Universitat de Girona
Escola Politècnica Superior

MASTER'S THESIS

Tuning a Generative Language Model for the Creation of Customized Chatbots for Public Administrations

Author:
Martí MAS FULLANA

Setembre 2024

Master's in Data Science

Supervisor:
Josep SUY FRANCH

CHAPTER 1

Summary

1.1 Introduction

This master's thesis explores the development of a customized chatbot system using GPT (Generative Pre-trained Transformer) and RAG (Retrieval Augmented Generation) technologies to improve the efficiency of public administration services in Catalonia, specifically targeting social rights and benefits. The project, conducted in collaboration with DXC Technology, focuses on creating a multilingual, accessible, and user-friendly interface while optimizing the chatbot's performance in terms of accuracy and cost.

Key components of the system include:

- **Frontend Interface:** Built with Angular to handle user interactions.
- **Backend System:** Manages the conversation flow, connects with Azure services, and employs a Vector Search API for information retrieval.
- **Data Preparation:** Involves sourcing information from public websites and processing it to improve chatbot responses.

1.2 Background

The thesis delves into the evolution of dialogue systems from rule-based models to advanced AI-driven systems. Key technologies and methods employed in this project include:

- **Generative Pre-trained Transformer (GPT):** A series of models from GPT-1 to GPT-4o that progressively improved natural language processing capabilities.
- **Retrieval Augmented Generation (RAG):** Integrates database retrieval with generative models to enhance response accuracy and relevance.

These technologies help in developing chatbots that are contextually aware and capable of providing precise information based on user queries.

1.3 Objectives

The main objective is to develop an advanced chatbot using GPT and RAG technologies that can provide accurate responses based on a database of social rights and benefits in Catalonia. The specific goals are:

- **Model Selection:** Choose an appropriate GPT model.
- **RAG Integration:** Implement a retrieval system for precise information access.
- **User Experience:** Design an accessible and user-friendly interface.
- **Evaluation and Validation:** Conduct user testing and performance analysis.

1.4 Methodology

The project employs the SCRUM methodology, an agile framework that supports iterative and incremental progress, ensuring adaptability and collaboration. Key roles and practices include:

- **SCRUM Roles:** Product Owner, SCRUM Master, Development Team.
- **Sprint Planning:** Organizes the workload in two-week cycles.
- **Daily Stand-ups:** Maintain team communication and transparency.
- **Sprint Review and Retrospective:** Collect feedback and continuously improve the process.

Data is collected and prepared from public websites, which is then indexed into a PostgreSQL database (with the PGVector plugin enabled) using the LlamaIndex Python library to facilitate efficient information retrieval.

1.5 Architecture

The chatbot system's architecture consists of several components:

- **Frontend:** Built with Angular, this handles the user interface, voice input, and interaction management.

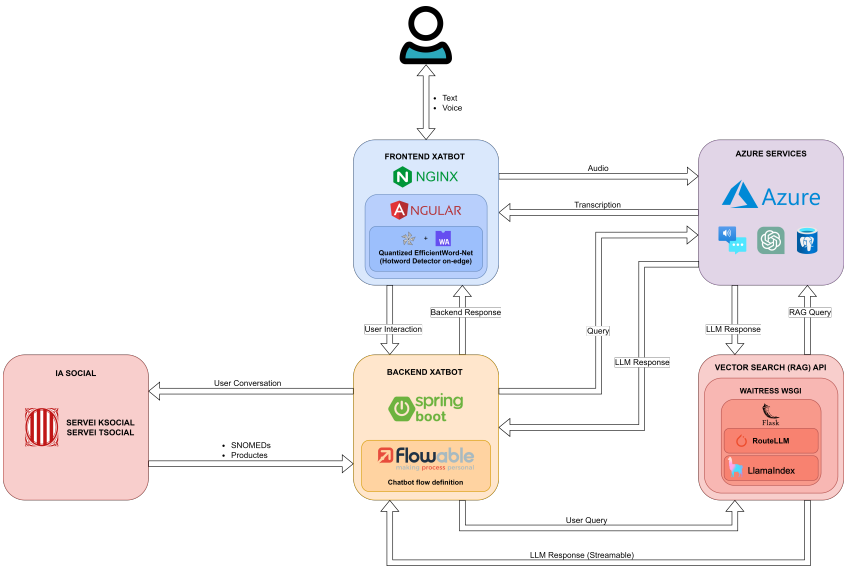

Figure 1.1: App Architecture

- **Backend:** Manages the conversation flow, error handling, and integrates with external services.
- **Azure Services:** Utilized for Speech-to-Text and Text-to-Speech functionalities.
- **Vector Search API:** A custom component for information retrieval that enhances response accuracy.

The system uses a modular design, enabling easy adaptation for future applications. The chatbot's top-level flow diagram is shown in Figure 1.2. Figure 1.1 illustrates the app architecture.

1.6 Results

The performance of the chatbot was evaluated using a benchmark of 35 questions, comparing the standard RAG and Small to Big Retrieval (STBR) methods. The STBR method demonstrated superior accuracy (0.79 average) compared to the standard RAG method (0.65 average), with a more consistent performance across different queries. However, STBR requires more computational resources, making it suitable for scenarios where accuracy is prioritized over speed.

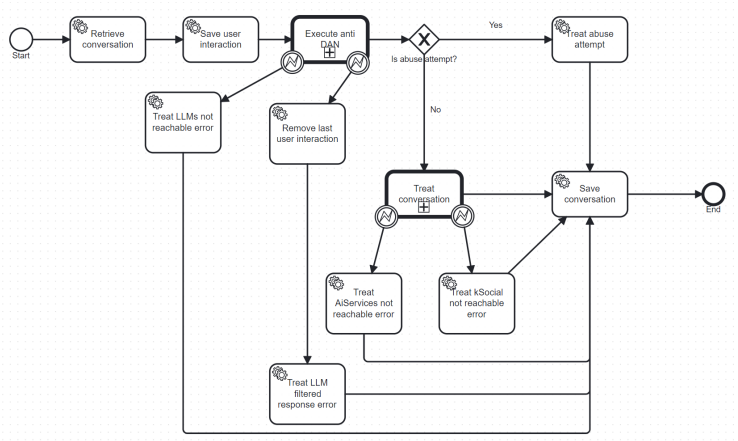

Figure 1.2: Top level chatbot flow diagram

Method	Accuracy ↑	Standard Deviation ↓
Normal RAG	0.65	0.031
STBR RAG	**0.79**	**0.024**

Table 1.1: Results of the benchmark. In bold we show the best result.

1.7 Conclusions

The project successfully developed a chatbot system that enhances the delivery of social rights and benefits information to users in Catalonia. The system's modular architecture, combined with advanced language models, provides a flexible and scalable solution for public administration needs. The use of RAG technology significantly improves the accuracy and relevance of the chatbot's responses.

1.8 Future Work

Future enhancements could include:

- **Multilingual Support:** Extend the system to support additional languages.
- **Enhanced Accessibility:** Improve features for users with disabilities.
- **Performance Optimization:** Reduce response times and improve scalability.
- **Continuous Improvement:** Regular user testing and iterative development.