



EUROPEAN CENTRAL BANK

EUROSYSTEM

# Large Language Models for ECB Data Portal

---

February 2025

Luca Petracca, Simone De Benedictis,  
Thomas Gottron, Zlatina Hofmeister



# Scope of the project

# The ECB Data Portal

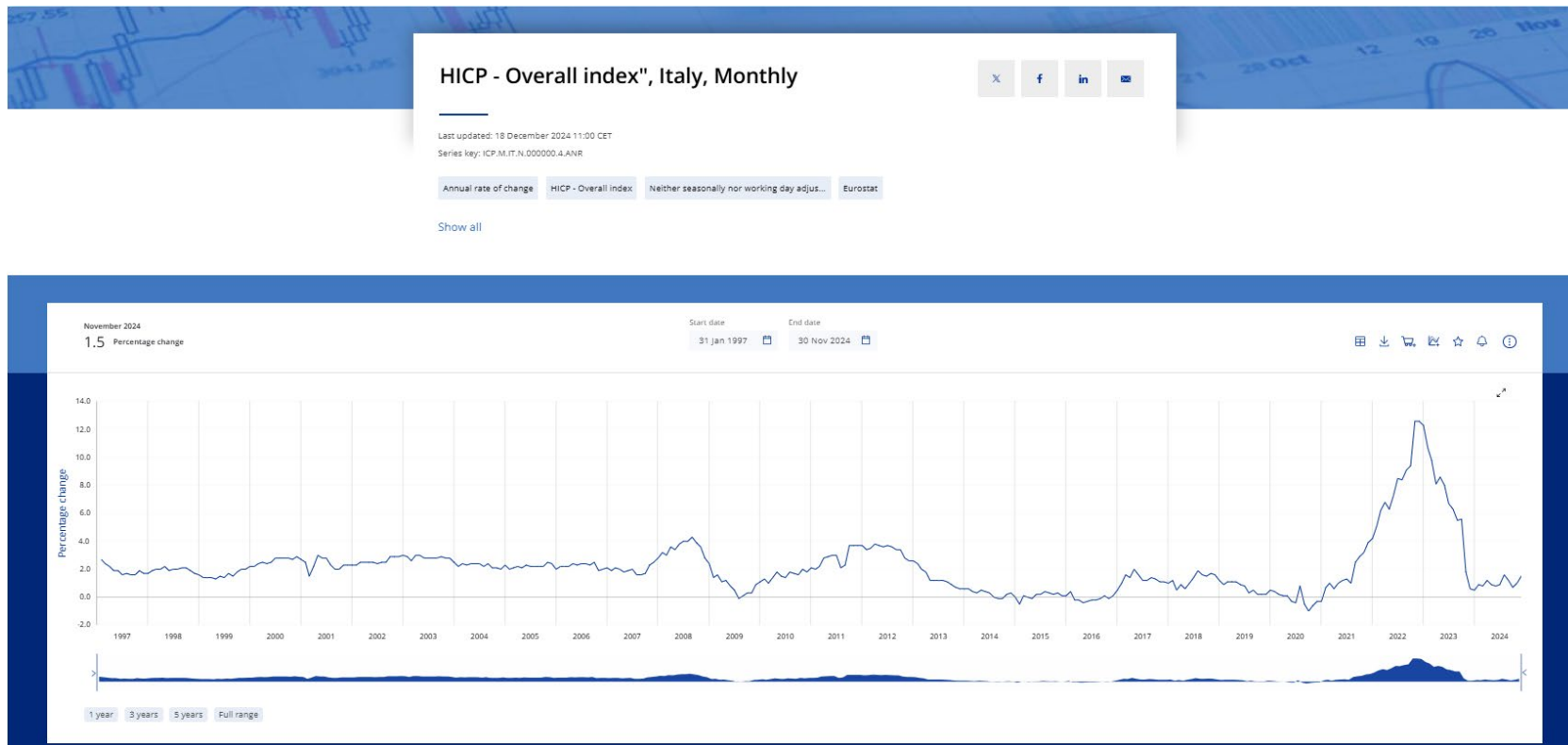
The screenshot displays the ECB Data Portal interface. At the top, the header includes the European Central Bank logo and the text "EUROPEAN CENTRAL BANK | EUROSISTEM". Below this, the "ECB Data Portal" title is centered, with navigation links for "DATA CART", "LOGIN", and "CONTACT US". A secondary navigation bar contains links for "Main figures", "Browse data", "Publications", "Blog", "Methodology", "Currency converter", and "Help".

The main content area features a search bar with the placeholder "Search for data" and a "Browse data by:" section with filters for "Category", "Concept", "Geographical area", and "Dataset". Below the search area, there are four tabs: "Euro area at a glance", "Financial developments", "External sector and exchange rates", and "Banking supervision".

The dashboard displays six data cards:

- Inflation rate** (November 2024): 2.3% ↑
- Monetary aggregate M3** (October 2024): 3.4% ↑
- US dollar** (29 November 2024): 1,0562 ↑
- Real GDP growth** (Q3 2024): [Line chart showing a dip in Q3 2024]
- Unemployment rate** (October 2024): [Line chart showing an upward trend]
- Government debt** (Q2 2024): [Line chart showing an upward trend]

# Example of Time Series data on the EDP



# The problem

## How can we help users finding information in the ECB Data Portal?

- Users report difficulties finding and identifying the right information

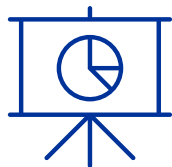
## The ECB Data Portal uses “traditional” ways of accessing information

- Browsing by categories, concepts, geographical areas and datasets
- Search by keywords
- Main indicators
- Interactive Publications
- Dashboards

## What about providing an interface where users can use their own “natural language”?

# Our Idea

- We are testing **Retrieval Augmented Generation (RAG)** techniques to explore how state-of-the-art **LLMs** can be leveraged to enable users to **find data** using **natural language** in a **chat** interface



**Understand the  
user queries**

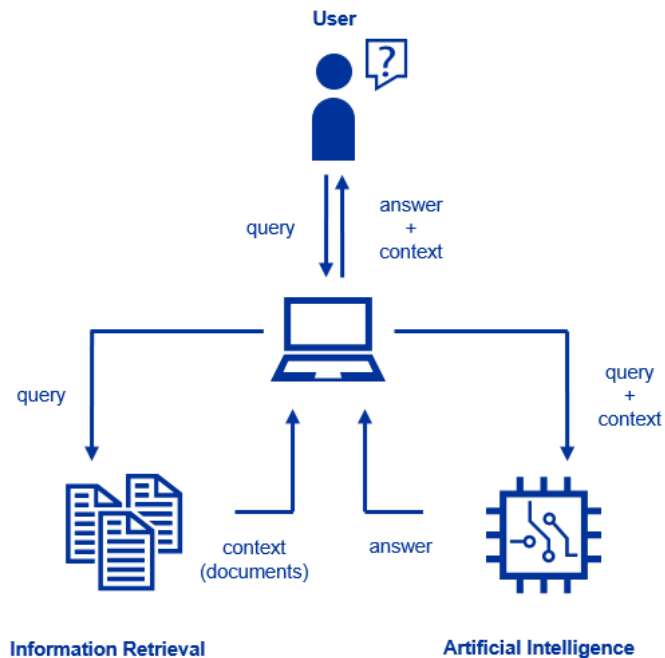


**Identify the right dataset  
and time series key**



**Retrieve and provide  
time series data**

# RAG – Retrieval Augmented Generation



**Problem:** Chatbot should give correct answers. Correct answers are in available data.

**Solution:**

1. **Split contents** into small packages
2. Upon question: **search** for relevant content package
3. Input found **content package** to an LLM to get final answer

**Works well on many problems!**

# RAG vs Text-to-SQL

We want to provide **two answers**: qualitative and quantitative answer

- **Qualitative Answer**: time series key, description and metadata
- **Quantitative Answer**: time series values

## Previous architecture: Simple RAG

Qualitative Answer: RAG  
Quantitative Answer: RAG

### **WEAKNESSES:**

- **Limited context window** for handling multiple years data queries.
- **Prone to hallucinations** due to reliance on LLM for interpreting numerical data.
- **Difficult to scale** since it requires the index to contain one row for each data observation

## Current Architecture: RAG + Text to SQL

Qualitative Answer: RAG  
Quantitative Answer: Text to SQL

### **STRENGTHS:**

- **Accuracy**: SQL ensures precise and deterministic retrieval of time series values, reducing the risk of hallucinations.
- **Scalability**: JSON-based metadata organization allows streamlined, scalable access to qualitative data.



How does it work?

# Chatbot Interface – Query Example

You: government debt and inflation in italy in 2024

**Assistant:** To find the government debt levels and inflation rate for Italy in 2024, you can refer to the following datasets and series keys:

## Government Debt in Italy

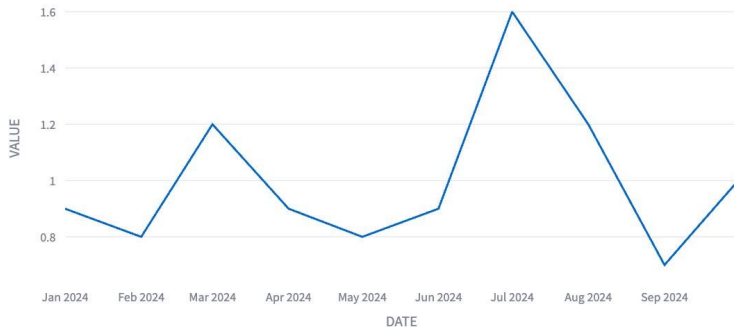
- **Series Key:** [GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N.\\_T](#)
- **Series Key Description:** This series represents the general government gross debt as a percentage of GDP for Italy. The data is available on a quarterly frequency.
- **Dataset:** [GFS](#) (Government Finance Statistics)
- **Dataset Description:** The GFS dataset provides a comprehensive overview of fiscal developments, including government debt, for the euro area, the European Union, and individual EU Member States. It includes data on deficit/surplus, revenue, expenditure, and government debt, among other fiscal indicators.
- **Dataset Link:** [GFS](#)
- **Series Key Link:** [GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N.\\_T](#)

## Inflation Rate in Italy

- **Series Key:** [ICP.M.IT.N.000000.4.ANR](#)
- **Series Key Description:** This series represents the Harmonised Index of Consumer Prices (HICP) for Italy, focusing on the overall index and its annual rate of change. The data is neither seasonally nor working day adjusted and is available on a monthly frequency.
- **Dataset:** [ICP](#) (Harmonised Index of Consumer Prices)
- **Dataset Description:** The HICP dataset measures changes in the prices of goods and services covered by final household monetary consumption expenditure. It provides a comparable measure of inflation across EU Member States.
- **Dataset Link:** [ICP](#)
- **Series Key Link:** [ICP.M.IT.N.000000.4.ANR](#)

You can explore the detailed data for Italy's government debt as a percentage of GDP and the inflation rate for 2024 by following the provided series key links.

**ICP.M.IT.N.000000.4.ANR**



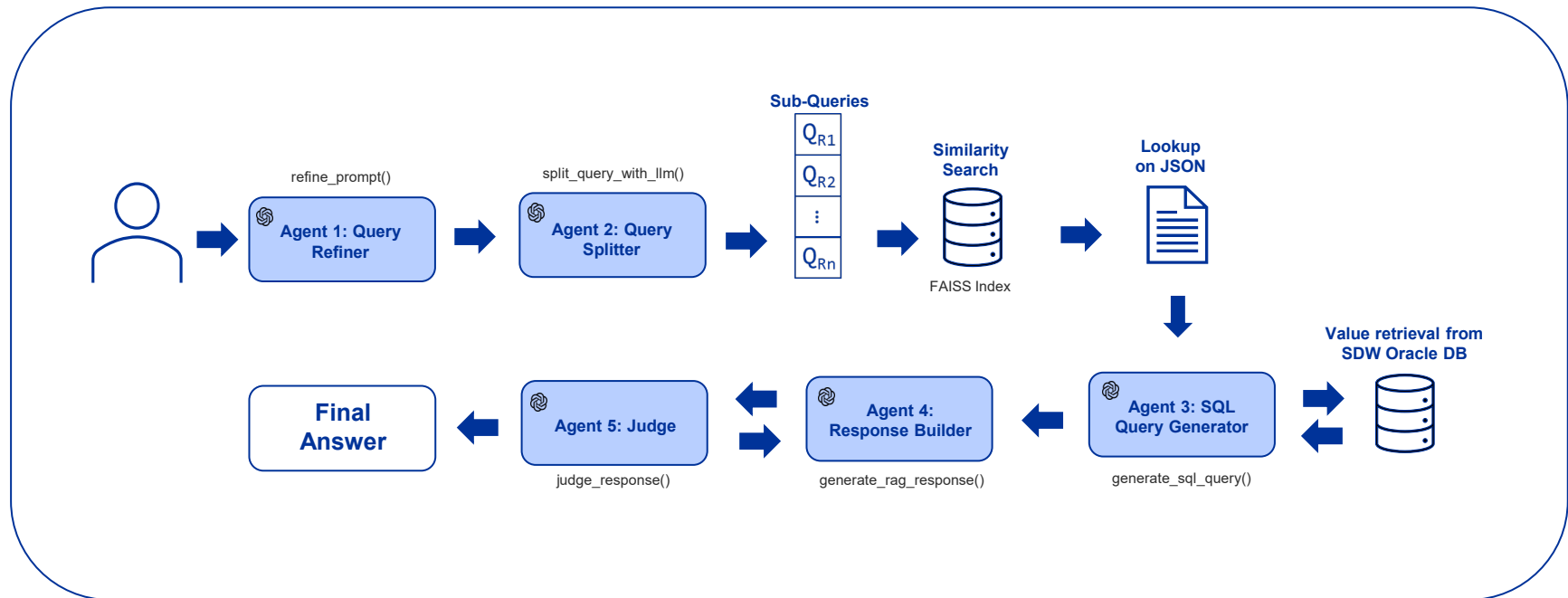
SERIES_KEY	DATASET	DATE	VALUE
<a href="#">GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N._T</a>	GFS	2024-01-01	135.2415
<a href="#">GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N._T</a>	GFS	2024-04-01	136.9999

# Our Solution

## Setup

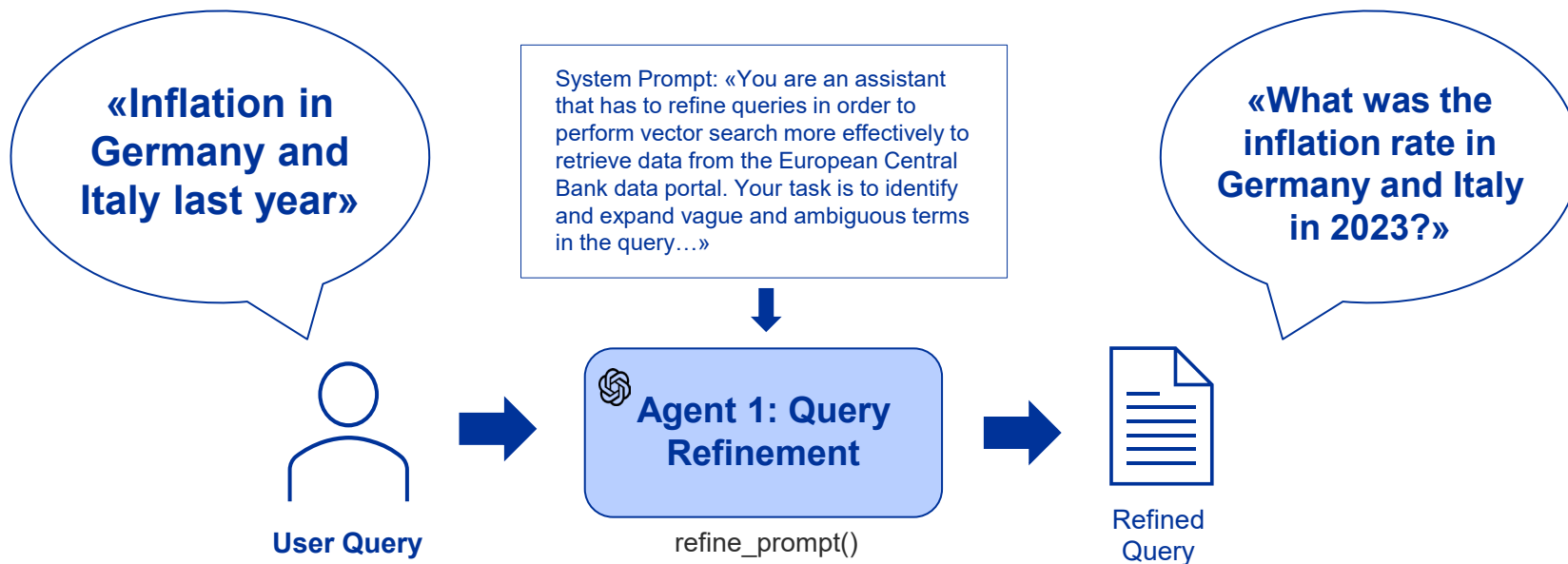
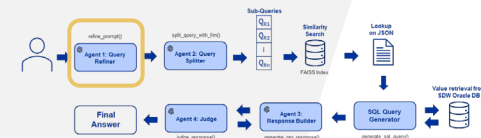
- Connection to **LLM via OpenAI API**
- Time Series Included:
  - **~300.000 Time Series on the EDP including all Main Figures and time series included in official publications, reports and articles**
- Specifics:
  - **FAISS (index)**
  - **Gpt-4o (chat completion)**
  - **Text-embedding-3-large (embeddings)**

# Current Architecture

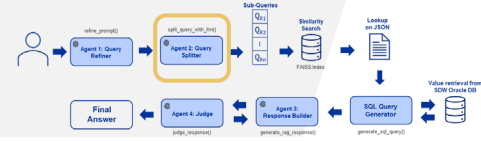


# Behind the Scenes

# Step 1: Query Refinement



# Step 2: Query Splitting

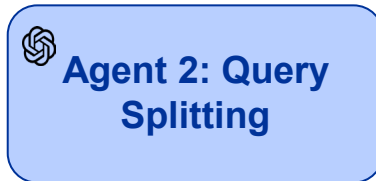


«What was the inflation rate in Germany and Italy in 2023?»

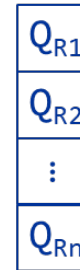


Refined Query

System Prompt: “You are an assistant specialized in breaking down complex queries into multiple simple queries. If a user query contains info about multiple countries, different compositions, datasets or time frames, you should split the queries and output a JSON object...”



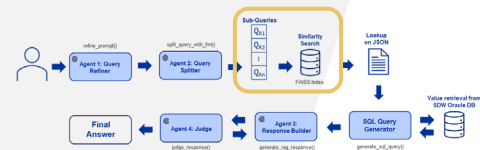
split\_query\_with\_llm()



N Sub-Queries

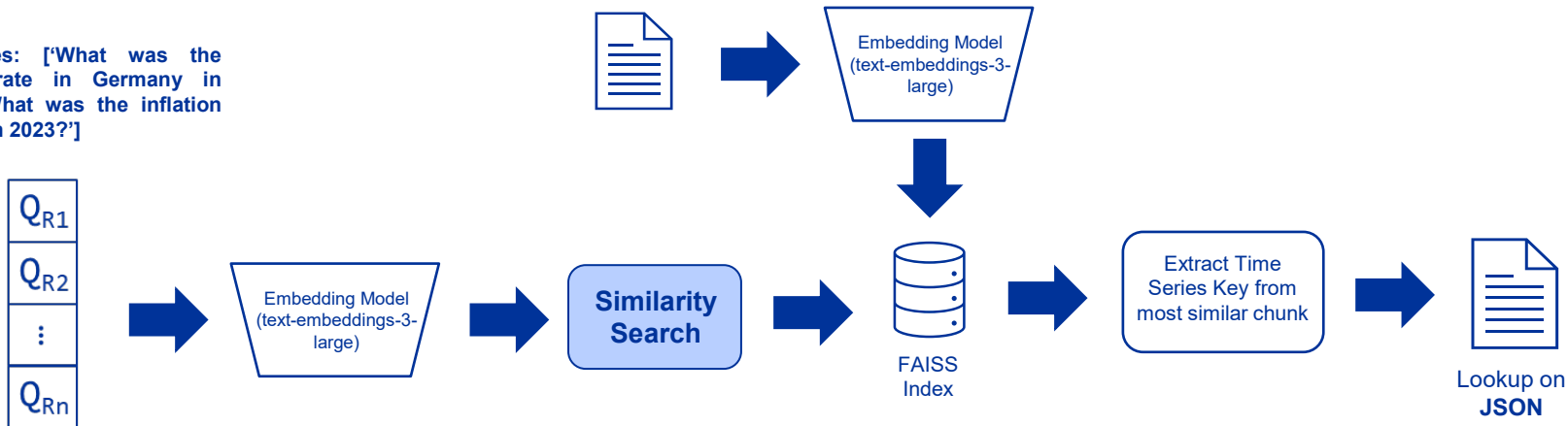
Sub-queries: [‘What was the inflation rate in Germany in 2023?’, ‘What was the inflation rate Italy in 2023?’]

# Step 3: Similarity Search



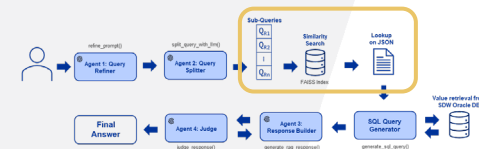
**RAG\_EDP.txt:** Text database containing dataset, time series key and time series complete title, description and metadata for each time series in the selected subset

Sub-queries: ['What was the inflation rate in Germany in 2023?', 'What was the inflation rate Italy in 2023?']





# Step 4: Information Retrieval



## Knowledge Base



## FAISS Index

We are looking for most similar chunk in the Index based on the content of the user query. Each line contains the dataset, key and description of a time series

## Metadata JSON

Given the extracted time series key, we look up on the JSON to extract the full metadata. **But we still need the values...**

Series Key: ICP.M.IT.N.000000.4.ANR, Dataset: ICP, Value for Italy - HICP - Overall index, Annual rate of change, Eurostat, Neither seasonally nor working day adjusted with Monthly frequency, in Italy

▼ ICP.M.IT.N.000000.4.ANR {6}

DATASET : ICP

DATASET : Scope: Data presentation - Summary description. The Harmonised Index of Consumer Prices (HICP) for the euro area is published by the European Commission (Eurostat) and generally available from 1996 onwards. Euro area results are obtained by aggregating indices for individual countries. The HICP is broken down following the European classification of individual consumption according to purpose (ECOICOP) and by goods and services special aggregates derived from it. The HICP covers monetary expenditure on final consumption by resident and non-resident households on the economic territory of the euro area. The seasonally adjusted HICP data are compiled by the ECB. Data presentation - Detailed description. HICPs measure changes in prices of goods and services covered by final household monetary consumption expenditure, including all indirect taxes paid by consumers. HICPs have a common coverage of goods and services across countries but country specific item lists and item weights. HICPs do not cover expenditure for owner occupied housing. For items fully or partly paid or refunded by the government, HICPs include only the share that is paid by the consumer (e.g. the "out-of-the-pocket" expenditure for health services). The HICPs are classified according to the ECOICOP. Additional compiled aggregates are also published. Methodological information. Time period: Monthly, Quarterly, Annual. Base period: 2015=100. Statistical concepts and definitions: For information about the naming convention (series key dimensions and metadata), refer to the ICP underlying DSD (ECB\_ICP1) maintained by the ECB. Statistical processing. Data compilation: Following the Maastricht Treaty the aim of the HICP is to measure inflation by means of the consumer price index on a comparable basis, taking into account differences in national definitions. Eurostat dedicated HICP website (which include short guide for users). Adjustment: X-12 ARIMA. Administrative Information. Title: ICP - Indices of Consumer prices. Data source: European Commission (Eurostat) and European Central Bank calculations based on Eurostat data. Quality. Timeliness: 17 working days after the end of the reference month. Legal and institutional environment. Legal acts and other agreements: Council Regulation (EC) No 2494/95 on HICPs and implementation regulations.

DATASET LINK : <https://data.ecb.europa.eu/data/datasets/ICP>

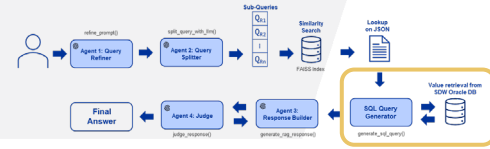
SERIES KEY LINK : <https://data.ecb.europa.eu/data/datasets/ICP/ICP.M.IT.N.000000.4.ANR>

FREQUENCY : Monthly

SERIES KEY : Value for Italy - HICP - Overall index, Annual rate of change, Eurostat, Neither seasonally nor working day adjusted with Monthly frequency, in Italy

DESCRIPTION

# Step 5: SQL Query Generation



**What was the inflation rate in Germany and Italy in 2023?,**

**'ICP.M.DE.N.000000.4.ANR',  
'ICP.M.IT.N.000000.4.ANR'**



Extracted Time Series Key(s)  
+ Refined Prompt

System Prompt: "You are a SQL query assistant. Your task is to create SQL queries to fetch data from the SDW SQL database. The data is stored in a table-like structure, and your goal is to generate queries that match the user's request..."

**Agent 3: SQL Query Generation**

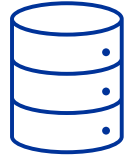
generate\_sql\_query()

```
SELECT *
FROM SDW_WEB.V_ICP_OBS

WHERE SERIES_KEY IN
('ICP.M.DE.N.000000.4.ANR',
'ICP.M.IT.N.000000.4.ANR')

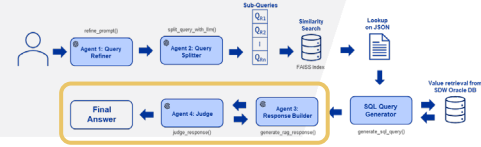
AND OBS_DATE BETWEEN
'2023-01-01' AND '2023-12-01'
```

Table with values

Value retrieval from  
SDW Oracle DB

# Step 6: Response Generation



System Prompt: "You are an assistant specialized in assisting users in finding the appropriate data and series key from the ECB Data Portal. Understand the user query, mention the datasets the information is coming from and retrieve and communicate the data. Always include the series key and the link to the EDP..."

System Prompt: "You are an expert judge tasked with evaluating the relevance and completeness of information extracted from the European Central Bank (ECB) data portal. Your task is to provide a 'total rating' scoring how well the system answer answers the user request expressed in the user question."

Best matching time series data and values



generate\_response\_rag()



judge\_response()



Final Answer



The first response is built using the **refined prompt**, the time series **metadata** extracted from the JSON, and the **values** extracted from the SQL query. If the answer is not satisfactory, the response builder provides a more **generic answer** to the user

# Evaluation

# How do we evaluate our solution?

## Main Ideas:

- What is the accuracy of the model in **extracting the right dataset and time series key?**
- What is the accuracy of the model in **extracting the right values given the right time series?**

## Test Set

- **250 queries divided in 3 categories** (specific, generic and off-topic)

# Results – Value Accuracy

## Simple RAG Approach

LLM helps..	Dataset Accuracy	Series Key Accuracy	Retrieved Values Accuracy	LLM wrong
..providing user with <b>metadata</b> and <b>values</b>	97.87 %	85.11%	27.66%	72.34%

## NL2SQL Extraction

LLM helps..	Dataset Accuracy	Series Key Accuracy	Retrieved Values Accuracy	LLM wrong
..providing user with <b>values</b>	100.00%	79.17%	79.17% ( <i>relatively 100.00%</i> )	0.00%

# Key Insights and Way Forward

# Key Insights and Way Forward

## Key Insights

- The PoC shows promising results with regards to our goal of helping users find data on the EDP. **The multi-agent with Text-to-SQL approach proved to be much more effective than vanilla RAG**

## Next Steps:

- **Improve Accuracy**
  - Exploring new retrieval techniques for the similarity search
  - Retrieving multiple time series keys instead of the top 1 most similar key
  - Limiting the set of series (only main indicators vs core series vs all)
  - Finding the right balance between cost-benefit-risk
- **Deploy the solution**



Thank you!