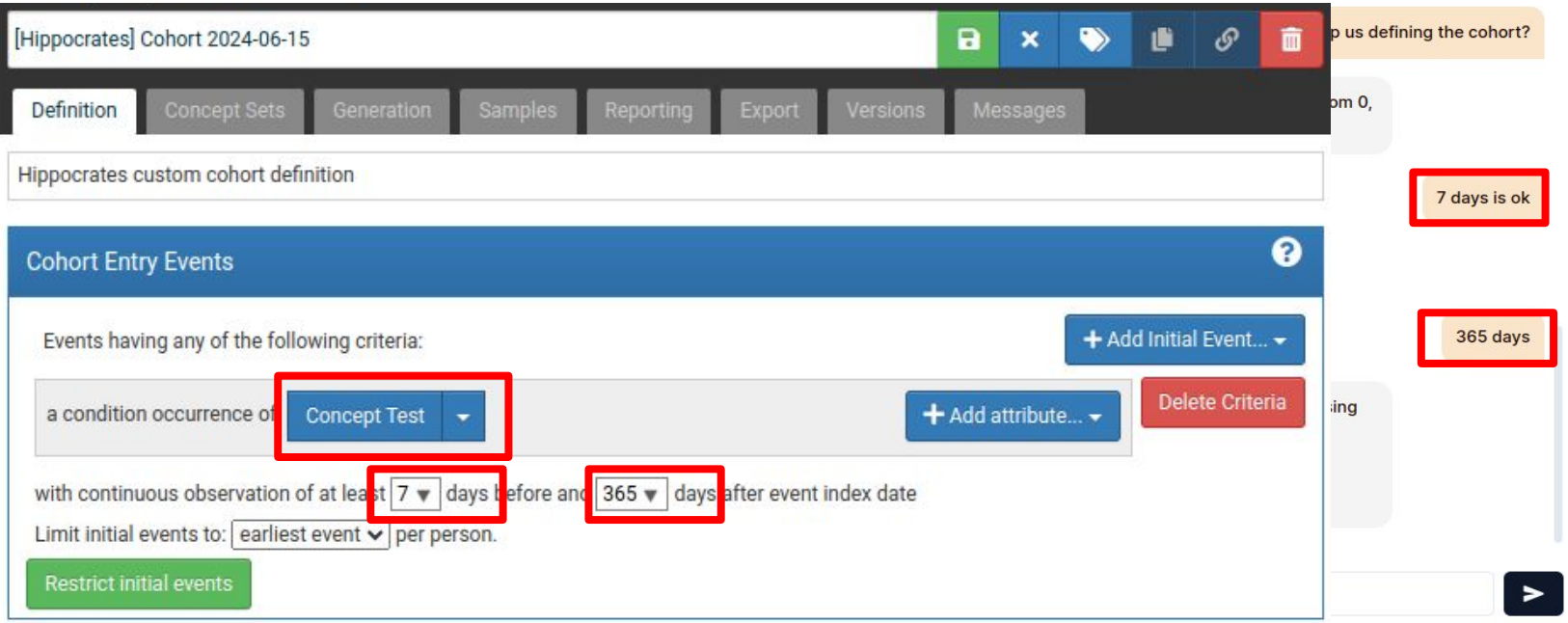
Submit

The concept set was successfully defined. You can either download it or send it to your ATLAS instance.

Download | Send to ATLAS


```
{
  "id": 0,
  "name": "<CONCEPT SET NAME>",
  "expression": {
      "items": [
          {
              "concept": {
                  "CONCEPT_ID": 231256,
                  "CONCEPT_NAME": "Covid-19",
                  "DOMAIN_ID": "Condition",
                  "VOCABULARY_ID": "ASG67HR",
                  "CONCEPT_CLASS_ID": "ASG67 code",
                  "CONCEPT_CODE": "953635",
                  "VALID_START_DATE": "2000-01-01",
                  "VALID_END_DATE": "2099-12-31"
              },
              "isExcluded": false,
              "includeDescendants": true,
              "includeMapped": false
          },
          {...}
      ]
  }
}
```

Cohort Definition



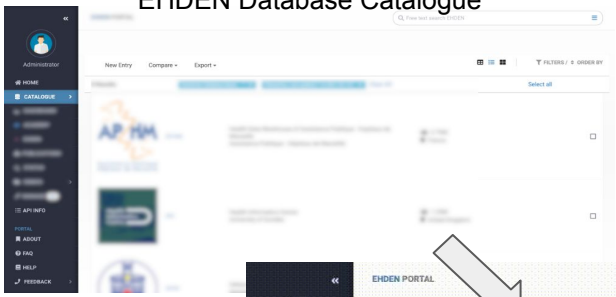
```
{
  "ConceptSets": "Are there any other concepts you'd like to add? If yes, please add them.
If no, simply respond with 'no'.",

  "PrimaryCriteria": {
    "ObservationWindow": {
      "PriorDays": "Regarding the observation window, what is the minimum number of days
needed before the continuous observation? You must choose from 0, 1, 7, 14, 21, 30, 60,
90, 120, 180, 365, 548, 730 or 1095.",
      "PostDays": "How many days after event index date? You must choose from 0, 1, 7,
14, 21, 30, 60, 90, 120, 180, 365, 548, 730 or 1095."
    }
  }
}
```

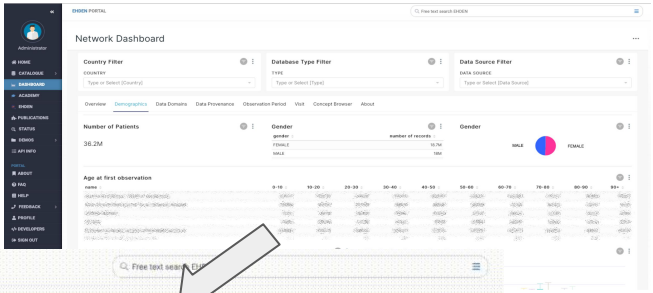
Processed by the LLM until it returns an acceptable answer

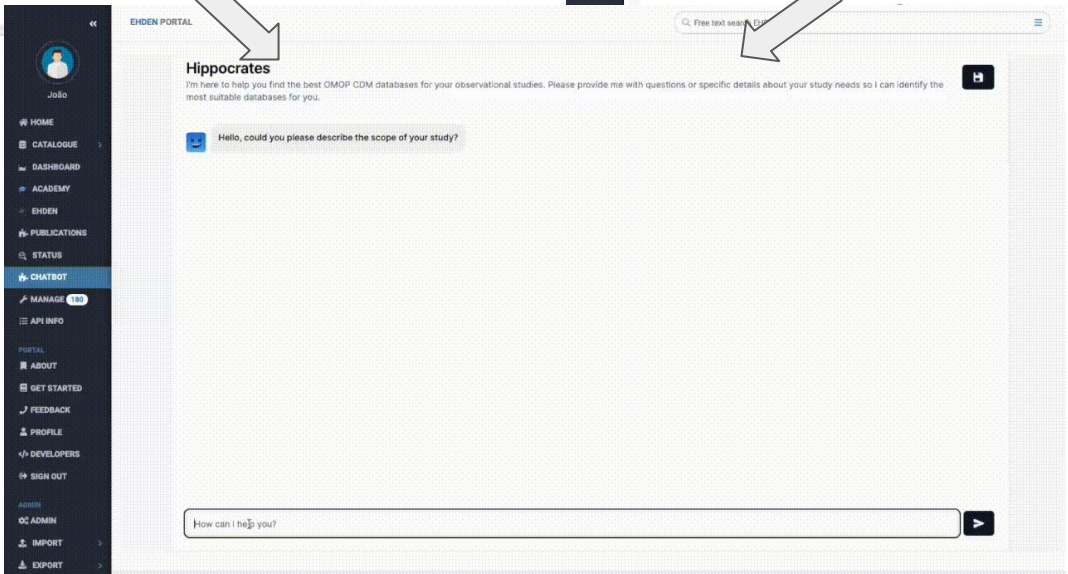
Database Discovery Tool | Query Builder

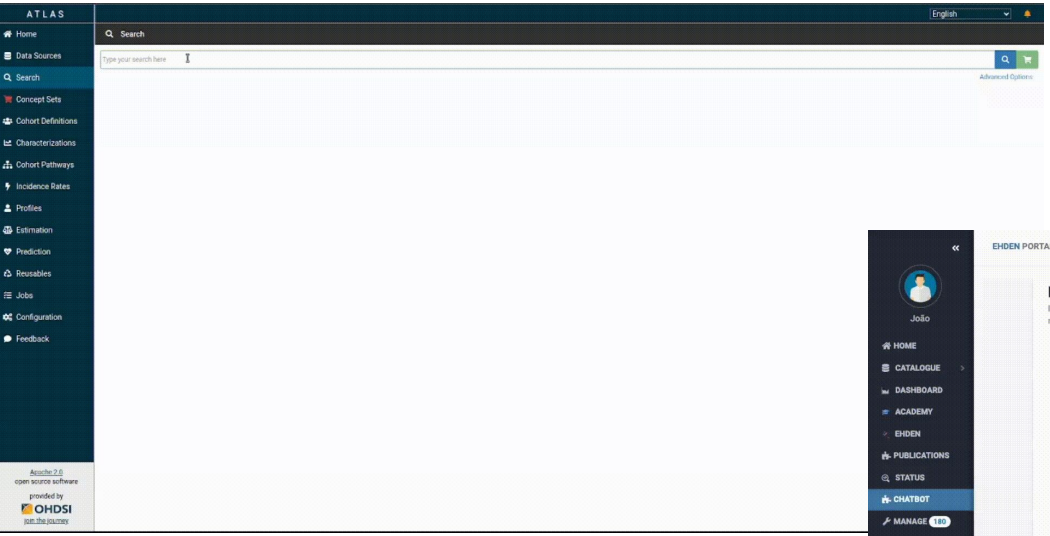
EHDEN Database Catalogue



EHDEN Network Dashboards

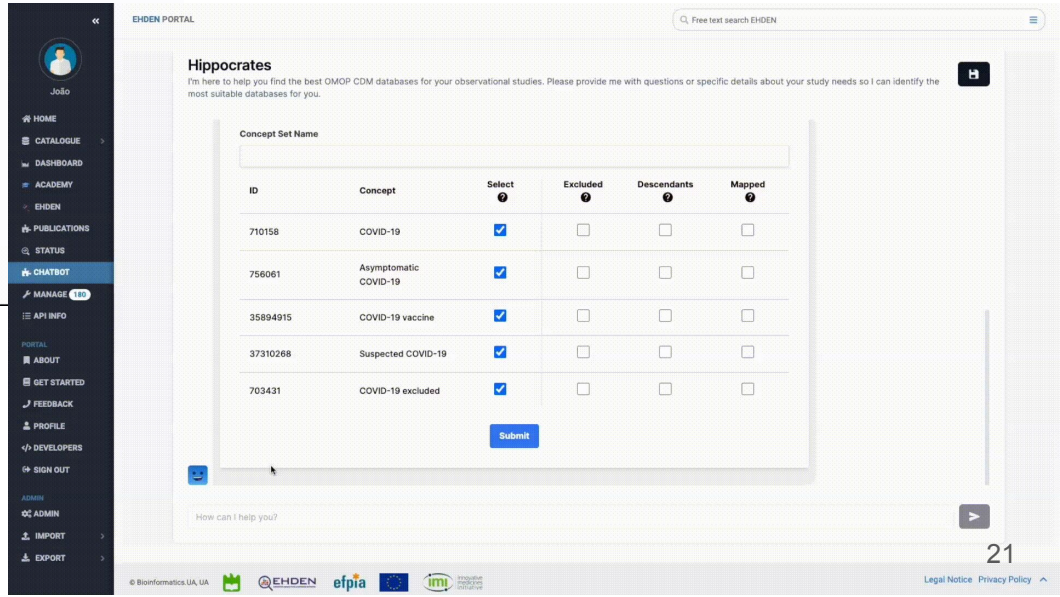




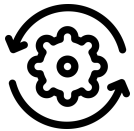


Concept set definition process in **ATLAS**

Concept set definition process using the **chatbot**



Streamline the process of medical database discovery



Impact on the speed and quality of research efforts

Innovation and advancement of the medical research field

- Enhance the query builder feature in order to provide the **full cohort definition**
- Creating **synthetic labeled dataset** to test and validate the Information Retrieval methods
- Enhance the Information Retrieval methods to **produce better search results**

Conferences

- *“Using Flowise to Streamline Biomedical Data Discovery and Analysis”*
  - **22nd IEEE Mediterranean Electrotechnical Conference** (MELECON), 2024
- *“HealthDBFinder: a question-answering task for health database discovery”*
  - **37th IEEE International Symposium on Computer-Based Medical Systems** (CBMS), 2024
- *“A chatbot-like platform to enhance the discovery of OMOP CDM databases”*
  - **34th Medical Informatics Europe Conference** (MIE), 2024
- *“BIT.UA at BioASQ 12: From Retrieval to Answer Generation” (submitted)*
  - **15th Conference and Labs of the Evaluation Forum** (CLEF), 2024

Poster / Software Demo

- *“A Chatbot to help discover OMOP DCM databases within EHDEN Network”*
  - **OHDSI Europe Symposium**, 2024.

Thank you!