UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

RAG programming on LLMs to improve the quality of life of elderly people.

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Yaser Haj, Diana

Tutor: Monserrat Aranda, Carlos

Cotutor: Ferri Ramírez, César

Experimental director: Garigliotti, Dario

ACADEMIC YEAR: 2023/2024

To my family, with a special dedication to my mother, who has been my guiding star and my main pillar. Her endless support, encouragement and love have brought me to where I am today. I am also very thankful to my brother for his constant companionship and support. To my beloved grandmother, whose memory and inspiration have been the driving force behind this project.

I would like to express my gratitude to my tutors, Carlos Monserrat Aranda and César Ferri Ramírez, for their invaluable guidance and support throughout this project. I am also grateful to the *Department of Applied Artificial Intelligence for Public Administration (Cátedra de Inteligencia Artificial aplicada a la Administración Pública)* for their collaboration and the resources they provided. This department is funded by the Council of Innovation, Industry, Commerce, and Tourism of the Valencian Government (GVA). Special thanks to my classmates who always supported me through out the degree.

Resumen

Esta tesis presenta el desarrollo de un asistente virtual personalizado dirigido a las necesidades de los usuarios mayores, abordando brechas críticas en la tecnología existente, como la personalización, escalabilidad, comprensión contextual, transparencia y ética. El asistente integra la Generación Aumentada por Recuperación (RAG) con Modelos de Lenguaje a Gran Escala (LLMs) y la infraestructura LangChain para proporcionar respuestas dinámicas, conscientes del contexto y altamente personalizadas. El diseño modular asegura la escalabilidad, permitiendo que el sistema evolucione y se adapte a medida que estén disponibles nuevas funciones.

Se empleó una metodología integral, comenzando con una revisión de la literatura para identificar las brechas existentes, seguida por la adquisición de datos mediante web scraping e integración de APIs. Se realizaron tareas de preprocesamiento y análisis de datos para optimizar el sistema para la recuperación eficaz durante las interacciones. En lugar de un entrenamiento tradicional de modelos, se centró en la construcción e integración del sistema RAG, que utiliza dinámicamente información relevante de bases de conocimiento preexistentes.

La evaluación del sistema incluyó métodos cualitativos y cuantitativos, utilizando métricas como Similitud de Coseno, BLEU, ROUGE y Similitud de Sentence-BERT en el caso del análisis cualitativo donde Sentence-BERT demostró ser la mejor opción, destacándose por su capacidad para capturar similitudes semánticas de manera precisa. Esta métrica fue usada para hacer la evaluación cuantitativa y los resultados indicaron que el modelo gpt-3.5-turbo fue, sorprendentemente, más efectivo y confiable que el modelo gpt-4.0 para esta aplicación, proporcionando alta precisión y consistencia.

El proyecto logró desarrollar un asistente multifuncional que incluye un módulo de asesoramiento médico confiable y ofrece recomendaciones personalizadas. La escalabilidad del sistema asegura su viabilidad a largo plazo, y su enfoque en la transparencia y consideraciones éticas mejora la confianza del usuario. Los hallazgos contribuyen al campo más amplio de la tecnología de atención a personas mayores, sentando una base para futuros avances.

Palabras clave: Modelos de Lenguaje Grande (LLM); Generación aumentada por pecuperación (RAG); Personas mayores; Calidad de vida; Interacción humano-computadora; Recomendación personalizada; Tecnologías asistivas; Inteligencia Artificial (IA); Recuperación de información.

Abstract

This thesis presents the development of a personalized virtual assistant tailored to the needs of elderly users, addressing critical gaps in existing technology, such as personalization, scalability, contextual understanding, transparency, and ethics. The assistant integrates Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) and LangChain infrastructure to provide dynamic, context-aware responses and highly personalized interactions. The modular design ensures scalability, allowing the system to evolve and adapt as new features become available.

A comprehensive methodology was employed, starting with a literature review to identify existing gaps, followed by data acquisition through web scraping and API integration. Data pre-processing and analysis were conducted to optimize the system for effective retrieval during interactions. Instead of traditional model training, the focus was on building and integrating the RAG system, which dynamically utilizes relevant information from pre-existing knowledge bases.

The evaluation of the system involved both qualitative and quantitative methods, using metrics such as Cosine Similarity, BLEU, ROUGE, and Sentence-BERT Similarity in the case of qualitative analysis, where Sentence-BERT proved to be the best option, standing out for its ability to capture semantic similarities accurately. This metric was used for the quantitative evaluation, and the results surprisingly indicated that the gpt-3.5-turbo model was more effective and reliable than the gpt-4.0 model for this application, providing high accuracy and consistency.

The project successfully developed a multifunctional assistant that includes a reliable medical advice module and offers personalized recommendations. The system's scalability ensures long-term viability, and its focus on transparency and ethical considerations enhances user trust. The findings contribute to the broader field of elderly care technology, setting a foundation for future advancements.

Key words: Large Language Models (LLM); Retrieval-augmented generation (RAG); Elderly People; Quality of Life; Human-Computer Interaction (HCI); Personalized recommendation; Assistive technologies; Artificial Intelligence (AI); Information Retrieval (IR).

Contents

Contents 7
List of Figures 9
List of Tables 9

1 Introduction 11
1.1 Challenges faced by the elderly . . . 12
1.2 Importance of technology for the elderly . . . 13
1.3 Motivation . . . 14
1.4 Objectives . . . 15
1.5 Expected impact . . . 15
1.6 Methodology . . . 16
1.7 Structure of the report . . . 17
2 State of the art 19
2.0.1 Retrieval-Augmented Generation (RAG) in Large Language Models . . . 21
2.1 Critique of the state of the art . . . 24
2.2 Proposal . . . 24
3 Analysis of the problem 27
3.1 Proposed solution . . . 28
3.1.1 Modules . . . 28
3.1.2 Medical module . . . 30
3.1.3 Personal module . . . 32
3.1.4 Reminders module . . . 34
3.1.5 Weather module . . . 35
3.1.6 Interface . . . 35
3.1.7 Phases of development and materials . . . 38
4 Preparation and understanding of the data 41
4.1 Data quality and transformation . . . 44
5 Knowledge extraction and model evaluation 47
5.1 Introduction . . . 47
5.1.1 Purpose and Scope . . . 47
5.1.2 Overview of Models . . . 47
5.2 Model evaluation . . . 48
5.2.1 Qualitative Evaluation . . . 48
5.2.2 Quantitative Evaluation . . . 49
5.3 Comparison with state of the art virtual assitants (Alexa vs Siri vs Elf) . . . 57
6 Conclusions 63
6.1 Legacy . . . 64

6.2 Relationship of the work carried out with the studies completed . . 65
6.3 Future work 66
Bibliography **69**
A Questionnaire for information collection in order to personalize the system **73**
B Extract of NHS JSON and Kaggle JSON **77**
C Human evaluation through Google Forms **87**

List of Figures

1.1 Population by age group, World [1] 11

3.1 Modules of the system 29
3.2 Functioning of the system 30
3.3 Medical module 32
3.4 Personal module 34
3.5 Elf Virtual Assitant interface in Streamlit 37

4.1 Most common words 41
4.2 Top 10 subcategories 42

5.1 Correlation of objective metrics with human mean evaluation . . . 50
5.2 Maximum Similarity per Model 51
5.3 Minimum Similarity per Model 52
5.4 Standard deviation per Model 53
5.5 Mean Similarity per Model 54
5.6 Mean Similarity vs Standard Deviation by Model 55
5.7 Model Performance Comparison 56
5.8 Elf Virtual Assitant interface in Streamlit 58
5.9 Cholesterol question example from ELF 60

C.1 Diabetes question evaluation example 87
C.2 Recipe question evaluation example 88
C.3 Coronary Heart Disease question evaluation example 89

List of Tables

4.1 Summary of Medical Conditions 43

C.1 Assessment of Sustainable Development Goals 91

CHAPTER 1
Introduction

The longevity trend is causing a significant demographic shift in the world since people are living longer than ever before. Due to the progress achieved in medicine and healthcare, improvements in general living conditions, better nutrition and healthier lifestyles, people enjoy life way beyond their sixties and seventies. Estimates show that the number of people aged 65 and over worldwide will double to 1.58 billion in 2050, growing from 808 million in 2023. As we can see in figure 1.1, it is the fastest growing age group and growth occurs globally, across all countries.

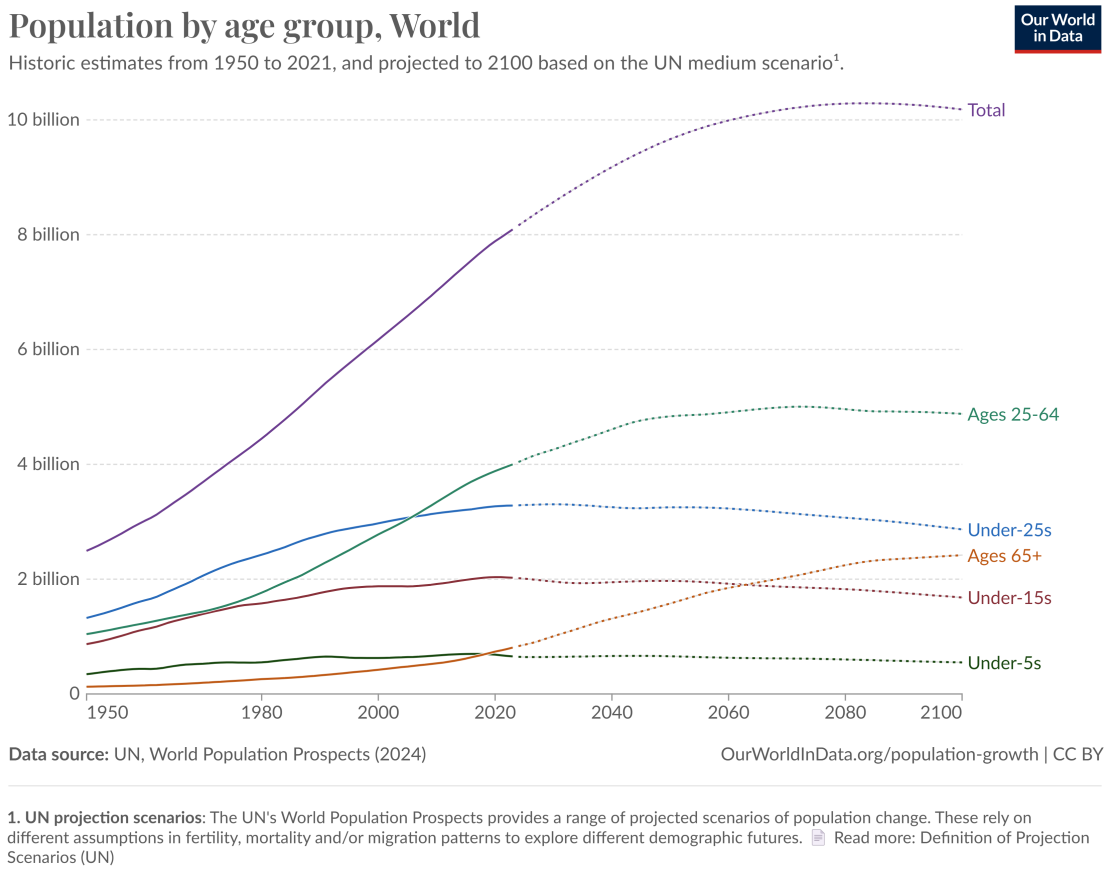


Figure 1.1: Population by age group, World [1]

It is predicted that by 2050, around 80% of the world's ageing population will be living in low- and middle- income countries [2] in contrast to current trends of significant growth in high-income nations located predominantly in Northern America and Europe. Northern Africa, Western Asia and sub-Saharan Africa are expected to experience the fastest growth over the next three decades. [3] With this rapid increase, unique challenges arise, as many of these countries may not have the necessary infrastructure, healthcare systems, or social services to adequately support their aging populations.

Spain will not remain behind and will also experience this demographic shift. By 2074, the senior population in Spain is estimated to rise to over 9.1 million women and 7.4 million men aged 65 and older. This represents an increase of almost seven million seniors compared to the 9.68 million registered in 2023. With 20 percent of its total population already aged 65 or more [4], Spain is considered an aging country, and it consistently ranks among the top 10 countries with the highest life expectancy. [5] The younger groups continue to lose weight across the population at large. At the beginning of the 20th century half the population was under 24; in 1999, half the population was under 36; and in 2050 half the population will be 50 or older (this will make Spain the country with the oldest population in the world, according to United Nations forecasts). [6]

The profound repercussions of this demographic shift are affecting healthcare systems, the dynamics of social interactions, and the existing economic structures on a global scale. This continuous rise in life expectancy highlights the necessity for effective strategies to address the challenges posed by an aging population. The focus lies in guaranteeing that this aging population can maintain a good quality of life by having access to necessary healthcare, fruitful and fulfilling social connections, and extensive support systems. These changes require a rethinking of how societies are structured and how resources are allocated to meet the needs of an aging world.

1.1 Challenges faced by the elderly

Age is not just a number. Aging brings inevitable changes in a person's life, affecting almost every aspect of it. One of the central issues faced by the elderly is the gradual decline in both physical and cognitive abilities. Common activities performed daily in the past can become heavy or even beyond the reach of a person's capacity including cooking, cleaning, shopping, or personal hygiene, among others. Over time, they suffer a reduction in muscle strength, balance, and coordination, leading to difficulties in mobility and an increased risk of falls. Chronic illnesses such as arthritis, osteoporosis, and cardiovascular diseases only add to these problems. On the other hand, cognitive deterioration including reduced memory and information processing or difficulties solving problems can cause the elderly person to lack the capacity to manage his personal affairs, including handling personal finances, appointments and medication schedules among others. Cognitive impairments such as dementia and Alzheimer's disease, which are more prevalent in older age, imply severe challenges to maintaining independence.

As a result of these changes carers, relatives and care institutions play a vital role due to the dependency of the elderly. This can result in a loss of independence and self-respect. In many cases it can lead to moving into assisted living facilities or nursing homes, causing a great emotional and financial strain on both the individuals and their loved ones.

Healthcare services are essential at any age, but they are even more indispensable at old age. Accessing these services can be quite challenging for many individuals, particularly when dealing with reduced mobility, requiring assistance from others. Most European countries for example have universal free healthcare for its citizens, but it is not the case in many low- and middle-income countries. Healthcare infrastructure is not adequately equipped to handle the increasing number of patients, reflected by insufficient medical facilities or a shortage of geriatric specialists, leaving the elders without the care they need.

Another common issue they struggle with is social isolation and loneliness due to changes in work status such as retirement, the loss of friends or a spouse or reduced mobility resulting in smaller social networks. Physical limitations make it harder for people in this age group to engage in social activities, further isolating them from family and community life. Social interaction is key to stimulating good mental health and its absence can have devastating effects. Loneliness and isolation have been linked to an increased risk of depression, anxiety, and cognitive decline. The stigma and shame surrounding mental health issues can prevent elderly individuals from reaching out for help, accentuating their feelings of isolation.

1.2 Importance of technology for the elderly

Technology has revolutionized every aspect of our lives and has become an indispensable tool to address the unique challenges faced by elders. Innovative solutions are a must to be able to support their independence, health, and social connections so that they can live a fruitful and prosperous elderhood.

At the lead of these technological advancements, we can find AI-driven virtual assistants. These virtual assistants can be utilized to support the independence of older adults by helping them manage daily tasks. For instance, they can set reminders or provide step-by-step guidance for routine activities like cooking or exercising, allowing the elderly to retain a level of autonomy that would otherwise be lost as they age. This provides a sense of tranquility for the family members.

Furthermore, technology can lessen social isolation. For instance, virtual assistants can facilitate communication with family and friends through messaging or video calls, helping users stay in touch.

As the proportion of elderly individuals in the global population continues to rise, the integration of AI into elderly care will likely become an essential component of societal strategies to ensure that older adults can age with dignity, health, and connection.

However, it is important to note that older adults often experience distrust towards digital technologies, particularly regarding privacy and confidence issues. A study by Knowles and Hanson has shown that many elderly individuals feel uneasy about how their personal data is used, likening their lives to "open books" vulnerable to unwanted scrutiny. Additionally, a lack of confidence in using new technologies can create anxiety, making older adults feel incompetent or "risky" when trying to engage with these tools. Addressing these concerns is crucial for ensuring that technological solutions are both effective and widely adopted by the aging population. [27]

While there are several tools and technologies currently available for elderly care, they often fall short of fully addressing the complex needs of the aging population.

1.3 Motivation

The motivation for this project arises from a deeply personal circumstance. Earlier this year, I lost my grandmother and it had a profound impact on me. Her passing was even harder to bear, knowing that she was living alone and was separated from my family and I by borders in a time of conflict. The sense of helplessness drove me to seek ways in which technology could bridge such gaps, especially for the elderly, who often face challenges in maintaining their independence. This project is a way for me to honor her memory by contributing to a field that could potentially improve the lives of others in similar situations.

On a technical level, my interest in this project is driven by a fascination with Large Language Models (LLMs) and the potential they hold for transforming various aspects of our lives. Last year, I initiated a project involving Retrieval-Augmented Generation (RAG), and it opened my eyes to the immense possibilities of combining AI with real-world applications. The idea of applying these advanced technologies to create solutions for elderly care provides a challenging and rewarding opportunity to deepen my knowledge in this area. I see this project as a chance to explore the cutting edge of AI while working on something that could have a real, positive impact on society.

From a professional standpoint, this project represents an opportunity to develop expertise in a rapidly growing field that is increasingly relevant in today's world. The aging population is a global phenomenon, and there is a critical need for innovative solutions that can support the elderly in maintaining their quality of life. By focusing on the integration of LLMs and RAG within elderly care, this project positions me at the intersection of AI and healthcare, two fields of personal interest. This experience will not only enhance my technical skills but also provide me with a strong foundation for a future career in AI-related roles.

In essence, this project is motivated by a combination of personal sorrow, technical curiosity, and professional ambition. It allows me to channel my grief into something meaningful while advancing my understanding of AI and contributing to a cause that could benefit many others.

1.4 Objectives

The primary objective of this project is to develop a virtual assistant for the elderly which will integrate multiple modules to address their diverse needs, including:

- Developing a comprehensive virtual assistant that seamlessly integrates various functionalities such as medical advice and information, personal recommendations, and daily reminders. The assistant will serve as a well-rounded support system that can cater to the complex and evolving needs of elderly individuals being able to maintain seamless conversations.
- Implementing a medical module capable of answering health-related questions and offering general health advice. This module will be designed to ensure that elderly users receive accurate medical information from reliable sources, contributing to better health management.
- Designing a system that feed on user data, allowing it to personalize suggestions related to everyday activities, such as meal planning, exercise routines, or social engagement. This individualized approach aims to adapt the assistant to the unique preferences and needs of each user.
- Creating a scalable architecture that allows for the easy integration of additional modules or functionalities in the future. This scalability will enable the system to evolve, incorporating new features as they become necessary or as new technologies emerge.
- Focusing on optimizing the retrieval-augmented generation (RAG) technique to reduce instances of system hallucinations. This involves refining the context provided to the system, ensuring that the virtual assistant delivers accurate and reliable information.
- Conducting a complete evaluation of different large-scale language models within a consistent context. The goal is to identify the most suitable model for the virtual assistant, balancing accuracy, responsiveness, and the ability to handle diverse queries.

By achieving these objectives, the project aims to deliver a robust and adaptable virtual assistant that significantly improves the daily lives of elderly users.

1.5 Expected impact

The virtual assistant developed through this project is expected to have a significant positive impact on several key areas of elderly care, directly benefiting elderly users, caregivers, and the broader community.

The primary beneficiaries of this system will be the elderly individuals who interact with the virtual assistant. By integrating multiple functionalities, the virtual assistant will empower users to maintain their independence for a longer

period. The system will be able to adapt to the individual preferences and needs of each user. Moreover, by providing continuous companionship and enabling easier access to healthcare information and daily life suggestions, the virtual assistant will help mitigate the feelings of isolation and loneliness that are prevalent among this age group.

Caregivers, whether professional or family members, will also benefit from using this virtual assistant. The system can answer questions about health conditions and help with daily activities, lightening the caregiving load and providing peace of mind. With the virtual assistant offering support, caregivers will have more time and energy to focus on the emotional and social well-being of their loved ones. Furthermore, the system's scalability allows for the addition of new modules over time.

At a broader level, the introduction of this virtual assistant aligns with contemporary efforts to address the challenges of an aging population, particularly in the context of the United Nations Sustainable Development Goals (SDGs). Specifically, the project directly contributes to SDG 3 (Good Health and Well-being) by promoting healthy aging and improving the accessibility and quality of care for the elderly. Additionally, by reducing the burden on caregivers, the project indirectly supports SDG 8 (Decent Work and Economic Growth) and SDG 10 (Reduced Inequalities), helping to create a more inclusive society where elderly individuals can continue to live with dignity and autonomy.

The virtual assistant developed in this project is expected to bring meaningful improvements by offering valuable support to caregivers and contributing to addressing broader societal challenges associated with aging.

1.6 Methodology

The project's methodology has been carefully designed to align with the objectives and the scope of the work, ensuring a systematic approach to achieving the desired outcomes. The steps outlined below are intended to guide the development process.

The first step involves a comprehensive literature review to understand the field's current state and identify gaps that this project could address. This review not only grounds the project in existing research but also helps to define the specific objectives that are realistic and achievable within the constraints of the project timeline and resources.

Following the literature review, the next step is data acquisition. Given the focus on creating a personalized virtual assistant for the elderly, there is an evident need to gather relevant data that will act as sources in the system. In particular, data will be extracted from publicly available information. Methods, such as web scraping and API integration, will be applied.

Once the data has been acquired, the next phase involves pre-processing and analyzing the data to prepare it for model evaluation. This includes steps such as data structuring and organizing the dataset to ensure it is suitable for integration with the Retrieval-Augmented Generation (RAG) system, which forms

the core of the conversational model. The data is structured to support the effective retrieval of relevant information during interactions, optimizing the system's performance.

Instead of traditional model training, the project focused on building the RAG system. This involved selecting appropriate frameworks to handle natural language processing tasks efficiently. The RAG system was designed to integrate pre-trained language models with a retrieval mechanism, allowing the model to access and use relevant information dynamically during conversations. This approach eliminates the need for extensive model training, as the system relies on existing knowledge bases and retrieval techniques.

The final step involves testing and validating the system to ensure it meets the established objectives. This includes both technical evaluations, such as assessing the accuracy and efficiency of the RAG system, and finally selecting the best model for integration. Once the optimal model is chosen, the next phase is to develop the user interface, which will serve as the primary point of interaction for users. This interface is designed to be intuitive and accessible, ensuring ease of use and seamless interaction with the conversational system.

By following this structured approach, the project aims to deliver a personalized conversational model that meets the technical requirements and provides meaningful and practical benefits to its intended users.

1.7 Structure of the report

The structure of this thesis is designed to guide the reader through a logical progression of ideas, from the introduction of the project to the final conclusions and future work. The document begins with the *Introduction* chapter, where the objectives, expected impact, methodology, and the overall structure of the report are outlined, setting the stage for the detailed exploration that follows. Next, the *State of the Art* chapter delves into the current landscape of virtual assistants for the ageing population and an overview of Retrieval-Augmented Generation (RAG) combined with Large Language Models, critically assessing existing approaches and laying the groundwork for the proposal presented later.

Following this, the *Analysis of the problem* chapter discusses the challenges the project aims to address, presenting the proposed solution and its various modules, including those focused on medical information, personal data, reminders, and weather updates. This chapter also describes the development phases, providing a roadmap of the project's implementation. The *Preparation and understanding of the data* chapter then takes the reader through the data acquisition, preprocessing, and analysis steps.

In the *Knowledge extraction and model evaluation* chapter, the thesis explores the evaluation and performance of the models employed, with a focus on both qualitative and quantitative measures.

Finally, the *Conclusions* chapter synthesizes the achievement of the objectives , discusses the legacy of the work, and reflects on how the work carried out relates to the studies completed. It also opens up possibilities for *Future Work*, suggest-

ing improvements and extensions that could be pursued in subsequent research. The thesis concludes with a *Bibliography* and *Appendix*, which provide additional resources and detailed information that supports the main text. The Appendix includes the questionnaire for collection of personal data of the elderly.

CHAPTER 2
State of the art

Virtual assistants, such as Alexa of Amazon and Siri of Apple, represent the forefront of artificial intelligence (AI) in consumer technology. These assistants utilize advanced natural language processing (NLP), generative AI models, and cloud computing to interact with users, perform tasks, and provide information in a seamless, conversational manner. The software behind these virtual assistants is constantly listening for a specific keyword to activate it. Once the keyword is detected, the user's voice is recorded and sent to a specialized server where it is processed and interpreted as a command. Depending on the nature of the command, the server provides the assistant with the necessary information to send back to the user, plays requested media, or executes tasks using connected services and devices. [7]

While each voice assistant has unique features, they share many common functionalities. They can provide information such as weather updates, facts from sources like Wikipedia or IMDb [8], set alarms, and manage to-do and shopping lists. Additionally, virtual assistants can send and read text messages and make phone calls. They are equipped to answer basic informational queries, set timers or reminders, and create calendar entries. Users can also control media playback from connected services such as Amazon Music, Google Play, iTunes, Netflix, and Spotify. Furthermore, these assistants are integrated with Internet-of-Things (IoT) devices, enabling control over smart home features such as thermostats, lights, alarms, locks, etc. Beyond these practical tasks, virtual assistants can even tell jokes and stories, adding to their interaction with users. [7]

Amazon Alexa was launched in 2014 and quickly became a leading virtual assistant due to its wide range of functionalities. Alexa is designed to respond to voice commands using a cloud-based voice service that is always listening for its wake word, typically "Alexa". Once activated, Alexa can perform a variety of tasks, including answering questions, for which the system collects information from multiple sources. In the background, this data is occasionally shared with various providers, such as WolframAlpha, IMDb, AccuWeather, Yelp, Wikipedia, and others, to produce accurate responses. [9] Alexa's functionality is further enhanced by "skills" which are third-party applications that users can enable to add specific capabilities. For example, Alexa can control home devices, order take-out food or play trivia. Alexa's voice responses are generated using text-to-speech (TTS) technology, which converts written text into spoken words, allowing it to respond in a human-like voice. [10]

Apple Siri was introduced in 2011 as one of the first virtual assistants embedded in a smartphone. Siri is integrated into Apple's ecosystem, available across devices like iPhones, iPads, Macs, and Apple Watches. Siri is designed to respond to voice commands initiated by the wake phrase "Hey Siri". Siri's functionalities are similar to Alexa's, but, has more limited capabilities when answering questions and it normally returns accessed information from the web. Siri is also capable of context-based understanding, which allows it to provide more accurate and relevant responses based on previous interactions and user preferences. For instance, Siri can suggest leaving earlier for an appointment if traffic conditions are unfavorable, demonstrating its ability to integrate contextual awareness into its functionalities. [11]

In recent years, numerous projects and products have emerged that seek to specifically address challenges faced by the elderly through the development of virtual assistants and social robots. These technologies are designed to support independent living, improve well-being, and combat loneliness among seniors.

One of the most prominent initiatives in this field is the SITIES project, an acronym for "Support System for the Elderly with Smart Devices." This research project is a collaboration between Gnomon Informatics SA and Aristotle University of Thessaloniki, aimed at developing a low-cost smart assistant tailored specifically to the needs of elderly users in Greece. The SITIES system is composed of two main components: a web platform (Cloud Platform SITIES) and a smart assistant device known as ELSA (Elderly-friendly Smart Assistant). The platform enables the development and management of services and applications, offering tools for both developers. These applications are designed to run on the ELSA device, providing a range of services that include emergency intervention, situational awareness, remote patient monitoring, social networking, and assistance with daily life activities.

The SITIES project is particularly noteworthy for its focus on creating a device that is both accessible and affordable, making it a viable option for elderly users across Greece. However, the project's regional focus on supporting only the Greek language and its relatively narrow scope of applications—primarily centered on basic assistance and monitoring—limit its broader applicability. [12]

Another important development in this field is ElliQ, a social robot developed by Intuition Robotics. ElliQ was launched in 2017 to address social isolation and loneliness among older adults. Unlike many other virtual assistants, ElliQ is designed to be proactive, empathetic, and emotionally engaging. The robot interacts with users by using environmental cues and a personalized AI program that adjusts to the user's behavior, preferences, and context. This way, ElliQ can build a trusting relationship with its users, which is important for maintaining consistent and meaningful engagement.

ElliQ offers a wide range of activities, including conversation, music, video calls, cognitive games, and well-being assessments, all designed to promote greater independence and social connectivity among elderly users. [13]

In addition to these projects, recent research has explored the use of large-scale language models, such as ChatGPT, as conversational companions for the elderly. A study conducted by Alessa and Al-Khalifa [14] investigated the potential of a ChatGPT-based system to provide companionship and reduce loneliness

among older adults. The study found that the system was capable of generating relevant and engaging responses, making it a promising tool for enhancing social interaction.

Another noteworthy development is the Charlie chatbot, which was introduced by Valtolina and Hu in 2021. [23] Charlie is intended to enhance the well-being of users by offering companionship and motivating them to remain active. The chatbot can assist with medication reminders, connect users with doctors and family members, and offer various forms of entertainment. Charlie also incorporates gamification and self-compassion strategies to promote mental health and prevent loneliness.

Existing systems reveal significant progress in the development of virtual assistants and social robots for elderly care. However, several gaps remain unaddressed and they will be discussed in the section *Critique of the state of the art*.

2.0.1. Retrieval-Augmented Generation (RAG) in Large Language Models

In recent years, Large Language Models (LLMs) have become powerful tools in natural language processing (NLP), demonstrating remarkable abilities to generate coherent and contextually relevant text across a wide range of tasks. These models, trained on extensive datasets, can store a vast amount of factual knowledge and have achieved state-of-the-art results in various applications. Despite these advancements, LLMs face significant challenges that researchers are actively working to address.

One major issue is the phenomenon of hallucinations, —where the LLMs may generate incorrect or fabricated information due to gaps in its training data. This can lead to the creation of imagined or inaccurate answers, undermining the reliability of the model. Another challenge is outdated knowledge; since LLMs are trained on data that becomes stale over time, they may lack awareness of recent developments or domain-specific information that emerged after their training period.

These challenges have prompted the research community to explore new approaches to enhance LLM capabilities, among which Retrieval-Augmented Generation (RAG) stands out as a notable innovation.

Retrieval-Augmented Generation (RAG) was introduced in 2020 through the paper titled "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" [24]. This technique aims to tackle the problems associated with outdated knowledge and hallucinations. RAG combines the strengths of pre-trained LLMs with an external retrieval mechanism, thereby enhancing the model's accuracy and factual consistency.

RAG systems are built around two key components: parametric memory and non-parametric memory.

Parametric memory refers to the internal knowledge that the LLM has acquired during its training, encapsulated by a pre-trained sequence-to-sequence (seq2seq) model. The term "parametric" signifies that the memory structure of

the model has a fixed number of parameters, which remain constant throughout both the retrieval and generation processes. Specifically, a pre-trained seq2seq model is trained on extensive datasets tailored for particular natural language processing (NLP) tasks, such as machine translation, summarization, or question answering. This model serves as the core component of the LLM's memory, housing the facts, patterns, and language structures that the model has learned during its training phase.

Non-parametric memory, on the other hand, involves an external database or knowledge base, such as Wikipedia. This is accessed via a pre-trained neural retriever, which searches the external database for relevant information based on the input query. The retrieved information acts as an external library that complements the LLM's internal memory. The neural retriever identifies and retrieves the most pertinent passages, which are then used to inform and enhance the language generation process.

There are two main formulations of RAG, each with a distinct approach to utilizing the retrieved passages. The first formulation conditions the entire generated sequence on the same set of retrieved passages, ensuring consistency throughout the text. The second formulation allows the model to use different passages for each token in the sequence, offering greater flexibility and potentially more nuanced responses. Both approaches have been shown to outperform traditional seq2seq models, particularly in open-domain question-answering tasks, where the accuracy and relevance of the generated content are crucial.

In more recent studies, the field of Retrieval-Augmented Generation (RAG) has seen significant advancements, some of which were presented in March 2024, in the paper "Retrieval-Augmented Generation for Large Language Models: A Survey". [25]

The paper highlights the primary issues with LLMs, such as hallucinations, outdated knowledge, and non-transparent reasoning and introduces new RAG approaches to tackle these problems.

RAG's development has progressed through several stages. *Naive RAG* represents the earliest approach, using a straightforward "Retrieve-Read" framework. This method indexes data, retrieves relevant chunks based on queries, and generates responses. However, it struggles with issues like inaccurate retrievals and challenges in integrating the retrieved information effectively.

To overcome these limitations, *Advanced RAG* incorporates improvements such as enhanced indexing, optimized query processing, and better post-retrieval techniques. These refinements focus on increasing retrieval accuracy and contextual relevance, thus making the system more reliable.

The most sophisticated form, *Modular RAG*, introduces flexibility by allowing the integration of specialized modules for various tasks. These modules can handle direct searches, manage memory, and route queries through different data sources. This adaptability improves retrieval precision and the overall quality of response generation.

Additionally, the paper compares RAG with Fine-Tuning (FT). While RAG excels in dynamic environments with real-time updates, FT is more suitable for deep customization. The choice between RAG and FT depends on the specific

needs of the application, and the two methods can be complementary in optimizing LLM performance.

The paper also covers various aspects of retrieval in RAG systems, including the use of different data structures (unstructured, semi-structured, and structured data) and methods for optimizing indexing and query processing. These advancements aim to enhance retrieval accuracy and efficiency, ensuring that LLMs generate more precise and relevant responses.

Future research directions highlighted in the paper include improving RAG's robustness against misinformation, exploring hybrid approaches that combine RAG with fine-tuning and small language models (SLMs), and expanding RAG's capabilities to multi-modal data, such as text, images, and audio. Addressing these challenges will be crucial for further enhancing the effectiveness and applicability of RAG in various domains.

Building on these advancements, the benefits of using RAG in the context of elderly care applications are particularly promising. Firstly, the ability to retrieve and incorporate up-to-date, factual information from a trusted knowledge base ensures that the virtual assistant can provide accurate and relevant responses to a wide range of queries, including medical advice, daily living tips, and social interaction prompts. This is particularly important in healthcare-related applications, where the accuracy of information can directly impact the well-being of the user.

Secondly, RAG offers a more transparent and interpretable approach to language generation. By providing explicit references to the sources of information used in the generation process, these models can offer explainability for their decisions, addressing one of the key challenges associated with traditional LLMs. This transparency is crucial in building trust between the elderly users and the virtual assistant, as users are more likely to rely on the system if they understand where the information is coming from and how it is being used.

Moreover, RAG is inherently more adaptable and updatable compared to traditional LLMs. Since the non-parametric memory can be updated independently of the LLM's parameters, the system can easily incorporate new information as it becomes available, ensuring that the virtual assistant remains current and relevant over time. This is particularly valuable in the rapidly evolving fields of healthcare and technology.

To summarize, the integration of Retrieval-Augmented Generation (RAG) into LLMs represents a significant advancement in the development of virtual assistants for senior care. By combining the generative capabilities of LLMs with the precision and factual accuracy of a retrieval-based system, the RAG technique offers a powerful and flexible solution. This approach not only addresses the limitations of traditional LLMs but also pave the way for more transparent, trustworthy, and adaptable AI-driven systems.

2.1 Critique of the state of the art

The development of virtual assistants designed to support elderly care has seen significant progress in recent years. Various projects and products, such as SITIES and ElliQ, have made notable contributions to this domain by providing smart assistants. However, despite these advancements, there are still considerable gaps, inefficiencies, and unaddressed issues within the current state of the art. These shortcomings underscore the necessity for the development of this project.

One of the primary limitations of current virtual assistants is their inability to offer a deeply personalized user experience. Systems such as ElliQ, despite using proprietary AI algorithms to adapt to user behavior, are constrained by a predefined set of functionalities. This results in a relatively static interaction that does not fully address the unique preferences and needs of each user.

Another significant challenge with existing solutions, such as SITIES, is their limited scalability and difficulty in integrating new technologies. These systems are often designed with a specific set of applications in mind, making it challenging to expand or introduce new features.

A major challenge faced by current LLM-based systems is their tendency to produce responses that are either factually incorrect or contextually irrelevant, a problem commonly referred to as "hallucination". While traditional LLMs have made progress in language generation, their ability to reliably access and utilize external knowledge remains limited. This shortcoming is particularly problematic in critical areas such as healthcare, where accurate and context-aware information is essential.

Building trust with elderly users is crucial, especially when the virtual assistant is handling sensitive topics such as health and well-being. Current systems often lack transparency in how they generate responses, leading to potential mistrust or reluctance to rely on the assistant. By fostering a more transparent and trustworthy interaction, the virtual assistant can become a more reliable companion.

Finally, while systems like ElliQ offer valuable companionship and support, they also raise ethical concerns, particularly regarding privacy and the potential for dependency. The use of AI-driven systems in such a personal and sensitive context necessitates robust privacy protections and careful consideration of ethical implications. However, as noted in recent research [14], these systems can introduce biases and misinformation, further complicating their deployment in vulnerable populations like the elderly. This underscores the importance of designing AI systems that not only prioritize ethical considerations but also actively mitigate potential biases."

2.2 Proposal

The proposed project aims to address significant gaps in current virtual assistants for the elderly, particularly in personalization, scalability, contextual understanding, transparency, and ethics. By integrating Retrieval-Augmented Generation

(RAG) with Large Language Models (LLMs) and LangChain infrastructures, the project will create a dynamic and adaptable virtual assistant that continuously contextualizes from user information and interactions, providing highly personalized and relevant suggestions.

This modular design ensures scalability, allowing the system to evolve alongside users by incorporating new features as they become available. The use of RAG enhances the system's ability to retrieve accurate, up-to-date information, reducing errors and improving reliability.

Transparency is a core focus, with the assistant providing clear references based on the sources of its information, building trust, and empowering users to verify the advice they receive. Additionally, the project places strong emphasis on ethical standards and privacy-preserving techniques, ensuring that the assistant interacts responsibly with elderly users, particularly in handling sensitive data. This approach positions the project as a responsible, future-proof solution for elderly care.

CHAPTER 3
Analysis of the problem

The development of a virtual assistant needs a thorough problem analysis to identify the innovation or business opportunities that can justify and guide the direction of this work. The proposed project seeks to overcome the limitations of existing virtual assistants. A systematic analysis, using appropriate techniques and methods, will help to clarify the project's objectives and its potential impact.

The mission of this project is to develop a next-generation virtual assistant that significantly contributes to the quality of life of seniors by offering personalized, scalable, and contextually accurate support. The assistant will leverage cutting-edge technologies, such as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), to create a reliable and ethical tool.

To be successful, the project must fulfill several key requirements. Personalization is crucial; the assistant must offer interactions that are tailored to the user's preferences and needs. Scalability is another requirement; the system must be designed in a way that will allow for the addition of new features without significant rework, for which a modular structure is chosen. Contextual accuracy is also essential; the assistant must be able to retrieve and incorporate accurate, up-to-date information in its interactions. Security is a fundamental requirement; user data must be securely stored locally and transmitted. Finally, the assistant must operate within the bounds of legal and ethical standards, particularly in terms of data protection and user interaction.

However, several constraints condition the project, including time and budget, which may limit the scope of development and testing. Technology shortages also play a role; the project is limited by the current state of LLM and RAG technologies. Additionally, user accessibility is a significant consideration; the assistant must be designed with the user in mind, ensuring that the design is easy to use for elderly individuals who may have varying levels of tech-savviness.

At the beginning of the project, several assumptions and initial conditions are taken into account. It is assumed that the elderly users will have some basic familiarity with digital devices, even though the assistant will be designed to be as user-friendly as possible. The project also assumes that sufficient data will be available to feed RAG systems.

The legal and ethical framework is another mandatory component of this problem analysis. Given that the project involves accessing and processing per-

sonal data, it is essential to outline how this data will be generated, stored and transferred.

Ethical considerations are equally important, especially when developing technology that interacts with vulnerable populations such as the elderly. The ethical implications of using AI-driven systems in such personal and sensitive contexts will be thoroughly considered, ensuring that the assistant not only respects user autonomy but also promotes their well-being.

The project will produce several key deliverables. The primary deliverable is a fully functional virtual assistant that meets the specified requirements deployed in Streamlit. In addition, comprehensive documentation outlining the design, implementation is reflected in this thesis.

The success of the project will be measured by several criteria. Performance metrics are one of the main evaluators; the assistant must demonstrate high accuracy high quality responses. Additionally, the accomplishment of all the initial objectives set forth at the beginning of the project will be another key measurement.

3.1 Proposed solution

The proposed solution is a virtual assistant tailored with ageing people in mind. The assistant is designed to be scalable, allowing for the integration of new features and updates as technology advances, ensuring that it remains relevant and useful in the long term. Moreover, the assistant will offer contextually accurate responses, drawing on curated data and personalized information to provide elderly users with reliable support in various aspects of their lives.

3.1.1. Modules

The virtual assistant is composed of several interconnected modules, each serving a specific function within the overall system. These modules work together to create a seamless user experience.

1. **Medical module with RAG**: This module is responsible for providing users with accurate and relevant medical information. By utilizing RAG, the assistant can retrieve information from trusted sources like the National Health Service (NHS) and other verified datasets, guaranteeing that the advice and support provided are both accurate and contextually appropriate. This module is crucial for helping elderly users manage their concerns regarding health matters, offering guidance on diseases, symptoms, and general well-being.

2. **Personal module with RAG**: The personal module focuses on tailoring the assistant's interactions to the specific preferences and necessities of the individual user. By analyzing data from a personalized questionnaire stored, this module enables the assistant to learn from previous interactions and adapt its responses accordingly. This level of personalization ensures that the assistant becomes more attuned to the user's habits, and lifestyle, making interactions more meaningful and relevant.

3. **Reminders module**: This module is designed to help users manage their daily routines by setting reminders for important tasks, such as taking medications, attending appointments, or engaging in social activities. The reminders module ensures that elderly users can stay organized and on top of their daily responsibilities, reducing the likelihood of missed appointments or forgotten tasks.

4. **Weather and location module**: This module provides real-time weather updates and location-based information to the user. By integrating weather APIs and location services, the assistant can offer advice and information that is specific to the user's geographic location, helping them make informed decisions about their daily activities, such as planning outings or dressing appropriately for the weather.

5. **User interface**: The user interface (UI) is the point of interaction between the user and the assistant. It is designed to be intuitive, user-friendly, and accessible. The UI facilitates easy navigation through the assistant's using voice-to-text and text-to-voice ensuring that users can interact with the system effortlessly, enhancing their overall experience.

Together, these modules form a comprehensive virtual assistant that is capable of addressing the multifaceted must-haves of elderly users.

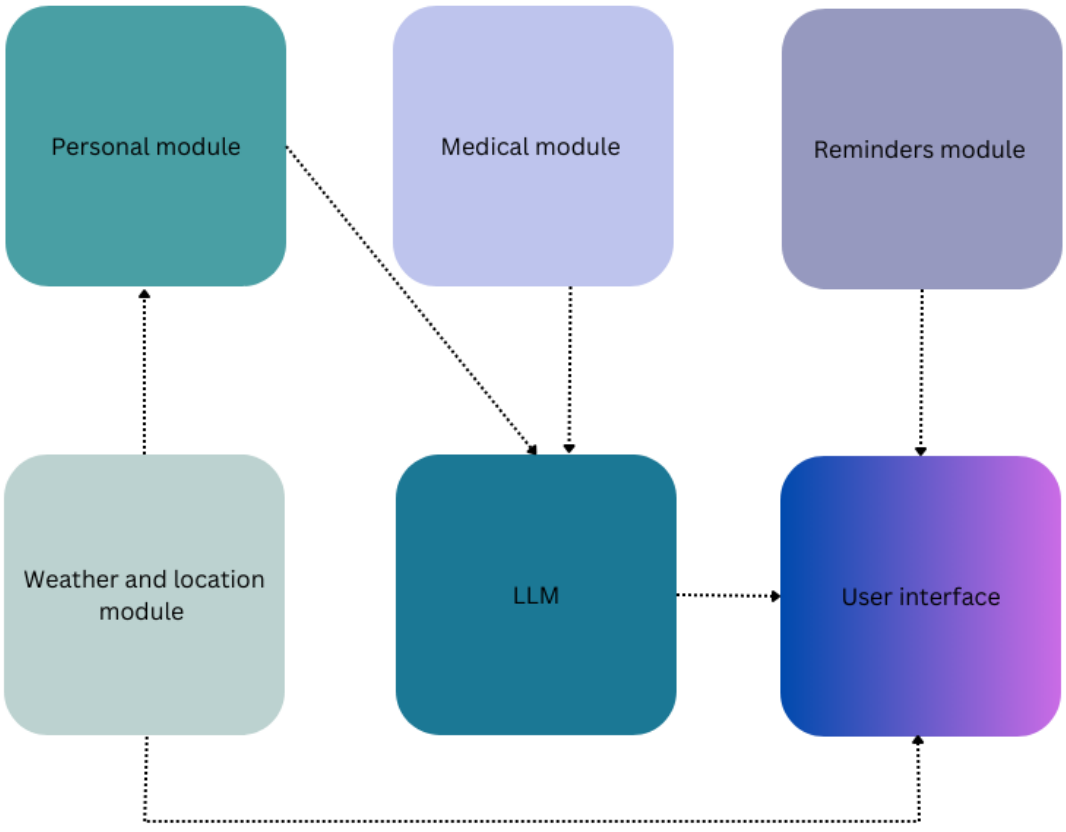


Figure 3.1: Modules of the system

In the next diagram, we can see in detail how the complete system works, from the moment the user starts talking, till he or she receives an answer, combining the modules presented in Figure 3.2.

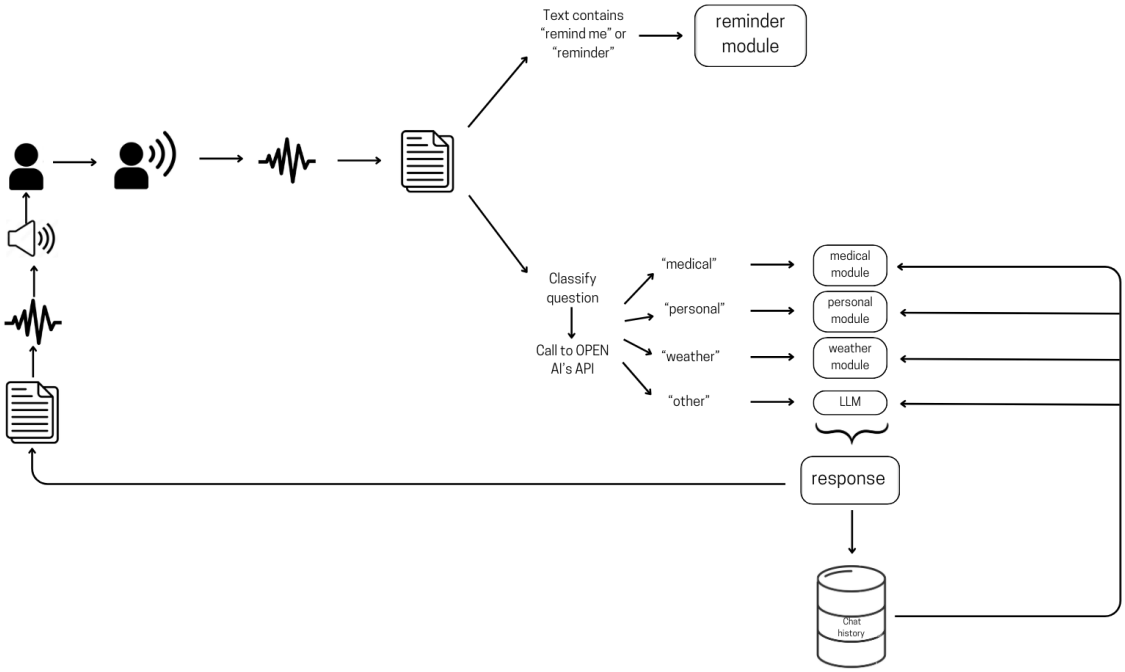


Figure 3.2: Functioning of the system

Once the user hits the button record, the voice is captured and the audio signal is translated from speech to text with a function from the Streamlit mic recorder library. Next, if the text contains the words *reminder* or *remind me* it will automatically invoke the reminders module to analyze the text and set the reminder. Otherwise, the question will be sent to the OPEN AI's API with a prompt instructing to distinguish whether the question is related to medical issues, personal ones, it is about the weather or none of the above categories. With the corresponding answer, the necessary module is invoked, except the case of the *other* category, in which case there is just a call to the Large Language Model used. After receiving the response, both the question and answer are stored in the chat history buffer, and the reply of the model is returned to the user in an audio format through the interface where the query was recorded.

In the next sections, we will analyze in detail how the different modules operate.

3.1.2. Medical module

The medical module is an essential part of the proposed virtual assistant and it is capable of delivering accurate medical information.

To establish a robust foundation of medical knowledge for this module, data was meticulously gathered from reliable sources. The primary source of this information was the National Health Service (NHS) website, a highly regarded provider of comprehensive healthcare information. Through sophisticated web scraping techniques, a vast array of medical data was extracted and subsequently compiled into a structured format. Additionally, a simpler JSON dataset, was integrated into the system. This dataset includes specific medical insights tailored for common conditions such as cuts, abrasions, and stings. Each entry in

the JSON file consists of a "tag" representing the condition, a set of "patterns" corresponding to possible user queries, and "responses" providing step-by-step guidance on how to address the condition. It was integrated into the system, significantly enriching the module's ability to respond effectively to a wide range of medical queries. This dual-sourced approach guarantees that the medical module is equipped with a broad and diverse spectrum of information, capable of addressing various health conditions and scenarios.

The functionality of the medical module is built upon the advanced capabilities of RAG. This enables the system to efficiently retrieve relevant information from the compiled medical datasets and generate responses that are finely tailored to the specific needs and inquiries. For example, if an elderly user inquires about managing diabetes, the module can swiftly retrieve the most relevant sections of the NHS data and generate a personalized response. The goal is to create an interaction that is both informative and empathetic.

The development of the medical module was carried out through several phases. The initial phase focused on data acquisition. This data was then processed and formatted into a JSON structure, facilitating easy access and manipulation within the system. To further enhance the module's knowledge base, an additional JSON dataset was integrated, providing a broader range of medical information.

Following data acquisition, the next phase involved processing and integrating the data using json-splitting techniques. They were employed to create manageable chunks of information, ensuring that the data could be effectively utilized by the RAG. This processing step was crucial for enabling efficient retrieval and accurate response generation, thereby enhancing the overall functionality of the module.

Once the data was processed, it was embedded using the OpenAI embeddings and stored in a vector store powered by Chroma. This vector store serves as the core of the retrieval system, allowing the module to quickly access the most relevant pieces of information based on user queries. This architecture ensures that the system can handle large volumes of data while maintaining high retrieval accuracy.

When a user interaction begins, the system considers additional context like chat history, besides just the user's query. This question is used to create a query embedding, which is then matched against the vector database to find the most relevant data chunks. These chunks, along with the contextual information, are then used to enhance the prompt sent to the Large Language Model (LLM). The model was instructed to interact with elderly users in a manner that is both accurate and empathetic, ensuring that the responses generated are not only informative but also considerate of the users' needs and feelings.

In the schema found below, the described functioning of the module is illustrated. We can see the indexing, augmenting and generating processes combined to be able to return a response to the main module and consequently, to the user.

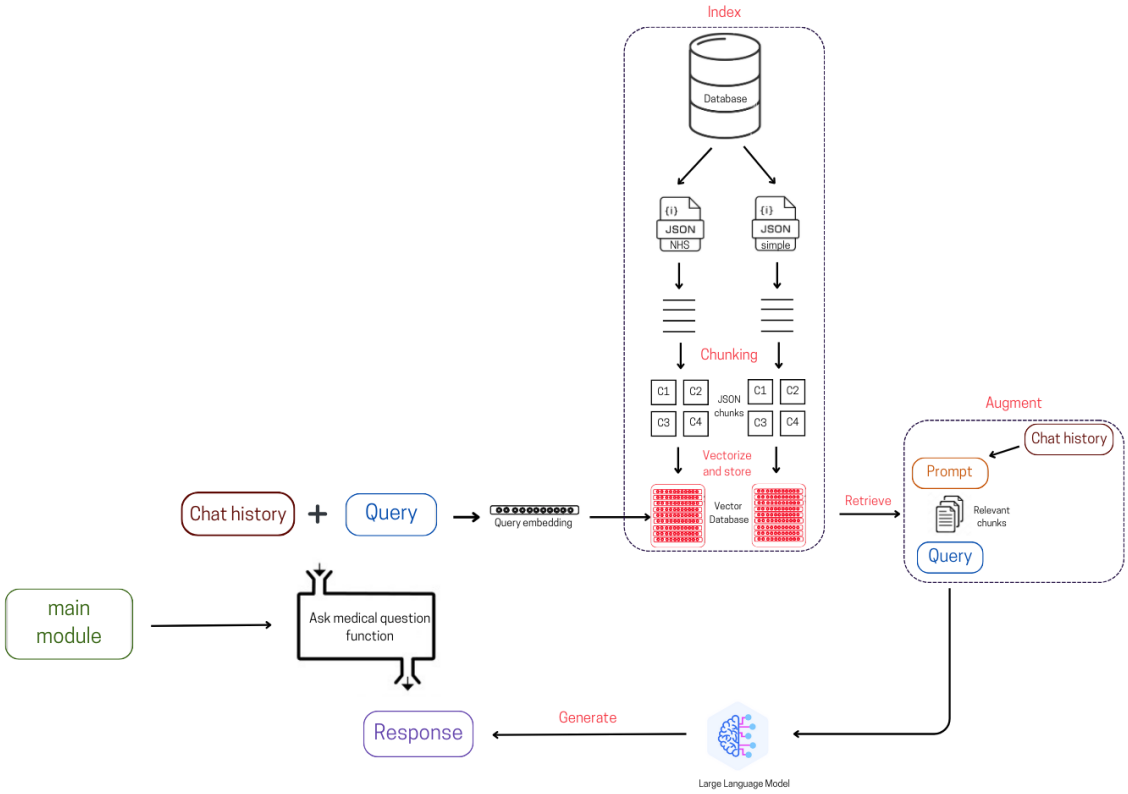


Figure 3.3: Medical module

3.1.3. Personal module

The personal module of the proposed virtual assistant is specifically designed to enhance user engagement by tailoring interactions to the unique preferences and needs of each elderly user.

The foundation of the personal module lies in the creation of a comprehensive JSON file, which was developed from a carefully designed personal questionnaire. This questionnaire is a crucial element of the module, as it captures detailed information comprising the individual preferences, habits, and needs. The questions are thoughtfully crafted to gather insights into various aspects of the user's life, such as their daily routines, hobbies, health concerns, and personal interests. These questions can be consulted in the annex to provide a clear understanding of the types of data collected and how they contribute to the personalized interaction.

A sample questionnaire was completed and the responses were compiled into a JSON file. This file serves as the primary data source for the personal module, enabling the virtual assistant to access and utilize the user's personal information during interactions. For instance, if the user asks for activity recommendations, the module can retrieve information about the user's preferences from the JSON file and generate suggestions that align with their interests and lifestyle. The query to the LLM also includes current weather conditions and location, provided by the weather module. The result is an interaction that feels personalized, relevant and contextual, enhancing the overall user experience.

The functionality of the personal module is very similar to the medical module. It is driven by RAG, enabling the module to retrieve relevant information from the user's personal data and generate responses that are customized to their specific needs and preferences.

The development of the personal module involved several key phases, beginning with the design of the questionnaire by a psychologist. This phase required careful consideration of the types of information that would be most valuable for creating a personalized experience. Questions were designed to be comprehensive yet user-friendly, ensuring that elderly users could easily provide the information needed without feeling overwhelmed. The questionnaire was completed with sample answers to be able to test the system.

Following the data phase, the next step involved processing the data and integrating it with the RAG. The system employs a process of chunking to break down the JSON file into smaller, more manageable pieces of information. This is done to improve the efficiency of the data retrieval process. Each chunk represents a specific aspect of the user's personal data, such as a particular hobby or health concern. These chunks are then vectorized and stored in a vector database. This allows the system to efficiently search through the data and retrieve the most relevant chunks based on the user's query.

When a user interaction is initiated, the system takes into account not only the user's query but also additional context such as chat history, location data, and current weather conditions. The user's query is used to generate an embedding, which is then compared against the vector database to retrieve the most relevant data chunks. The retrieved chunks, along with the additional contextual data, are used to augment the prompt sent to the Large Language Model (LLM).

The LLM then generates a response which is sent back to the main module.

Below, we can find the personal module illustrated in detail.

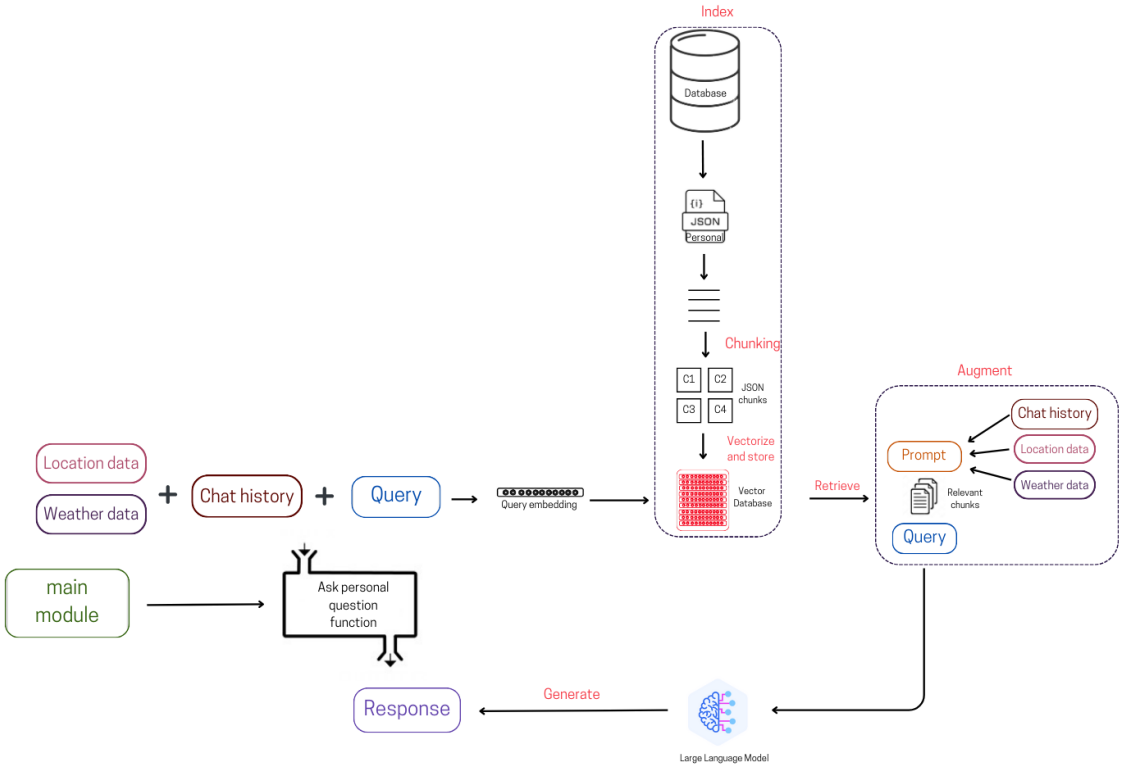


Figure 3.4: Personal module

3.1.4. Reminders module

The design of the reminders modules is set to assist users in managing their daily routines by setting and announcing reminders. This module offers a practical way of keeping track of important tasks, appointments, and other time-sensitive activities.

The primary functionality of the reminders module revolves around its capability to receive user input, parse this input to extract relevant date and time information, and then schedule reminders accordingly. Once the reminder is set, the module ensures that the user is notified at the appropriate time. This makes it easier to manage daily schedules without the stress of remembering every detail.

The development of the reminders module involved several key phases, beginning with the design of a parsing function. This function is designed to identify and convert phrases like "on August 10th, 2024 at 3:30 PM," into a standardized datetime format that the system can work with. The accuracy of this parsing function is crucial, as it directly impacts the reliability of the reminders set by the system.

Once the user input is successfully parsed, the system proceeds to the scheduling phase, where the reminder is added to a list of scheduled events. This list is continuously monitored by the module, which checks for any reminders that are due. The system uses a time-based checking mechanism that operates at regular intervals, typically every minute, to verify whether any reminders need to be activated. When the time for a reminder arrives, the system alerts the user with the relevant message, fulfilling the reminder request.

To facilitate the persistence of reminders, the module includes functionality to save and load reminders from a file. This ensures that reminders are not lost if the system is restarted or shut down. Each reminder, along with its scheduled time, is stored in a JSON file, which can be reloaded when the system is restarted. This feature is particularly important for maintaining the continuity of service, as it allows the assistant to manage long-term reminders, such as those set for days or weeks in advance.

The module offers a straightforward yet powerful tool for elderly users to manage their daily schedules, ultimately contributing to their well-being and independence.

3.1.5. Weather module

The weather and location module is designed to provide users with real-time, location-based weather updates. This module leverages the Open-Meteo API and location services to offer personalized weather forecasts.

The core functionality of the weather and location module lies in its ability to gather accurate weather data based on the user's location and present this information in a user-friendly manner.

The development of the weather and location module involved several phases. First it was necessary to determine the user's location. This was achieved by integrating the ipinfo.io API, which allows the system to retrieve the user's location based on their IP address. The API returns detailed information, including the user's city, region, country, and geographical coordinates. The precision of this data is essential to ensure that the weather forecasts provided are relevant to the user's exact location.

Once the location data was secured, the next phase involved integrating the Open-Meteo API to fetch current and forecasted weather data based on the user's coordinates. The API returns a comprehensive dataset that includes current weather conditions, such as temperature and wind speed, as well as hourly forecasts.

Finally, the data is then parsed to extract the most relevant information, which is then used to generate both immediate and short-term weather summaries for the user.

3.1.6. Interface

The ELF Virtual Assistant interface is designed with simplicity and ease of use in mind, tailored specifically for elderly users. The main components of the interface are as follows:

1. **Title and Branding:** The top section features a prominent title, *ELF Virtual Assistant*, styled with a green color and an elf emoji. This gives the interface a friendly and approachable feel, helping users recognize the assistant's name and purpose immediately.
2. **Welcome Message:** Below the title, a warm welcome message is displayed: *"Welcome! I am the Elderly Life Friend. Here's how I can help you:"* This phrase

clearly sets the tone for the interaction, emphasizing the assistant's role as a supportive tool for seniors.

3. **Key Services Provided:** A list of key services is presented, highlighting different categories of help the assistant offers. These services are easy to read and understand, and they are accompanied by a brief explanation of what each service involves. The categories include:

    - **Medical Help:** Users can ask about illnesses or treatments.
    - **Personal Questions:** The assistant provides personalized recommendations.
    - **Set Reminders:** A quick and easy way to set reminders for tasks or events.
    - **Weather Updates:** Instant access to weather information.
    - **General Questions:** A catch-all category for any other queries users may have.

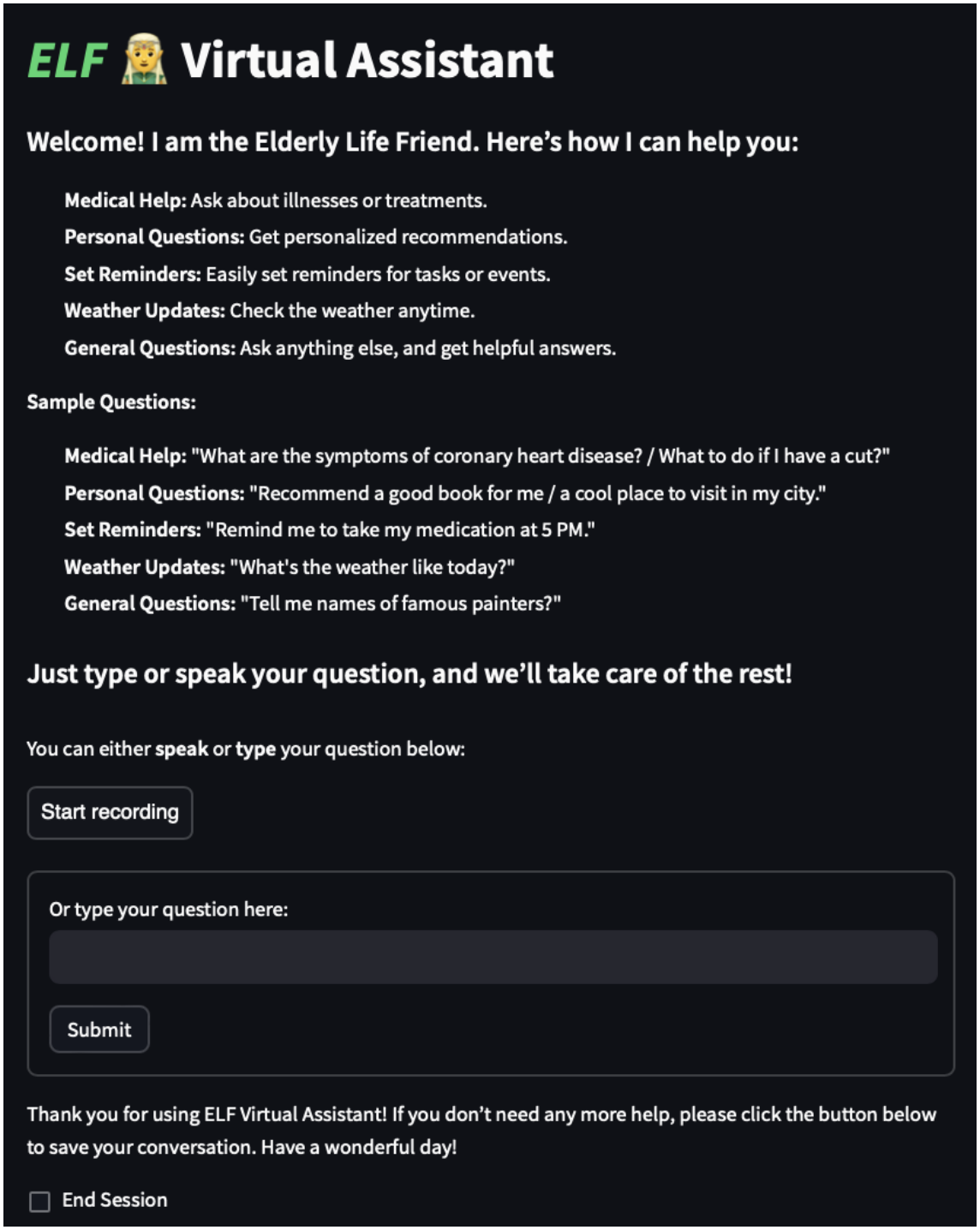


Figure 3.5: Elf Virtual Assitant interface in Streamlit

4. **Sample Questions:** Directly below the service descriptions, a few *sample questions* are provided to guide users in framing their own queries. These sample questions are relevant to elderly users, such as asking about medical conditions, setting medication reminders, or inquiring about the weather. This section encourages users to explore different types of questions they can ask the assistant.

5. **Interaction Methods:** The interface allows users to communicate with the assistant either by *speaking* or *typing* their questions. This dual input option

ensures accessibility for users with different preferences or abilities, even being able to assist people with certain disabilities like blindness or deafness:

- A button labeled **"Start recording"** is available for users who prefer speaking.
- Below that, there is a text box where users can *type their questions* if they feel more comfortable doing so.

6. **Submit and Session End Options:**

   - If they recorded their question, they will automatically get the answer once it is ready, with a previous print which reflects what they introduced, the answer text and an audio recording of the same answer so there is no need to read it.
   - If users have typed their questions, they can submit them using the **Submit** button. The result and answer are identical to the speaking version
   - At the bottom of the interface, there is an **End Session** checkbox that allows users to conclude their conversation with the assistant. A friendly reminder encourages users to click this button if they don't need more help and want to save their conversation for future reference.

7. **Design Elements:** The interface uses a dark background with white text, ensuring high contrast for readability. The use of bold text, large buttons, clear instructions and labels enhances the user experience, making it straightforward for elderly users to navigate.

3.1.7. Phases of development and materials

The development of this complex system is a multifaceted process that unfolds across several distinct phases. It starts with research and data collection. Next is the module development, where individual components are created, using the collected data with RAG and an LLM to generate relevant responses. In the integration phase, these modules are combined into a unified system, ensuring smooth interaction between them. Subsequently comes testing and validation, the system undergoes rigorous model evaluation to ensure that it is reliable, functional, and has the best possible performance. Finally the interface was built and deployed.

The majority of the work was conducted using Python [15], specifically in version 3.12. Depending on the task, different programming environments were utilized. Basic data handling was performed locally on a desktop computer, using Visual Studio and Sublime Text, as these tasks did not require extensive computational resources. For the modeling process, Google Colaboratory was used to leverage cloud-based computing power, avoiding reliance on the desktop's capabilities.

A variety of free Python libraries were employed throughout the project, including:

- **Data handling:** `pandas` [16], `numpy` [17]
- **Machine Learning and text processing:** `sklearn` [18], `sentence-transformers`, `transformers`
- **Visualization:** `matplotlib` [19], `seaborn` [20], `plotly` [21]
- **Text analysis and evaluation:** `nltk`, `rouge_score`, `openai`
- **Interface:** `streamlit`
- **Additional libraries:** `json`, `beautifulsoup4` [22]

CHAPTER 4

Preparation and understanding of the data

In data science projects, meticulous data preparation and understanding are of great importance. This involves several processes including data creation, integration, transformation, and preparation. Each step plays a crucial role in ensuring that the data is accurate, relevant, and ready for analysis. This chapter will explore the processes involved in preparing and understanding data, with a focus on three primary sources:

- Medical JSON from the NHS: This JSON was obtained through web scraping from the NHS website, specifically from the Health A-Z page. It is an online Health Encyclopedia with information related to diferent health conditions, symptoms and treatments, including what to do and when to get help. The data is organized by health condition, with each condition having multiple attributes detailing different aspects of the condition such as symptoms, diagnosis, medical treatments, prevention, and causes.

  In total there are 922 medical conditions. In the following word cloud we can see that some of the most common words are *symptom*, *may*, *treatment*, *people*, *help* and *cause*.

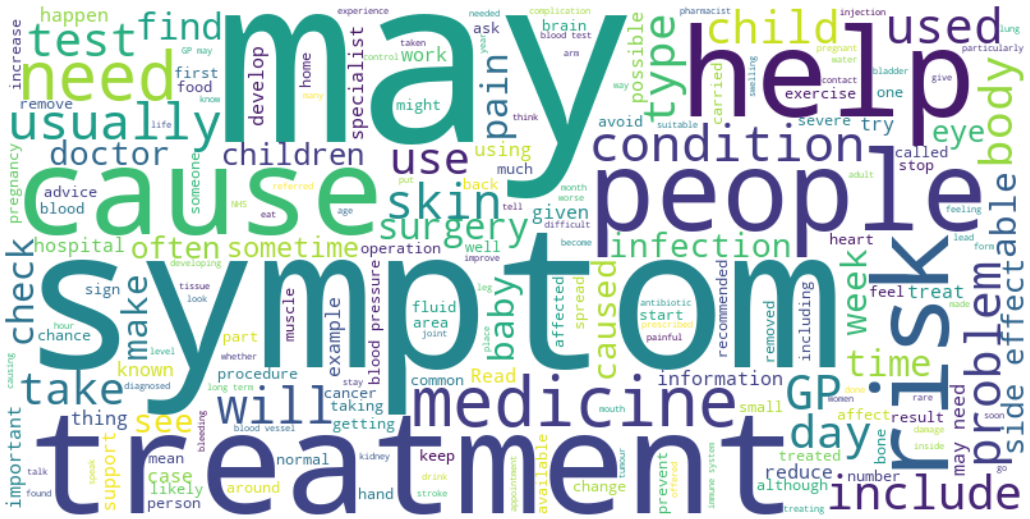


Figure 4.1: Most common words

Each medical condition has a various number of subcategories. In the next bar chart we can see which are the most common ones. The top three are comprised by the subclasses *symptoms* (coinciding with one of the most frequent words displayed in the word cloud), *self care* and *medical treatments*.

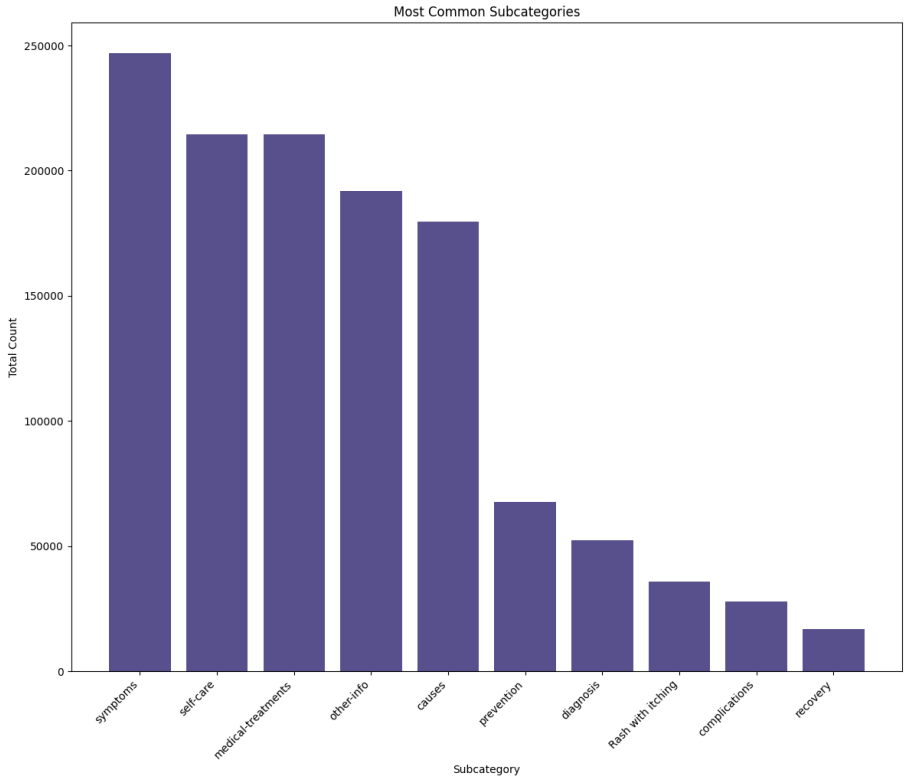


Figure 4.2: Top 10 subcategories

- Kaggle medical JSON with treatment advice: This second JSON, is a Kaggle dataset called *Dataset for medical-related chatbots* [28] and it is shorter and more simple than the NHS JSON. It provides practical advice for treating common issues like cuts, abrasions, and stings. The data is categorized by tags and includes patterns and responses that offer treatment advice. The Kaggle JSON is structured to provide direct answers to specific questions related to these conditions, making it useful for quick reference.

  The table below summarizes the 44 conditions included, categorized into relevant groups to help identify and manage different types of health issues. The categories include injuries, skin problems, abdominal issues, and more, covering a comprehensive array of medical conditions.

Category	Conditions
Injuries	Eye injury, head injury, broken toe, pulled muscle, fracture, chemical burn, sun burn, frost bite, snake bite, animal bite, bruises, wound
Skin problems	Rash, wound, sun burn, chemical burn, frost bite, insect bites, snake bite, animal bite
Abdominal issues	Abdominal pain, diarrhea, rectal bleeding, gastrointestinal problems
Heat and cold related conditions	Heat exhaustion, heat stroke, frost bite
Bleeding disorders	Nose bleed, normal bleeding, rectal bleeding
Poisoning and toxic exposure	Poison, chemical burn
Neurological issues	Seizure, vertigo, fainting, headache
Respiratory issues	Nasal congestion, cough, sore throat, cold
General health issues	Fever, teeth, choking, drowning, CPR
Muscle and joint issues	Sprains, strains, strained muscle, testicle pain
Skin conditions	Abrasions, stings, splinter

Table 4.1: Summary of Medical Conditions

- Personal JSON from a Questionnaire:

  The personal questionnaire data was collected to tailor a conversational model specifically to the preferences of older adults. This questionnaire was carefully crafted by a psychologist to capture various aspects of an individual's life, including personal information, family background, interests, work history, health, etc.

  The questionnaire is divided into several sections:

  - **Basic Information:** Includes details such as name, age, and place of residence.
  - **Family Background:** Questions about family roles and anecdotes.
  - **Interests and Hobbies:** Covers leisure activities, hobbies, and travel experiences.
  - **Work and Education History:** Inquires about professions, significant jobs, and educational background.
  - **Life Experiences:** Focuses on important life events and achievements.
  - **Preferences:** Includes favorite foods, movies, and other personal preferences.
  - **Health and Well-being:** Questions on health concerns, exercise habits, and mental well-being.
  - **Aspirations and Goals:** Looks into future goals and unfulfilled dreams.
  - **Memories and Nostalgia:** Captures cherished memories and childhood stories.
  - **Opinions and Values:** Explores current concerns, values, and self-perception.

- **Self-esteem and Self-Perception:** Asks about positive self-descriptions and skills.

To facilitate testing and integration with the retrieval-augmented generation implementation, the collected responses were converted into a JSON format. This conversion allows for structured and systematic handling of the data, enabling efficient testing.

The JSON structure generated from this questionnaire includes key-value pairs where each question is a key, and the individual's responses are the corresponding values. This structured format allows for easy parsing and manipulation of data. Here is a sample of the Personal JSON data structure based on the collected responses:

```
{"question": "What is your favorite food?",
"answer": "I love pizza, especially pepperoni with extra cheese."},
{"question": "What type of movies do you like?",
"answer": "I really enjoy science fiction and adventure movies."},
{"question": "If you have a favorite movie, what is it?",
"answer": "My favorite movie is 'Inception' by Christopher Nolan."},
{"question": "Can you tell me a book that you liked?",
"answer": "A book that impacted me was '1984' by George Orwell."}
```

4.1 Data quality and transformation

The quality of data is a critical factor in data science. For the NHS JSON, the quality is generally high due to its authoritative source. However, web scraping can sometimes lead to incomplete or inconsistent data due to changes in the website structure or scraping errors. To address this, the code includes error handling and validation steps to ensure that the data collected is accurate and complete.

The Kaggle medical JSON with treatment advice is straightforward, but its simplicity means that it may not cover all scenarios comprehensively. It provides a general overview of treatment options, which may need to be supplemented with more detailed medical guidelines for complete coverage.

The Personal JSON's quality depends on the accuracy of the responses collected and how well the questionnaire was designed. Personal data often requires careful handling to ensure privacy and accuracy, for that reason it is stored locally, avoiding cloud storage solutions.

Not much transformation of the data was involved, since the structure was thought of and implemented during the extraction of the data in the case of NHS and during the conversion from plain text to json in the case of the personal questionnaire. For instance, the NHS JSON data is structured into a hierarchical format where each health condition is a key, and its associated information is stored as nested dictionaries. Similarly, the Kaggle JSON is organized by tags and the personal data by questions, making it easy to retrieve specific treatment information based on keywords. Refer to the appendices to view a snippet of both JSONs.

Data preparation and understanding are fundamental to the success of data science projects. By carefully managing the processes of data creation, integration, transformation, and preparation, it can be ensured that the data is of high quality and relevant to the project's objectives.

CHAPTER 5
Knowledge extraction and model evaluation

5.1 Introduction

5.1.1. Purpose and Scope

This section is dedicated to the comprehensive analysis and evaluation of the models developed using the data extracted from our study. The primary objective is to assess the effectiveness, accuracy, and practical utility of these models. By systematically evaluating each model, we aim to determine their performance across various metrics, identify their strengths and weaknesses, and provide actionable insights for future refinement and application.

First, we will use qualitative ratings, comparing human feedback with objective metrics such as cosine similarity, BLEU, and ROUGE. This will help determine which model aligns best with human evaluations and performs optimally across different scenarios. Finally, a comparative analysis will be conducted to determine which models excels overall.

Finally, we will highlight any limitations or areas for improvement, offering recommendations for enhancing model performance.

5.1.2. Overview of Models

In this evaluation, we examine a diverse set of advanced language models, each with unique features and capabilities. The models under consideration include gemma-7b-it, mixtral-8x7b-32768, llama-3.1-8b-instant, GPT-4o, GPT-3.5-turbo, and llama-3.1-70b-versatile. Each of these models brings distinct strengths to the table, and their evaluation provides valuable insights into their performance and applicability.

The gemma-7b-it model, with its 7 billion parameters, is designed for intermediate-level text processing tasks. It excels in providing detailed and contextually accurate responses across a range of language tasks. This model is particularly effective in applications requiring a moderate level of information synthesis and

contextual awareness, such as information retrieval systems and conversational agents.

The mixtral-8x7b-32768 model integrates multiple architectures and employs 7 billion parameters with an extensive context window size of 32,768 tokens. This design enhances its capability to handle large-scale data and complex queries. Its ability to manage lengthy documents and multi-turn conversations makes it ideal for advanced natural language understanding and document summarization.

The llama-3.1-8b-instant model features 8 billion parameters and is optimized for real-time processing and rapid response generation. Its emphasis on speed and efficiency allows it to deliver quick, contextually relevant answers, making it well-suited for interactive applications such as chatbots and virtual assistants.

The GPT-4o model represents an advanced iteration of the GPT series, known for its powerful language capabilities and large-scale text generation. With its advanced architecture, GPT-4o provides highly coherent and contextually appropriate text, making it effective for creative writing, content generation, and complex question-answering systems.

The GPT-3.5-turbo model, a streamlined version of GPT-3.5, balances efficiency and performance while maintaining high output quality. This model is designed to provide high-quality text generation with reduced computational demands, making it suitable for general-purpose text generation and interactive applications where cost and performance need to be balanced.

Finally, the llama-3.1-70b-versatile model boasts an impressive 70 billion parameters, offering exceptional versatility and depth in text understanding and generation. Its large parameter size enables it to capture and apply extensive knowledge across diverse domains. This model excels in providing comprehensive and contextually rich responses, making it ideal for advanced research applications, comprehensive content creation, and in-depth analysis tasks.

Each of these models contributes unique strengths and capabilities, and their evaluation will provide a thorough understanding of their performance across different contexts. The following sections will delve into specific aspects of model performance, including evaluation metrics, and comparative analysis to offer a complete picture of each model's capabilities.

5.2 Model evaluation

The evaluation of models is a critical phase that involves both qualitative and quantitative assessment methods. This multifaceted approach ensures a comprehensive understanding of model performance and reliability.

5.2.1. Qualitative Evaluation

To capture the human perspective on model performance, a Google Form was created to facilitate the evaluation of model responses. Participants were asked to assess 12 questions—six personal and six medical—using a Likert scale ranging from 1 to 5. Here, 1 represented "poor" and 5 denoted "excellent." This scale

allows for a nuanced evaluation of the responses generated by each model, with the aim of obtaining a reliable measure of quality from human evaluators. Multiple individuals rated each response to mitigate biases and derive a more balanced average score. This "human" evaluation serves as a benchmark for understanding how effectively each model replicates human-like responses and insights. Examples from the evaluation questionnaire with a summary of results can be found in appendix C.

5.2.2. Quantitative Evaluation

The quantitative assessment involved several objective metrics to evaluate model performance against the human ratings. These metrics included cosine similarity, token overlap, BERT embedding similarity, Sentence-BERT similarity, BLEU score, ROUGE score, and METEOR score. Each metric offers a different perspective on how well the model's responses align with the expected answers, and understanding these metrics is crucial to the evaluation of model performance.

- **Cosine Similarity**: This metric measures the cosine of the angle between vectors of model responses and expected answers, providing an indication of textual similarity.
- **Sentence-BERT Similarity**: Utilizes sentence embeddings to calculate similarity, capturing semantic relationships between responses and expected answers.
- **BLEU Score**: Evaluates the overlap of n-grams between model responses and reference answers, assessing the quality of generated text.
- **ROUGE Score**: Measures the overlap of n-grams, words, and word sequences, focusing on recall and the quality of summary-like responses.
- **METEOR Score**: Computes alignment between the generated response and the reference answer, considering synonyms and stemming.
- **BERT Embedding Cosine Similarity**: Analyzes the similarity between BERT-generated embeddings of the responses and expected answers.
- **Token Overlap**: Calculates the proportion of overlapping tokens between the model's response and the expected answer.

Since the scales of these objective metrics differ from the human ratings, they were rescaled to match the 1-5 Likert scale using Min-Max scaling. This adjustment allows for direct comparison between human evaluations and objective measures. The correlation between each rescaled metric and the human ratings was then analyzed to identify which metric most accurately reflects human judgments.

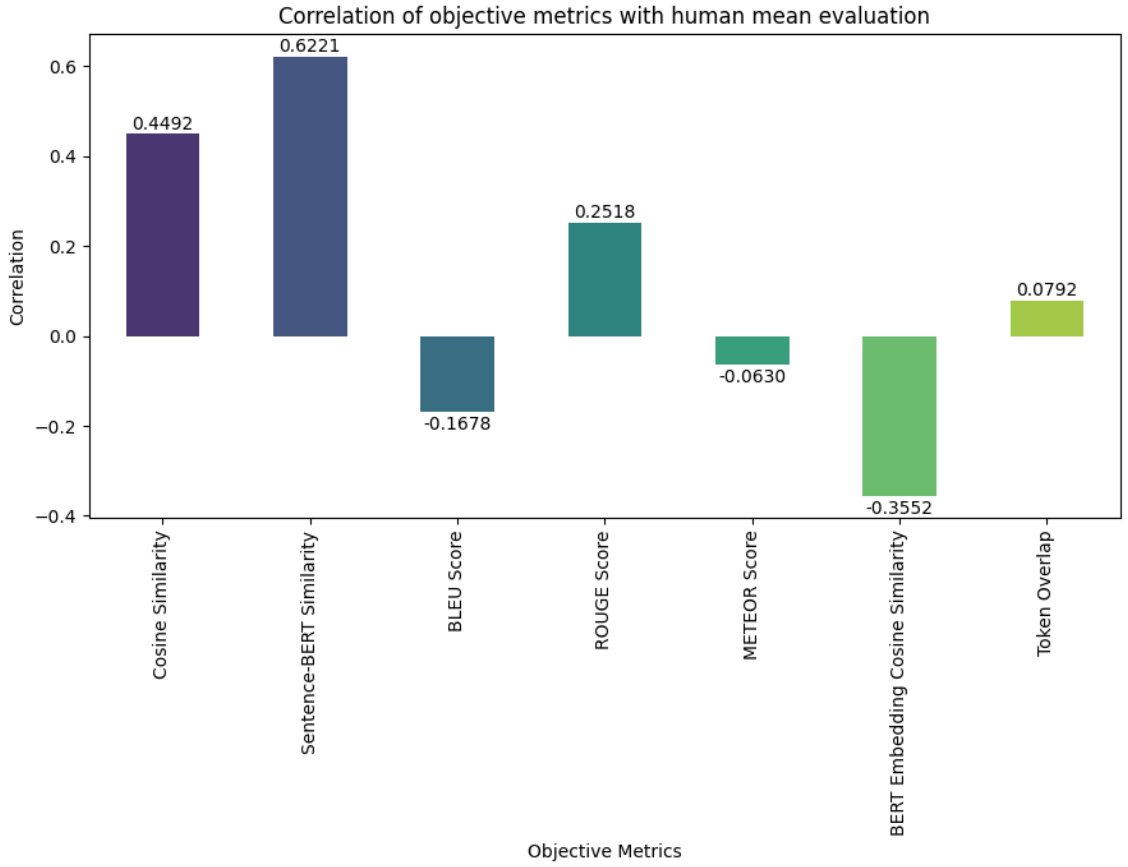


Figure 5.1: Correlation of objective metrics with human mean evaluation

The above plot depicts the correlation between objective metrics and human mean evaluation showing which metrics align most closely with human judgments. This visual representation helps in understanding which metrics are most reliable for evaluating model responses. The plot clearly identifies that the metric with the highest correlation to the human evaluation is the Sentence-BERT Similarity (which will be used to evaluate the chosen models), followed by the Cosine Similarity metric.

With the optimal metric identified through correlation analysis, the next step involved generating multiple responses from each model using the same set of questions that were originally employed in the Google Forms questionnaire for human evaluation. This approach ensured consistency across both the qualitative and quantitative assessments, allowing for a direct comparison of the models' outputs.

By applying the selected metric to these responses, it was possible to conduct a comprehensive analysis of each model's performance. The comparison involved evaluating how closely each model's responses aligned with the expected answers, as determined by the metric. This process highlighted the relative strengths and weaknesses of each model, offering valuable insights into their effectiveness in various contexts.

In evaluating the performance of various language models, several key metrics are crucial for determining the best-suited model for a given application. The analysis focuses on a multi-metric assessment, including Mean Similarity, Maxi-

mum Similarity, Minimum Similarity, and Standard Deviation, providing a comprehensive view of each model's strengths and weaknesses. Through the use of radar charts, bar charts, and scatter plots, we can break down the performance of each model to conclude which one is the most effective.

Maximum Similarity per Model

The maximum similarity metric demonstrates the peak performance a model can achieve. It reflects the model's ability to produce outputs that are highly aligned with the expected responses.

According to the bar chart, **gemma-7b-it** slightly outperforms the other models, achieving a maximum similarity score of **91.53%**. This suggests that **gemma-7b-it** has the potential to generate responses with very high fidelity to the expected answers, making it a strong candidate for tasks where peak performance is critical. However, other models like **mixtral-8x7b-32768** and **gpt-3.5-turbo** are close behind, indicating that they too are capable of producing highly accurate responses. **llama-3.1-70b-versatile**, on the other hand, records the lowest maximum similarity score, implying that even at its best, it struggles to match the output quality of the other models.

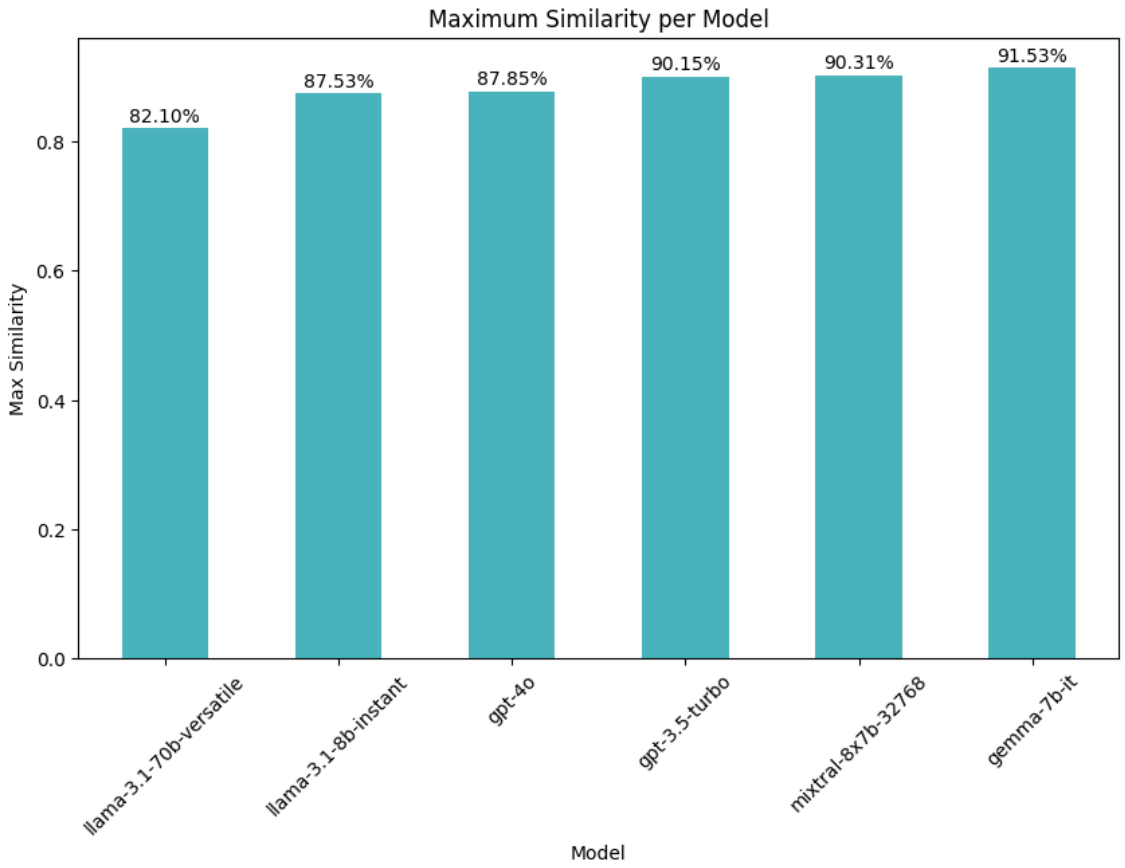


Figure 5.2: Maximum Similarity per Model

Minimum Similarity per Model

The minimum similarity metric reveals the lowest similarity score achieved by each model, highlighting their worst-case performance. This is particularly important in applications where even the worst output needs to meet a certain quality threshold.

In this metric, **gpt-3.5-turbo** and **gpt-4.0** distinguish themselves, with minimum similarities above **60%**. This indicates that even their least accurate outputs are still relatively close to the expected answers, making them reliable choices for applications where poor performance cannot be tolerated. **llama-3.1-70b-versatile** stands out negatively here, not only having the lowest minimum similarity but also recording a negative similarity (**-6.62%**). This suggests that in some cases, the model's output could be entirely misaligned with the expected answer, further cementing its unsuitability for tasks requiring dependable performance.

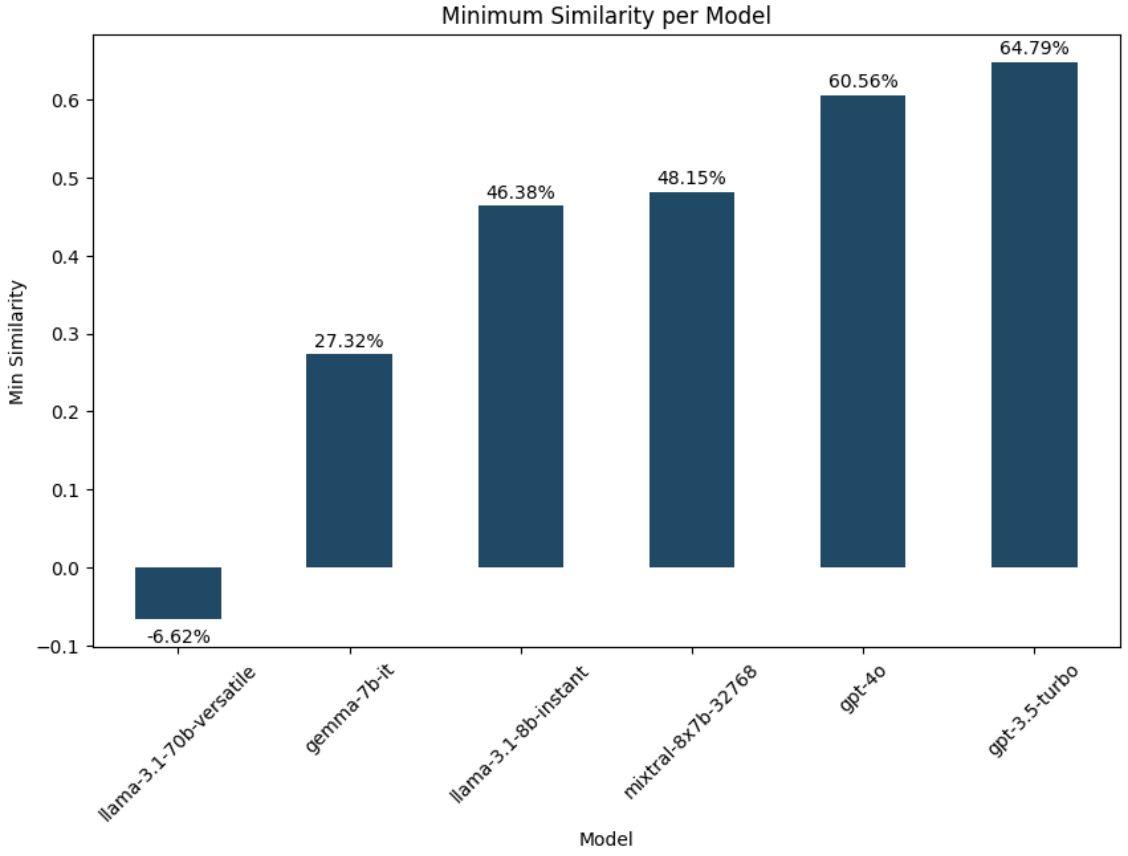


Figure 5.3: Minimum Similarity per Model

Standard Deviation per Model

The standard deviation provides a crucial insight into the consistency of each evaluated language model. Essentially, a lower standard deviation indicates that a model delivers consistent results, while a higher standard deviation suggests that a model's responses are more variable and thus less reliable.

From the chart, we observe that gpt-3.5-turbo stands out as the most consistent model, with the lowest standard deviation of 8.27%. This suggests that gpt-3.5-turbo performs well on average with minimal variation across different inputs. Following gpt-3.5-turbo is gpt-4.0, with a slightly higher but still low standard deviation of 9.15%. This model also demonstrates a high degree of consistency, making it another strong candidate for tasks that require stable and dependable performance. The close standard deviation scores between both models indicate that they can be trusted to deliver consistently accurate results.

Llama-3.1-70b-instant and Mixtral-8x7b-32768 show moderate variability with standard deviations of 13.42% and 14.75%, making them fairly consistent but less predictable than top performers. Gemma-7b-it, with a higher standard deviation of 21.00%, offers potential for high scores but with more inconsistency, which could be problematic in applications requiring uniform results leading to less reliable outputs. The most concerning is Llama-3.1-70b-versatile, which has the highest standard deviation at 37.92%, indicating significant variability and making it the most inconsistent and unreliable model in this comparison, leading to inconsistent results that may not align with expected outcomes.

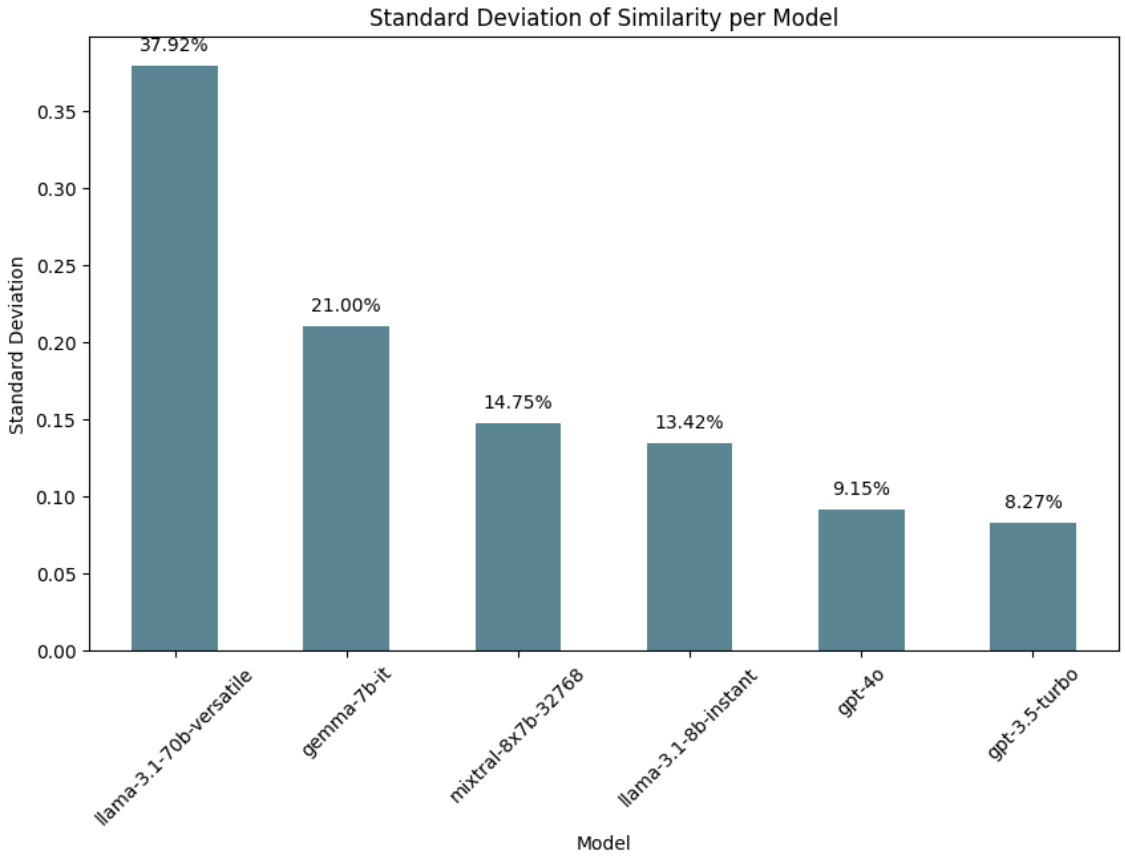


Figure 5.4: Standard deviation per Model

Mean Similarity per Model

The mean similarity metric provides insight into the average performance of the models across all test cases. It reflects how consistently a model can produce responses that are close to the expected outcomes.

In this aspect, once again **gpt-3.5-turbo** and **gpt-4.0** lead the pack, with mean similarity scores around **78-79%**. This high mean score indicates that these models are not only capable of generating high-quality responses occasionally but do so consistently across various inputs. Such consistency is crucial for applications where reliable performance is required across diverse scenarios, specially when dealing with such a vulnerable population.

Llama-3.1-70b-instant and Mixtral-8x7b-32768 show slightly lower performance, with scores ranging around 72-74%, yet they still deliver solid results. These models maintain a good level of reliability, making them viable options for many applications, though they might occasionally fall short in more demanding contexts. Gemma-7b-it follows with a score of 65.69%, indicating a noticeable drop in performance but still within a range that could be acceptable depending on the use case.

In contrast, llama-3.1-70b-versatile has a mean similarity of just 33.12%, which is significantly lower than the others. This low score suggests that it struggles to maintain quality across different contexts, making it less suitable for tasks that demand uniform performance.

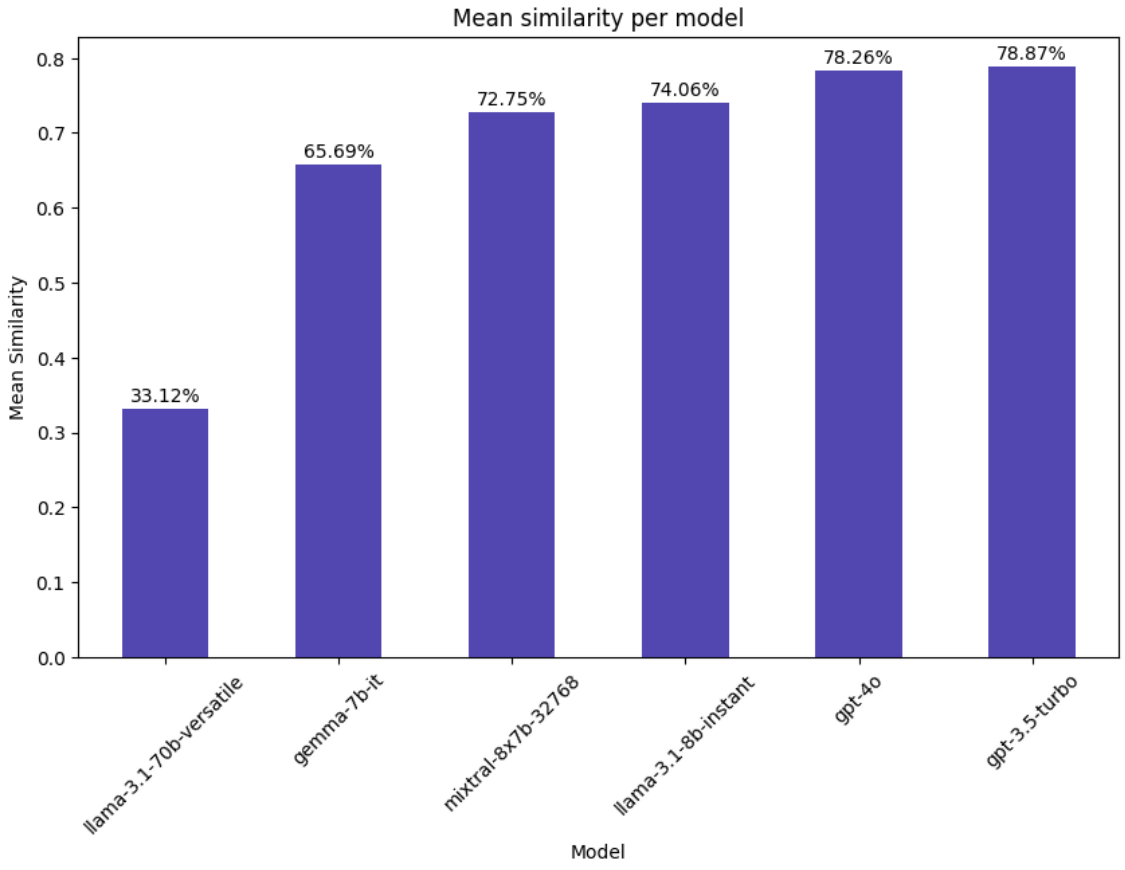


Figure 5.5: Mean Similarity per Model

Mean Similarity vs. Standard Deviation

The scatter plot comparing mean similarity to standard deviation is essential for understanding how well a model performs on average and how consistent that performance is. Even though we analyzed already both metrics separately in the previous sections, viewing both in the same plot gives us a clear vision of the best models. A model with high mean similarity and low standard deviation is ideal, as it suggests that the model performs well and does so consistently.

Both **gpt-4.0** and **gpt-3.5-turbo** excel in this area, demonstrating high mean similarity coupled with low standard deviation. This combination highlights their reliability; they not only produce high-quality responses but do so with little variation, ensuring consistent output quality. gemma-7b-it, while still strong, shows a slightly higher standard deviation, indicating that although it often performs well, it may occasionally produce outlier results. llama-3.1-70b-versatile, with its high standard deviation and low mean similarity, further reinforces its position as the least reliable model in this analysis, exhibiting both poor average performance and high variability.

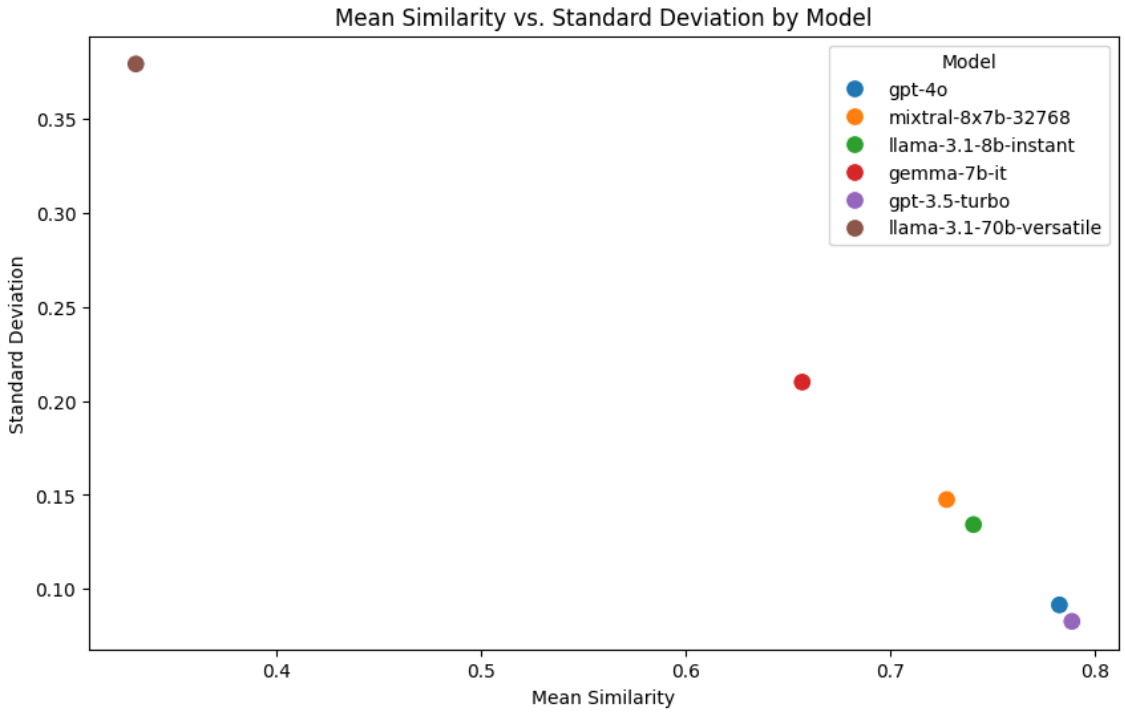


Figure 5.6: Mean Similarity vs Standard Deviation by Model

Model Performance Comparison

The radar chart serves as a visual summary of the models' performance across multiple metrics, each represented by a different colored polygon. This approach allows for a holistic comparison.

In this visualization, **gpt-3.5-turbo** and **gemma-7b-it** emerge as the frontrunners, displaying balanced and consistent performance across all evaluated metrics. Their polygons are more symmetrical and expansive, indicating strong performance in Mean, Median, Max, and Min Similarity metrics. In contrast, **llama-3.1-70b-versatile** shows significant deviations, particularly in Mean and Min Similarity, suggesting that while it may perform well in certain areas, it lacks the reliability and consistency of the leading models. The significant variation in its polygon indicates potential instability, making it less reliable for tasks requiring consistent output.

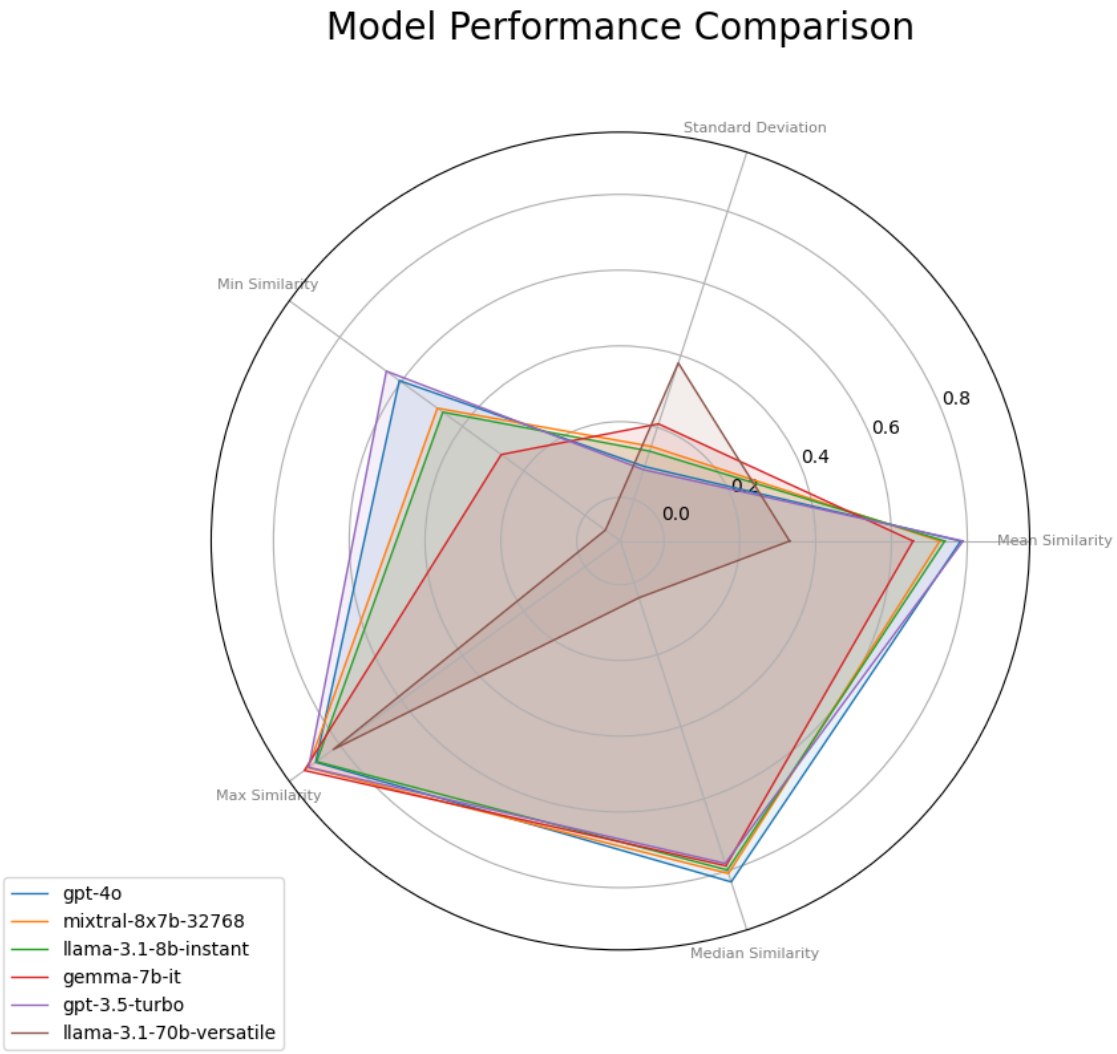


Figure 5.7: Model Performance Comparison

Overall Recap and Best Model Conclusion

After a thorough analysis of all the metrics, **gpt-3.5-turbo** emerges as the overall best model. It consistently scores high across all similarity metrics (Maximum, Mean, Minimum) while maintaining a low standard deviation, suggesting it provides both high accuracy and consistency. This balance makes it the most reliable and effective model for a wide range of applications. gemma-7b-it also performs exceptionally well, particularly in terms of maximum similarity, but its slightly higher standard deviation compared to gpt-3.5-turbo suggests that it might have less consistent performance across all tasks. gpt-4.0 is another strong contender, with performance closely aligned to gpt-3.5-turbo, though it slightly lags in certain metrics like maximum similarity.

In contrast, llama-3.1-70b-versatile consistently underperforms across all metrics, indicating it is not as well-suited for this task compared to the other models. Its low mean similarity, high standard deviation, and negative minimum similarity underscore its unreliability.

In conclusion, **gpt-3.5-turbo** is the most reliable and effective model according to the evaluated metrics, making it the best choice for applications that require consistent and accurate natural language processing capabilities.

5.3 Comparison with state of the art virtual assitants (Alexa vs Siri vs Elf)

In this section, we compare the responses of the **ELF Virtual Assistant** with those provided by popular virtual assistants, **Alexa** and **Siri**, to common questions that elderly users may ask. The goal is to highlight how ELF is designed specifically with the needs of older adults in mind, providing more tailored, accessible, and detailed responses compared to mainstream assistants.

Below, we present a sample of questions aleatory questions asked to each of the three systems, with screenshots showing how they respond.

By looking at these examples, we can see how ELF better addresses specific concerns of elderly users, such as clarity, user-friendliness, and focus on relevant details.

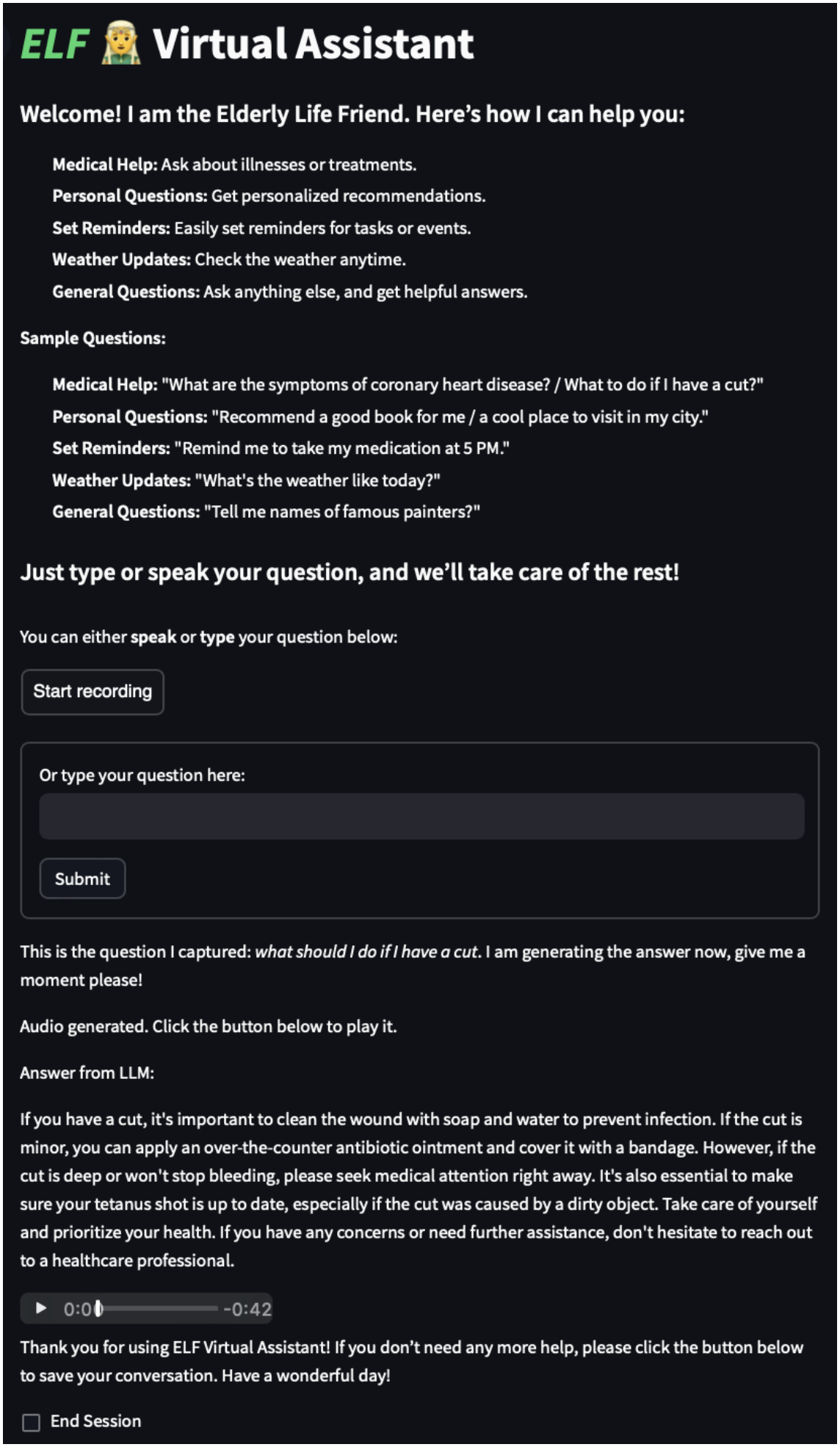
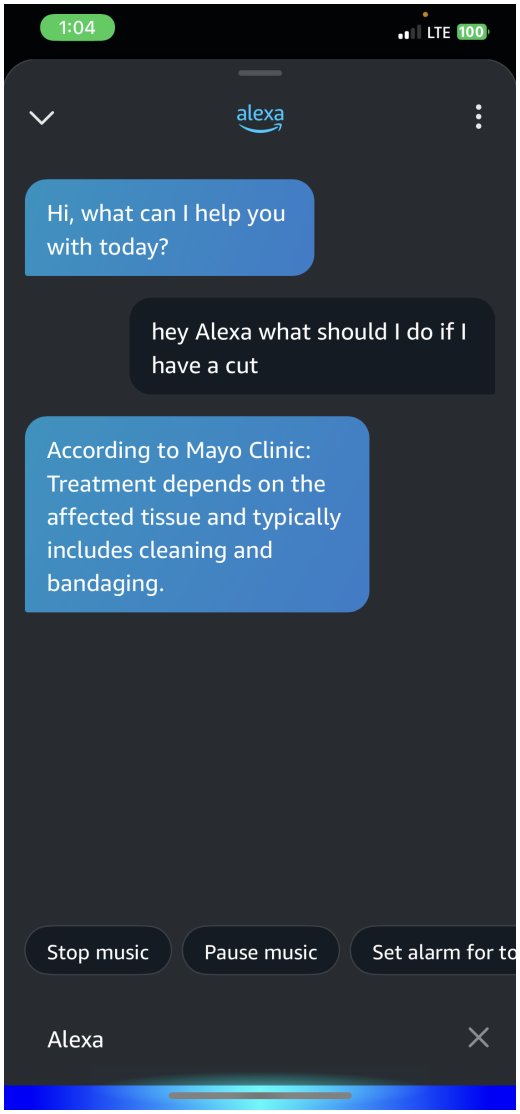
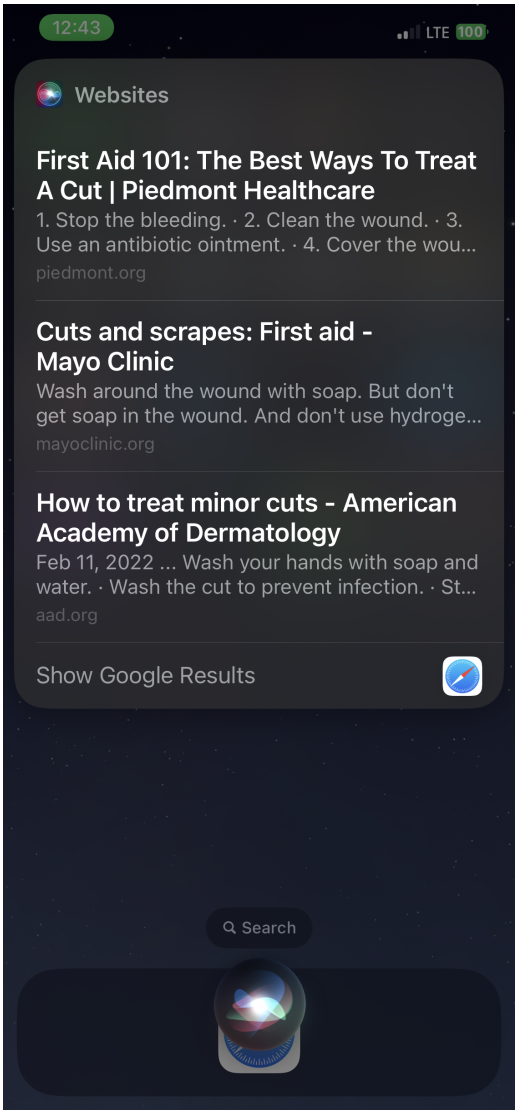


Figure 5.8: Elf Virtual Assitant interface in Streamlit


1:04
LTE 100
alexa
Hi, what can I help you with today?
hey Alexa what should I do if I have a cut
According to Mayo Clinic: Treatment depends on the affected tissue and typically includes cleaning and bandaging.
Stop music
Pause music
Set alarm for to
Alexa


12:43
LTE 100
Websites
First Aid 101: The Best Ways To Treat A Cut | Piedmont Healthcare
1. Stop the bleeding. · 2. Clean the wound. · 3. Use an antibiotic ointment. · 4. Cover the wou...
piedmont.org
Cuts and scrapes: First aid - Mayo Clinic
Wash around the wound with soap. But don't get soap in the wound. And don't use hydroge...
mayoclinic.org
How to treat minor cuts - American Academy of Dermatology
Feb 11, 2022 ... Wash your hands with soap and water. · Wash the cut to prevent infection. · St...
aad.org
Show Google Results
Search

In this example, only the answer is displayed since the top instructions remain constant in each question.

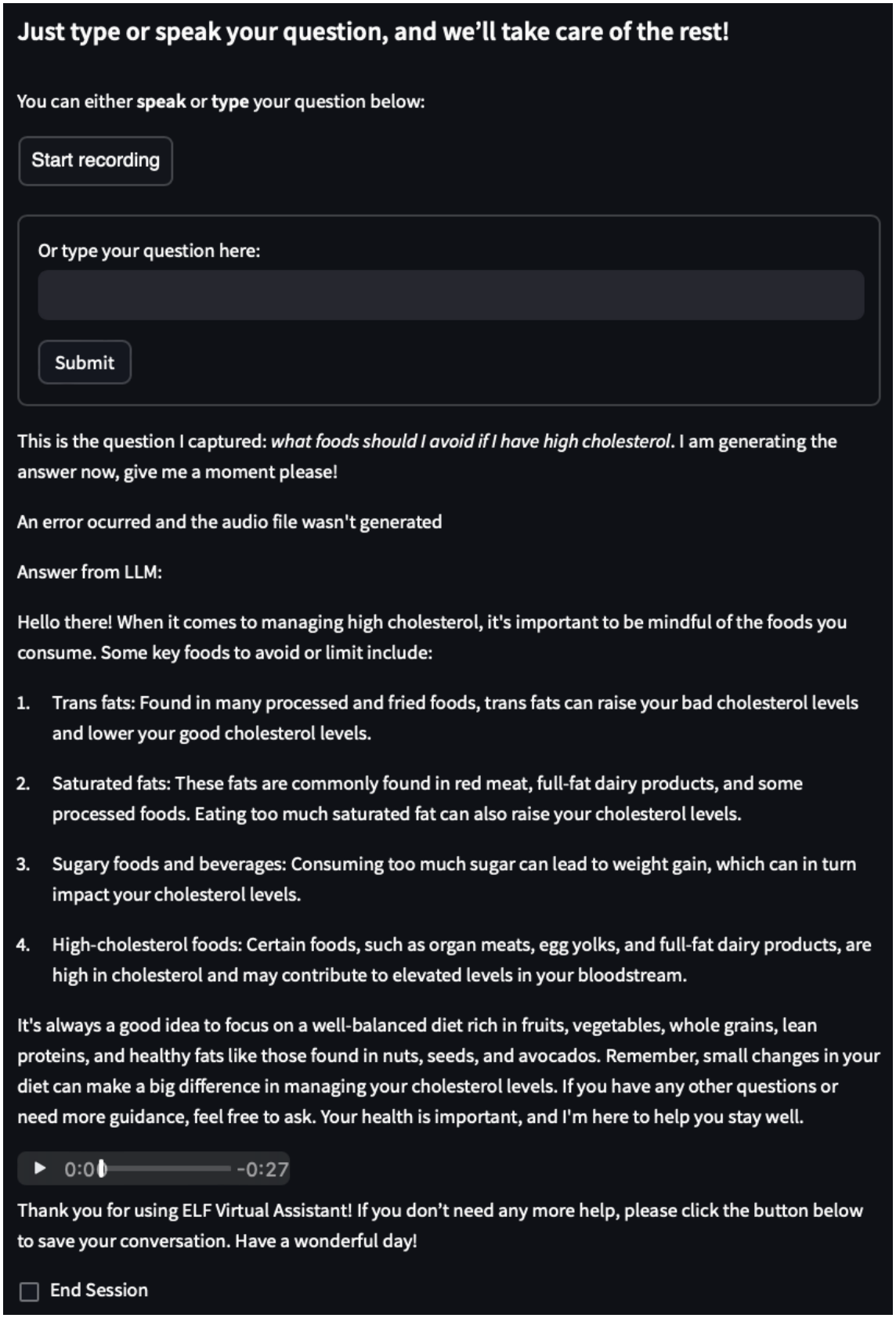
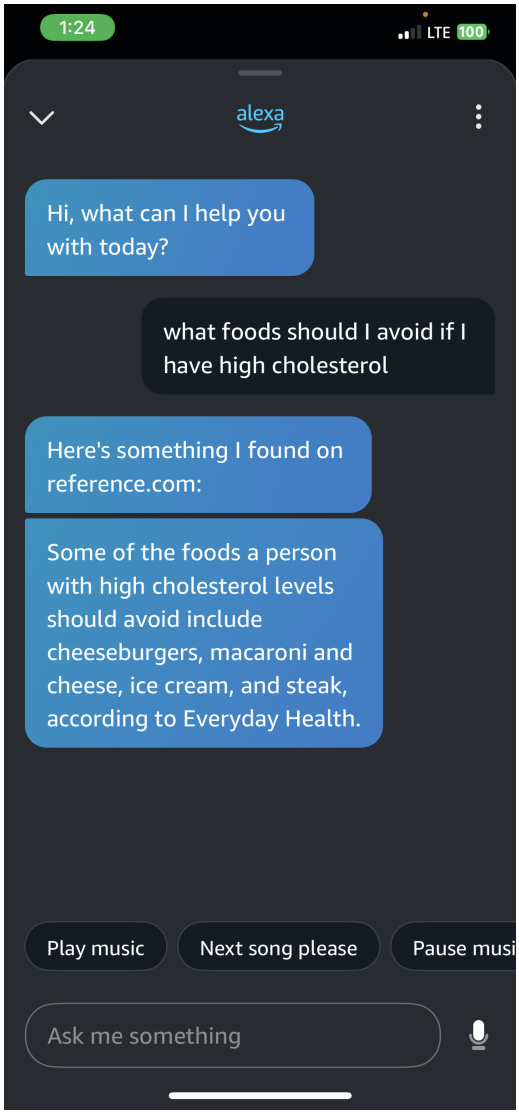
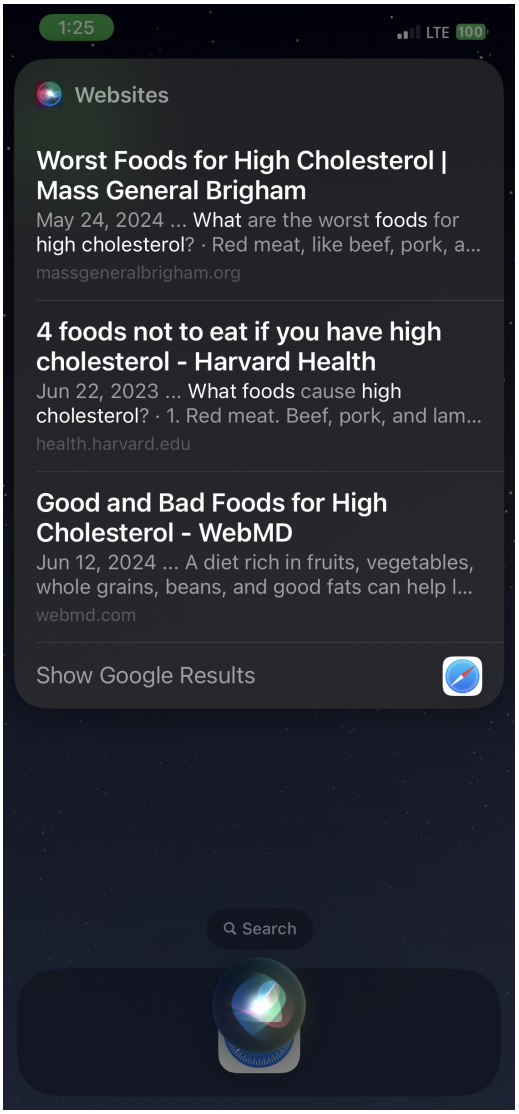


Figure 5.9: Cholesterol question example from ELF





This comparison highlights key differences between Siri, Alexa, and Elf, particularly in their handling of medical queries and personalized responses for elderly users. Siri frequently resorts to Google searches without providing direct answers, offering general search results rather than specific guidance. Similarly, Alexa gives basic responses, often lacking depth in medical information. Neither Siri nor Alexa offers personalized responses to individual queries, instead providing generic suggestions.

Elf, on the other hand, stands out by offering detailed, tailored answers that are specific to elderly care. Its medical responses are more in-depth and targeted compared to the general information provided by Siri and Alexa. Elf's personalization is also superior, adapting responses based on the user's preferences, lifestyle, and needs, making it particularly suited for elderly users who benefit from tailored interactions.

CHAPTER 6
Conclusions

This project embarked on an ambitious journey to develop a virtual assistant specifically designed to cater to the needs of elderly users. The primary objectives were to create a multifunctional assistant which incorporates a reliable medical advice module, enables personalized interactions, design a scalable architecture, optimize the retrieval-augmented generation (RAG) technique used, and evaluate various of the implemented large-scale language models. Each of these goals have been addressed with a significant level of success, reflecting a well-rounded approach to solving the problem outlined in the introduction.

First and foremost, the development of a comprehensive virtual assistant was achieved. The assistant integrates a range of functionalities including medical advice, personal recommendations, and reminders. This integration has proven to be highly effective in creating a robust support system tailored to the diverse needs of elderly users.

A key component of the system was the medical advice module, which was designed to offer accurate and timely health information. This module has been developed with a strong emphasis on reliability, ensuring that users receive sound advice from credible sources. The successful implementation of this feature has been instrumental in supporting better health management for elderly users.

Personalization was another critical objective achieved. The system's ability to offer tailored suggestions for various aspects of daily life ensured that the assistant could adapt to individual preferences and needs effectively.

In terms of system design, the focus was on creating a scalable architecture that allows for future growth and enhancement. The design implemented ensures that new features and functionalities can be integrated as technology evolves or as new needs emerge. The scalability of the system is a testament to its long-term viability and adaptability.

Evaluating various large-scale language models was another major task. Through evaluation, the most suitable model that balanced accuracy, responsiveness, and versatility was identified. This process was essential in ensuring that the virtual assistant could handle a wide range of queries effectively.

Reflecting on the project's development, several challenges emerged, particularly in integrating multiple functionalities into a single, cohesive system. To

overcome these challenges, an iterative approach was employed, involving frequent testing and refinement of each module.

One notable mistake encountered was the initial underestimation of the complexity involved in integrating diverse functionalities. This oversight led to delays and necessitated a reevaluation of the system architecture. Future projects would benefit from a more detailed initial planning phase to better anticipate and address integration challenges.

This project has been a profound learning experience. Professionally, it has provided me with valuable expertise in developing scalable systems and optimizing advanced language models. Personally, I have learned the importance of iterative testing and adaptability in problem-solving. The project required me to acquire new skills and knowledge in areas not covered in my degree, such as system architecture design.

On a final note, this project has successfully met its objectives by delivering a versatile and adaptable virtual assistant for elderly users. The experience has provided significant insights into both the technological and practical aspects of system development, setting the stage for future advancements and improvements. The project not only achieved its goals but also contributed to a deeper understanding of the integration of cutting-edge technologies in solving real-world problems.

6.1 Legacy

The legacy of this project encompasses a multifaceted contribution to both the field of technology and the quality of life for elderly users. By developing a virtual assistant specifically designed to meet the diverse needs of seniors, this project has achieved its objectives and established a framework for future advancements in elderly care.

At the heart of this project is the enhanced quality of life it offers to its users. The virtual assistant is designed to address a wide range of needs, significantly improves the daily lives of elderly individuals. Such comprehensive support is crucial for seniors, enabling them to navigate their daily lives with greater ease and confidence.

One of the project's key contributions is its commitment to open access. All project data, code, and supplementary documentation have been made available through Github. By providing access to these resources, the project not only allows other researchers and developers to replicate the system and analysis, but also encourages them to build upon the work. This commitment to reproducibility is vital for validating the project's findings and promoting further research in the field. It underscores the project's dedication to scientific rigor and its role in advancing knowledge within the domain of virtual assistants. The project's repository, hosted at , includes detailed documentation to facilitate understanding and use, ensuring that the project's findings and methodologies can be effectively utilized and extended.

In terms of impact, the project promises several benefits in the short, medium, and long term. In the immediate future, the virtual assistant provides elderly users with a valuable tool for managing their daily activities and health. The system's personalized recommendations and timely reminders have the potential to improve user satisfaction and well-being significantly. Over the medium term, the open-access resources from this project will facilitate further research and development in the field. Insights gained from this project can inspire the creation of similar systems, leading to enhanced features and broader applications. In the long term, the project has the potential to influence the broader landscape of elderly care technology. As the virtual assistant evolves and integrates with new technologies, it could become a central component of a comprehensive care system for seniors, fostering further innovations and advancements in assistive technologies.

6.2 Relationship of the work carried out with the studies completed

The development of my Final Degree Project has been a deeply reflective exercise in synthesizing and applying the extensive knowledge acquired throughout my Data Science Bachelor's program. This project embodies the practical application of the theoretical and technical skills gained from a comprehensive curriculum designed to provide a robust understanding of the field.

Courses such as Programming Fundamentals, Programming, and Data Structures were pivotal in equipping me with the technical skills necessary for developing and implementing software solutions. The ability to write efficient code and manage data structures effectively was indispensable for handling various aspects of the virtual assistant project, from data processing to system integration. These skills ensured that the project's software components were both robust and adaptable.

The project also benefited significantly from the knowledge gained in Statistical Models for Decision Making I and II. These courses provided a strong foundation in both theoretical and applied statistics, which was crucial for building and validating the virtual assistant's system.

In particular, the coursework on Data Acquisition and Transmission equipped me with practical skills such as web scraping and working with APIs. The ability to effectively scrape web data and utilize APIs enabled me to source information and acquire the necessary data for the project and the functioning of the system.

Moreover, the knowledge gained from Natural Language and Information Retrieval, as well as Deep Learning for Text Data, provided me with a deep understanding of large language models (LLMs). These courses offered insights into the mechanisms behind natural language processing.

The experience gained from the practical courses Projects I, II, and III was particularly relevant. These courses offered hands-on experience in understanding data, integrating and preparing data, and conducting comprehensive data analysis. They simulated real-world scenarios, helping me develop essential skills

in project management, data integration, and the application of analytical and machine learning techniques. This practical background was crucial for the successful execution of the project, allowing me to effectively manage and coordinate various aspects of it.

Throughout my studies, I have sharpened my critical problem-solving and analytical skills. The complexity of the virtual assistant project required innovative solutions, reflecting my capacity to apply these competences effectively. Additionally, managing the project involved meticulous planning, organization, and execution of various development phases. This experience demonstrated my ability to oversee a project from conception to completion, showcasing the skills emphasized in courses like Visualization and Project Management.

The alignment between my studies and the work developed in the Final Degree Project is evident in the comprehensive application of knowledge across various areas of Data Science. By successfully applying a broad range of skills and addressing real-world challenges, the project stands as a testament to the depth and breadth of my education and the ability to translate academic knowledge into impactful solutions.

6.3 Future work

As with any complex project, there are always areas that could benefit from further development and refinement. While the virtual assistant project has achieved significant milestones, several aspects remain open for future exploration and enhancement. This section outlines potential avenues for future work, including additional functionalities, refinements, and extensions, as well as considerations for avoiding certain paths.

One notable area for future improvement is the incorporation of additional modules to enhance the virtual assistant's capabilities. For instance, integrating cognitive games specifically designed for elderly users could provide valuable mental stimulation and promote cognitive health. Such games could be tailored to the user's cognitive level and interests, offering both entertainment and therapeutic benefits. Adding this functionality would support the overall well-being of elderly users.

Another potential avenue for development is the fine-tuning of the model using specialized datasets. The Corpus of Interactions between Seniors and an Empathic Virtual Coach, developed within the EMPATHIC project and funded by the European Union's Horizon 2020 Research and Innovation program, presents a promising resource. This corpus includes video recordings of conversations with both a Wizard of Oz (WOZ) and an automatic dialogue system, covering a range of topics in Spanish, French, and Norwegian. Fine-tuning the virtual assistant with this dataset could enhance its ability to engage empathetically and effectively in multiple languages. However, it is important to note that this dataset is costly and involves high training expenses, which were beyond the scope of the current project. Despite the potential benefits, such an investment should be carefully considered against budget constraints and the anticipated return on investment.

Additionally, there is scope to expand the virtual assistant's functionality by incorporating features related to medication management. This could include tracking medication schedules, providing reminders for medication intake, checking for potential drug interactions and include detailed information about medicine compounds from the NHS Medicines A-Z page[26]. Integrating such a module would enhance the assistant's utility by potentially preventing adverse drug interactions and improving adherence to prescribed treatments.

While these future directions are promising, there are also areas where it may be prudent to exercise caution. For example, pursuing overly ambitious technological integrations or expanding functionalities without adequate testing could compromise the system's stability and user experience. It is essential to ensure that any new features or enhancements are thoroughly evaluated and validated before deployment to maintain the system's reliability and effectiveness.

Bibliography

[1] Ritchie, H., Roser, M. *Population by Age Group with Projections*. Our World in Data. Available at: https://ourworldindata.org/grapher/population-by-age-group-with-projections. Accessed August 28, 2024.

[2] World Health Organization. *Ageing and Health Fact Sheets*. Available at: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed August 28, 2024.

[3] United Nations Department of Economic and Social Affairs. *World Social Report 2023: Leaving no one behind in an ageing world*. Available at: https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2023/01/WSR_2023_Chapter_Key_Messages.pdf. Accessed August 28, 2024.

[4] Statista Research Department. *Seniors in Spain - Statistics and facts*. Available at: https://www.statista.com/topics/8117/seniors-in-spain/#:~:text=This%20represents%20an%20increase%20of,for%20both%20men%20and%20women.. Accessed August 28, 2024.

[5] Worldometers. *Life Expectancy of the World Population*. Available at: https://www.worldometers.info/demographics/life-expectancy/. Accessed August 28, 2024.

[6] Antonio Abellán, Spanish National Research Council (CSIC) . *Demographic Indicators*. Available at: http://envejecimiento.csic.es/documentos/documentos/imserso-agespacap1-01.pdf. Accessed August 28, 2024.

[7] Matthew Hoy. *Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants*. Medical Reference Services Quarterly, 2018, 37(1), 81-88. DOI 10.1080/02763869.2018.1404391.

[8] **Virtual Assistant**. *Wikipedia, The Free Encyclopedia* [online]. Last modified August 25, 2024 [viewed August 31, 2024]. Available from: https://en.wikipedia.org/wiki/Virtual_assistant

[9] **Amazon Alexa**. *Wikipedia, La enciclopedia libre* [online]. Last modified August 30, 2024 [viewed August 31, 2024]. Available from: https://es.wikipedia.org/wiki/Amazon_Alexa

[10] Wallis, Jerry. *The Technology In Amazon Alexa – The Tech Behind Series* [online]. 2024. Available from: https://intuji.com/the-tech-behind-amazon-alexa/ [Accessed: 2024-08-31].

[11] Siri. *Wikipedia, The Free Encyclopedia* [online]. Last modified August 25, 2024 [viewed August 31, 2024]. Available from: https://en.wikipedia.org/wiki/Siri.

[12] ECHAlliance. *Elsa, a Smart Virtual Assistant for a Specific Age Group of People* [online]. 2024. Available from: https://echalliance.com/news/elsa-a-smart-virtual-assistant-for-a-specific-age-group-of-people/ [Accessed: 2024-08-31].

[13] Broadbent, E., Loveys, K., Ilan, G., Chen, G., Chilukuri, M. M., Boardman, S. G., Doraiswamy, P. M., & Skuler, D. (2024). ElliQ, an AI-Driven Social Robot to Alleviate Loneliness: Progress and Lessons Learned. *JAR Life*, 13, 22–28. doi:10.14283/jarlife.2024.2. PMID: 38449726; PMCID: PMC10917141.

[14] Abeer Alessa and Hend Al-Khalifa, *Towards Designing a ChatGPT Conversational Companion for Elderly People*, arXiv:2304.09866 [cs.HC], submitted on 18 April 2023. Available at: https://doi.org/10.48550/arXiv.2304.09866.

[15] Python Documentation, https://www.python.org/doc/

[16] `pandas` Documentation, https://pandas.pydata.org/docs/

[17] `numpy` Documentation, https://numpy.org/doc/

[18] `scikit-learn` Documentation, https://scikit-learn.org/stable/

[19] `matplotlib` Documentation, https://matplotlib.org/stable/contents.html

[20] `seaborn` Documentation, https://seaborn.pydata.org/

[21] `plotly` Documentation, https://plotly.com/python/

[22] `beautifulsoup4` Documentation, https://pypi.org/project/beautifulsoup4/

[23] Valtolina, Stefano and Hu, Lili. *Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness*. In *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 2021. Available at: https://api.semanticscholar.org/CorpusID:235812785.

[24] Lewis, Patrick; Perez, Ethan; Piktus, Aleksandra; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Naman; Küttler, Heinrich; Lewis, Mike; Yih, Wen-tau; Rocktäschel, Tim; Riedel, Sebastian; Kiela, Douwe. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Accepted at NeurIPS 2020. [Online] Available at: https://doi.org/10.48550/arXiv.2005.11401. [Accessed 7 August 2024].

[25] Gao, Yunfan; Xiong, Yun; Gao, Xinyu; Jia, Kangxiang; Pan, Jinliu; Bi, Yuxi; Dai, Yi; Sun, Jiawei; Wang, Meng; Wang, Haofen. *Retrieval-Augmented Generation for Large Language Models: A Survey*. [Online] Available at: https://doi.org/10.48550/arXiv.2312.10997. [Accessed 7 August 2024].

[26] NHS, "List of medicines," UK Government. [Online]. Available: https://www.nhs.uk/medicines/. [Accessed: 07-Aug-2024].

[27] B. Knowles and V. L. Hanson, "Older Adults' Deployment of 'Distrust"', ACM Transactions on Computer-Human Interaction, vol. 25, no. 4, Article 21, pp. 1-25, Aug. 2018. [Online]. Available: https://doi.org/10.1145/3196490. [Accessed: 07-Aug-2024].

[28] T. Khete, "Dataset for medical-related chatbots," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/tusharkhete/dataset-for-medical-related-chatbots. [Accessed: 07-Aug-2024].

APPENDIX A
Questionnaire for information collection in order to personalize the system

1. **Basic Information:**
    - Full Name
    - Age
    - Date of Birth
    - Gender
    - Place of Residence
    - Place of Birth
    - Marital Status
    - Number of children/Grandchildren (if applicable)
2. **Family Background:**
    - What role do your family members play in your life?
    - Do you have any special anecdotes about your family that you would like to share?
3. **Interests and Hobbies:**
    - What do you like to do in your free time?
    - Do you have any favorite hobbies or pastimes?
    - What has been your most memorable trip or experience?
    - What trip would you like to take in the future?
4. **Work and Education History:**
    - What was your profession?
    - If you were a homemaker, what was the best part of that role?
    - What was your most significant job?

- What did you enjoy about that job?
- Do you have any interesting memories or stories from your student or work life that you would like to share?
- What did you study?

5. **Life Experiences:**
   - Is there any major event in your life that had a significant impact on you?
   - What experiences in your life have been the most important to you?
   - What do you consider to be your greatest achievements?
6. **Tastes and Preferences:**
   - What is your favorite food?
   - What kind of movies do you like?
   - If you have a favorite movie, please share it.
   - Can you tell me about a book you enjoyed?
   - Tell me about a song you like.
   - What is your favorite place in the world?
   - Do you play any musical instruments?
7. **Health and Well-being:**
   - Do you have any health concerns that you would like to share?
   - Do you practice regular exercise?
   - What is your favorite sport?
   - How do you manage stress or worries?
   - How do you take care of your emotional well-being?
   - Do you often engage in activities with others, such as going to the movies, theater, excursions, trips, etc.?
   - Do you consider your diet to be healthy?
   - Do you usually sleep well?
   - Do you regularly engage in activities that keep your mind stimulated, such as puzzles, crosswords, reading, workshops, etc.?
8. **Aspirations and Goals:**
   - What would you like to achieve in the future?
   - Is there anything you haven't done yet but have always wanted to do?
9. **Memories and Nostalgia:**
   - What is your most treasured memory?
   - Do you have any stories or anecdotes from your childhood that you would like to share?

10. **Opinions and Values:**
    - Is there a current issue that particularly concerns you?
    - What are your thoughts on this issue?
    - What values are most important to you?
    - Which of these values do you believe you possess?
11. **Self-Esteem and Self-Perception:**
    - What positive adjectives do you think define you?
    - What are you good at doing?

APPENDIX B
Extract of NHS JSON and Kaggle JSON

NHS JSON


```
{
  "dengue": {
    "other-info": "Information: Check before you travel. You
       can check the risks for a country you're travelling to
       on the TravelHealthPro website.",
    "causes": "Check if you're at risk of dengue. You can get
       dengue if you're bitten by an infected mosquito. The
       mosquitoes that carry the dengue virus bite during the
       day. Dengue is very common in certain parts of the
       world. It's often found in tropical areas including:
       parts of Africa and Asia, Central and South America,
       the Caribbean, the Pacific islands, some southern areas
        of North America. There's also a risk of getting
       dengue at certain times of the year (spring to November
       ) in parts of southern Europe. European countries where
        dengue has been found include: Croatia, France, Italy,
        Spain, Portugal and Madeira. Dengue is not found in
       the UK and you cannot catch it from another person.",
    "symptoms": "Symptoms of dengue. Dengue does not always
       cause symptoms. If you do have symptoms, they usually
       start 4 to 10 days after being bitten by an infected
       mosquito. Some dengue symptoms are similar to flu. They
        include: a high temperature, a severe headache, pain
       behind your eyes, muscle and joint pain, feeling or
       being sick, swollen glands, a blotchy rash made up of
       flat or slightly raised spots     this can affect large
        areas of your body.",
    "complications": "Severe dengue. Some people get a more
       severe type of dengue a few days after they first
       started feeling ill, but this is rare. You may start to
        feel better with your temperature returning to normal,
        but about 24 to 48 hours later you may get more
       serious symptoms. Symptoms of severe dengue include:
```

```
      severe tummy pain, repeatedly being sick, fast
      breathing, bleeding gums or nose, extreme tiredness (
      fatigue), being unable to relax (restlessness), blood
      in your vomit or poo.",
    "medical-treatments": "Treatments for dengue. Most people
      with dengue feel better in a few days. There is no
      treatment for dengue, but you can help ease your
      symptoms by: resting, drinking plenty of fluids, taking
       paracetamol to help bring down your temperature and
      ease any pain. Do not take anti-inflammatory
      painkillers like ibuprofen or aspirin. These can cause
      bleeding problems if you have dengue. If you have
      severe dengue, you'll need to stay in hospital until
      you recover.",
    "prevention": "How to prevent dengue. There's no vaccine
      available in the UK that you can have to prevent dengue
       before travelling to a country where there's a risk of
       infection. If you're at increased risk of dengue, you
      should avoid travelling to countries where the
      infection is found. At-risk groups include those who:
      are very young, are over 65, are pregnant, have a
      weakened immune system. If  y o u re  in a country where
      dengue is found, the best way to prevent infection is
      to avoid being bitten by mosquitoes. Do: wear long-
      sleeved clothing and trousers to cover your arms and
      legs, particularly during early morning and early
      evening; use insect repellent on your skin (ideally one
       that contains the ingredient DEET); close windows and
      doors whenever possible, or use blinds or screens;
      sleep under a mosquito net treated with insecticide,
      including during the day.",
  },
  "dental-abscess": {
    "other-info": "",
    "symptoms": "Check if you have a dental abscess. Signs of
      a dental abscess include: intense toothache or pain in
      your gums, redness inside the mouth, or outside the
      mouth on the face or jaw, sensitivity to hot or cold
      food and drink in the affected area, a bad taste in
      your mouth, difficulty opening your mouth and chewing
      food, a swollen face or jaw, a high temperature.",
    "self-care": "How to ease the pain caused by a dental
      abscess. There are some things you can do to help with
      the pain until you can get dental treatment. Do: take
      painkillers, such as ibuprofen or paracetamol (children
       under 16 should not take aspirin); eat soft foods like
       soup, scrambled eggs, mashed potatoes and yoghurt; use
       a soft toothbrush to clean your teeth; avoid having
      sugary, or very hot or cold, foods and drinks.",
    "medical-treatments": "Treatment for a dental abscess.
      Dental abscesses are usually treated by a dentist. The
      dentist will drain away the pus. If a problem with your
```

```
        tooth has caused the abscess, you may need root canal
       treatment, or the tooth may be removed. You'll be given
        a local anaesthetic, so you do not feel any pain. You
       may be offered painkillers to take for a few days after
        treatment and may also be given antibiotics.",
15    "causes": "Causes of a dental abscess. Dental abscesses
       usually happen when pus builds up beneath your teeth or
        gums when you have an infection in your mouth. You may
        get a dental abscess if: you have a tooth that has not
        grown out of your gums properly (an impacted tooth),
       you have tooth decay or gum disease, you have injured
       your teeth, gums or mouth, you're having radiotherapy
       or chemotherapy.",
16    "prevention": "How to prevent dental abscesses. There are
       some things you can do to help prevent dental abscesses
       . Do: brush your teeth with fluoride toothpaste at
       least twice a day - spit after brushing, do not rinse;
       clean in between your teeth every day using floss or
       interdental brushes; replace your toothbrush every 1 to
        3 months; have regular dental check-ups so problems
       can be spotted early.  D o n t : do not use mouthwash
       straight after brushing your teeth; do not smoke."
17  },
18  "toothache": {
19    "If you need to see a dentist": "If you need to see a
       dentist. To see a dentist in an emergency or out of
       hours: call your dentist     if they're closed, their
       answerphone may tell you what to do. If you do not have
        a dentist or cannot get an emergency appointment: call
        111 or visit 111 online     they can advise you what
       to do; find a dentist near you     ask if you can have
       an emergency appointment. You may have to pay for your
       appointment. Find out more about NHS dental charges.",
20    "self-care": "How to ease toothache while waiting for an
       appointment. Do: take painkillers, like ibuprofen or
       paracetamol (children under 16 should not take aspirin)
            a pharmacist can advise you; use a pain-relieving
       gel for your mouth     this can be bought from
       pharmacies or supermarkets; try rinsing your mouth with
        salt water (children should not try this); eat soft
       foods, like yoghurt or scrambled eggs, and try to avoid
        chewing with the sore tooth; use a soft toothbrush and
        avoid flossing around the sore tooth.  D o n t : do not
       eat foods that are sweet, very hot or very cold; do not
        smoke     it can make some dental problems worse. How
       to rinse with salt water: Dissolve half a teaspoon of
       salt in a glass of warm water     warm water helps salt
        dissolve. Rinse your mouth with the solution, then
       spit it out     do not swallow it. Repeat as often as
       you like. Children should not try rinsing their mouth
       with salt water in case they swallow it.",
```

```
    "causes": "Causes of toothache. Toothache can be caused by
       : tooth decay, a dental abscess, a cracked or damaged
       tooth, a loose or broken filling, an infection     this
        often happens when a tooth (such as a wisdom tooth)
       has broken the skin, but does not have enough room to
       fully come through, problems with your braces, gum
       disease, grinding your teeth. Tooth pain can also be
       caused by having sensitive teeth.",
    "other-info": "The Oral Health Foundation has more
       information about sensitive teeth.",
    "prevention": "How to prevent toothache. The best way to
       prevent toothache is to keep your teeth and gums as
       healthy as possible. To do this: have regular dental
       check-ups; cut down on sugary foods and drinks     only
        have them as an occasional treat at mealtimes; brush
       your teeth twice a day for about 2 minutes with a
       fluoride toothpaste; clean between your teeth using
       floss or an interdental brush every day to remove food,
        debris and plaque.",
    "More information": "Information: More information. Taking
        care of your teeth and gums. Looking after your
       children's teeth. Looking after your baby's teeth."
  },
  "dentures": {
    "other-info": "Dentures are removable false teeth made of
       acrylic (plastic), nylon or metal. They fit snugly over
        the gums to replace missing teeth and eliminate
       potential problems caused by gaps. Gaps left by missing
        teeth can cause problems with eating and speech, and
       teeth either side of the gap may grow into the space at
        an angle. Sometimes all the teeth need to be removed
       and replaced. You may therefore need either: complete
       dentures (a full set)     which replace all your upper
       or lower teeth, or partial dentures     which replace
       just 1 tooth or a few missing teeth. Dentures may help
       prevent problems with eating and speech. If you need
       complete dentures, they may also improve the appearance
        of your smile and give you confidence. It's also
       possible that dentures might not give you the result
       you hope for. Discuss plans openly with your dentist
       before you agree to go ahead.",
    "How dentures are fitted": "How dentures are fitted.
       Complete dentures: A full denture will be fitted if all
        your upper or lower teeth need to be removed or you're
        having an old complete denture replaced. The denture
       will usually be fitted as soon as your teeth are
       removed, which means you won't be without teeth. The
       denture will fit snugly over your gums and jawbone. But
        if you have dentures fitted immediately after the
       removal of several teeth, the gums and bone will alter
       in shape fairly quickly and the dentures will probably
       need relining or remaking after a few months.
```

```
    Occasionally, your gums may need to be left to heal and
     alter in shape for several months before dentures can
    be fitted. You can either see a dentist or a qualified
    clinical dental technician to have your dentures made
    and fitted. The difference between them is that a:
    dentist will take measurements and impressions (moulds)
     of your mouth, and then order your full or partial
    dentures from a dental technician; clinical dental
    technician will provide a full set of dentures directly
     without you having to see your dentist (although you
    should still have regular dental check-ups with your
    dentist). A trial denture will be created from the
    impressions taken of your mouth. The dentist or
    clinical dental technician will try this in your mouth
    to assess the fit and for you to assess the appearance.
     The shape and colour may be adjusted before the final
    denture is produced. Partial dentures: A partial
    denture is designed to fill in the gaps left by one or
    more missing teeth. It's a plastic, nylon or metal
    plate with a number of false teeth attached to it. It
    usually clips onto some of your natural teeth via metal
     clasps, which hold it securely in place in your mouth.
     It can easily be unclipped and removed. Occasionally,
    the clips can be made of a tooth- or gum-coloured
    material, although this type of clip isn't always
    suitable because it tends to be more brittle than metal
    . Your dentist can measure your mouth and order a
    partial denture for you, or you can see a qualified
    clinical dental technician, who can provide a partial
    denture for you directly after you have first seen your
     dentist for a treatment plan and certificate of oral
    health. The Oral Health Foundation website has more
    information and advice about bridges and partial
    dentures, including which type of denture (metal or
    plastic) is best for you. A fixed bridge is an
    alternative to a partial denture and may be suitable
    for some people. Crowns are put on the teeth either
    side of the gap and joined together by a false tooth
    that's put in the gap.",
29  "Looking after your dentures": "Looking after your
    dentures. Dentures may feel a bit strange to begin with
    , but you'll soon get used to wearing them. At first,
    you may need to wear your dentures all the time,
    including while sleeping. Your dentist or clinical
    dental technician will advise you on whether you should
     remove your dentures before you go to sleep. It isn't
    always necessary to remove your dentures at night, but
    doing so can allow your gums to rest as you sleep. If
    you remove your dentures, they should be kept moist
     for example, in water or a polythene bag with some
    dampened cotton wool in it, or in a suitable overnight
    denture-cleaning solution. This will stop the denture
```

material drying out and changing shape. Dental hygiene: Keeping your mouth clean is just as important when you wear dentures. You should brush your remaining teeth, gums and tongue every morning and evening with fluoride toothpaste to prevent tooth decay, gum disease and other dental problems. Read more about how to keep your teeth clean. Cleaning dentures: It's important to regularly remove plaque and food deposits from your dentures. This is because unclean dentures can also lead to problems, such as bad breath, gum disease, tooth decay and oral thrush. Clean your dentures as often as you would normal teeth (at least twice a day: every morning and night). You should: brush your dentures with toothpaste or soap and water before soaking them to remove food particles; soak them in a fizzy solution of denture-cleaning tablets to remove stains and bacteria (follow the manufacturer's instructions); brush them again as you would your normal teeth (but don't scrub them too hard). Dentures may break if you drop them, so you should clean them over a bowl or sink filled with water, or something soft like a folded towel. Find more information on denture cleaning, on the Oral Health Foundation website. Eating with dentures: When you first start wearing dentures, you should eat soft foods cut into small pieces and chew slowly, using both sides of your mouth. Avoid chewing gum and any food that's sticky, hard or has sharp edges. You can gradually start to eat other types of food until you're back to your old diet. Never use toothpicks. Denture adhesive: If your dentures fit properly, you shouldn't necessarily need to use denture fixative (adhesive). But if your jawbone has shrunk significantly, adhesive may be the only way to help retain your dentures. Your dentist or clinical dental technician will advise you if this is the case. At first, some people feel more confident with their dentures if they use adhesive. Follow the manufacturer's instructions and avoid using excessive amounts. Adhesive can be removed from the denture by brushing with soap and water. Remnants of adhesive left in the mouth may need to be removed with some damp kitchen roll or a clean damp flannel.",
"When to see your dentist": "When to see your dentist. You should continue to see your dentist regularly if you have dentures (even if you have complete dentures) so they can check for any problems. Your dentures should last several years if you take good care of them. But your gums and jawbone will eventually shrink, which means the dentures may not fit as well as they used to and can become loose, or they may become worn. See your dentist as soon as possible if: your dentures click when you're talking; your dentures tend to slip, or you

```
        feel they no longer fit properly; your dentures feel
       uncomfortable; your dentures are visibly worn; you have
        signs of gum disease or tooth decay, such as bleeding
       gums or bad breath. If poorly fitting or worn dentures
       aren't replaced, they can cause great discomfort and
       lead to mouth sores, infections or problems eating and
       speaking.",
31    "How much dentures cost on the NHS": "How much dentures
       cost on the NHS. Having dentures fitted is a band 3
       treatment. Read about understanding NHS dental charges
       for the different bands and how to get help with dental
        costs."
32  },
33  "detached-retina-retinal-detachment": {
34    "other-info": "Find out more about posterior vitreous
       detachment from RNIB.",
35    "symptoms": "Check if you have a detached retina. Symptoms
        of a detached retina include: floaters (dots and lines
       ) or flashes of light in your eye, a dark \"curtain\"
       or shadow in your vision, changes to your eyesight,
       such as blurred vision.",
36    "medical-treatments": "Treatment for a detached retina.
       You'll be referred to hospital for surgery if tests
       show your retina may be detached or has started to come
        away (retinal tear). Surgery will usually stop your
       vision getting worse. What happens during surgery for a
        detached retina or tear: Surgery to re-attach the
       retina or fix a retinal tear may involve: removing and
       replacing the jelly inside your eye (vitrectomy);
       attaching a small band around your eye to push the wall
        of your eye and retina closer together (scleral
       buckling); injecting a bubble of gas into your eye to
       push the retina against the back of your eye (pneumatic
        retinopexy); sealing the tear in your retina with
       laser or freezing treatment (cryotherapy). It's usually
        done with local anaesthetic, so you're awake but your
       eye is numbed. You do not normally need to stay in
       hospital overnight. You may be asked to lie or sit in a
        particular position for up to 7 days after the surgery
       . This is so that your retina is in the correct
       position to help it heal.",
37    "recovery": "Recovering from a detached retina. Recovery
       time after surgery for a detached retina varies. But as
        a general guide, for 2 to 6 weeks after surgery: your
       vision may be blurry; your eye may be sore and red
       take paracetamol if you need to; you may need to take
       time off work; you may not be able to drive; you may
       need to avoid flying (if you've had a bubble of gas put
        into your eye). Most people are eventually able to
       return to all their normal activities. Important: Call
       the hospital or go to A&E if the pain, redness or
```

```
        blurriness gets worse after surgery. You may need
        further treatment.",
38      "causes": "Causes of a detached retina. A detached retina
        is usually caused by changes to the jelly inside your
        eye, which can happen as you get older. This is called
        posterior vitreous detachment (PVD). It's not clear
        exactly why PVD can lead to retinal detachment in some
        people and there's nothing you can do to prevent it.
        But it's more likely to happen if you: are short-
        sighted; have had an eye operation (such as cataract
        surgery); have had an eye injury; have a family history
         of retinal detachment."
39    }
40  }
```

Kaggle JSON Snippet

```
1   {
2       "intents": [
3           {
4               "tag": "Cuts",
5               "patterns": ["What to do if Cuts?", "How to cure
                   Cuts?", "Which medicine to apply for Cuts?", "
                   what to apply on cuts?", "Cuts"],
6               "responses": ["Wash the cut properly to prevent
                   infection and stop the bleeding by applying
                   pressure for 1-2 minutes until bleeding stops.
                   Apply Petroleum Jelly to make sure that the
                   wound is moist for quick healing. Finally cover
                    the cut with a sterile bandage. Pain relievers
                    such as acetaminophen can be applied."],
7               "context_set": ""
8           },
9           {
10              "tag": "Abrasions",
11              "patterns": ["how do you treat abrasions?", "Do
                   Abrasions cause scars?", "Abrasions", "what to
                   do if abrasions?", "Which medicine to apply for
                    abrasions?", "How to cure abrasions?"],
12              "responses": ["Begin with washed hands. Gently
                   clean the area with cool to lukewarm water and
                   mild soap. Remove dirt or other particles from
                   the wound using sterilized tweezers. For a mild
                    scrape thats not bleeding, leave the wound
                   uncovered. If the wound is bleeding, use a
                   clean cloth or bandage, and apply gentle
                   pressure to the area to stop any bleeding.
                   Cover a wound that bled with a thin layer of
                   topical antibiotic ointment, like Bacitracin,
                   or a sterile moisture barrier ointment, like
```

```
                Aquaphor. Cover it with a clean bandage or
                gauze. Gently clean the wound and change the
                ointment and bandage once per day. Watch the
                area for signs of infection, like pain or
                redness and swelling. See your doctor if you
                suspect infection."],
            "context_set": ""
        },
        {
            "tag": "stings",
            "patterns": ["How do you treat Sting?", "Stings",
                "What to do if you get a sting?", "Which
                medicine to apply if sting?"],
            "responses": ["Remove any stingers immediately.
                Some experts recommend scraping out the stinger
                 with a credit card. Applying ice to the site
                may provide some mild relief. Apply ice for 20
                minutes once every hour as needed. Wrap the ice
                 in a towel or keep a cloth between the ice and
                 skin to keep from freezing the skin. Taking an
                 antihistamine such as diphenhydramine (
                Benadryl) or a nonsedating one such as
                loratadine (Claritin) will help with itching
                and swelling. Take acetaminophen (Tylenol) or
                ibuprofen (Motrin) for pain relief as needed.
                Wash the sting site with soap and water.
                Placing hydrocortisone cream on the sting can
                help relieve redness, itching, and swelling."],
            "context_set": ""
        },
        {
            "tag": "Splinter",
            "patterns": ["How to remove Splinters", "How to
                cure Splinters?", "What to do if I have
                splinters?", "How do you bring a splinter to
                the surface?"],
            "responses": ["1. SOAK IT IN EPSOM SALTS. Dissolve
                 a cup of the salts into a warm bath and soak
                whatever part of the body has the splinter.
                Failing that, you can also put some of the
                salts onto a bandage pad and leave it covered
                for a day; this will eventually help bring the
                splinter to the surface. 2. VINEGAR OR OIL.
                Another simple way to draw out that stubborn
                splinter is to soak the affected area in oil (
                olive or corn) or white vinegar. Just pour some
                 in a bowl and soak the area for around 20 to
                30 minutes."],
            "context_set": ""
        }
    ... ]
}
```

APPENDIX C
Human evaluation through Google Forms

Question: How can I manage my diabetes better?

Model's Response:

To manage your diabetes better, you should eat healthily, exercise regularly, and have regular check-ups, including blood tests. It's important to quit smoking, cut down on alcohol, and maintain a healthy weight. For type 2 diabetes, consider lifestyle changes like losing weight and seeking medical advice for possible remission.

Expected Ideal Answer:

To better manage your diabetes, focus on maintaining a healthy lifestyle, which includes eating a balanced diet, staying physically active, and regularly monitoring your blood sugar levels. Keep up with routine medical check-ups to catch any potential complications early, and take medications as prescribed by your healthcare provider. If you have type 2 diabetes, losing weight can help improve your condition, and in some cases, may lead to remission.

8 responses

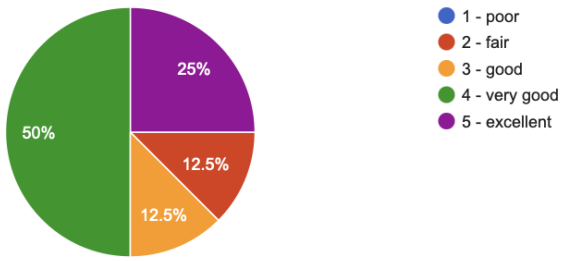


Figure C.1: Diabetes question evaluation example

Question: What's a good recipe I can try at home?

Model's Response: *A delicious homemade pizza recipe you can try is a Pepperoni and Extra Cheese Pizza. You can buy pre-made pizza dough or make your own, spread tomato sauce on the dough, top it with mozzarella cheese, pepperoni slices, and then sprinkle some extra cheese on top. Bake in the oven until the crust is golden brown and the cheese is melted and bubbly. Enjoy your homemade Pepperoni and Extra Cheese Pizza!*

Expected Ideal Answer: *I know you love pizza, so if you feel really inspired, you can try to make everything from scratch. Homemade dough and tomato sauce taste amazing. Top it with your favorite toppings like mushrooms and ham, bake and enjoy!*

8 responses

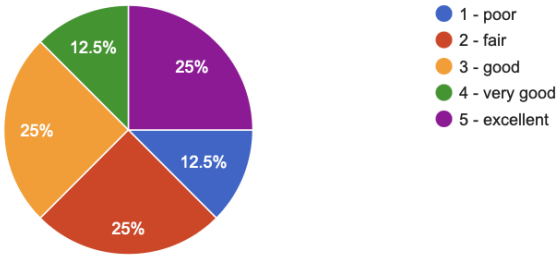


Figure C.2: Recipe question evaluation example

Question: What are the primary symptoms of coronary heart disease (CHD)?

Model's Response:

Coronary heart disease (CHD) can present with a variety of symptoms, especially in older adults. Some common symptoms to watch out for include chest pain or discomfort (angina), shortness of breath, fatigue, weakness, and even sudden dizziness or fainting. It's important to note that some individuals may not experience any symptoms at all, which is why regular check-ups with your healthcare provider are crucial. If you ever have concerns about your heart health or experience any of these symptoms, please seek medical attention promptly to ensure your well-being and heart health.

Expected Ideal Answer:

The primary symptoms of coronary heart disease are chest pain (angina) and breathlessness. Angina is often a mild, uncomfortable feeling, but can become severe and spread to other areas such as the arms, neck, jaw, back, or stomach. In some cases, CHD may not show any symptoms before diagnosis.

8 responses

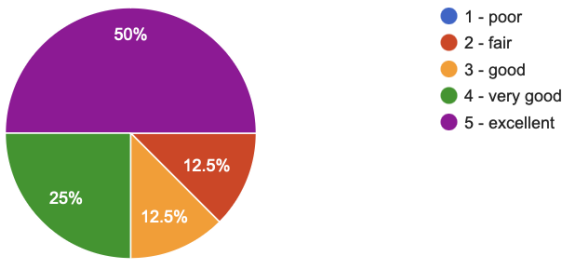


Figure C.3: Coronary Heart Disease question evaluation example

ANNEX: SUSTAINABLE DEVELOPMENT GOALS (SDGs)

Degree of Relationship of the Work with the Sustainable Development Goals (SDGs)

On September 25, 2015, world leaders adopted a set of global goals to eradicate poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development agenda. Each goal has specific targets to be achieved over the next 15 years.

Sustainable Development Goals	**High**	**Medium**	**Low**	**Not Applicable**
SDG 1. No Poverty				X
SDG 2. Zero Hunger				X
SDG 3. Good Health and Well-being	X			
SDG 4. Quality Education				X
SDG 5. Gender Equality				X
SDG 6. Clean Water and Sanitation				X
SDG 7. Affordable and Clean Energy				X
SDG 8. Decent Work and Economic Growth	X			
SDG 9. Industry, Innovation, and Infrastructure	X			
SDG 10. Reduced Inequality	X			
SDG 11. Sustainable Cities and Communities		X		
SDG 12. Responsible Consumption and Production				X
SDG 13. Climate Action				X
SDG 14. Life Below Water				X
SDG 15. Life on Land				X
SDG 16. Peace, Justice, and Strong Institutions				X
SDG 17. Partnerships for the Goals				X

Table C.1: Assessment of Sustainable Development Goals

Reflection on the Relationship Between the Thesis and the SDGs

Of the previously mentioned Sustainable Development Goals, the related project is associated with: (SDGs).

SDG 3: Good Health and Well-being
The project directly contributes to SDG 3 by promoting healthy aging and enhancing the accessibility and quality of care for the elderly. The virtual assistant is designed to provide continuous support and companionship, which helps in maintaining the well-being of elderly users. By offering access to healthcare information and daily life suggestions, the system aids in mangining health issues and encourages a healthier lifestyle, thus supporting the overall goal of good health and well-being.

SDG 8: Decent Work and Economic Growth
Indirectly, the virtual assistant supports SDG 8 by reducing the burden on caregivers. This alleviation of caregiving responsibilities can lead to more efficient use of time and resources, allowing them to focus on their professional responsibilities and personal well-being. As the system helps streamline care processes, it contributes to a more balanced work-life scenario.

SDG 10: Reduced Inequalities
The virtual assistant also aligns with SDG 10 by addressing inequalities faced by elderly individuals. The technology helps bridge the gap in access to quality care and support services, which can be particularly beneficial for those in underserved or disadvantaged communities. By enabling elderly individuals to live with greater independence and dignity, the system plays a role in reducing social inequalities and improving their overall quality of life.

SDG 9: Industry, Innovation, and Infrastructure
The project is in line with SDG 9 through its use of advanced technologies such as Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). These innovations are integral to developing a sophisticated virtual assistant capable of addressing the specific needs of an aging population. The scalable and modular design of the system ensures it can evolve with emerging technologies, making it a forward-thinking solution that supports ongoing technological progress and infrastructure development.

SDG 17: Partnerships for the Goals
Lastly, the project embodies SDG 17 by fostering collaborations among technology developers, healthcare providers, and caregiving communities. By working together across various sectors, the initiative not only enhances the development of the virtual assistant but also strengthens the collective effort to address the challenges of an aging society. This collaborative approach ensures that the solution is well-integrated and maximally effective.