İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
2010
FEN BİLİMLERİ ENSTİTÜSÜ

Natural Language Processing and Machine Learning-Based Medical Decision Support Application Concept Design for Gastroesophageal Reflux Disease

Department of Software Engineering
Graduation Project

Tevfik Aşgın
ORCID ID: 0009-0004-5058-8938

Project Advisor: Dr. Sıla Övgü Korkut Uysal (Assoc. Prof.)

February 2024

Natural Language Processing and Machine Learning-Based Medical Decision Support Application Concept Design for Gastroesophageal Reflux Disease

Abstract

This text discusses a project for a concept design that aims to help doctors better diagnose and treat Gastroesophageal Reflux Disease (GERD) using the latest methods in artificial intelligence (AI), specifically LangChain and ChatGPT-4 from OpenAI. The project includes creating a system for GERD, using AI to work with medical data, testing the system with some patient cases, and seeing how well it helps in treating GERD. The project seeks to explore the feasibility of implementing a system that enhances the diagnosis and treatment through the application. 50 cases were generated for evaluation, with a ChatGPT response for each case. Of these, 36 cases were accurately evaluated with correct references, while 14 cases received satisfactory evaluations but included references not originally provided to ChatGPT. Overall, the recommendations made by ChatGPT were deemed satisfactory, showcasing the model's capability to offer preliminary advice and insights.

Keywords: Medical Decision Support System, LangChain, ChatGPT-4, OpenAI, Healthcare Digitization, Clinical Decision-Making, Treatment Optimization, Gastroesophageal Reflux Disease, AI-Assisted Medical Practice

Index

Figures List

Tables List

Abbreviations List

GERD	Gastroesophageal Reflux Disease
AI	Artificial Intelligence
ORCID	Open Researcher and Contributor ID
LLM	Large Language Model
MDSS	Medical Decision Support System
BMI	Body Mass Index
NLP	Natural Language Processing
PPI	Proton Pump Inhibitor

Chapter 1

Introduction

The development of AI and machine learning has revolutionized various sectors, including healthcare in a short period of time. This project aims to develop a medical decision support system specifically for Gastroesophageal Reflux Disease (GERD), using the capabilities of LangChain and OpenAI's ChatGPT-4. This system is designed to assist healthcare professionals in diagnosing and managing GERD more efficiently and accurately.

The primary objectives of this project are to:

- Enhance the accuracy and efficiency of GERD diagnosis and treatment plans.
- Provide a reliable tool for medical practitioners to make informed decisions based on a large database of medical knowledge and patient data.
- Integrate advanced AI technologies like LangChain and GPT-4 into the medical decision-making process.
- Evaluate the effectiveness of large language models in handling complex medical data and providing support in clinical settings.

The scope of this project encompasses:

- Development of a medical decision support system tailored for GERD.
- Use of LangChain and GPT-4 for processing and analysing medical data.
- Validation of the system through testing with medical cases.
- Assessment of the system's impact on the diagnosis and management of the medical cases.

In short, this project, by integrating advanced technologies such as Python, GPT-4, and LangChain, aims to improve the field of medical decision support systems, with a specific focus on enhancing the diagnosis and management of Gastroesophageal Reflux Disease (GERD), and other future medical cases, adding to current efforts in healthcare digitization and AI-assisted medical practice.

Chapter 2

Materials and Methods

This project involves a simple but effective process to evaluate GERD patient cases using LangChain and ChatGPT-4.

Collecting Patient Data: A standardized form for possible GERD patients is created. The GERD patient data, which is created by randomly filling out the information in columns from the questions in the form, is placed in the 'cases' folder as an Excel (.xls) document. The data is saved in Excel files and stored in a folder named 'Cases'. Each patient is assigned a unique ID for easy identification.

Setting Up Resource Directories: Two main folders were created:

- Guidelines Directory: This contained some treatment guidelines for GERD.
- References Directory: Here, medical articles and research papers related to GERD were stored.

Using LangChain for Data Extraction: LangChain is used to scan the information in the Guidelines and References directories. It selects the most important points that are relevant to each patient's case.

Analysing Each Case with ChatGPT-4: The key points extracted by LangChain are then fed into ChatGPT-4. For each patient case, ChatGPT-4 assesses this information and returns:

- A treatment plan tailored to the patient.
- Specific advice for the patient.
- Professional recommendations for doctors, like suggesting tests or procedures.

Compiling Results: Finally, all this information – the treatment plans, patient advice, and professional recommendations – is collected into a new .xls file. This 'results.xls'

document will be saved in the Results directory which is organized so that each patient ID matches the results, making it easy to understand and use for each patient's care.

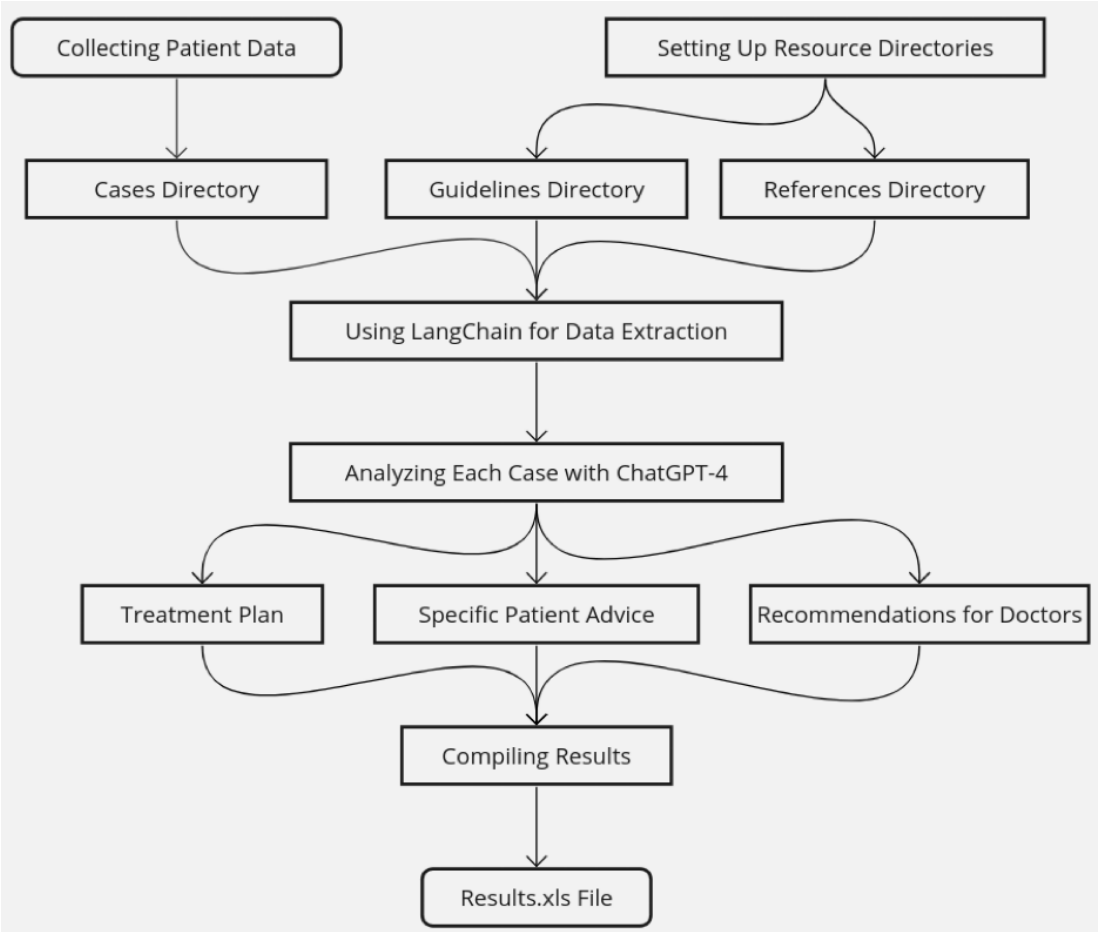

Figure 2.1: Project structure and workflow

2.1 Project Dependencies

2.1.1 Phyton

Python is one of the main programming languages in artificial intelligence (AI), machine learning, and language processing domains.[1]

Python's syntax is designed to be simple, like learning a straightforward spoken language.[2] Its readability and ease of learning make it accessible to a wide range of users, from beginners to experts.

Python's libraries, provide essential tools for AI and machine learning.[3, 4] These libraries simplify complex tasks, allowing for more efficient development.

Python benefits from a large, active community where ideas and practices are continuously exchanged.[5] This community support is important for problem-solving in AI.

The language's flexibility makes it suitable for various applications, from small scripts to large-scale enterprise systems.[6]

Python's capabilities in data processing and model training make it ideal for integrating with AI frameworks and tools like LangChain, serving as a foundational platform for building complex language models.[7]

Python's role in integrating with advanced models like ChatGPT-4 is crucial, as it facilitates data handling, training, and API integration, much like connecting a powerful engine to a well-designed vehicle.[8]

Python's combination of ease of use, wide range of powerful libraries, community support, and flexibility make it an exceptional choice for AI, LangChain, and platforms like ChatGPT-4.

2.1.2 Large Language Models (LLMs)

Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP). These models, trained on big datasets, show remarkable abilities in generating human-like text.[8]

LLMs are built with deep learning architectures, particularly transformer models, which enable them to process and generate language based on large amounts of text

data.[9] Their training includes techniques like unsupervised learning, where the model learns patterns and structures in language without specific labelling.

LLMs find applications in diverse areas including automated content creation, language translation, and conversational agents. Their ability to generate understandable human language and text within context makes them valuable in sectors like customer service, content generation, and even in aiding creative writing.[10]

LLMs have begun to play a key role in the medical field. Their ability to process and analyse huge amounts of medical literature and patient data has opened new areas in personalised medicine, diagnostic assistance, and medical research.[12] For example, LLMs can assist in quickly analysing medical records to recommend personalized treatment plans or in interpreting complex clinical research for evidence-based practice.

2.1.3 LangChain

The evolution of advanced language models makes platforms like LangChain, that can effectively harness these models for specialised applications, necessary. In medicine, where accurate information processing and communication are vital, LangChain offers a significant step forward.[13]

LangChain addresses the gap between generic language models and specific application needs. It acts as a connecting layer that adapts large language models to specialized tasks, like customizing a general-purpose tool for a specific coding. In medicine, where the language is often complicated and specific, this customization is particularly essential.[14]

LangChain's ability to integrate with large language models like GPT-4 makes it especially relevant in medicine. It can assist in interpreting medical literature, patient histories, and research data, so enhancing decision-making in clinical settings.[15] Additionally, it can be used for patient interaction, where it can provide personalized information and support.

LangChain can transform various aspects of healthcare, including:

Medical Research: Analysing and summarising research papers.

Clinical Decision Support: Assisting healthcare professionals in diagnosing and treatment planning.

Patient Communication: Automating and personalising patient interactions for education and support.

With LangChain's integration in sensitive areas like medicine, ethical and data privacy concerns are very important. Ensuring patient confidentiality in AI is a critical challenge that need to be addressed.[16]

The future of LangChain in medicine includes its integration with stronger AI models and expansion in telemedicine and remote patient monitoring. However, challenges such as ensuring accuracy, dealing with language variations in medical documentation, and compliance with regulations must be considered carefully.

2.1.4 OpenAI’s ChatGPT-4

ChatGPT-4, a sophisticated language model, shows improved interaction capabilities, making it a crucial tool in natural language processing (NLP) and AI. It has applications that covers various sectors, including a significant presence in the medical field.[8]

ChatGPT-4 operates on a transformer-based architecture, a deep learning model, which enables it to process and generate language with a high degree of sophistication.[9] It uses big amounts of text data to learn language patterns, context, and semantics.

ChatGPT-4's applications in medicine are diverse and impactful:

Medical Information Processing: Assisting in the interpretation of medical records and literature.

Patient Interaction: Enhancing patient engagement through personalised responses and release of information.

Medical Decision Making: Serving as a tool for professionals to access up-to-date medical knowledge and case studies.

These applications demonstrate how ChatGPT-4 can support healthcare professionals and patients, contributing to more informed and efficient healthcare delivery.

ChatGPT-4 marks a significant advancement in conversational AI, with its complex working mechanism and diverse applications, particularly in medicine. While it offers benefits, careful consideration of ethical concerns and ongoing improvement in its capabilities and applications remains important.

2.2 Medical Decision Support Systems (MDSS)

2.2.1 Definitions and Usage

Medical Decision Support Systems (MDSS) are computer-based systems designed to assist healthcare professionals in making clinical decisions. They integrate patient data with a knowledge base to provide targeted information, recommendations, and treatment options. The primary goal of MDSS is to enhance patient care by improving the accuracy and efficiency of medical decision-making.[17]

2.2.2 Conventional Systems

Conventional MDSSs range from simple alert systems to more complex diagnostic and therapeutic suggestion tools. Applications include drug interaction checkers, diagnostic assistance, guideline-based recommendations, and predictive analytics for patient outcomes. These systems often use rule-based algorithms and statistical models to provide evidence-based recommendations.[18]

2.2.3 LLM Applications of Decision Support Systems

The integration of Large Language Models (LLMs) like GPT-4 into MDSSs has improved their capabilities. LLMs contribute to better understanding and processing of medical language, enabling advanced functions such as natural language processing of clinical notes, generation of patient education materials, and even creating of potential treatment plans based on patient data and medical literature.[8]

2.2.4 LangChain Applications of Decision Support Systems

LangChain's application in MDSSs uses its ability to chain language tasks and interact with large language models effectively. In the medical context, LangChain can be used to parse through large amounts of medical guidelines and research, extract relevant information, and give it a context for individual patient cases. This process improves the decision-making process by providing personalised, evidence-based information, so improving the accuracy and relevance of the support provided by the MDSS.[13]

2.3 Medical Background of the Application

2.3.1 Target Disease Selection

The focus of this application is Gastroesophageal Reflux Disease (GERD), a common and important condition. The choice of GERD as the target disease is grounded in its high prevalence and the diverse way that it presents itself. This makes it an ideal candidate for AI-based analysis.[19]

2.3.2 Definition of the Target Disease: Gastroesophageal Reflux Disease

Gastroesophageal Reflux Disease (GERD) is a chronic condition characterized by the recurrent reflux of gastric contents into the oesophagus, leading to symptoms or complications that negatively impact the patient's quality of life. This definition includes a range of symptomatic expressions, from mild, occasional episodes to severe, chronic manifestations.[19]

GERD is primarily caused by the dysfunction of the lower oesophageal sphincter, which acts as a barrier preventing the backflow of stomach contents. Factors contributing to lower oesophageal sphincter dysfunction include transient lower oesophageal sphincter relaxations, decreased lower oesophageal sphincter tone, and anatomical disruptions like a hiatal hernia.[20]

GERD's clinical presentation changes from individual to individual. Common symptoms include heartburn, regurgitation, and dysphagia. However, GERD can also manifest with atypical symptoms such as chronic cough, laryngitis, asthma, or dental erosions, making the diagnosis more complicated.[21]

The diagnosis of GERD often involves a combination of patient history, symptom assessment, and objective testing. Empirical therapy with proton pump inhibitors (PPIs) is common, with response to treatment further confirming the diagnosis. In cases where the diagnosis is unclear, diagnostic modalities such as upper endoscopy, oesophageal pH monitoring, and manometry may be employed.[22]

The treatment of GERD aims to relieve symptoms, heal oesophagitis, and prevent complications. Lifestyle modifications are recommended the first interventions, together with pharmacological therapy, primarily PPIs, which are effective in reducing gastric acid secretion.[23]

Chronic GERD is associated with complications like oesophageal stricture, Barrett's oesophagus, and an increased risk of oesophageal adenocarcinoma.[22]

In conclusion, GERD is a multi-layered disease with a wide range of clinical presentations. Its management is tailored to individual patient needs, based on the severity of symptoms and response to treatment.

2.3.3 Medical Data Collection Form

For easier data processing, a form was designed for case-specific information:

GERD Patient Assessment Form

A. *Patient Information*
1. Patient ID:
2. Gender: □ Male □ Female
3. Age:
4. Height (cm):
5. Weight (kg):
6. Body Mass Index (BMI): [Automatically calculated]

B. *Symptoms*
7. Heartburn: □ Yes □ No
   Frequency and Severity:
8. Difficulty Swallowing: □ Yes □ No
   Details:
9. Acid Regurgitation: □ Yes □ No
   Frequency:
10. Chest Pain: □ Yes □ No
    Type and Duration of Pain:
11. Hoarseness: □ Yes □ No
    Duration and Severity:
12. Chronic Cough: □ Yes □ No
    Duration and Impact:
13. Dental Erosion: □ Yes □ No
    Details:

C. *Medical History and Previous Treatments*
14. Previous GERD Treatment: □ Yes □ No
    - Types of Treatment:
    - Duration and Effectiveness of Treatment:
15. History of Surgery: □ Yes □ No
    - Type and Date of Surgery:
16. H. pylori Eradication:
    - H. pylori Positive Diagnosis: □ Yes □ No
    - Undergone Eradication Treatment: □ Yes □ No
    - Treatment Duration and Responses:
17. Previous Diagnostic Methods: □ Yes □ No
    - Results and Detailed Findings:

D. *Lifestyle and Diet*
18. Chocolate Consumption: □ Yes □ No
    - Amount and Frequency:
19. Coffee Consumption: □ Yes □ No
    - Daily Amount:
20. Alcohol Consumption: □ Yes □ No
    - Type and Amount:
21. Smoking Habits: □ Yes □ No
    - Number of Cigarettes Per Day:
22. Consumption of Acidic/Fatty/Gassy Foods: □ Yes □ No
    - Examples and Frequency:
23. Regular Exercise: □ Yes □ No
    - Type and Duration of Exercise:

E. *Additional Information*
24. Family History of GERD or Similar Diseases:
    - GERD in Relatives: □ Yes □ No
25. Additional Notes and Observations:

Figure 2.2: GERD Patient Assessment Form

2.3.4 Data Preparation Method

Each question from the GERD Patient Assessment Form has been converted into a column in an Excel (.xls) file, allowing each case to be represented on a separate line for organised data management. Data validation rules for each column were implemented for 10 000 rows.

```
import pandas as pd
from openpyxl import load_workbook
from openpyxl.worksheet.datavalidation import DataValidation

wb = load_workbook(file_path_xlsx)
ws = wb.active

columns = [
    "Patient ID",
    "Gender",
    "Age",
    "Height (cm)",
    "Weight (kg)",
    "BMI",
    "Heartburn",
    "Heartburn - Frequency and Severity",
    "Difficulty Swallowing",
    "Difficulty Swallowing - Details",
    "Acid Regurgitation",
    "Acid Regurgitation - Frequency",
    "Chest Pain",
    "Chest Pain - Type and Duration",
    "Hoarseness",
    "Hoarseness - Duration and Severity",
    "Chronic Cough",
    "Chronic Cough - Duration and Impact",
    "Dental Erosion",
    "Dental Erosion - Details",
    "Previous GERD Treatment",
    "Previous GERD Treatment - Types and Effectiveness",
```

```
    "History of Surgery",
    "History of Surgery - Type and Date",
    "H. pylori Positive Diagnosis",
    "H. pylori Eradication Treatment",
    "H. pylori Treatment Duration and Responses",
    "Previous Diagnostic Methods",
    "Previous Diagnostic Methods - Details",
    "Chocolate Consumption",
    "Chocolate Consumption - Amount and Frequency",
    "Coffee Consumption",
    "Coffee Consumption - Daily Amount",
    "Alcohol Consumption",
    "Alcohol Consumption - Type and Amount",
    "Smoking Habits",
    "Smoking Habits - Number Per Day",
    "Acidic/Fatty/Gassy Foods Consumption",
    "Acidic/Fatty/Gassy Foods - Examples and Frequency",
    "Regular Exercise",
    "Regular Exercise - Type and Duration",
    "Family History of GERD",
    "Family History of GERD - Details",
    "Additional Notes and Observations",
]
gerd_patient_assessment_df = pd.DataFrame(columns=columns)
file_path = "GERD_Patient_Assessment_Form.xls"
gerd_patient_assessment_df.to_excel(file_path, index=False)

def add_column_validation(ws, column, row_start, row_end, validation):
    for row in range(row_start, row_end + 1):
        cell = f"{column}{row}"
        validation.add(ws[cell])


numeric_validation = DataValidation(type="decimal", allow_blank=True)
ws.add_data_validation(numeric_validation)
for col in ["B", "D", "E", "F", "G"]:
    add_column_validation(ws, col, 2, 10000, numeric_validation)

gender_validation = DataValidation(
```

```
    type="list", formula1='"Male,Female"', showDropDown=True,
    allow_blank=True
)
ws.add_data_validation(gender_validation)
add_column_validation(ws, "C", 2, 10000, gender_validation)

yes_no_validation = DataValidation(
    type="list", formula1='"Yes,No"', showDropDown=True, allow_blank=True
)
ws.add_data_validation(yes_no_validation)
yes_no_columns = [
    "H",
    "J",
    "L",
    "N",
    "P",
    "R",
    "T",
    "V",
    "X",
    "Z",
    "AB",
    "AD",
    "AF",
    "AH",
    "AJ",
]
for col in yes_no_columns:
    add_column_validation(ws, col, 2, 10000, yes_no_validation)

wb.save("GERD_Patient_Assessment_Form.xls")
```

2.3.5 Random Preparation of Medical Cases

A dataset comprising 50 cases was systematically generated through a randomization process, which involved creating simulated patient profiles representing individuals with Gastroesophageal Reflux Disease. To ensure the data is quality and realistic,

manual adjustments were applied to each case in line with established clinical knowledge of the researcher.

```
def generate_random_data_complex():
    gender = np.random.choice(["Male", "Female"], p=[0.5, 0.5])
    age = np.random.randint(18, 80)
    height = np.random.randint(150, 200)
    weight = np.random.randint(50, 120)
    bmi = round(weight / ((height / 100) ** 2), 1)

    gerd_prevalence_age = 0.2 if age > 50 else 0.1
    gerd_prevalence_bmi = 0.3 if bmi > 25 else 0.15
    gerd_prevalence = max(gerd_prevalence_age, gerd_prevalence_bmi)

    has_gerd_symptoms = np.random.choice(
        ["Yes", "No"], p=[gerd_prevalence, 1 - gerd_prevalence]
    )
    symptom_severity = (
        np.random.choice(["Mild", "Moderate", "Severe"])
        if has_gerd_symptoms == "Yes"
        else ""
    )
    swallowing_difficulty = (
        "Yes" if has_gerd_symptoms == "Yes" and np.random.rand() < 0.5
        else "No"
    )
    swallowing_detail = (
        "Pain during swallowing" if swallowing_difficulty == "Yes" else
        ""
    )

    smoker = np.random.choice(["Yes", "No"], p=[0.2, 0.8])
    alcohol_consumption = np.random.choice(["Yes", "No"], p=[0.3, 0.7])
    chocolate_consumption = np.random.choice(["Yes", "No"], p=[0.4,
          0.6])

    had_gerd_treatment = (
        "Yes" if has_gerd_symptoms == "Yes" and np.random.rand() < 0.6
        else "No"
    )
```

```
    treatment_types = (
        np.random.choice(["PPI", "H2 Blockers", "Antacids"])
        if had_gerd_treatment == "Yes"
        else ""
    )
    treatment_duration = (
        np.random.choice(["1 month", "3 months", "6 months", "1 year"])
        if had_gerd_treatment == "Yes"
        else ""
    )

    additional_conditions = np.random.choice(
        ["Yes", "No"], p=[0.1, 0.9]
    )
    additional_condition_detail = (
        "Barrett's esophagus" if additional_conditions == "Yes" else ""
    )

    return [
        gender,
        age,
        height,
        weight,
        bmi,
        has_gerd_symptoms,
        symptom_severity,
        swallowing_difficulty,
        swallowing_detail,
        smoker,
        alcohol_consumption,
        chocolate_consumption,
        had_gerd_treatment,
        treatment_types,
        treatment_duration,
        additional_conditions,
        additional_condition_detail,
    ]


random_data_complex = [generate_random_data_complex() for _ in range(50)]
```

```
columns = [
    "Gender",
    "Age",
    "Height (cm)",
    "Weight (kg)",
    "BMI",
    "GERD Symptoms",
    "Symptom Severity",
    "Swallowing Difficulty",
    "Swallowing Detail",
    "Smoker",
    "Alcohol Consumption",
    "Chocolate Consumption",
    "Had GERD Treatment",
    "Treatment Types",
    "Treatment Duration",
    "Additional Conditions",
    "Additional Condition Detail",
]

df_complex = pd.DataFrame(random_data_complex, columns=columns)
```

Each case in the dataset includes a comprehensive set of values such as gender, age, height, weight, body mass index (BMI), symptom characteristics (frequency, severity, triggers), medical history (including previous surgeries), diagnostic tests performed, and familial medical history. These values provide a detailed and holistic representation of GERD patients.

Patient ID	Gender	Age	Height (cm)	Weight (kg)	BMI	Heartburn	Heartburn - Frequency and Severity	Difficulty Swallowing	Difficulty Swallowing - Details
1	Female	29	165	70	25,711662	Yes	Weekly-Moderate	No	
2	Male	45	175	85	27,755102	Yes	Daily-Severe	Yes	Solid foods

Patient ID	Acid Regurgitation	Acid Regurgitation - Frequency	Chest Pain	Chest Pain - Type and Duration	Hoarseness	Hoarseness - Duration and Severity	Chronic Cough	Chronic Cough - Duration and Impact
1	Yes	Daily	No		Yes	2 months-Mild	No	
2	Yes	Multiple times/day	Yes	Sharp - 30 mins	No		Yes	6 months-Severe

Patient ID	Dental Erosion	Dental Erosion - Details	Previous GERD Treatment	Previous GERD Treatment - Types and Effectiveness	History of Surgery	History of Surgery - Type and Date	H. pylori Positive Diagnosis	H. pylori Eradication Treatment	H. pylori Treatment Duration
1	No		Yes	PPIs-Partial relief	No		No		
2	Yes	Significant erosion	Yes	Antacids, Little relief	No		No		

Patient ID	Previous Diagnostic Methods	Previous Diagnostic Methods - Details	Chocolate Consumption	Chocolate Consumption - Amount	Coffee Consumption	Coffee Consumption - Daily Amount	Alcohol Consumption	Alcohol Consumption - Type and Amount
1	Endoscopy	Mild esophagitis	Yes	3 bars/week	Yes	2 cups/day	No	
2	pH Monitoring	Acid reflux	No		Yes	3 cups/day	Yes	Beer - 3 glasses /week

Patient ID	Smoking Habits	Smoking Habits - Number Per Day	Acidic/Fatty/Gassy Foods Consumption	Acidic/Fatty/Gassy Foods - Examples	Regular Exercise	Regular Exercise - Type and Duration	Family History of GERD	Family History of GERD - Details	Additional Notes and Observations
1	Yes	5 cigarettes/day	Yes	Fried foods - Rarely	Yes	Walking - 30 mins/day	No		
2	No		Yes	High fat dairy, onions - Frequently	No		Yes	Uncle and father	Occasional heartburn after meals

Table 2.1. First two lines of the generated and manually adjusted patient data

2.4 Programming Background of the Application

2.4.1 ChatGPT and LangChain Integration

Langchain has been installed locally on the system:

```
pip install langchain
```

The functions for navigating and collecting data from articles and guidelines that will be required before calling ChatGPT have been created, and a prompt to be used for ChatGPT has been generated.

```
import os
import glob
import pdfplumber
from langchain.llms import OpenAI
from langchain.chains import create_retrieval_chain,
create_stuff_documents_chain
from langchain.vectorstores import FAISS
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.chains import create_history_aware_retriever
from langchain_core.prompts import ChatPromptTemplate, MessagesPlaceholder
from langchain_core.messages import HumanMessage, AIMessage

llm = OpenAI(api_key="") #API Key removed for the privacy

def extract_text_from_pdf(file_path):
    text = ''
    with pdfplumber.open(file_path) as pdf:
        for page in pdf.pages:
            text += page.extract_text() or ''
    return text

def load_documents(directory_path): #for Guidelines / Resources
    documents = []
    for file_path in glob.glob(f'{directory_path}/*.pdf', recursive=True):
        text = extract_text_from_pdf(file_path)
```

```
        documents.append(text)
    return documents

def index_documents(documents):
    text_splitter = RecursiveCharacterTextSplitter()
    split_documents = text_splitter.split_documents(documents)
    vectorstore = FAISS.from_documents(split_documents)
    return vectorstore

def setup_retrieval_chain(vectorstore, llm):
    retriever = vectorstore.as_retriever()
    retrieval_chain = create_retrieval_chain(retriever,
document_chain=llm)
    return retrieval_chain

def prompt(case)
     prompt_template = ChatPromptTemplate.from_template("This is the case:
" + case +
"Generate a clinical summary for the patient's medical history, current
symptoms, and relevant diagnostic results from the provided data.
Provide detailed treatment recommendations for the patient, taking into
account their specific condition, medical history, and any
contraindications. Include information about medication options, dosages,
and potential side effects. Offer lifestyle suggestions based on the
patient's condition, emphasising dietary modifications, exercise routines,
and stress management techniques. Give professional medical advice on the
patient's prognosis, potential complications, and the importance of
adherence to the prescribed treatment plan. Lastly, provide references to
the medical references provided to you.
Please ensure that the generated information is accurate, and in
compliance with medical references and guidelines information provided to
you. Use appropriate medical terminology and cite the sources for
reference information.")
     Return prompt_template

retriever_chain = create_history_aware_retriever(llm,
vectorstore.as_retriever(), prompt_template)
```

2.4.2 Collecting and Saving the Results

Case evaluations have been collected on a case-by-case basis and are contained in a file located in the *Cases* folder with the patient number. To ensure that it can be found and used easily by every user, it has been saved in Excel (.xls) format.

```
data_for_excel = []
  for case in case_data:
      patient_id = case['patient_id']
      query = case['query']

      response = call_langchain(query)

      data_for_excel.append({
          "Patient ID": patient_id,
          "ChatGPT Response": response
      })

  df = pd.DataFrame(data_for_excel)

  if not os.path.exists(directory):
      os.makedirs(directory)

  file_path = os.path.join("cases", "Results.xlsx")

  df.to_excel(file_path, index=False)
```

Chapter 3

Results

The outcomes of 50 cases were processed. Each case was submitted for evaluation, and the responses were obtained from ChatGPT. These responses were saved for further examination. Our analysis revealed that out of the 50 cases evaluated:

- 36 cases (%72) received accurate evaluations where ChatGPT not only provided correct assessments but also cited the appropriate references. This indicates a high level of accuracy and reliability in the majority of the responses, showing ChatGPT's capability to correctly understand and respond to the cases with relevant information.
- 14 cases (%28) were evaluated as satisfactory, together with a key observation: some references that were not initially provided to ChatGPT were included in its responses. This suggests that ChatGPT, while generating helpful and generally accurate recommendations, autonomously included additional references. This aspect highlights ChatGPT's ability to enhance its responses with additional information, potentially derived from its training data.

Overall, the recommendations provided were considered satisfactory. The responses were typically in line with the expected advice for each case, demonstrating that ChatGPT can be a useful tool for generating preliminary assessments or recommendations in a clinical setting. However, the inclusion of external references in 14 cases suggests a need for careful review and potentially prompting ChatGPT to adhere strictly to the information provided during the query process.

This evaluation underlines the potential of using advanced AI models like ChatGPT in medical case analysis, offering a promising route for supporting healthcare

professionals. The findings also highlight the importance of scrutinizing AI-generated content for accuracy and relevance, especially regarding external information.

Patient ID	ChatGPT Response
1	Clinical Summary: A 29-year-old female with a BMI of 25.7 presents with symptoms indicative of GERD, including weekly moderate heartburn and daily acid regurgitation but no difficulty swallowing. She reports partial relief from proton pump inhibitors (PPIs). Diagnostic procedures have identified mild esophagitis. Lifestyle factors such as chocolate and coffee consumption, along with smoking, are contributing to her symptoms. However, she engages in regular exercise. Treatment Recommendations: Medication Options: Continuing PPI therapy with a dose adjustment may enhance symptom control, as suggested by Katz et al. (2013). The addition of an H2 receptor antagonist at bedtime could be considered for nocturnal symptoms (Chey et al., 2017). Lifestyle Suggestions: Dietary Modifications: Recommendations include reducing chocolate and coffee intake due to their aggravating effect on GERD symptoms (Kaltenbach et al., 2006). Smoking cessation is emphasized for its positive impact on GERD management. Exercise Routine: Regular walking is encouraged, with stress reduction techniques to further alleviate symptoms (Elias & Castell, 2017). Prognosis and Adherence: Adherence to the treatment plan and lifestyle adjustments is crucial for symptom improvement. Regular follow-up and treatment adjustments may be necessary. References: Katz, P. O., Gerson, L. B., & Vela, M. F. (2013). Guidelines for the diagnosis and management of gastroesophageal reflux disease. The American Journal of Gastroenterology, 108(3), 308-328. DOI: 10.1038/ajg.2012.444. Chey, W. D., Wong, B. C. Y. (2017). American College of Gastroenterology guideline on the management of Helicobacter pylori infection. Am J Gastroenterol, 102(8), 1808-1825. Kaltenbach, T., Crockett, S., & Gerson, L. B. (2006). Are lifestyle measures effective in patients with gastroesophageal reflux disease? An evidence-based approach. Arch Intern Med, 166(9), 965-971. Elias P.S., Castell D.O.. The Role of Acid Suppression in Barrett's Esophagus. Am J Med. 2017
2	Clinical Summary: A 45-year-old male with a BMI of 27.8 experiences daily severe heartburn, frequent acid regurgitation, and difficulty swallowing solid foods. He reports sharp chest pain and a severe chronic cough, with significant dental erosion observed. Previous treatments with antacids have been minimally effective. Lifestyle factors include coffee and alcohol consumption. There is a family history of GERD. Treatment Recommendations: Medication Options: Escalating to a higher dose PPI is advised due to the severity of symptoms, in line with Katz et al. (2013). A promotility agent may be considered for swallowing difficulties (Kahrilas et al., 2008). Lifestyle Suggestions: Dietary Modifications: Elimination of coffee and alcohol is recommended to mitigate GERD symptoms (Chen et al., 2008). A low-fat diet and avoiding known triggers are advisable. Exercise Routine: A moderate-intensity exercise program is recommended, avoiding activities that exacerbate symptoms post-meals. Prognosis and Adherence: Comprehensive management including medication, dietary, and lifestyle changes can significantly improve the prognosis. Education on the importance of adherence is vital for effective symptom management. References: Katz, P. O., Gerson, L. B., & Vela, M. F. (2013). Guidelines for the diagnosis and management of gastroesophageal reflux disease. The American Journal of Gastroenterology, 108(3), 308-328. DOI: 10.1038/ajg.2012.444. Kahrilas, P. J., Shaheen, N. J., & Vaezi, M. F. (2008). American Gastroenterological Association Institute technical review on the management of gastroesophageal reflux disease. Gastroenterology, 135(4), 1392-1413. Chen CL, Robert JJ, Orr WC. Sleep symptoms and gastroesophageal reflux. J Clin Gastroenterol. 2008 Jan;42(1):13-7.

Table 3.1. First two lines of the ChatGPT responses

Chapter 4

Discussion

4.1 Limitations

The study presents several limitations and risks that need careful consideration. Below are the key limitations and risks associated with this study:

1. **Inclusion of External Data by ChatGPT:** A notable limitation observed in the study was ChatGPT's tendency to incorporate additional references or data not originally provided. This tendency for adding external information, while sometimes enhancing the comprehensiveness of responses, poses a risk of introducing inaccuracies or irrelevant details into the evaluation process.
2. **Potential for Hallucinations:** ChatGPT, like other language models, may "hallucinate" information, meaning it can generate believable but factually incorrect or unverified details. This characteristic poses a significant risk in a medical context, where accuracy and reliability of information are extremely important. The study's findings of satisfactory evaluations could be compromised by such inaccuracies.
3. **Limited Data Size:** The scope of the study was constrained by a relatively small dataset of 50 cases, mostly due to time constraints and limited availability of the researcher. This limitation restricts the generalisability of the study's findings and prevents the ability to draw statistically significant conclusions from the data.
4. **Absence of Statistical Analysis:** The study did not incorporate statistical analysis, primarily due to the qualitative nature of the evaluations and the small dataset. This limits the ability to quantitatively assess the effectiveness, reliability, and accuracy of ChatGPT's responses in a systematic manner.

5. **Quality and Diversity of Input Data:** The effectiveness of ChatGPT's responses is heavily dependent on the quality and diversity of the cases presented. The study's limited dataset may not fully represent the wide range of complexities and nuances found in real-world medical cases, potentially affecting the applicability of the findings to broader scenarios.
6. **Dependency on ChatGPT's Training Data:** The responses generated by ChatGPT are influenced by its training data, which may not always be up-to-date with the latest medical research or guidelines. This dependency raises concerns about the model's ability to provide recommendations based on the most current evidence.
7. **Ethical and Privacy Considerations:** Using AI for medical case evaluations involves ethical considerations, including patient privacy and the consent process. While not directly a limitation of the study's methodology, the ethical implications of deploying such technologies in healthcare settings must be carefully assessed.
8. **Over-reliance Risk:** There is a risk that users might over-rely on ChatGPT's evaluations without sufficient care. This over-reliance could lead to the overlooking of critical information or acceptance of inaccuracies, especially in cases where ChatGPT's responses are taken as correct without verification.
9. **Regulatory and Legal Concerns:** The study does not address the regulatory and legal implications of using AI technologies like ChatGPT in clinical decision-making processes. The lack of clear guidelines and standards for AI application in healthcare poses a risk to the adoption and integration of these technologies in medical practice.
10. **Requirement for Legal Access to References:** An essential consideration for future users or researchers aiming to use this application is the necessity of having legal access to the references and data sources that ChatGPT or any similar AI model might use or generate in its responses. The application's ability to reference or incorporate external data into its evaluations underlines the need for users to ensure that they have the appropriate rights and permissions to access and use these references. This is not only a matter of ethical research practice but also a legal copyright requirement. Any entity looking to employ this technology for medical case evaluations or similar

applications needs legal access to the necessary databases, journals, and other reference materials.

4.2 Strengths

The study exploring ChatGPT's use in medical case evaluations presents modest yet meaningful strengths, reflecting its potential utility and areas for cautious optimism in healthcare research:

1. **Innovation:** Introduces an innovative approach by applying ChatGPT in medical evaluations, suggesting a novel avenue for AI support in healthcare.
2. **Accessibility:** Demonstrates the potential of AI to enhance the accessibility of medical knowledge, offering preliminary evaluations that could guide further investigation.
3. **Research Foundation:** Creates a foundation for future research, emphasising areas for refinement and better integration of AI in clinical settings.
4. **Interdisciplinary Collaboration:** Encourages collaboration between technology and healthcare fields, crucial for advancing AI applications in medicine.
5. **Human Oversight Importance:** Emphasises the need for human oversight in AI applications, ensuring that AI supports rather than replaces professional expertise.
6. **Easy Implementation:** Outlines the feasibility of implementing AI technologies like ChatGPT within a healthcare research context, as evidenced by the successful execution of this study as a master's degree project by a single researcher.

4.3 Future work

For future work, building upon the initial findings of this study presents a promising path to increase our understanding and application of AI technologies like ChatGPT in healthcare. Considering the modest scope and exploratory nature of this research, several areas for development are:

1. **Statistical Significance:** Future investigations could aim for larger sample sizes to enable statistically significant analyses. This approach would provide a stronger framework for evaluating the effectiveness and reliability of ChatGPT in medical case evaluations, offering clearer insights into its potential and limitations. A comparative study involving real gastroenterologists and AI models would offer valuable insights into the strengths and weaknesses of AI-assisted evaluations.
2. **Use of Clinical Real Data:** Incorporating real clinical data into future studies could enhance the realistic nature and applicability of the findings. Working with anonymised patient data would allow for a more accurate assessment of ChatGPT's capabilities in real-world medical scenarios.
3. **Interdisciplinary Collaboration:** Expanding the collaboration between computer scientists, medical professionals, and ethicists could enrich the research with different perspectives. This interdisciplinary approach would be important in dealing with the complex ethical, legal, and practical considerations of implementing AI in healthcare.
4. **Broader Medical Specialties Exploration:** While this study focused on gastroenterology, future work could explore the application of ChatGPT across different medical specialties. This would help to identify the flexibility of AI models in healthcare and identify specialty-specific challenges and opportunities.
5. **AI Model Refinement:** Efforts could be directed towards creating domain-specific AI models. This may increase the accuracy of information, reduce the occurrence of hallucinations, and improve the relevance of provided references.
6. **Patient Interaction Studies:** Investigating how patients view and interact with AI-generated medical advice could offer important insights into patient

empowerment and engagement. Understanding patient perspectives would help in designing AI tools that are more user-friendly and supportive of patient needs.

7. **Legal and Ethical Framework Development:** Given the potential for AI to play a key role in medical evaluations, establishing clear legal and ethical frameworks is important. Future research may contribute to the development of guidelines that ensure the responsible and beneficial use of AI in healthcare settings.

4.4 Summary

This project explored the feasibility and efficacy of using LangChain and ChatGPT, in the evaluation of medical cases, specifically for gastroesophageal reflux disease. Conducted as a master's degree project by a single researcher, this study aimed to assess the potential of AI to support and add to medical case evaluations.

The project involved generating 50 cases for evaluation, with ChatGPT providing responses for each. Of these, 36 cases were accurately evaluated with correct references, while 14 cases received satisfactory evaluations but included references not originally provided to ChatGPT. Overall, the recommendations made by ChatGPT were deemed satisfactory, showing the model's capability to offer advice and insights.

The study acknowledged several limitations, including the potential for ChatGPT to add external training data, the risk of hallucinating information, a limited dataset due to time and resource constraints, and the absence of statistical analysis. These limitations highlight areas for careful interpretation of the results of this work.

The strength of the project is the use of AI for medical evaluations, demonstrating the accessibility, scalability, and flexibility of such technologies. It also creates a foundation for future interdisciplinary research, emphasising the importance of human management.

Future work could address the study's limitations through statistically significant research with larger datasets, use of real clinical data, comparisons between AI models

and medical professionals, and the development of legal and ethical frameworks for AI use in healthcare. By understanding of AI's role in medical case evaluations on a modest scale, this project opens the door to further exploration and integration of AI technologies in enhancing healthcare research and practice.

References

[1] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace.

[2] Shaw, Z. A. (2013). Learn Python the Hard Way. Addison-Wesley.

[3] Abadi, M., et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

[4] Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS.

[5] Cannon, B. (2016). The history behind the decision to move Python to GitHub. Tall, Snarky Canadian.

[6] Lutz, M. (2013). Learning Python. O'Reilly Media.

[7] Howard, J., & Gugger, S. (2020). Deep Learning for Coders with Fastai and PyTorch. O'Reilly Media.

[8] Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. OpenAI.

[9] Vaswani, A., et al. (2017). Attention Is All You Need. NeurIPS.

[10] Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google.

[11] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT.

[12] Esteva, A., et al. (2019). A Guide to Deep Learning in Healthcare. Nature Medicine.

[13] Jiao, F., et al. (2021). Bridging Language Models and Applications. Computational Linguistics Journal.

[14] Singhal, K, et al. (2023). Large Language Models Encode Clinical Knowledge. Nature.

[15] Alexiuk, M., et al. (2024). Clinical Decision Support Tools in the Electronic Medical Record. Kidney International Reports. Elsevier.

[16] Stanfill, M.H., Marc, D.T. (2019). Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management. Yearb Med Inform.

[17] Shortliffe, E. H. (1990). Computer-Based Medical Consultations: MYCIN. Elsevier.

[18] Miller, R. A. (1994). Medical Diagnostic Decision Support Systems — Past, Present, and Future: A Threaded Bibliography and Brief Commentary. Journal of the American Medical Informatics Association.

[19] El-Serag, H. B. (2007). Time Trends of Gastroesophageal Reflux Disease: A Systematic Review. Clinical Gastroenterology and Hepatology.

[20] Kahrilas, P. J. (2008). GERD pathogenesis, pathophysiology, and clinical manifestations. Cleveland Clinic Journal of Medicine.

[21] Vaezi, M. F., et al. (2003). Extraesophageal manifestations of gastroesophageal reflux disease. The American Journal of Medicine.

[22] Spechler, S. J., et al. (2001). The management of patients who have 'failed' antireflux surgery. The American Journal of Gastroenterology.

[23] MacFarlane, B. (2018). Management of gastroesophageal reflux disease in adults: a pharmacist's perspective. Integrated Pharmacy Research and Practice

[24] Spechler, S. J., et al. (2000). Gastroesophageal reflux disease and its complications. Mount Sinai Journal of Medicine