



EUROPEAN CENTRAL BANK

EUROSYSTEM

Supporting users in seeking data on the ECB data portal

A use case for RAG

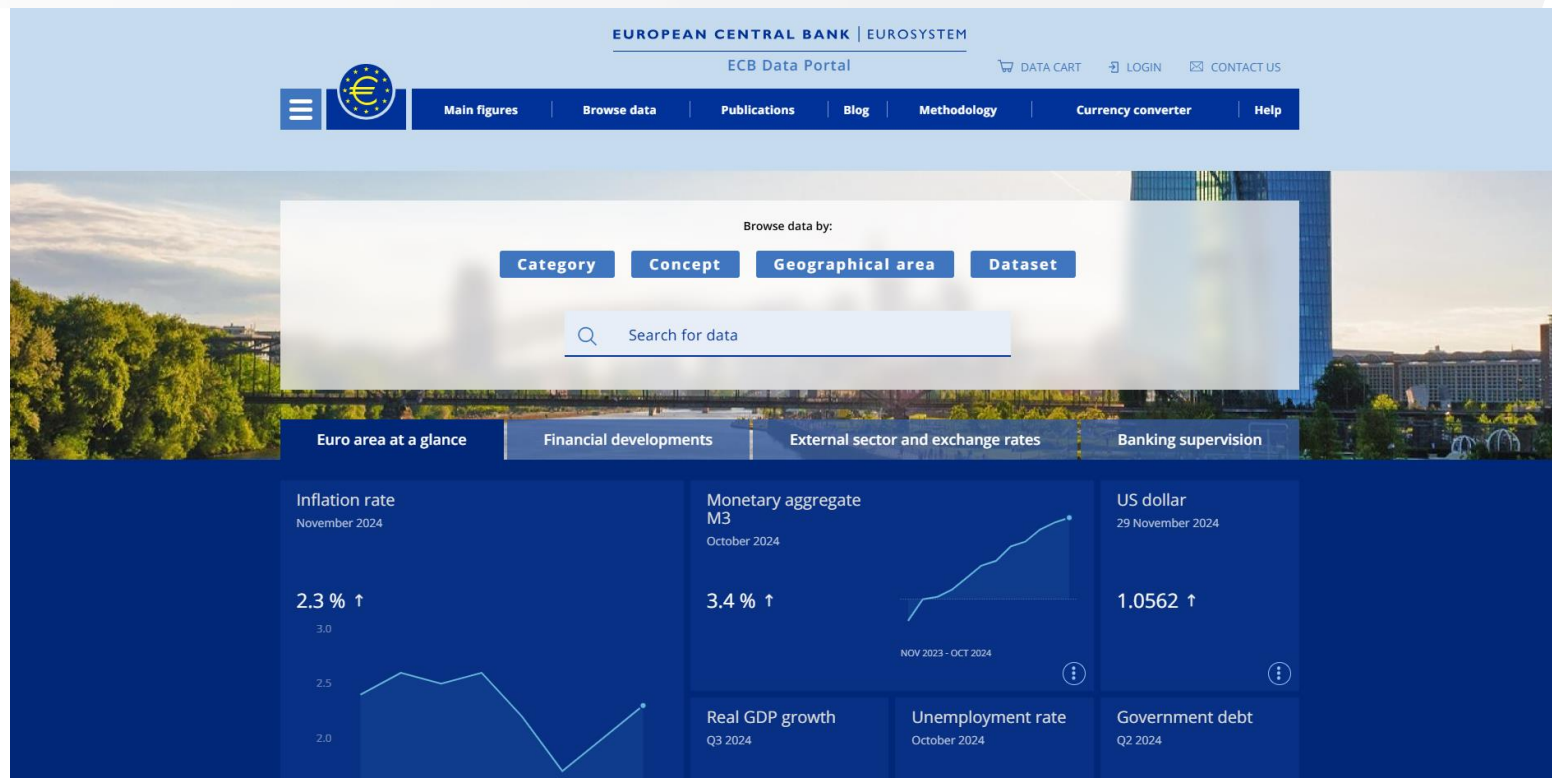
The views expressed are those of the authors and do not necessarily reflect those of the ECB.

February 2025

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



The ECB Data Portal



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

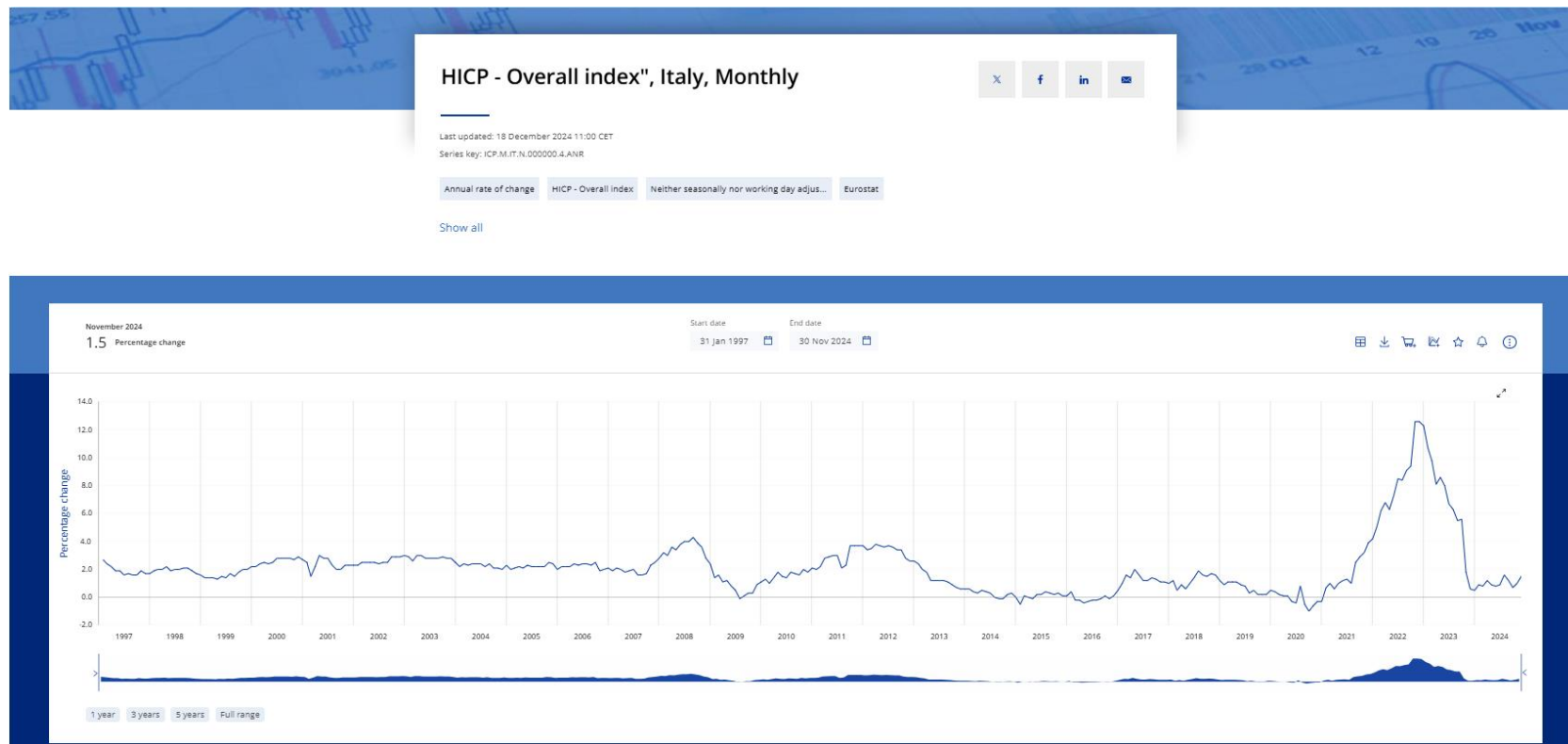
The main content area features a "Browse data by:" section with four buttons: "Category", "Concept", "Geographical area", and "Dataset". Below these is a search bar with the placeholder text "Search for data".

The interface also includes a horizontal menu with four categories: "Euro area at a glance", "Financial developments", "External sector and exchange rates", and "Banking supervision".

Below the menu, there are six data widgets arranged in a 2x3 grid:

- Inflation rate** (November 2024): 2.3 % ↑. A line chart shows the trend from 2.0 to 3.0.
- Monetary aggregate M3** (October 2024): 3.4 % ↑. A line chart shows the trend from NOV 2023 to OCT 2024.
- US dollar** (29 November 2024): 1.0562 ↑. A line chart shows the trend.
- Real GDP growth** (Q3 2024): A line chart shows the trend.
- Unemployment rate** (October 2024): A line chart shows the trend.
- Government debt** (Q2 2024): A line chart shows the 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

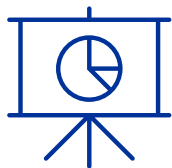
The ECB Data Portal uses “traditional” ways of accessing information

- 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

- **PoC:** testing **Retrieval Augmented Generation (RAG)** and **Text-to-SQL** 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

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.Z.S.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.

Additional Series Keys for Government Debt in Italy:

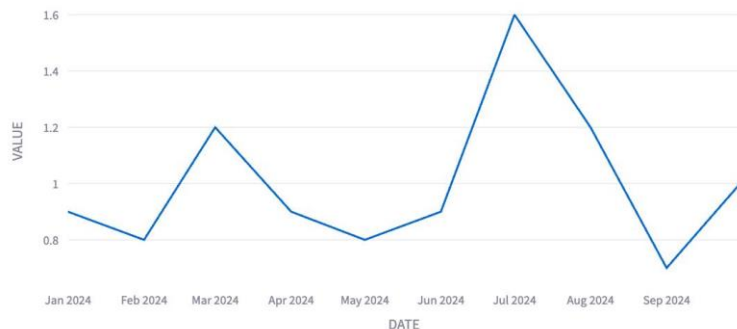
- [GFS.Q.N.IT.W0.S13.S1.Z.B.B9P.Z.Z.XDC.R.B1GQ.Z.S.V.N.T](#) - [Link](#)
Quarterly frequency, in Italy.
- [GFS.A.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.T.F.V.N.T](#) - [Link](#)
Annual frequency, in Italy.
- [GFS.A.N.IT.W0.S13.S1.Z.B.B9P.Z.Z.XDC.R.B1GQ.Z.S.V.N.T](#) - [Link](#)
Annual frequency, in Italy.
- [GFS.Q.N.I9.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N.T](#) - [Link](#)
Quarterly frequency, in Euro Area.

Additional Series Keys for Inflation Rate in Italy:

- [ICP.M.IT.N.000000.4.INX](#) - [Link](#)
Overall index, Index, Monthly frequency, in Italy.
- [MPD.A.IT.HIC.A.G24.0000](#) - [Link](#)
Annual growth rate - June 2024 staff macroeconomic projections, Annual frequency, in Italy.
- [MPD.A.IT.HIC.A.A23.0000](#) - [Link](#)
Annual growth rate - December 2023 staff macroeconomic projections, Annual frequency, in Italy.
- [ICP.M.U2.N.000000.4.ANR](#) - [Link](#)
Overall index, Annual rate of change, Monthly frequency, in Euro area (changing composition).

You can explore these datasets and series keys further using the provided links.

ICP.M.IT.N.000000.4.ANR



| SERIES_KEY | DATASET | DATE | VALUE |
|--|---------|------------|----------|
| GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N.T | GFS | 2024-01-01 | 135.2415 |
| GFS.Q.N.IT.W0.S13.S1.C.L.LE.GD.T.Z.XDC.R.B1GQ.CY.T.F.V.N.T | GFS | 2024-04-01 | 136.9999 |

PoC – Setup

Scope:

- **Time Series Included:**

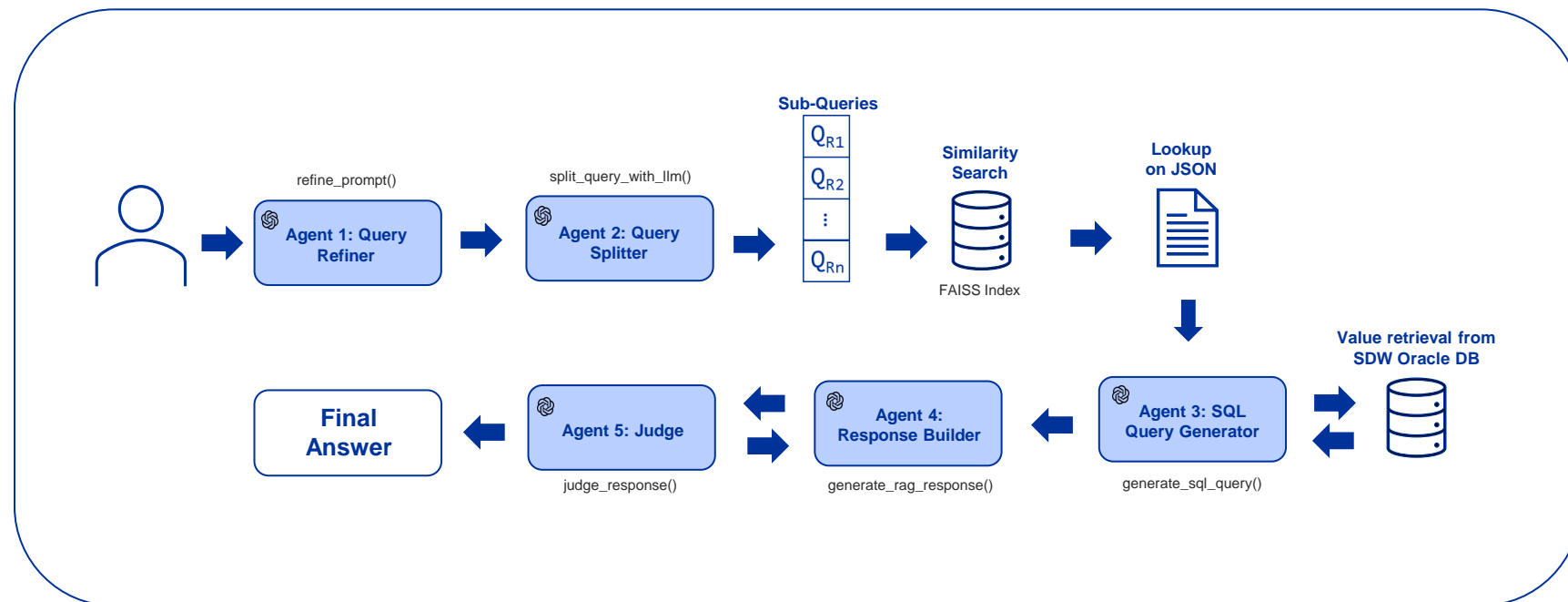
- ~300.000 TS are included (~10% of the EDP)
- including all TS either in Main Figures* or included in official publications, reports and articles

Technologies & APIs:

- Connection to LLM via **OpenAI API**
- Technologies:
 - **FAISS** (For RAG semantic search)
 - **GPT-4o** (Chat completion model)
 - **Text-embedding-3-large** (Embedder for RAG semantic search)

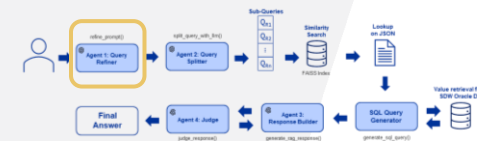
*Main Figures: <https://data.ecb.europa.eu/main-figures>

Current Architecture

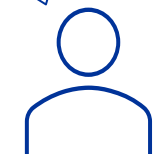


Behind the Scenes

Step 1: Query Refinement

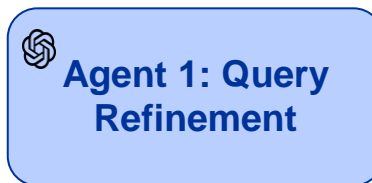


« What was the Inflation in Germany and Italy last year »



User Query

System Prompt: «You are an assistant that has to refine queries in order to perform vector search more effectively to retrieve data from the European Central Bank data portal. Your task is to identify and expand vague and ambiguous terms in the query...»



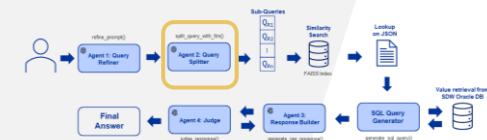
refine_prompt()



Refined Query

«Inflation rate in Germany and Italy in 2024»

Step 2: Query Splitting

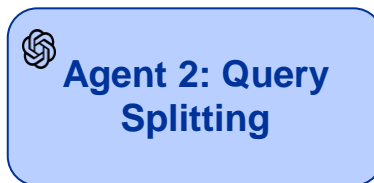


«Inflation rate in Germany and Italy in 2024»



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()

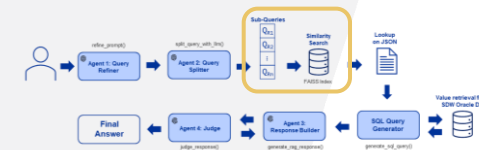


Q_{R1}
 Q_{R2}
:
 Q_{Rn}

N Sub-Queries

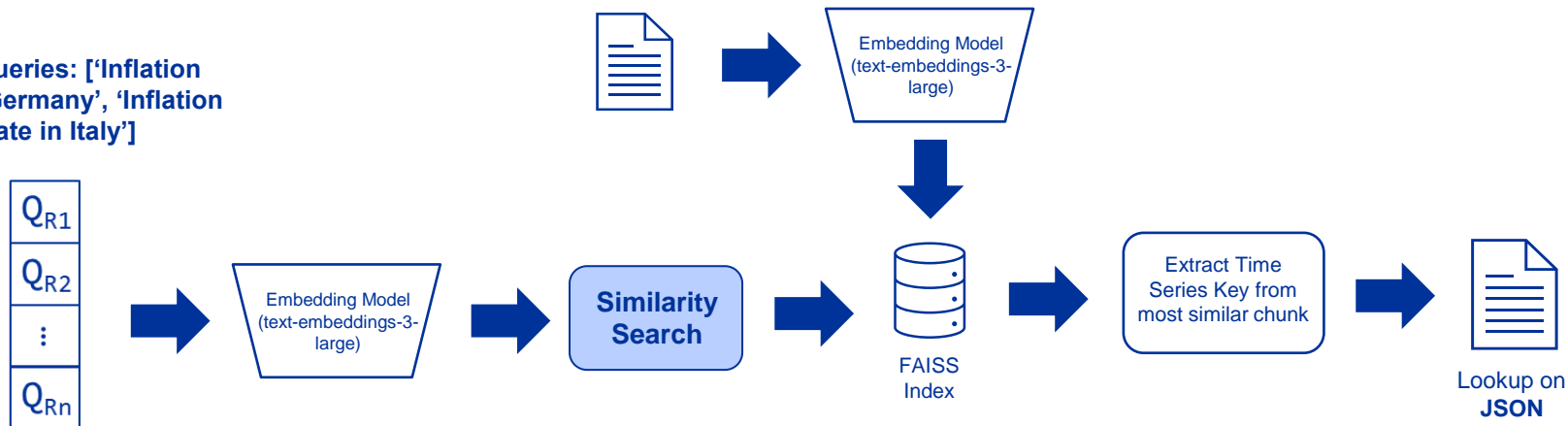
Sub-queries:
['Inflation rate in Germany',
'Inflation rate in Italy']

Step 3: Similarity Search

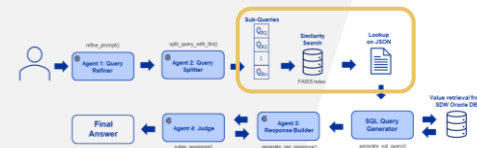


RAG_EDP.txt: Text database containing dataset, time series key and time series complete title, description and metadata for each time series in the selected subset

Sub-queries: ['Inflation rate in Germany', 'Inflation rate in Italy']



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>

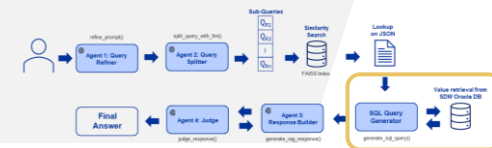
SERIES KEY LINK : <https://data.ecb.europa.eu/data/datasets/ICP/ICP.M.IT.N.000000.4.ANR>

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

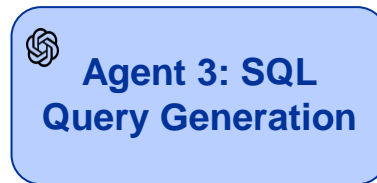


**Inflation rate in Germany
and Italy in 2024?,**

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



Extracted Time Series Key(s)
+ Refined Prompt



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
'2024-01-01' AND '2024-12-01'
```



Value retrieval from
SDW Oracle DB



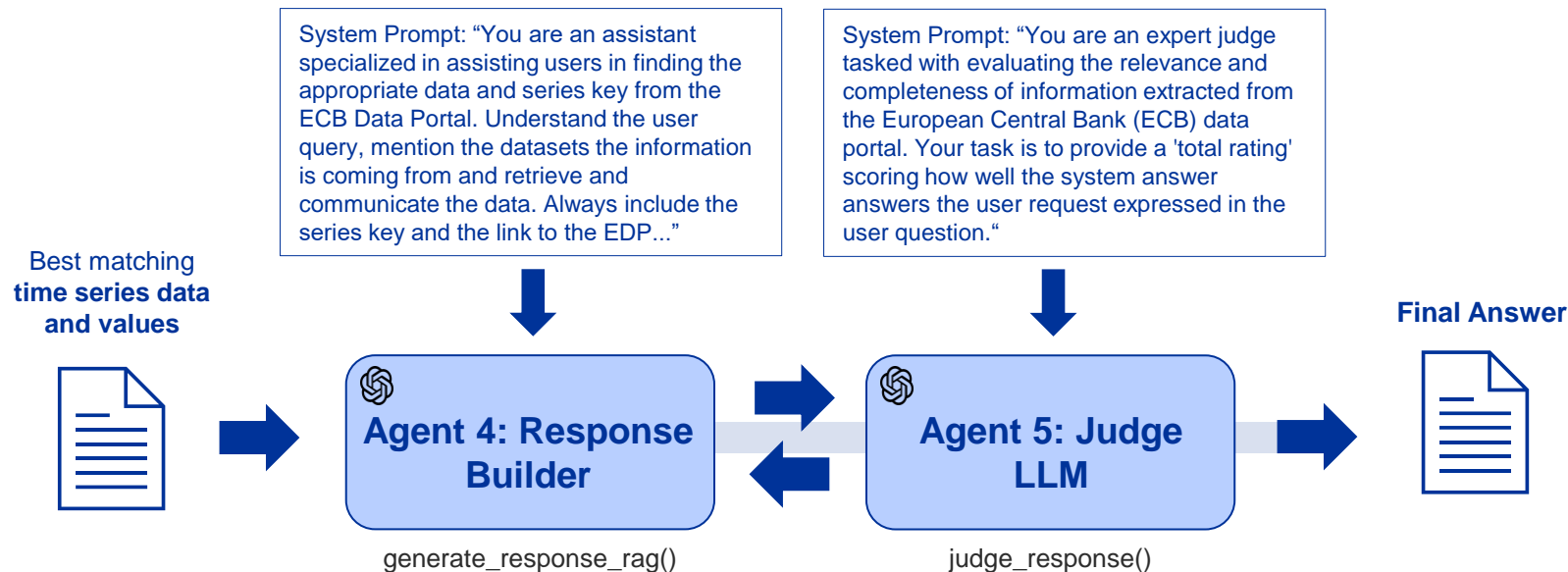
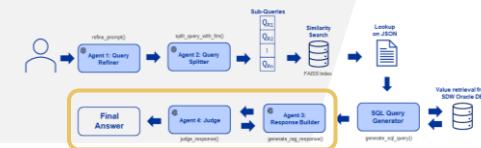
| | | |
|--|--|--|
| | | |
| | | |
| | | |

Table with values

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..."



Step 6: Response Generation



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

Our Approach

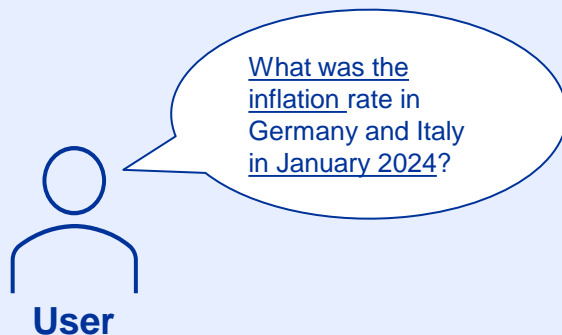
We leverage a mixture of state-of-the-art AI methods based on LLMs (**RAG and TEXT2SQL**).

Our chat assistant supports users with 2 types of requests in natural language:

- ✓ **Type 1 - Retrieving Values:** We retrieve the specific values a user in a time series a user is looking for.
- ✓ **Type 2 - Finding Data:** We identify the correct datasets and time series users are looking for.

Example: Retrieving Values Request

(later referred to as Type 1 user request)



Example: Finding Data Request

(later referred to as Type 2 user request)



Results – Value Accuracy

Test Set: 525 Q&A labelled pairs

Results (overall):

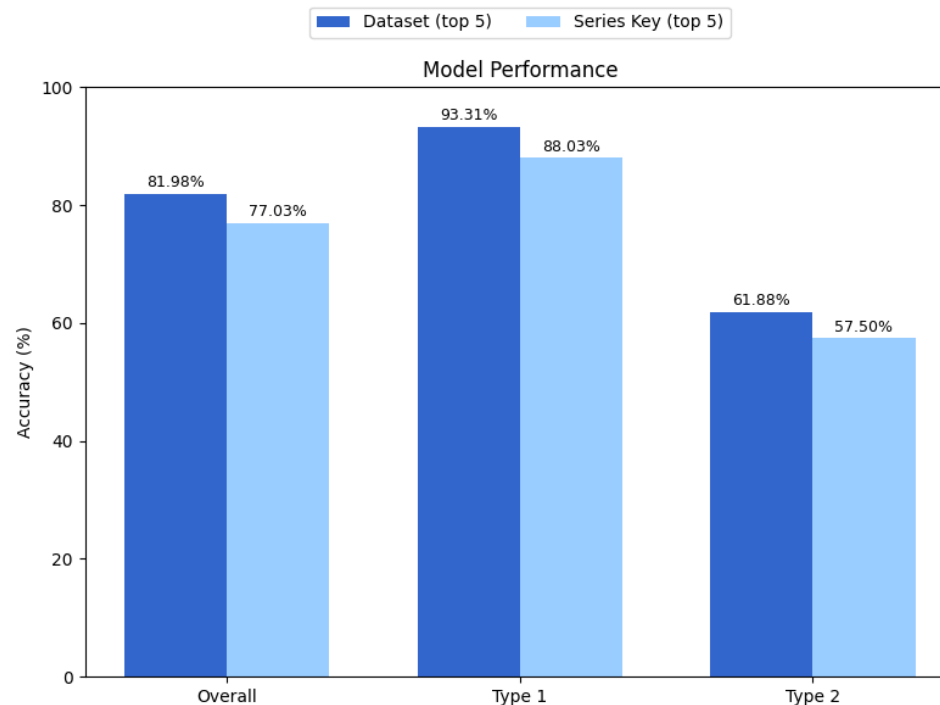
- Dataset Accuracy - **81.98%**
- Series Key Accuracy - **77.03%**

Type 1:

- Dataset Accuracy – **93.31%**
- Series Key Accuracy - **88.03%**

Type 2:

- Dataset Accuracy – **61.88%**
- Series Key Accuracy – **57.50%**



Key Insights and Way Forward

Key Insights and Way Forward

In a nutshell...

- The PoC shows promising results with regards to our goal of **helping users find data on the EDP**
- The only way to increase accuracy and reliability while decreasing costs is... **to avoid LLMs**
 - **Vector search** and fixed **SQL generation** instead of RAG for search and value extraction was crucial for reducing hallucination
 - Judge LLM as the “core” AI-based agent
- **Top-K multiple series key** retrieval drastically improved model accuracy

Next Steps:

- **Improve Accuracy**
 - Exploring new retrieval techniques for the similarity search (**clustering embeddings, hierarchical search...**)
 - Test a **conversational approach** building the SQL query field-by-field with the user
 - Increase the number of questions-answers to better measure accuracy
- **Finding the right balance between cost-benefit-risk**
- **Deploy the first version internally for a selected group of users with the objective of better understanding how users interact with the chatbot**

Thank you!