Can Large Language Models Revolutionalize Open Government Data Portals? A Case of Using ChatGPT in statistics.gov.scot

Marios Mamalis, Evangelos Kalampokis, Areti Karamanou, Petros Brimos and Konstantinos Tarabanis

PCI 2023

Information Systems Lab
University of Macedonia - Greece

HELLENIC REPUBLIC
UNIVERSITY OF MACEDONIA

Motivation

- New emerging technologies: Conversational large language models (LLMs), possessing strong natural language understanding and generation capacities.
- Vast amounts of already existing public sector data: Open government data, that offer a way for high-quality information to be publicly available.
- New opportunities to revolutionalize already existing open government data systems with the use of LLMs, in order to make them more accessible to the end users.

Large Language Models

- Part of the broader family of generative AI
- Demonstrating natural language understanding and generation capacities
- Conversational LLMs are LLMs that generate natural language in the form of a question and an answer.
- One such model is ChatGPT, created by OpenAI.
- Conversational LLMs rely on knowledge that comes as a result of their training procedure.

- The problem with LLMs: They lack factually correct knowledge at times, and tend to “make up” the facts as if they were real (hallucinating).
- The Solution: Enhance the LLMs’ knowledge by providing context along with the question.

Retrieval Augmented Generation

- The process of retrieving and supplying factual information as context to the conversational LLM is known as retrieval augmented generation (RAG).
- RAG components transfer knowledge to the model, relying on the retriever. Simple retrievers such as embedding cosine similarity based ones tend to underperform compared to more complex ones but have the added benefit of requiring less resources.
- RAG enhanced LLM applications can answer truthfully to the user as long as the external data used are factually correct.

Instruction Learning

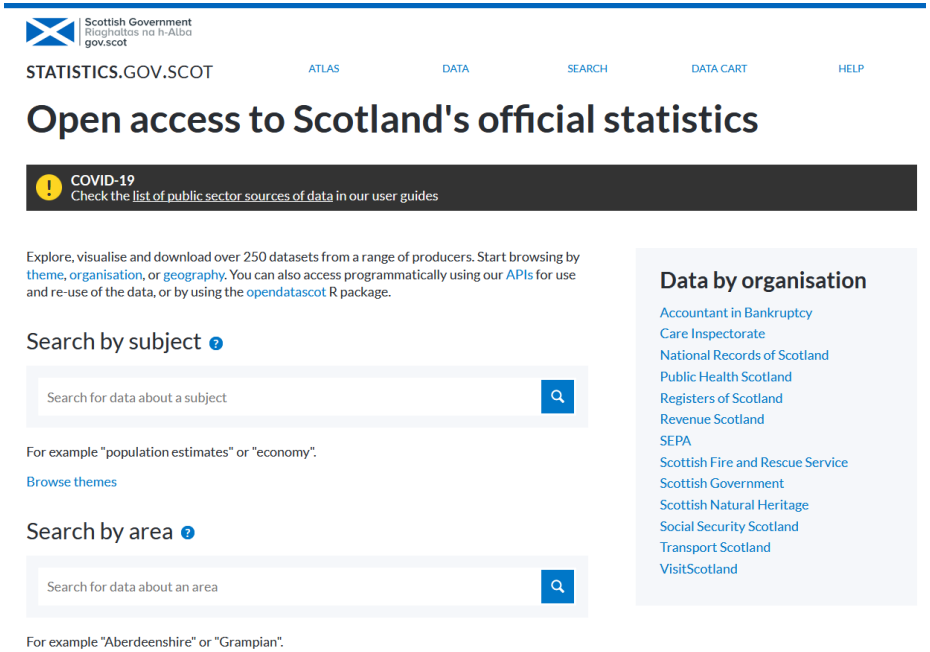
- Conversational LLMs of significant parameter size have demonstrated the ability of following written instructions that influence their response.
- Instruction learning refers to the cases were the user prompts the LLM with instructions to be followed and the LLM responds in a way that satisfies the rules set in the instruction.
- Through instruction learning LLMs can accomplice many downstream tasks such as structuring text in a certain way or extracting entities.

Open Government Data

- Published in the official data portals of governments in order to be freely accessible to the public.
- A political priority in the last decade in many countries.
- Enhance evidence-based policy making and stimulate economic growth.
- A large part of OGD are statistical data that:
  - Commonly regard aggregated demographic, social, and business indicators.
  - Are multidimensional, meaning that a measure is described based on multiple dimensions.
- A large number of official OGD portals are publishing OGD as linked data, which facilitate interoperability.

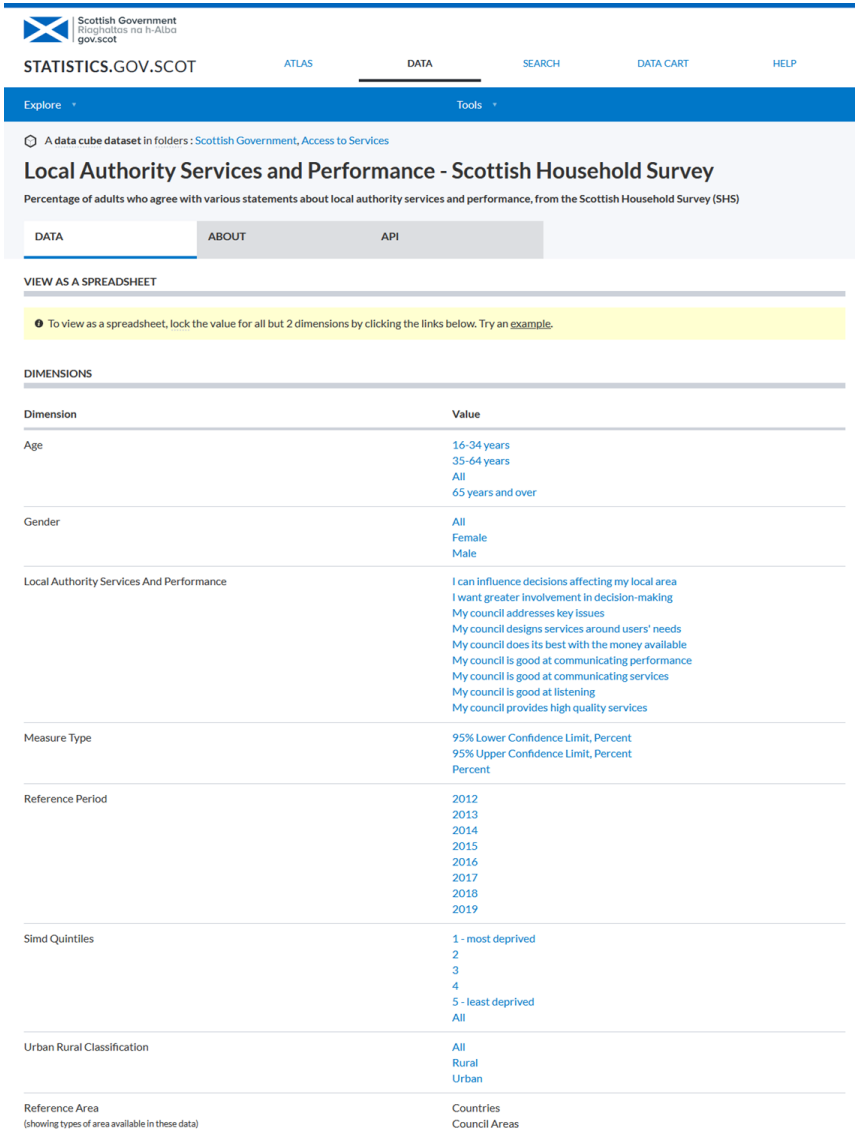
The case of the Scottish open statistics portal

- Hosts more than 250 linked datasets.
- Covers various societal and business aspects of Scotland classified into 18 themes.



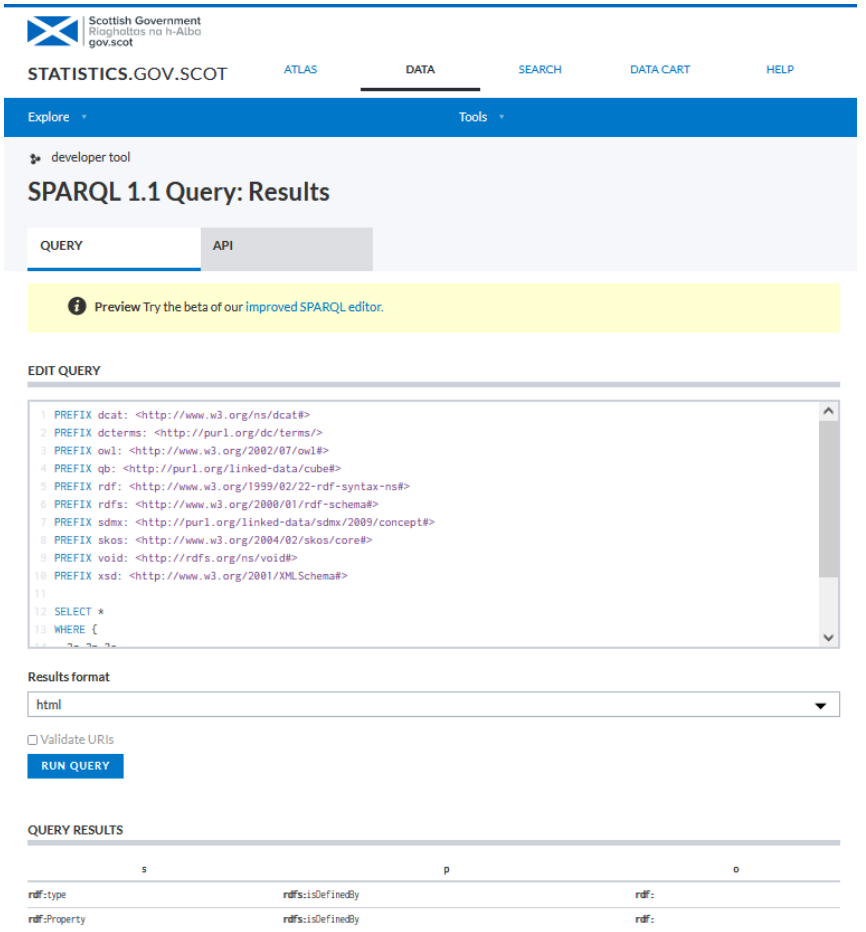
The interface

- Users can view and retrieve data as tables, maps, and charts or download them in various formats (e.g., html, json, csv).



The SPARQL query language API

- Alternatively, they can retrieve them as linked data by submitting flexible queries to the SPARQL endpoint released by the portal.
- The application is based on the declarative SPARQL query language.
- The endpoint is also available through an API.



Our approach

Our Goal: Create a proof of concept application with LLMs that answers user questions regarding the portal's data.

Our approach: Design the system as an integration of several components performing different tasks.

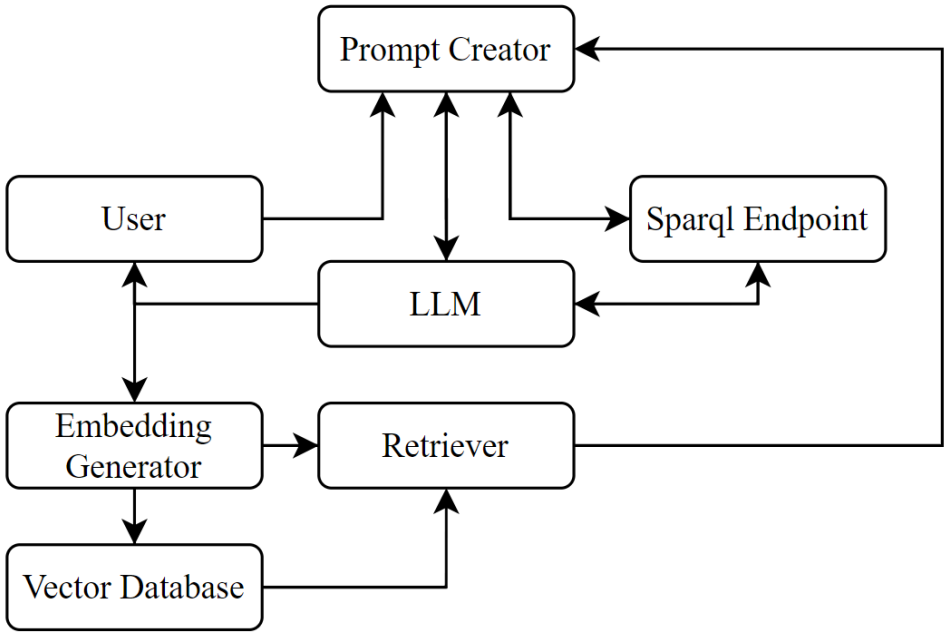
The components are six:

- The prompt creator
- The SPARQL endpoint
- The LLM
- The embedding generator
- The vector database

and

- The retriever

By working in conjunction with one another, the application can insert and filter information as needed to answer the user question truthfully.

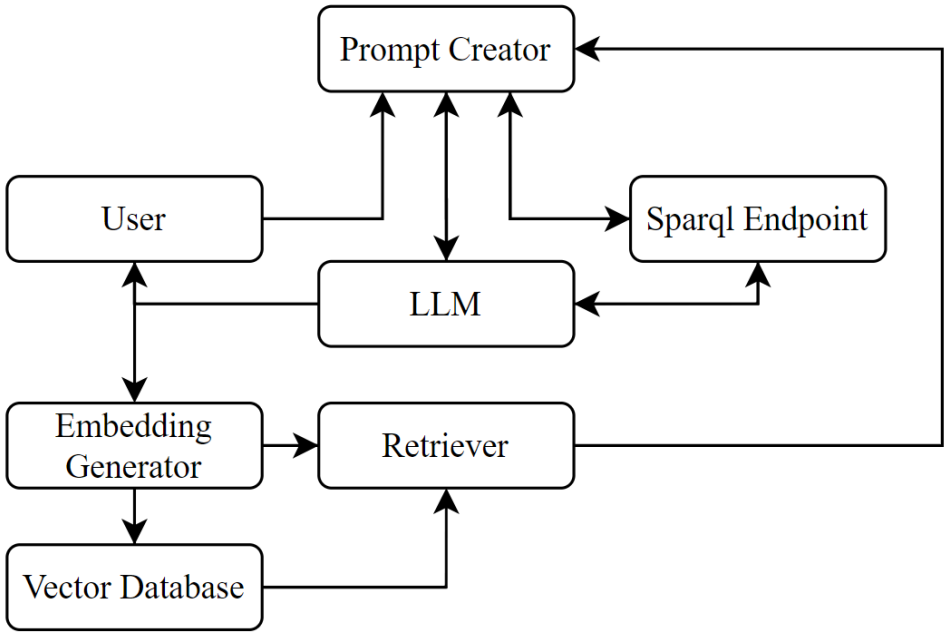


Our approach

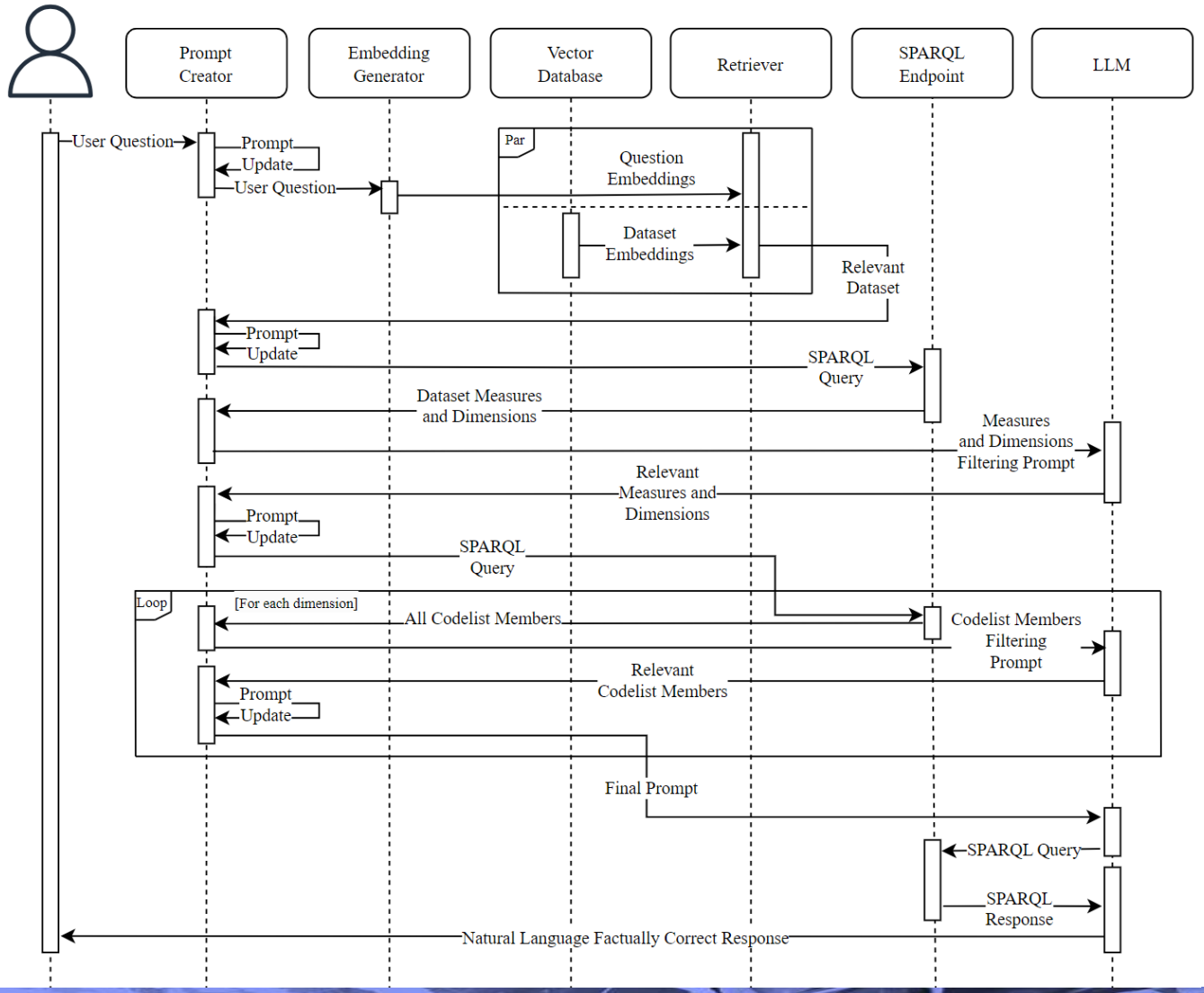
Components used:

- LLM: gpt-3.5-turbo-0613 and gpt-3.5-turbo-16k0613 offered through the OpenAI API.
- Embedding generator: text-embedding-ada-002 embedding generator model, offered through the OpenAI API.
- Vector database: ChromaDB vector store.
- Retriever: Cosine distance.

Limitations: The application can currently only support a certain type of question, more specifically, questions whose answer is a single value and not a list or a table.



The designed application



Evaluation

In order to evaluate the application in a qualitative manner, three sample questions were created for three different datasets with different properties, and passed as initial questions. The questions and answers follow:

- Q1: How many HMO licences were in force in Dundee City in 2014?
- A1: In 2014, there were a total of 1,577 HMO licences in force in Dundee City.

- Q2: How much vacant urban land was there in East Ayrshire in 2007?
- A2: In 2007, there were 37 units of vacant urban land in East Ayrshire.

- Q3: What was the count of international exports in 2002 in Scotland for education?
- A3: In 2002, Scotland had a remarkable count of 170 international exports in the field of education. This signifies the significant global recognition and demand for Scottish educational expertise and resources.

Discussion

- The proof of concept application functions as the connective link between user and open government data portals.
- It accepts queries in natural language, and through a series of retrievals, returns a factually correct, natural language answer.
- The potential for such an application are massive:
  - It eliminates barriers that currently restrict access to users with no technical knowledge from accessing available information.
  - Paves the road towards the democratization of information through the use of AI.
  - Maintains a high standard in terms of accessed data quality.
- However:
  - The proof of concept is restricted to a certain type of questions, and is not yet generalized.
  - It currently relies on models created and offered by OpenAI through APIs, making it intransparent, especially considering the nature of the data.

Thank you for your attention!