Integration of Large Language Models in the Public Sector

Martijn de Bruin[1,2], Marcio Fuckner[1], and Ahmed Nait Aicha[2]

[1] Amsterdam University of Applied Sciences (Responsible IT Lab), Amsterdam, The Netherlands
martijn.de.bruin@hva.nl, marcio.fuckner@hva.nl
[2] Municipality of Amsterdam, Amsterdam, The Netherlands
m.debruin@amsterdam.nl, a.naitaicha@amsterdam.nl

1 Extended Abstract

The integration of Large Language Models (LLMs) in public sector applications has gained attention due to their advanced natural language processing (NLP) capabilities. Despite their proficiency in generating text, LLMs are prone to generating inaccurate or unverified information, known as hallucinations, which limits their utility in domains requiring precise and reliable data.

This research focuses on improving information retrieval within the Finance and Procurement department of the municipality of Amsterdam. The current information system, F-desk, comprises predominantly unstructured data in PDF format, posing significant challenges for conventional retrieval systems.

LLMs offer superior language understanding but are not inherently designed to work independently for information retrieval. They are not supposed to function as reliable databases, often failing, coming up with incorrect, mismatched, and unverifiable sources.

To address these challenges, this study proposes implementing a Retrieval-Augmented Generation (RAG) approach, which combines the linguistic strengths of LLMs with a retrieval mechanism that sources information from authoritative databases. The hypothesis is that this integration can improve the accuracy and reliability of the information retrieval process, thereby enhancing operational efficiency within the public sector.

A Q&A Chatbot was developed and tested within the municipality of Amsterdam, using Azure OpenAI's ChatGPT-3.5-Turbo to manage natural language requests posed by users. The Fragments of the F-Desk documentation were transformed into low-dimensional vectors (embeddings) to capture semantic meanings and relationships. For this version, the text-embedding-ada-002 was used. A vector database was used for embedding storage and retrieval.

The chatbot's performance was evaluated using quantitative metrics, including BERTScore, Word Mover's Distance (WMD), and Rouge-L, as well as qualitative human assessments. Preliminary results demonstrated superior performance in maintaining answer accuracy compared to other data retrieval methods, particularly in handling unstructured data. With a chunk size of 1600 tokens, 88% of the generated answers were factually accurate. Although a graph database offered the potential for enhanced contextualization of domain-specific terms, the vector database was more effective in this context.

In summary, the RAG-based Q&A chatbot has the potential to reduce response times and increase productivity within the Finance and Procurement department by providing more accurate and accessible information. Future research should aim to expand the scope of source data: Performance should be assessed for different scenarios, such as different fields of knowledge and different document types.

To further enhance the PoC, the study recommends expanding the amount of source data available to the system to compare the impact of different types of documents. It also suggests improving evaluation mechanisms to ensure the ongoing reliability of generated answers, refining the graph infrastructure to provide more nuanced context, and splitting user input to handle multiple questions simultaneously rather than relying solely on the LLM.

Keywords: LLM · RAG · Chatbot · Public Sector.