Large Language Model Powered Chatbot for Comprehensive Citizens Information Services in Ireland

MSc Research Project
Data Analytics

Reinaldo Zanello Klostermann
Student ID: x21133018

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Reinaldo Zanello Klostermann		
Student ID:	x21133018		
Programme:	Masters in Data Analytics	**Year:**	2023
Module:	MSc Research Project		
Supervisor:	Vladimir Milosavljevic		
Submission Due Date:	14/08/2023		
Project Title:	Large Language Model Powered Chatbot for Comprehensive Citizens Information Services in Ireland		
Word Count:	8379	**Page Count:**	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Large Language Model Powered Chatbot for Comprehensive Citizens Information Services in Ireland

Reinaldo Zanello Klostermann
x21133018

Abstract
This study explores the development and evaluation of a large language model (LLM)-powered chatbot that employs GPT-4 to improve public information accessibility through WhatsApp. The chatbot uses the Citizens Information Ireland website as a data source. Creating the evaluation dataset is an automated process that ensures that the sample is unbiased and diverse. The chatbot was evaluated using ROUGE, BLEU, BERTScore, and UNIEVAL metrics based on 114 distinct question-answer pairs. A comprehensive analysis of the chatbot's performance was obtained based on mean scores and standard deviations. It demonstrates its ability to provide relevant, accurate, and context-appropriate responses. With a particular focus on improving public information accessibility, this study provides valuable contributions to the literature on LLM-enabled customer support, demonstrating the broader commercial potential of LLM-powered chatbots.

1 Introduction

1.1 Background & Motivation

Information plays a critical role in society, especially in the digital era. It shapes communication, decision-making, and interaction. Digital resources and online platforms have made information more accessible and enabled people to make informed decisions.

Information availability does not guarantee effectiveness. The recipient must be able to comprehend and interpret the information to transform it into knowledge that can assist them. Access to relevant information may also be limited by language barriers. Further complicating interpretation is the use of formal, technical language in official communications issued by public institutions. It is possible to divide the pursuit of knowledge and understanding in a particular subject into some tasks: finding relevant information, interpreting, and applying it effectively. It is evident that these challenges exist when considering most companies offer customer service to resolve inquiries. This is despite the fact that the information can also be found written on their websites in Frequently Asked Questions (FAQ). In many cases, customer service acts as an intermediary, locating and providing information to customers in an understandable way. Representatives can rephrase answers based on client knowledge. This is done by taking into account diverse educational backgrounds and varying degrees of subject expertise among individuals.

Technological transformation transitions human customer service to chatbots. Despite their ability to locate information quickly, chatbots still have difficulty maintaining fluid conversations and adjusting their tone to meet customers' specific needs. These challenges are addressed by Artificial Intelligence in Large Language Models (LLMs), such as a Generative Pre-training Transformer (GPT). By using algorithms, they produce text that closely resembles human text. ChatGPT presents this ability and has been trained on extensive textual datasets. Consequently, it can perform a wide range of language-related tasks effectively (Brown *et al.*, 2020).

Citizens Information[1] is a website that provides relevant information on public services and citizens' entitlements in Ireland. Users can access government information through this website. The website is only available in English and Gaelic. This can be a problem for an estimated 12% of non-Irish nationals residing in Ireland[2] who are not native English speakers. In addition, the website does not offer online customer support. As a result of these considerations, the implementation of a chatbot on the website would be a valuable addition. This would improve communication and support for users.

WhatsApp has over 2 billion users around the world, making it one of the biggest communication platforms. Because of its popularity, many businesses have started using it as a customer support platform.

The primary limitation of existing research lies in the technology employed. Current technology struggles with understanding queries, processing information, and responding in a conversational manner that mirrors human interaction. Given its novelty, the use of Large Language Models in chatbot development is not widespread. This project seeks to address these limitations by implementing the state-of-the-art GPT-4 model, which became available to paid users on July 10, 2023. This model overcomes the aforementioned limitations and represents a significant advancement in the field. During the literature review, it was observed that there seems to be a scarcity of published academic work that implements this technology in a chatbot designed for information retrieval. This observation suggests the innovative nature of this project and its potential contribution to the field.

The hypothesis is that by using a large language model, such as the GPT-4, it is possible to build a chatbot that will assist customers in obtaining information from the Citizens Information website, presenting it in a clear and concise manner, translating it to the user language, if necessary, on a platform such as WhatsApp that is widely used around the globe.

1.2 Research Question & Objective

1.2.1 Research Question

How effective is a Large Language Model powered AI chatbot in providing comprehensive citizens information services on WhatsApp?

[1] Citizens Information: https://www.citizensinformation.ie/en/

[2] Central Statistics Office: https://www.cso.ie/

1.2.2 Research Objectives

The following objectives were set to deal the research question posted above:

1. To conduct a critical analysis of recent literature on LLMs, information store and retrieval.
2. To preprocess the data for compatibility with the LLM.
3. To implement an AI chatbot capable of interacting with users via WhatsApp to answer queries based on the Citizens Information website.
4. To evaluate the performance of the implemented model.

2 Related Work

2.1 Chatbots

By automating tasks, artificial intelligence has revolutionized many industries. Chatbots have become one of the most popular artificial intelligence applications. Chatbots are applications that provide information and perform specific tasks based on human interaction and understanding (Rane et al., 2022).

In 1950, Alan Turing published a paper on machine thinking, which sparked interest in the concept of chatbots. Eliza, the first chatbot, was developed by Joseph Weizenbaum in 1966. Despite Eliza's limited conversational ability, it operated as a starting point for further advances in the field (Weizenbaum, 1966). Since then, many tech companies have implemented conversational chatbots, such as Facebook, Google, Amazon and Microsoft. Through Natural Language Processing (NLP) and continuous learning, these chatbots provide a more human-like experience, enabling them to comprehend meanings, have natural conversations, and improve over time (Chhabria and Damle, 2022). They have evolved from basic conversational models to sophisticated systems capable of understanding the context and providing personalized responses to user queries.

The development of chatbots has been extensively researched. Their review (Chhabria and Damle, 2022) emphasises the importance of NLP and semantics. (Gonsalves and Deshmukh, 2023) discuss the evolution of chatbots, highlighting their impact on businesses and the growing interest in using them for customer communication. These studies contribute to the understanding of chatbot advancements and potential applications in various domains.

User interactions have been changed by AI and NLP. In particular, Weizenbaum's Eliza set the stage for their subsequent development. Currently, GPT-4 is a cutting-edge LLM that is gaining a great deal of popularity among users. Major corporations are utilizing conversational chatbots to improve user experiences and streamline operations through advances in NLP. Based on given instructions, ChatGPT is an advanced conversational AI that provides comprehensive and tailored responses. Within the Artificial Intelligence Generated Content (AIGC) community, ChatGPT has shown impressive capabilities in understanding and generating language (Wu *et al.*, 2023).

2.2 Natural Language Processing

Natural language (NL) is the primary mean of communication. However, the digital world operates on binary data, a format inherently incompatible with human language. In order to resolve this discrepancy, NLP is required, a sophisticated multidisciplinary domain that integrates computer science, AI, and linguistics. NLP is employed to develop systems that comprehend, interpret, and generate human language (Vajjala *et al.*, 2020). NLP is used to build a conversational agents to be used in question-answering applications. anguage ambiguity and creativity are two of the reasons that make NLP such a challenging field.

In NLP, Machine Learning (ML) and Deep Learning (DL) have emerged as pivotal tools for pattern recognition, semantic analysis, and linguistic interpretation. Their role is especially crucial in an era of increasing data, where traditional linguistic approaches may not be sufficient. Despite their capabilities, they also have limitations, which further illustrate the field's complexities.

2.3 Deep Learning

Deep learning is a branch of machine learning that involves training artificial neural networks with multiple layers to learn hierarchical representations of data (MINAEE *et al.*, 2022). A significant advancement in NLP has been made by DL, which allows neural networks to learn directly from raw text data, eliminating the need for manual feature engineering. Modelling sequential data in NLP tasks has been successful using recurrent neural networks (RNNs) and their variants (Yin *et al.*, 2017). Transformers, which are based on self-attention mechanisms, have recently become a powerful architecture for NLP. A number of Language Models (LMs) have been developed using transformers, including Bidirectional Encoder Representations from Transformers (BERT) and GPT, which have achieved state-of-the-art performance in areas such as language translation, question answering, and text classification (Gong, 2022). By enabling neural networks to automatically learn and extract meaningful features from text data, DL has revolutionized NLP. In particular, transformers have contributed significantly to NLP advancement.

2.4 Transformers

In 2017, Google researchers introduced Transformers, a breakthrough in NLP, which solves limitations previously encountered with RNNs. By employing a mechanism called self-attention, Transformers can efficiently overcome traditional challenges associated with RNNs, such as parallelization difficulties and the vanishing/exploding gradient problem. By doing so, the AI model can selectively focus on pertinent portions of the input, thus enhancing its ability to deal with sequences of varying lengths. To determine the importance of different parts of the input, the self-attention mechanism utilizes three matrices - Query (Q), Key (K), and Value (V). This focus on relevant information and the employment of semantic and positional encodings contribute to Transformers' impressive performance in NLP tasks like translation and summarization (Vaswani *et al.*, 2017). Figure 1 presents the transformer architecture.

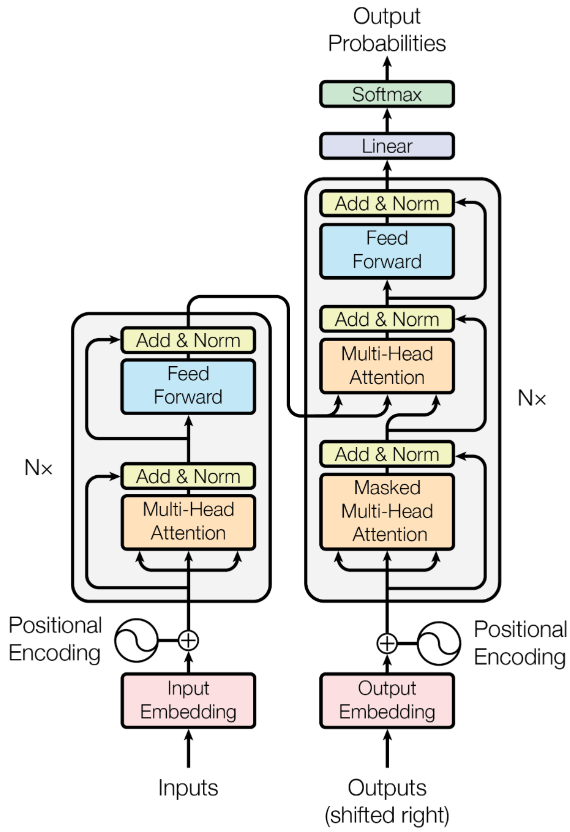

Figure 1: Transformer model architecture (Vaswani *et al.*, 2017).

Transformers are crucial for their ability to generate and translate text, as well as writing computer code. The most advanced models are based on them, including GPT-4. Unlike RNNs, Transformers process language non sequentially, enabling them to recall the context of long text sequences. These systems are characterized by three key innovations: positional encoding, attention, and self-awareness. Word order is represented by positional encodings. Attention is used to allow a text model to consider all words in a sentence when translating it. Self-attention enables the neural network to understand a word in its context, which helps disambiguate words, recognize parts of speech, and identifying the tense.

Commonly used in popular models like BERT, which powers Google Search and Google Cloud's NLP tools, Transformers are highly adaptable to a variety of tasks, including text summarization, question answering, and classification. As a result, they manage large text sequences and offer enhanced training efficiency. Since being developed in 2017, it has revolutionized NLP.

2.5 Large Language Models

NLP has been transformed by LLMs enabling a wide range of tasks. For them to be effective, concepts such as prompt design, prompt engineering, and chain of thought reasoning are essential. In prompt design, prompts are tailored specifically to the task at hand, while in prompt engineering, prompts are optimized to improve their performance. Chain of thought reasoning improves problem solving, where models provide reasoning before presenting an answer.

AIGC is an emerging technology that allows artificial intelligence to automatically create personalized content, such as images, text, and videos (Raffel *et al.*, 2020). A major advancement in GPU technology has led to the development of large-scale deep networks that are capable of processing downstream tasks more effectively (Gusak *et al.*, 2022). The release of OpenAI's ChatGPT at the end of 2022 marked a significant milestone in this field, as it successfully integrated a variety of technologies, including DL, unsupervised learning, instruction fine-tuning, multi-task learning, in-context learning, and reinforcement learning, accumulating over 100 million active users within two months of its initial release.

LMs are pre-trained models, which use statistics to calculate NL probability (Zhai, 2007). Together with self-supervised learning over large-scale texts (Qiu *et al.*, 2020), it has paved the way for a innovative two-stage learning paradigm of pre-training and fine-tuning, transforming the landscape of NLP tasks. Several large models, such as GPT, BERT, and T5, have demonstrated the capability of learning complex features from raw data. This has improved the generality and generalization of large-scale pre-trained language models. An example of this advancement is the use of In-Context Learning in the GPT-3 series models, which exhibits impressive results in tasks such as article generation and code writing without requiring specific training or fine-tuning of NLP tasks (Wu *et al.*, 2023).

Even though LLMs are powerful, they can still struggle with tasks that require knowledge beyond their training data.

2.6 Prompt Engineering

For a LLM to be effective, prompt engineering is crucial. It can improve the quality and accuracy of the answer. Few-shot learning allows these models to comprehend and apply new concepts with minimal examples, whereas zero-shot learning enables them to recognize unseen categories or concepts without prior examples. Prompting is a popular technique in NLP research, which requires models to be conditioned on specific examples or instructions.

The capacity of trained LLMs to learn few-shots is known, however (Kojima *et al.*, 2023) demonstrate that LLMs have significant zero-shot reasoning skills. They present Zero-shot-CoT (Chain of Thought), a single prompt template that significantly improves zero-shot LLM performance on a variety of reasoning tasks. In various benchmark reasoning tasks, such as arithmetic, symbolic reasoning, and logical reasoning, experimental results demonstrate substantial accuracy improvements. The study emphasizes the importance of exploring and analysing the untapped zero-shot knowledge hidden within LLMs before relying on fine-tuning datasets or few-shot examples.

As highlighted in Kojima's research impressive results can be achieved when the appropriate prompt is utilized. The model was tested on a variety of benchmark datasets related to reasoning, and it demonstrated a significant improvement in accuracy. The research suggests that in certain instances, crafting specific templates may not be necessary. A zero-shot approach, prompting the model to "think step by step," can be sufficient to obtain an accurate response. This method simplifies the process while still yielding satisfactory results, underscoring its significance.

However, the examples provided in Kojima's work do not encompass complex tasks such as income tax calculation, which require consideration of numerous intermediate steps. Moreover, the time needed to test this capability might surpass the time required to implement tailored examples for few-shot learning. Despite these findings, this area of study deserves more in-depth research.

(Wei *et al.*, 2023) investigates the use of chain of thought prompts to improve language models' reasoning abilities. Their proposal is to generate a series of short sentences in response to a question to induce a coherent chain of thought, simulating human reasoning. A chain of thought prompting significantly increased accuracy in their experiment. By decomposing multi-step problems into intermediate steps, language models can solve individual components rather than tackling the entire problem at once.

In an effort to develop the performance of LLMs for language understanding and interactive decision-making, (Yao *et al.*, 2023) studied their reasoning and acting capabilities. The proposed method, called ReAct, enables LLMs to generate both reasoning traces and task-specific actions simultaneously. Inducing, tracking, and updating action plans are made possible by reasoning traces. Actions are utilized by the model to interface with external sources and collect additional data. Experimental results indicate that ReAct is more interpretable and trustworthy than state-of-the-art baselines on a variety of language and decision-making tasks. The method effectively addresses issues of hallucination and error propagation in chain-of-thought reasoning, producing human-like task-solving paths.

To conclude, prompt engineering is crucial to improving reasoning capabilities in large language models. As discussed in this literature review, different prompt engineering techniques have been demonstrated to be effective. These techniques include zero-shot prompting, chain of thought prompting, and integrating reasoning with action. These approaches have demonstrated promising results in improving reasoning performance and expanding the number of tasks LMs can perform. With prompt engineering, there is an opportunity for the development of more versatile and broader cognitive abilities in language models.

2.7 Embeddings

Word embeddings in NLP are numerical vectors in a continuous space that represent words. It is designed to capture the semantic and syntactic relationships between words, allowing machines to comprehend and reason about NL. By representing words as vectors in which each dimension represents a specific feature or property, embeddings can capture a variety of aspects of word meaning and context (Neelakantan *et al.*, 2022). Words with similar semantic or contextual meanings have similar vector representations, and consequently, are clustered together in space. Because of this proximity, algorithms can perform word similarity comparison. Unsupervised ML techniques are normally used to learn word embeddings. To learn vector representations, these algorithms use the co-occurrence statistics of words in the text (Devlin *et al.*, 2019).

Word embeddings are not limited to words. It can represent phrases, sentences, or even entire documents.

2.8 Vector Database

High-dimensional vector representations are stored and retrieved efficiently in vector databases, also called similarity search databases. Similarity searches are performed on large vector collections using this tool. This is not limited to text; it can include audio and images. In the database, these vectors are organized and indexed in a way that facilitates efficient search operations using similarity or nearest neighbour criteria (Johnson, Douze and Jégou, 2017). Vector databases' indexing structure is crucial for fast search operations. Indexing techniques aim to reduce search space and enable quick retrieval of vectors similar to or nearest to a given query vector. Vector databases are commonly used for similarity searches and their applications include recommendation systems (Cheng et al., 2016) NLP (Devlin et al., 2019), multimedia search, anomaly detection, Question Answering (QA) systems and many others. Using vector databases, these applications can efficiently search and retrieve relevant data points based on their similarity or proximity to a given query vector. This enables tasks such as question and answer, information retrieval, and data analysis.

2.9 Evaluation

Evaluation of QA systems is a complex process that involves the construction of the system, and also its quantitative formalization (Farea *et al.*, 2022). A variety of methods and metrics have been used to evaluate QA systems.

According to the literature review the most used metrics are Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), F1, and BERTScore (Chen *et al.*, 2019). These metrics, which have been widely used in a variety of NL generation tasks are primarily based on n-gram similarity. They are key to evaluating QA, however, they have significant limitations. This is primarily because they do not account for the lexical and compositional diversity that preserves meaning. BLEU, for instance, is a precision-based metric that scores a candidate by computing the number of n-grams that also appear in a reference. Based on the longest common subsequence, ROUGE is another F-measure metric designed to evaluate translation and summarization (Chen *et al.*, 2019).

In general, the F1 score has been used to evaluate span-based question answering (Rajpurkar et al., 2016). It gives a partial F1 score to a potential answer when its tokens match those of the correct answer. The ultimate measure is determined by taking the average of F1 scores from the test dataset (Kamalloo *et al.*, 2023). As BERTScore uses word representations rather than exact matches, it captures paraphrases more accurately than existing metrics and correlates better with human judgment (Chen et al., 2019). In certain situations, automated assessment systems can replace lexical comparison, but not for the extended responses produced by LLMs. Human judgement improves evaluation, however it has high cost and scalability limitations (Kamalloo *et al.*, 2023).

Most QA studies in the NLP field predominantly use two token-level measurements, Exact Match (EM) and Token F1 (F1). Unfortunately, neither of these metrics distinguish between major and minor span discrepancies (Bulian *et al.*, 2022). An evaluation of end-to-end answer retrieval models has also been introduced with the benchmark Retrieval Question Answering (ReQA) (Ahmad *et al.*, 2019). This benchmark assesses a model's ability to efficiently extract pertinent answers from an extensive set of documents. The BERTScore evaluation metric has been proposed as an improvement to the existing n-gram-based evaluation metrics. It has been demonstrated that BERTScore correlates highly with human evaluations due to its ability to match paraphrases and capture distant dependencies (Zhang *et al.*, 2020). Verifiability of generative search engines, which produce answers to user questions accompanied by embedded references, is also evaluated. This includes their citation recall and precision, providing a comprehensive and accurate picture of their performance (Liu, Zhang and Liang, 2023). Unlike traditional metrics, UNIEVAL evaluates generated text across multiple explainable dimensions rather than focusing solely on n-gram similarity. In this way, it is possible to provide a more nuanced and comprehensive assessment of the quality assurance system outputs. Having an intermediate learning phase, UNIEVAL can also incorporate external knowledge. Using this approach, it provides a more human-like evaluation, addressing some traditional metrics shortcomings. Compared to other metrics, experimental results indicate that UNIEVAL has a higher connection to human results. Thus, UNIEVAL represents a significant development when finding more reliable QA metrics (Zhong *et al.*, 2022). Overall, research on the evaluation of QA systems and LLMs continues to evolve aiming to overcome current limitations.

3 Research Methodology

The framework employed in this project is an adaptation of the Cross-Industry Standard Process for Data Mining (CRISP-DM). This methodology contains six interactive stages. The progression through these stages may not always be linear and may require cyclical movement between stages depending on the needs of the project. Figure 2 presents the stages involved in CRISP-DM methodology. This approach offers a standardized and industry-accepted guideline to ensure the project stays on track and yields beneficial results.

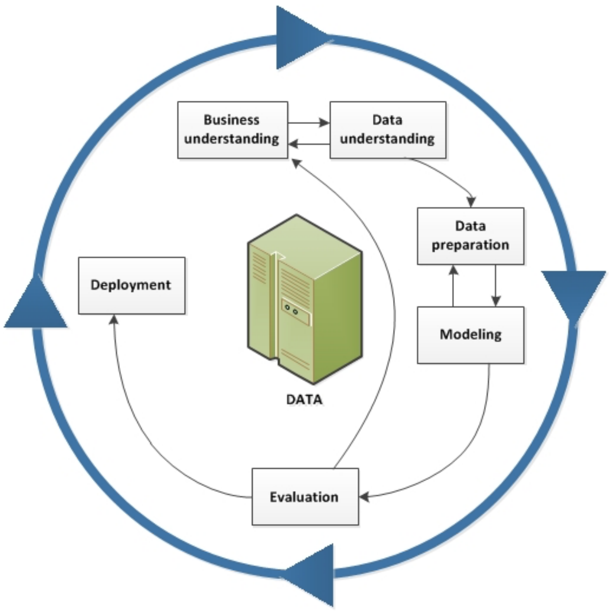

Figure 2: CRISP-DM methodology[3]

[3]CRISP-DM Overview: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

3.1 Business Understanding

The main goal at this stage is to understand the project objectives and requirements. The key objective is to create a customer-oriented chatbot that leverages a LLM which is GPT-4. This chatbot will provide customers with easy access to pertinent information from a Citizens Information website. The tool will translate this information if necessary, enabling users to understand and interact with the content easily.

As stated on the Citizens Information website, it was developed to meet customer demands for quick, easy access to comprehensive information on rights and entitlements. The website covers a broad range of subjects, including social welfare, employment rights, home buying, moving abroad, education, and much more. The information, available in both English and Irish, is sourced from official entities such as the Revenue and Department of Justice. It provides step-by-step calculations and simulations for taxpayers, mortgage applicants, and others. The site serves as a vital information resource for all residents in Ireland, particularly for foreigners who may not be familiar with the country's obligations and entitlements. While this project primarily focuses on the methodology, which can be applied to any domain, it is important to note the potential impact of such a tool on the accessibility and understanding of crucial information for a wide range of users.

The output should be a sophisticated language model capable of understanding user queries, sourcing the correct information, processing it into a clear and concise response, and presenting this in a chat interface. The model should also be capable of processing user input for simulation requests and performing the necessary calculations to provide relevant results. By establishing these objectives and requirements, it can now proceed to the next stage in the CRISP-DM process, which is the data understanding stage.

3.2 Data Understanding

This project utilizes various types of data from citizensinformation.ie. It includes text, numbers, tables, and calculations. Various aspects of citizens' rights, state services, financial aid, taxes, and benefits are explored in the data.

Even though Citizens Information is an official source, quality verification is crucial for data accuracy and consistency. This phase prepares the data for the subsequent data preparation phase. This is where data will be pre-processed and transformed to be inserted into the LLM that powers the chatbot.

3.3 Data Preparation

Prior to modelling, raw data must be pre-processed, cleaned, and transformed as part of the Data Preparation phase of the CRISP-DM methodology. The LLM that powers the chatbot has a limit on the number of tokens it can accept as input for this research project. Consequently, smaller, more manageable chunks of data must be prepared.

- **Chunking and Embedding**

The first step is to break down the text data from citizensinformation.ie into smaller chunks. The size of these chunks is determined by the LLM token limit. Through a process known as embedding, these chunks of text are transformed into numerical representations. Word embedding converts words into numerical vectors. This technique encapsulates semantic and syntactic relationships between words so that a machine can understand NL. As stated in (Neelakantan *et al.*, 2022), words are represented as dense vectors in which each vector dimension represents an aspect of the meaning or context of the word. As a result of this transformation, storage and processing are more efficient.

- **Vector Database**

After the data is embedded, it is added to a vector database. The database stores and retrieves numerical representations of data chunks. The vectors within the database maintain the context and semantics of the original text data, which is crucial to the chatbot's operation.

- **Similarity Search**

Using this method, data chunks relevant to a user query can be efficiently retrieved. Similarity searches are used to find vectors and data chunks most similar to the query vector.

The preparation of the data is crucial to ensuring that the LLM can effectively process the information from the citizensinformation.ie website. This will enable the chatbot to provide accurate and useful responses to users.

3.4 Modelling

During the modelling phase of the CRISP-DM methodology, the prepared data is utilized to build the model that powers the chatbot. This step leverages the power of the LLM, specifically the GPT-4 model, which is the latest version of the Generative Pretrained Transformer series. In this study, prompt engineering is the primary technique used to condition the LLM to perform a wide range of tasks.

This modelling phase sets the stage for the evaluation stage, where the model's effectiveness will be assessed based on its performance on a variety of reasoning tasks and its ability to provide accurate and useful responses to user queries.

3.5 Evaluation

Despite these developments, determining how effective QA systems are remains a challenge. The most reliable evaluations are currently based on human judgment, which is fundamentally qualitative and individual. However, human evaluation cannot be scaled due to the impracticality and inefficiency of having individuals annotate thousands or even hundreds of thousands of data points. A number of benchmark datasets have been developed to evaluate QA systems to address this issue. However, a model with high accuracy on a benchmark dataset may not achieve the same level in Citizens Information data. The significance of this consideration lies in the fact that evaluation should not be limited to generic benchmark datasets but should also incorporate specific datasets that reflect real-world scenarios. Therefore, creating a dataset based on Citizens Information data is necessary for this research project. Human annotators may frame questions based on areas where they know the QA system's performance. To mitigate these challenges, this project employs LLM to generate questions and answers which will be presented in the implementation section.

3.6 Deployment

In the deployment phase of the CRISP-DM methodology, the final model is made available to the intended users. This study successfully integrates the GPT-4 model with WhatsApp. Using Twilio and ngrok[4], the chatbot was connected to WhatsApp's large user base. The Twilio API enabled the chatbot to send and receive WhatsApp messages, however it required a publicly accessible URL to communicate with the local servers. By creating a secure connection to the local Python environment, ngrok enabled Twilio to establish a connection with the chatbot. Deploying AI systems involves not only programming AI models, but also managing integrations within environments. Despite the challenges, these integrations are essential for allowing the chatbot to function efficiently across a variety of platforms. Through this successful integration, the chatbot solution has proven its ability to provide high-quality, AI-based customer service via WhatsApp.

4 Design Specification

This project aims to build a sophisticated chatbot powered by a LLM. This chatbot retrieves, processes, and interacts with data from a Citizens Information website, as shown in Figure 3. This chatbot engages users on the WhatsApp platform. Firstly, the website content is mapped and transformed into textual format. Split into smaller chunks. Numerical vectors are generated from these text chunks during the embedding phase. The vectors are then stored in a vector database. The database is designed to perform high-speed vector operations, enabling quick retrieval and manipulation of embedded text. Based on these queries, the model extracts relevant information from the vector database and inserts it into the LLM alongside the question. This generates an appropriate response. User queries are answered by a chatbot powered by LLM on the WhatsApp platform.

[4] ngrok: https://ngrok.com/

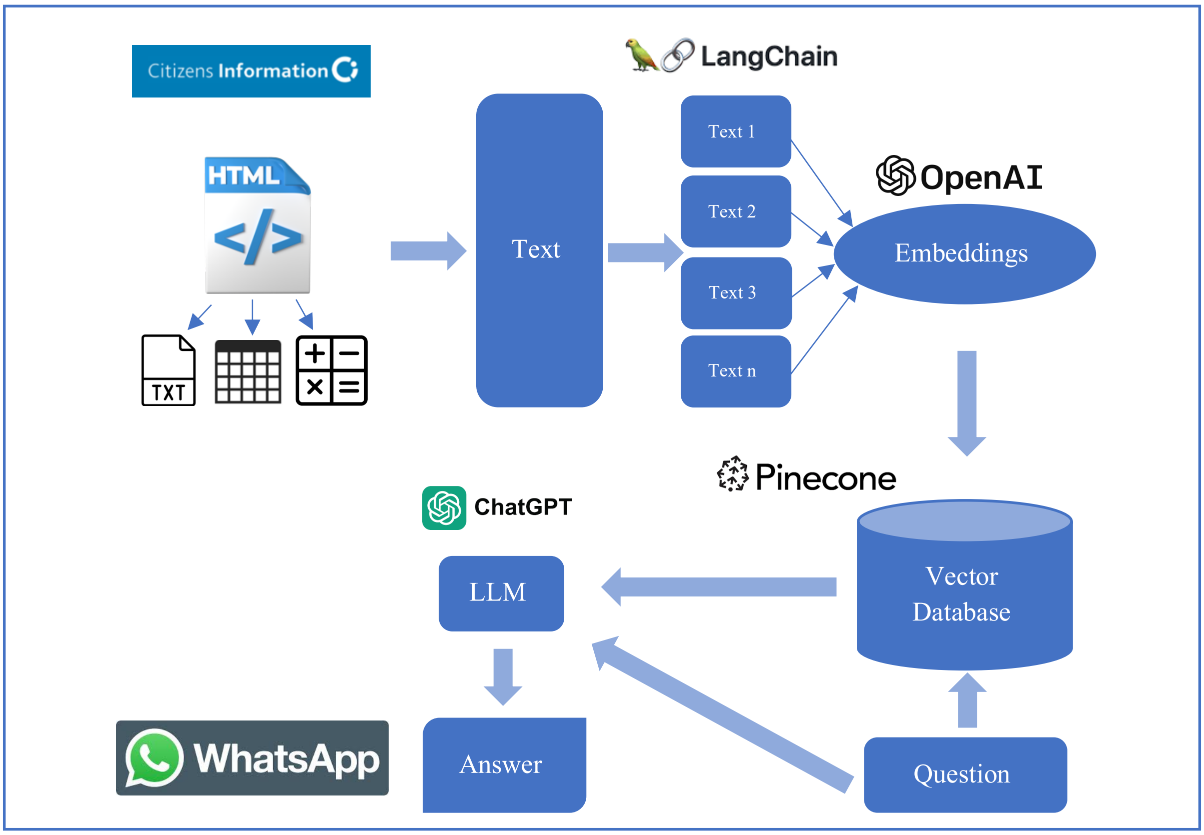

Figure 3: Research project design framework

5 Implementation

This study employs a systematic and sequential methodology to develop a chatbot that interacts with the Citizens Information webpage. A GitHub repository for the project will be made public to encourage replication and peer review. This will promote collaboration and transparency in research.

5.1 Environment Setup

The implementation process starts with the definition of a computational environment. Visual Studio was the Integrated Development Environment (IDE) used. Due to its vast array of NLP-specific libraries, Python 3.11 has been chosen as the coding language for this project.

The solution employs GPT-4 as the LLM, Pinecone as the vector database, and WhatsApp as the user interface. Python interacts with GPT-4 and Pinecone via API calls. To generate these API keys, an account with Pinecone[5] and OpenAI[6] is required. LangChain[7] is the framework used in this research project. LangChain provides a framework for developing applications based on language models.

5 Pinecone: https://www.pinecone.io/

6 OpenAI: https://openai.com/

7 LangChain: https://python.langchain.com/en/latest/

5.2 Data

As shown in Figures 4, 5 and 6, the site information is unstructured, consisting of text, numerical figures, and tables. Calculations illustrating various concepts, such as tax simulation, are included to help the user understand these concepts.

The Universal Social Charge (USC) is a tax on your income. It is charged on your gross income before any pension contributions or PRSI. You cannot use tax credits or tax relief (except for certain capital allowances) to reduce the amount you must pay. Find out more in our document on the Universal Social Charge.

Figure 4: Citizens Information text

If one person is earning €50,000 and their spouse or civil partner is earning €33,000:

The standard rate cut off point for the couple is €49,000 plus €31,000. The increase in the standard rate band is not transferable between spouses or civil partners, so the first spouse or civil partner's tax bands would be calculated as €49,000 @ 20% = €9,800 and €1,000 @ 40% = €400. The second spouse or civil partner's tax bands would be calculated as €31,000 @ 20% = €6,200 and €2,000 @ 40% = €800.

Figure 5: Citizens Information numbers and calculations

Standard rate cut-off points

	2023		2022		2021, 2020 and 2019	
	20%	40%	20%	40%	20%	40%
Single person	€40,000	Balance	€36,800	Balance	€35,300	Balance
Married couple/civil partners, one income	€49,000	Balance	€45,800	Balance	€44,300	Balance

Figure 6: Citizens Information table

5.3 Data Pre-processing

A Python library specifically designed to handle the loading of unstructured documents was used to load the Citizens Information, which is in HTML format[8]. To divide the large Citizens Information text into manageable chunks, the LangChain tool utilizes the Recursive Character Text Splitter method. With this method, the user can specify a list of characters for splitting the text. Chunks of 500 tokens are created based on GPT-4 model constraints. The overlap between chunks ensures context is retained. Considering the token-based input limitations of the model, the length function employed is a token counter. The metadata includes the starting index of each chunk within the original document for tracking and referencing purposes. As a

[8] Unstructured library documentation: https://pypi.org/project/unstructured/

result of this methodical splitting process, large text inputs are effectively processed while maintaining semantic coherence[9].

A foundation for semantic search is created by using the OpenAI Embedding API to generate language embeddings from processed chunks of Citizens Information text. These embeddings serve as high-dimensional vector representations of the text, encapsulating their semantic and syntactic content.

In the Pinecone vector database, embeddings are uploaded once created. Pinecone's high scalability makes it possible to store and index vector embeddings, regardless of their number. Additionally, Pinecone provides ultra-low latency search, which is critical for handling large text corpora. Through the OpenAI Embedding API, the question is passed through the semantic search process. Pinecone receives this query and generates a vector embedding for it. Pinecone returns documents semantically connected to the question. Despite the fact that the documents do not share any explicit keywords with the query, these results demonstrate the effectiveness of the semantic approach.

By integrating vector embeddings with semantic search, the methodology provides the foundation for building applications such as question-answering systems and chatbots, which are the focus of this study. This method allows the system to navigate through the document, understand user questions more effectively, and provide accurate, contextually appropriate answers.

5.4 Deployment and Testing

The chatbot interacts with customers during deployment and testing. The Twilio API for WhatsApp provides this integration. Twilio enables the exchange of messages between the user and the chatbot. Using Ngrok, local servers can be exposed to the Internet. To ensure the effective interaction between the chatbot server and Twilio API, the chatbot server runs on localhost and is securely connected to the internet via ngrok.

The result is an efficient, and user-friendly, information access and interaction platform for citizens provided by OpenAI's GPT-4, Pinecone, Twilio, and ngrok technologies.

To answer the user's query effectively, the chatbot must be able to navigate the information embedded in 25577 vectors. There is a possibility that the correct response could span multiple segments and webpages. To allow the user to verify the information provided, the chatbot provides the corresponding URLs along with the answer.

Figure 7a shows ChatGPT answering "What is USC?". The answer can be considered correct, however, it does not match the user’s intent. Figure 7b shows ChatGPT provides a more relevant answer when provided with a more detailed question. However, because it is restricted to the training data accumulated, it still tends to be general. When asked about 2023, as illustrated in Figure 7c, ChatGPT replies that the cut-off is September 2021, therefore information cannot be provided. In Figure 7d, responses to the posed questions are presented in a table.

[9] LangChain Text Splitters: https://python.langchain.com/en/latest/modules/indexes/text_splitters.html

The same experiment was performed using this project chatbot, as shown in Figure 8. In response to the question "What is USC?" the model successfully finds the relevant information on three different pages and provides the source to the user for further verification if needed. The chatbot provides the relevant information about 2023 rates. The USC rate table contains additional information about self-employment, which was not found in the initial response. The chatbot allows the user to ask a follow-up question: "I am self-employed, my income is €120,000, what is my USC rate?". An 11% rate applies because €120,000 is higher than €100,000. It is noteworthy that this detail is not explicitly presented elsewhere. This confirms that the model determines that €120,000 exceeds €100,000, and so the 11% rate is applicable. This demonstrates the model's reasoning capacity.

It was noted in the introductory section that linguistic barriers may block the user's ability to interpret information. The LLM is proficient at content translation. Figure 9 shows users can request responses in their preferred language. Multilingual functionality increases tool accessibility to a worldwide audience.

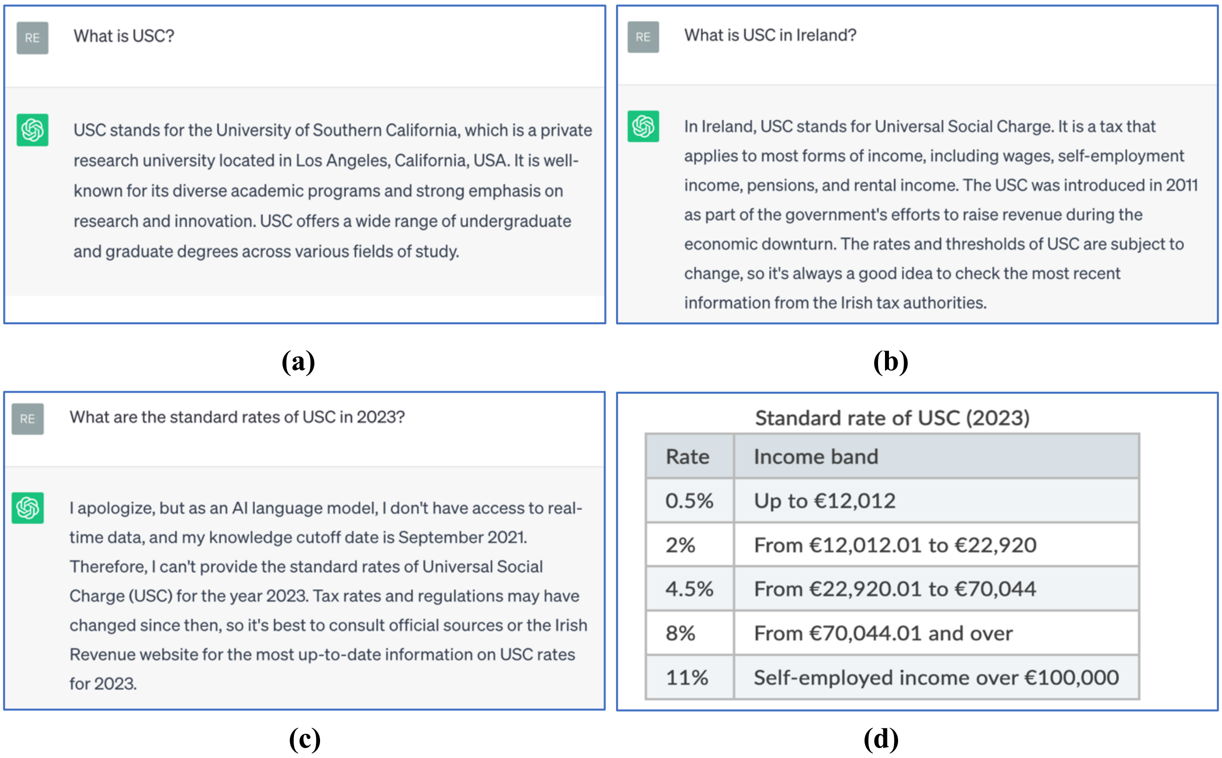

Figure 7: Chat GPT (a) Simple Question (b) Specific Question (c) Current Affairs (d) Citizens Information Website USC rate

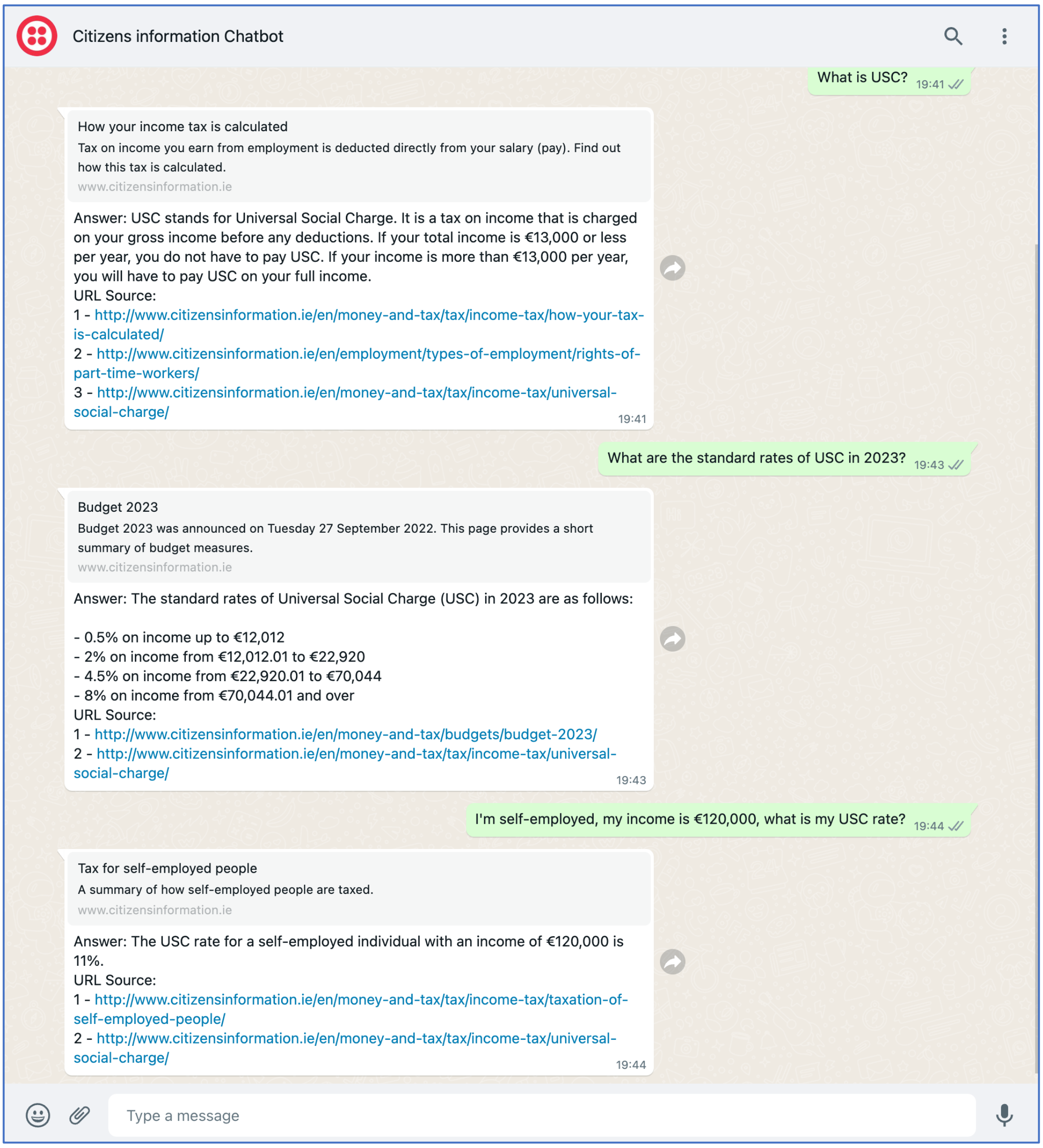

Figure 8: Citizens Information Chatbot

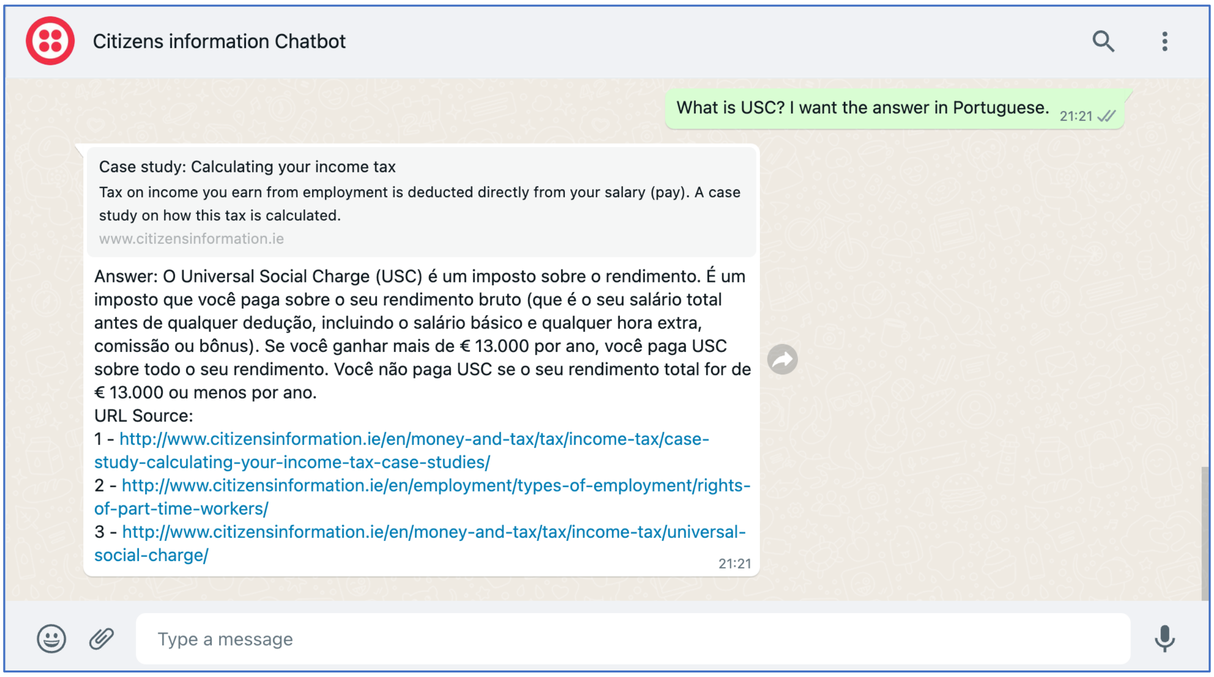

Figure 9: Citizens Information Chatbot Portuguese Answer

It was verified during the implementation process that the chatbot could overcome ChatGPT hallucinations. It was also demonstrated that the chatbot can retrieve relevant information and providing the source to the user. Moreover, the chatbot is equipped with reasoning capabilities that enable it to take decisions and perform calculations in addition to simple questions.

Due to token limitations, it is not possible to include all 1567 pages of Citizen Information in the LLM prompt. Ideally, if these restrictions did not exist, we could incorporate all this information into subsequent querying. Only relevant information from these pages must be included in the prompt. Table 1 illustrates the prompt format sent to the LLM after the prompt engineering process.

Table 1 – Prompting Formatting Example

Question	What is USC?
Prompt after formatting	System: Use the following pieces of context to answer the users question. If you don't know the answer, just say that you don't know, don't try to make up an answer. ---------------- Home Money and Tax Tax Income tax > Universal Social Charge (USC) Universal Social Charge (USC) Introduction Rules Rates Administration of the Universal Social Charge

	Where to apply Introduction The Universal Social Charge (USC) is a tax on income. USC The Universal Social Charge (USC) is tax you pay on gross income (this is your total pay before any money is deducted, including your basic salary and any overtime, commission or bonus). If you earn more than €13,000 a year, you pay USC on your full income. You do not pay any USC if your total income is €13,000 or less a year. Do I have a right to be paid overtime? You are entitled to overtime pay if your full-time co-workers are paid overtime. The Universal Social Charge (USC) is a tax on your income. It is charged on your gross income before any pension contributions or PRSI. You cannot use tax credits or tax relief (except for certain capital allowances) to reduce the amount you must pay. Find out more in our document on the Universal Social Charge. Income that is assessed for tax Rules Income liable for the USC The Universal Social Charge is tax you pay on gross income, including notional pay (notional pay is a non-cash benefit, such as benefit-in-kind), after any relief for certain capital allowances. The USC is payable on pension contributions. If your total income is €13,000 or less per year, you do not pay any USC. If it is more than €13,000 per year, you pay USC on your full income. Human: What is USC?

Semantic search retrieves relevant information from three different pages. This information is then compiled into the final prompt format and sent to the LLM along with the initial instructions and the query. There is no need to pre-process the information, even if it contains repetition, since the LLM handles such scenarios.

Table 2 illustrates the chain of thought, reasoning, and actions described in the literature review. For the bot to perform more advanced tasks, such as calculations, this is critical. In this instance, the question asked was, "*I'm single, I live in Ireland and my income is €75,000. How much is my total income tax?*" Prior to providing the user with the final answer, the Chatbot follows multiple steps outlined in Table 2. Consequently, the user receives only the conclusive answer, which states, *"Based on the calculations, if you're single, living in Ireland, and your income is €75,000, your total income tax would be €22,245.28. Therefore, your net income after tax would be €52,754.72."*

Table 2 – Prompt Engineering

~~~
> Entering new AgentExecutor chain...
```json
{
  "action": "Calculation",
  "action_input": "Calculate income tax for a single person living in Ireland with an income of €75,000"
}
```
> Entering new RetrievalQA chain...
~~~

```
Number of requested results 4 is greater than number of elements in index 2, updating n_results = 2
> Finished chain.
Observation: To calculate the income tax for a single person living in Ireland with an income of €75,000, we need to follow the same steps as in the example provided.
First, calculate the standard and higher rate tax:
Standard rate: The first part of the income, up to €40,000, is taxed at 20%.
€40,000 * 20% = €8,000
Higher rate: The amount that exceeds the standard rate is taxed at 40%.
(€75,000 - €40,000) * 40% = €14,000

Next, calculate the USC:
0.5% on the first €12,012:
€12,012 * 0.5% = €60.06
2% on the amount from €12,012.01 to €22,920:
(€22,920 - €12,012) * 2% = €218.16
4.5% on the amount from €22,920.01 to €70,044:
(€70,044 - €22,920) * 4.5% = €2,120.58
8% on the amount from €70,044.01 to €75,000:
(€75,000 - €70,044) * 8% = €396.48

Total USC = €60.06 + €218.16 + €2,120.58 + €396.48 = €2,795.28

Next, calculate the PRSI:
PRSI is 4% of the total income
€75,000 * 4% = €3,000

Finally, subtract the total tax credit from the sum of the standard rate, higher rate, USC, and PRSI:
Total tax credit = €3550

Total income tax = standard rate + higher rate + USC + PRSI - Total tax credit
Total income tax on €75,000 = €8,000 + €14,000 + €2,795.28 + €3,000 - €3550 = €22,245.28

So, if you earn €75,000, your net income after tax would be €75,000 - €22,245.28 = €52,754.72.
Thought:```json
{
   "action": "Final Answer",
   "action_input": "Based on the calculations, if you're single, living in Ireland, and your income is €75,000, your total income tax would be €22,245.28. Therefore, your net income after tax would be €52,754.72."
}
```
> Finished chain.
```

# 6 Evaluation

To create an evaluation dataset, the project utilized LLM to generate questions and answers. This automated approach significantly reduced human involvement, ensuring diverse and unbiased queries and responses. Citizens Information website text chunks were inserted into the LLM. It then generated questions and answers from the chunks. Question-answer pairs were created covering a variety of topics and contexts on the website.
```

6.1 Evaluation Dataset

Dataset creation is a complex task because defining the optimal size of a dataset to ensure model efficiency is uncertain. The project strategy involved engaging with the Citizens Information website, which contains 1568 pages. As a result of dividing these pages into sections, 25577 vectors were generated. Random function in Python was used to select 15 pages from the website. 114 distinct questions were produced by segmenting these pages into chunks and embedding them. Each of these sections was inserted into the LLM system, which automatically generated question-answer pairs. In this manner, the need for human intervention was significantly reduced and an unbiased result was achieved. The result is a diverse dataset comprising 114 questions and answers. It contains a random sample of the website as illustrated in Table 3.

Table 3 - Question Answer Dataset with Scores

Question	Answer
What are the requirements for qualifying for Jobseeker's Benefit?	To qualify for Jobseeker's Benefit (JB) you must be aged under 66 and have social insurance (PRSI) contributions.
What is the duty of care that teachers must follow?	Teachers must follow good standards and approved practice. Risky classes and activities require a greater degree of supervision. If goggles, protective clothing, etc., are supplied for certain activities, it is the teacher's duty to ensure that they are worn.

6.2 Results

With the dataset completed, it can be used for querying the chatbot. It is then possible to compare the predicted answers generated by the chatbot with the actual or real answers. Table 4 illustrates this process.

Table 4 – Predicted Answer Compared with Real Answer and Scores

Question	Real Answer	Predict Answer	Score
What are the requirements for qualifying for Jobseeker's Benefit?	To qualify for Jobseeker's Benefit (JB) you must be aged under 66 and have social insurance (PRSI) contributions.	To qualify for Jobseeker's Benefit (JB), you must be aged under 66, unemployed (fully unemployed or unemployed for at least 4 days out of seven), capable of work, available for work and genuinely seeking work, and have enough social insurance (PRSI) contributions	rouge=0.6129, bert_score=0.7913, uni_eval=0.9140, bleu=0.2312
What is the duty of care that teachers must follow?	Teachers must follow good standards and approved practice. Risky classes and activities require a greater degree of supervision. If goggles, protective clothing, etc., are	Teachers must take reasonable care to ensure that their students do not meet with foreseeable injury. They have a duty to protect the children against foreseeable risks of personal injury or harm, to the standard of a reasonably careful parent, and	rouge=0.2124, bert_score=0.6456, uni_eval=0.9298, bleu=0.0839

	supplied for certain activities, it is the teacher's duty to ensure that they are worn.	must have an effective system of supervision in operation. They must also follow good standards and approved practice, and ensure that any necessary protective equipment is worn during risky classes and activities.	
What is the phone number to contact the Citizens Information Phone Service?	0818 07 4000	The phone number to contact the Citizens Information Phone Service is 0818 07 4000.	rouge=0.3529, bert_score=0.5162, uni_eval=0.9071, bleu=0.0000
What type of training is available for people with disabilities?	Training for people with disabilities is available through apprenticeships, in-company training, and specialist training providers. Specialist training includes adapted equipment, a more individual approach, and longer training sessions.	People with disabilities can access specific skills training, traineeships, apprenticeships, and in-company training.	rouge=0.2727, bert_score=0.6471, uni_eval=0.9557, bleu=0.0000

The results are evaluated using ROUGE, BLEU, BERTScore and UNIEVAL. Each question has its own score. Table 3 and Figure 10 show the results of all questions.

Table 3 – Results

Metric	Mean	Standard Deviation
BLEU	0.2287	0.2889
ROUGE	0.5144	0.3122
BERTScore	0.7140	0.1858
UNIEVAL	0.7754	0.2343

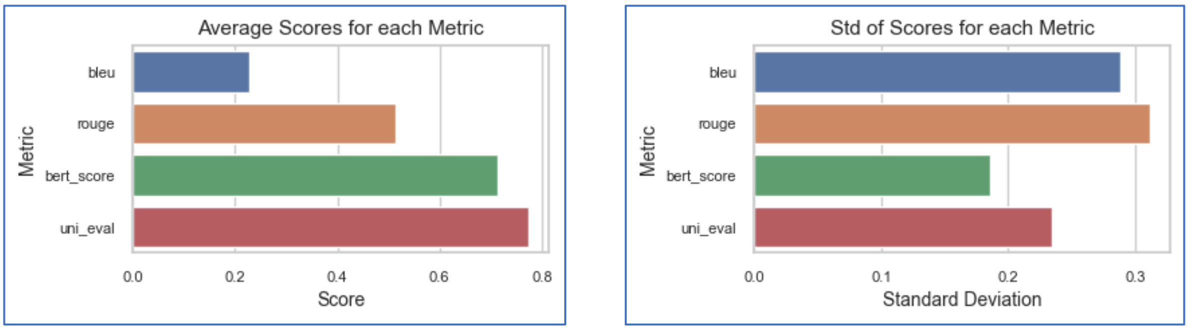

Figure 10: Average and Standard Deviation of Each Metric

For he examples in Table 4, a human evaluator would deem all the answers correct. It is pertinent to note, however, that the scoring system is not binary. BLEU, which received the lowest score, relies on n-gram models. The bot provided a complete and accurate response to the phone number inquiry, however BLEU scored it zero, highlighting the inherent limitations of its evaluation system. In the literature, BERTScore and UNIEVAL were presented as having the highest correlation with human judgments.

The dataset contained one hundred pairs after eliminating duplicate queries and questions with high levels of ambiguity. A human reviewed and found 93% correct answers.

7 Discussion

Using LLM, this research project focuses on the development of an AI-powered chatbot designed to enhance the accessibility and comprehension of information on the Citizens Information website for users interacting via WhatsApp. To ensure frictionless communication between users and the chatbot, the project sought to overcome language and comprehension barriers. There are numerous QA benchmark datasets available, such as the Stanford Question Answering Dataset SQuAD2.0 (Rajpurkar, Jia and Liang, 2018). Using these datasets, one can evaluate the general performance of a model, for example, by comparing GPT-3 and GPT-4. On a benchmark dataset, a model may perform well, but poorly on a specific dataset. Therefore, it is worthwhile to test the model on production data. In addition to providing information about the model's performance, the achieved scores provide a broader context beyond its use with a particular benchmark dataset. The chatbot was tested with a variety of questions. The chatbot was evaluated using ROUGE, BLEU, BERTScore, and UNIEVAL. The mean scores and standard deviations calculated for each metric gave a holistic assessment of the chatbot's performance. As evidenced by the metrics scores, the chatbot provided relevant, accurate, and contextually appropriate responses. This demonstrated the successful integration and application of the GPT-4 model. Despite positive results, there are some limitations and challenges. As highlighted in the referenced papers within this thesis, evaluation metrics, particularly n-gram-based ones like BLEU and ROUGE, have their limitations. While they serve as important benchmarks for comparing different models or solutions, they cannot solely determine the correctness of an answer. Metrics with a stronger correlation to human evaluation, such as BERTScore and UNIEVAL, could potentially be tested using a threshold to determine the correctness of information. As language models evolve, so too does the evaluation of AI using LLMs. Despite the implementation and testing of some of these novel approaches, no published work using them was found, hence their exclusion from this project. In this project, a small dataset was created to supplement automatic metrics with human evaluation. However, having a human annotator is not feasible for solutions intended for production deployment. Some systems use a thumbs up/down feature or ask the question, "did you find this information useful?" to gather customer feedback and use this information for model evaluation and improvement. A similar approach could be employed for the Citizens Information chatbot. Given that the website covers critical domains such as finance and legal matters, incorrect information could mislead users into making erroneous decisions with potential legal implications. To mitigate these risks, the chatbot provides the URL source of the information for verification. In such cases, it is crucial to have a human-in-the-loop for manual reviews during testing stages and for subsets after deployment. It is also vital to closely monitor the development of new AI evaluation tools. A disclaimer could be added to each message, highlighting the associated risks. It is worth noting that all customer support, whether human or chatbot, carries the risk of providing incorrect information. In the chatbot implemented for this project, where information retrieval is employed, the bot is limited to the facts present in the written content. Humans, on the other hand, can be biased, rely on their memory, and consequently provide incorrect information. The research also aimed to overcome language barriers using the LLM's translation capabilities. However, this study did not check the accuracy of these translations, which could affect the chatbot's effectiveness for non-English speakers. The cost of API calls to OpenAI was ignored. For the implementation of this project, the costs were relatively low. However, large scale deployments in production could rapidly escalate these costs. In addition, the study did not compare different LLMs, but rather focused exclusively on GPT-4. The chatbot setup, which utilizes ngrok, is not suitable for production environments. A more robust and scalable setup would be required for real-world applications, signalling a critical transition from experimental settings to scalable deployment environments. A tailored template containing step-by-step income tax calculations was created to support the LLM when performing this task. Due to its chained thought processes, the LLM can perform these calculations when prompted accordingly. Therefore, there is an opportunity to map all the calculations and simulations on the website and create prompt templates.

8 Conclusion and Future Work

This study assessed the effectiveness of an artificial intelligence chatbot powered by an LLM providing domain-specific customer support via WhatsApp. As a result of the development and testing of this chatbot, it has been demonstrated that LLM-powered chatbots can improve public information accessibility and comprehension. This study also uncovered some opportunities for future work. By investigating cost-optimization strategies, conducting comparative analyses of different LLMs, establishing a more robust and scalable setup, refining chatbot response generation, and investigating translation accuracy. There is also potential for future work on leveraging LLM capabilities to handle complex calculations and simulations. This research could be extended by mapping and creating prompt templates for these tasks. The LLM-powered chatbot demonstrates considerable commercial potential across a wide range of sectors, including businesses and government agencies wishing to enhance customer service or improve access to public information. It is however critical to consider the cost implications and the need for a robust, scalable setup prior to deployment. This study contributes to the field of LLM-powered chatbots. As a result of the insights gained and limitations identified, a compelling roadmap for future research and innovation has been established.

References

Ahmad, A. *et al.* (2019) ‘ReQA: An Evaluation for End-to-End Answer Retrieval Models’, in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 137–146. Available at: https://doi.org/10.18653/v1/D19-5819.

Brown, T.B. *et al.* (2020) ‘Language Models are Few-Shot Learners’. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14165.

Bulian, J. *et al.* (2022) ‘Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation’, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 291–305. Available at: https://aclanthology.org/2022.emnlp-main.20 (Accessed: 13 July 2023).

Chen, A. *et al.* (2019) ‘Evaluating Question Answering Evaluation’, in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, pp. 119–124. Available at: https://doi.org/10.18653/v1/D19-5817.

Cheng, H.-T. *et al.* (2016) ‘Wide & Deep Learning for Recommender Systems’. arXiv. Available at: https://doi.org/10.48550/arXiv.1606.07792.

Chhabria, K. and Damle, M. (2022) ‘Evolving Journey of Chatbots: Insights into Business Decisions’, in *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*. *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, pp. 102–107. Available at: https://doi.org/10.1109/IIHC55949.2022.10060780.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.

Farea, A. *et al.* (2022) ‘Evaluation of Question Answering Systems: Complexity of judging a natural language’. arXiv. Available at: http://arxiv.org/abs/2209.12617 (Accessed: 14 July 2023).

Gong, H. (2022) ‘Implements of Transformer in NLP and DKT’, in *2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*. *2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pp. 805–807. Available at: https://doi.org/10.1109/AIAM57466.2022.00163.

Gonsalves, X. and Deshmukh, S. (2023) ‘Designing an Interactive Chatbot for Educational Assistance using Rasa Framework’, in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 68–74. Available at: https://doi.org/10.1109/IDCIoT56793.2023.10053457.

Gusak, J. *et al.* (2022) ‘Survey on Large Scale Neural Network Training’. arXiv. Available at: http://arxiv.org/abs/2202.10435 (Accessed: 3 June 2023).

Johnson, J., Douze, M. and Jégou, H. (2017) ‘Billion-scale similarity search with GPUs’. arXiv. Available at: https://doi.org/10.48550/arXiv.1702.08734.

Kamalloo, E. *et al.* (2023) ‘Evaluating Open-Domain Question Answering in the Era of Large Language Models’. arXiv. Available at: https://doi.org/10.48550/arXiv.2305.06984.

Kojima, T. *et al.* (2023) ‘Large Language Models are Zero-Shot Reasoners’. arXiv. Available at: http://arxiv.org/abs/2205.11916 (Accessed: 3 June 2023).

Liu, N.F., Zhang, T. and Liang, P. (2023) ‘Evaluating Verifiability in Generative Search Engines’. arXiv. Available at: http://arxiv.org/abs/2304.09848 (Accessed: 10 July 2023).

MINAEE, S. *et al.* (2022) ‘Deep Learning-based Text Classification: A Comprehensive Review.’, *ACM Computing Surveys*, 54(3), pp. 1–40. Available at: https://doi.org/10.1145/3439726.

Neelakantan, A. *et al.* (2022) ‘Text and Code Embeddings by Contrastive Pre-Training’. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.10005.

Qiu, X. *et al.* (2020) ‘Pre-trained Models for Natural Language Processing: A Survey’, *Science China Technological Sciences*, 63(10), pp. 1872–1897. Available at: https://doi.org/10.1007/s11431-020-1647-3.

Raffel, C. *et al.* (2020) ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’, *Journal of Machine Learning Research*, 21(140), pp. 1–67.

Rajpurkar, P., Jia, R. and Liang, P. (2018) ‘Know What You Don’t Know: Unanswerable Questions for SQuAD’. arXiv. Available at: http://arxiv.org/abs/1806.03822 (Accessed: 24 July 2023).

Rane, A. *et al.* (2022) ‘AI driven Chatbot and its Evolution’, in *2022 5th International Conference on Advances in Science and Technology (ICAST)*. *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pp. 170–173. Available at: https://doi.org/10.1109/ICAST55766.2022.10039515.

Vajjala, S. *et al.* (2020) *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O’Reilly Media, Inc.

Vaswani, A. *et al.* (2017) ‘Attention Is All You Need’. arXiv. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 3 June 2023).

Wei, J. *et al.* (2023) ‘Chain-of-Thought Prompting Elicits Reasoning in Large Language Models’. arXiv. Available at: http://arxiv.org/abs/2201.11903 (Accessed: 3 June 2023).

Weizenbaum, J. (1966) ‘ELIZA—a computer program for the study of natural language communication between man and machine’, *Communications of the ACM*, 9(1), pp. 36–45. Available at: https://doi.org/10.1145/365153.365168.

Wu, T. *et al.* (2023) ‘A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development’, *IEEE/CAA Journal of Automatica Sinica*, 10(5), pp. 1122–1136. Available at: https://doi.org/10.1109/JAS.2023.123618.

Yao, S. *et al.* (2023) ‘ReAct: Synergizing Reasoning and Acting in Language Models’. arXiv. Available at: http://arxiv.org/abs/2210.03629 (Accessed: 3 June 2023).

Yin, W. *et al.* (2017) 'Comparative Study of CNN and RNN for Natural Language Processing'. arXiv. Available at: http://arxiv.org/abs/1702.01923 (Accessed: 4 June 2023).

Zhai, C. (2007) 'Statistical Language Models for Information Retrieval A Critical Review', *Foundations and Trends® in Information Retrieval*, 2(3), pp. 137–213. Available at: https://doi.org/10.1561/1500000008.

Zhang, T. *et al.* (2020) 'BERTScore: Evaluating Text Generation with BERT'. arXiv. Available at: http://arxiv.org/abs/1904.09675 (Accessed: 15 July 2023).

Zhong, M. *et al.* (2022) 'Towards a Unified Multi-Dimensional Evaluator for Text Generation'. arXiv. Available at: http://arxiv.org/abs/2210.07197 (Accessed: 15 July 2023).