An

Industrial Training Report

Submitted to

UKA TARSADIA UNIVERSITY

In partial fulfillment of the requirements for the degree of

Bachelor of Technology

In

Information Technology

By

Rajkumar V. Munjapara (202003103510281)

Under the Guidance

of

Ms. Divya Patel

Department of Information Technology and Cyber Security

Chhotubhai Gopalbhai Patel Institute of Technology

Uka Tarsadia University

Bardoli – 394350

June 2024

CERTIFICATE

This is to certify that project work embodied in this report entitled **Industrial Training** as carried out by **Rajkumar Vinubhai Munjapara (202003103510281)** under my guidance in partial fulfilment of the degree of Bachelor of Technology in Information Technology, Chhotubhai Gopalbhai Patel Institute of Technology, UTU, Bardoli during the academic year 2023-24.

Date:
Place:

Guided By:

Ms. Divya Patel Assistant Professor, Department of IT & CS CGPIT, UTU, Bardoli.	Ms. Purvi H. Tandel Head of the Dept., Department of IT & CS CGPIT, UTU, Bardoli.

Signature of Examiner

Chhotubhai Gopalbhai Patel Institute of Technology
Uka Tarsadia University
Bardoli – 394350

GENESIS
WEB

Website
genesisweb.co.in

TO:
Rajkumar Munjapara

DATE:
16/06/2024

Dear Mr. Rajkumar Munjapara

I am pleased to confirm that as **AI/ML Intern**, Mr. Rajkumar Munjapara has completed a **6 months** internship successfully at **Genesis Web** from December 18, 2023, to June 14, 2024.

During this period, he demonstrated exceptional programming, project development, and problem-solving skills, contributing significantly to our team's success.

Raj consistently delivered high-quality work, showing strong analytical skills and a keen eye for detail. He worked effectively within our team, displaying excellent communication and collaboration abilities. We are confident that he will excel in his future endeavors and be a valuable asset to any organization.

Please feel free to contact me at **8980344746** or **info@genesisweb.co.in** if you require any further information.

Regards,
Uttam Rabadiya.
For GENESIS WEB
Partner

Uttam Rabadiya
CEO

Phone
9909343988
8980344746

Email
info@genesisweb.co.in

Address
227, Ambika Pinnacle,
Lajamni Chowk, Surat

ABOUT COMPANY

Genesis web is best known for its Creativity, Design, Website development, Mobile development, artificial intelligence, software development, machine learning and IOT. With an innovative, creative and pragmatic approach, software development at no less than an exciting and engaging process. We have an in-built knack for delivering high-performance websites, mobile and software applications - of all size and complexities. In more simple terms, we make compelling products that scale to your business needs.

Vision :

Genesis web envisions becoming a global leader in providing innovative technology solutions that empower businesses and individuals to thrive in the digital age. Our vision is to constantly push the boundaries of technology to enhance the way people work, connect, and live.

Mission :

Our mission at Genesis web is to deliver high-quality software and technology services that drive value for our clients. We are dedicated to fostering innovation, fostering a culture of continuous learning, and maintaining the highest standards of integrity and professionalism.

Products :

Genesis web offers a range of cutting-edge products and solutions, including web and mobile applications, cloud-based software, and custom software development services tailored to meet the unique needs of our clients.

Partners :

Genesis web collaborates with a network of strategic partners, including technology providers, industry associations, and research institutions. These partnerships enable us to stay at the forefront of technological advancements and deliver comprehensive solutions to our clients.

Customers :

Our customer base spans across industries, including healthcare, finance,e-commerce, and education. Some notable clients include Fortune 500 companies, startups, and government agencies, who trust us to deliver high-quality software solutions to meet their objectives.

Achievements :

Genesis web has achieved several significant milestones, including Successful completion of projects that have streamlined operations and enhanced customer experiences for our clients. Expansion into international markets, with a growing presence in France.

Certificates :

Developers with 3 to 8+ years of experience.

Expert Laravel 5 developer for different tasks.

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this project would be incomplete without mentioning the people who made it possible, without whose constant guidance and encouragement would have made efforts go in vain. I consider myself privileged to express gratitude and respect towards all those who has guided through the completion of projects.

I convey thanks to my project guide **Ms. Divya Patel**, Information Technology and Cyber Security department, CGPIT for providing encouragement, constant support and guidance which was of a great help to complete this project work successfully.

I am grateful to my external guide **Mr. Uttam Rabadiya, CEO** in **Genesis web** for giving me the support and encouragement that was necessary for the completion of this project.

I am grateful to **Ms. Purvi H. Tandel**, Head of the Department, Information Technology and Cyber Security, CGPIT for giving us the support and encouragement that was necessary for the completion of this project.

I would also like to express my gratitude to **Prof. B. M. Vadher**, Director, Chhotubhai Gopalbhai Patel Institute of Technology for providing us congenial environment to work in.

I would like to thank all the faculty members for their patience, understanding and guidance that gave me strength and will power to work through the long tedious hours for developing a project and preparing the report.

Last but not the least, I would also like to thank my colleagues, who have co-operated during the preparation of our report and without them this project has not been possible. Their ideas helped me a lot to improve my project report.

Rajkumar Munjapara (202003103510281)

ABSTRACT

This abstract encapsulates the transformative journey of an internship at GenesisWeb, undertaken from December 18th, 2023, to June 15th, 2024. As a Software developer Developer, the experience delved into the intricacies of new technologies to serve a seamless software development. The internship provided a comprehensive understanding of agile development methodologies, collaborative problem-solving, and the utilization of cutting-edge tools and frameworks to architect robust digital solutions. Working within GenesisWeb's dynamic environment offered firsthand exposure to real-world challenges, fostering adaptability and innovation.

From conceptualization to deployment, each phase of the development lifecycle was meticulously explored, honing technical expertise and project management skills. Engaging with cross-functional teams enabled the cultivation of effective communication and teamwork abilities, essential for navigating complex projects. Beyond the technical realm, the internship instilled a profound sense of accountability and professionalism, emphasizing the significance of deadlines and client satisfaction. The hands-on experience garnered at GenesisWeb serves as a springboard for future endeavors, equipping with invaluable skills and insights that transcend the boundaries of a single internship. As I embark on my professional journey, the lessons learned and relationships forged during this internship will undoubtedly shape my career trajectory, paving the way for success in the ever-evolving landscape of technology.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

1.1 OVERVIEW

Background:

The project revolutionizes the way users interact with website and PDF document content through the innovative implementation of a Chatbot interface. By seamlessly integrating this technology, users gain the ability to engage in natural language conversations directly with website or PDF content. Accessible through URL inputs or PDF uploads, this groundbreaking app empowers users to query and extract information with ease. Utilizing a sophisticated language model, the Chatbot ensures accurate and relevant responses to user inquiries. It's important to note that the app exclusively responds to questions pertaining to the provided website URL, ensuring focused and targeted interactions. This transformative approach streamlines information retrieval processes, enhancing user productivity and efficiency. Through its conversational interface, the app bridges the gap between users and digital content, fostering a more intuitive and engaging user experience. With its emphasis on natural language understanding, the app caters to diverse user preferences and communication styles. By leveraging cutting-edge technology, the project sets a new standard for website and PDF document interaction, paving the way for enhanced accessibility and usability. The app's intuitive design facilitates seamless navigation and interaction, even for users with limited technical proficiency. Through continuous refinement and optimization, the app strives to deliver unparalleled accuracy and responsiveness in addressing user queries. As a testament to its innovation, the project represents a significant leap forward in the field of information retrieval and user-centric design. Overall, the project's visionary approach promises to redefine the way users engage with digital content, unlocking new possibilities for productivity and knowledge discovery.

Prerequisites:

To interact with the software, users must provide either a PDF document or a website URL containing relevant data such as articles, novels, or business reports. The system's functionality relies on this input to generate accurate responses. Users are encouraged to pose questions directly related to the uploaded PDF or website URL. This ensures that the system can provide tailored answers that are pertinent to the provided data. By adhering to this requirement, users can make the most of the software's capabilities and receive meaningful insights based on the content they've supplied. This approach streamlines the user experience, focusing on delivering targeted information aligned with the user's specific query. Through this user-centric design, the software aims to optimize user engagement and satisfaction by providing relevant and contextually appropriate responses. It's imperative for users to understand that the software's ability to deliver accurate answers hinges on the relevance of the provided data. By adhering to these guidelines, users can unlock the full potential of the software and harness its capabilities for efficient information retrieval and analysis.

Expected Results:

Upon successful completion, the conversational interface empowers users to effortlessly engage with the application by inputting a website URL or uploading a PDF document. This intuitive interface serves as a gateway for users to pose questions directly related to the content of the specified website or PDF. By facilitating seamless interaction, users can navigate through information with ease, leveraging natural language queries to retrieve relevant insights. This innovative approach simplifies the process of information retrieval, eliminating the need for complex search queries or manual navigation. Through its conversational nature, the interface bridges the gap between users and digital content, fostering a more intuitive and accessible user experience. Users can expect a fluid and responsive interaction, where their queries are met with accurate and contextually appropriate responses. By harnessing the power of natural language processing, the interface ensures that users can extract valuable insights efficiently and effectively. Ultimately, the conversational interface revolutionizes the way users interact with

digital content, offering a seamless and intuitive experience for information retrieval and exploration.

1.2 PROBLEM DEFINITION

The Chat App represents a pioneering Python application designed to facilitate seamless communication with multiple PDF documents. Through its intuitive interface, users can engage in natural language conversations, posing questions about both PDFs and websites alike. Leveraging advanced language models, the application generates precise and contextually relevant responses tailored to user queries. It functions as a sophisticated tool for information retrieval, offering users the ability to extract insights from loaded documents with ease. It's important to note that the application's functionality is limited to responding to inquiries directly related to the PDFs that have been loaded into the system. By adhering to this focus, users can maximize the effectiveness of their interactions with the application, ensuring that they receive accurate and pertinent information. This emphasis on relevance enhances the user experience, streamlining the process of accessing and comprehending document content. Overall, the Chat App represents a powerful solution for engaging with PDF documents in a conversational manner, unlocking new possibilities for efficient information retrieval and analysis.

1.3 Scope

- **ChatAPP Compatibility:** The effectiveness of the chatbot may be limited by the complexity and structure of the websites it interacts with. Some websites may have intricate layouts or dynamic content that the chatbot struggles to parse accurately.

- **Content Accuracy:** The accuracy of information retrieved by the chatbot depends on the quality and consistency of the website content. Inaccuracies or inconsistencies within the website may lead to incorrect responses.

- **User Expectations:** Users may have varying expectations regarding the chatbot's capabilities and may encounter limitations in its ability to provide comprehensive answers to all types of queries.

- **Language Support:** The chatbot's ability to understand and respond to user queries may be restricted by its language capabilities. It may not perform optimally with websites in languages other than those it is trained on.

- **Dynamic Content Handling**: Websites with dynamic content, such as real-time updates or user-generated content, may pose challenges for the chatbot in accurately parsing and interpreting information. Ensuring robust handling of dynamic content is essential for maintaining accuracy and relevance.

- **Contextual Understanding**: The chatbot's ability to understand and maintain context during conversations is critical for providing accurate and coherent responses. Enhancing its contextual understanding capabilities can help mitigate misunderstandings and improve overall user satisfaction.

CHAPTER 2 TRAINING ACTIVITIES

- ## Learning Activities of Week 1 :

  I recently began my journey into learning Python, a versatile and powerful programming language. As part of my learning process, I completed numerous coding questions on LeetCode, a popular platform for practicing coding skills and preparing for technical interviews. This experience helped me develop a strong foundation in Python syntax, problem-solving techniques, and algorithmic thinking. In addition to honing my coding skills, I delved into key Python libraries that are essential for data science and machine learning. I explored NumPy, which is indispensable for numerical computing due to its efficient handling of arrays and matrices. I also learned Pandas, a library that provides data structures and functions needed to manipulate and analyze structured data with ease. Furthermore, I familiarized myself with Matplotlib, a powerful plotting library used for creating static, animated, and interactive visualizations in Python.These libraries form the backbone of data preprocessing in machine learning. By mastering them, I can efficiently clean, transform, and visualize data, which are critical steps in any data science project. My newfound skills in Python and its libraries have equipped me with the tools necessary to tackle complex data-driven challenges and advance further into the field of machine learning.

- ## Learning Activities of Week 2 & 3 :

  In next two weeks I have started learning about what is machine learning and how it's works. Learn about the different types of machine learning like supervised learning, Unsupervise learning and reinforcement learning. I have done preprocessing steps on different dataset and learn about all machine learning models for instance on which dataset which type of model can be used to get the efficient result and in 3rd week I have learning about fundamental of nueral network by learning the pytorch library.

- **Learning Activities of Week 4 & 5 :**

  In the fourth week, I deepened my understanding of machine learning regression, focusing on predicting continuous outputs based on input data. I learned about the architecture of a regression neural network and how to convert data into tensors, a crucial step for efficient computation in PyTorch. I coded a neural network for regression tasks using the torch.nn.Sequential module, which simplifies model construction by stacking layers sequentially.I explored essential components like loss functions, specifically Mean Squared Error (MSE), and various optimizers like Stochastic Gradient Descent (SGD) and Adam, which help minimize the loss and improve model performance. I also learned how to handle model logits and convert them into prediction probabilities, crucial for making accurate predictions.To evaluate the model, I implemented training and testing loops, allowing for iterative training and performance assessment on test data. I created a simple straight-line dataset to understand how the regression model predicts outputs and evaluated the model's predictions against actual values.Lastly, I delved into the concept of non-linearity by using non-linear activation functions, which enable neural networks to model complex relationships in data. This comprehensive approach provided a solid foundation in building, training, and evaluating regression models using PyTorch.

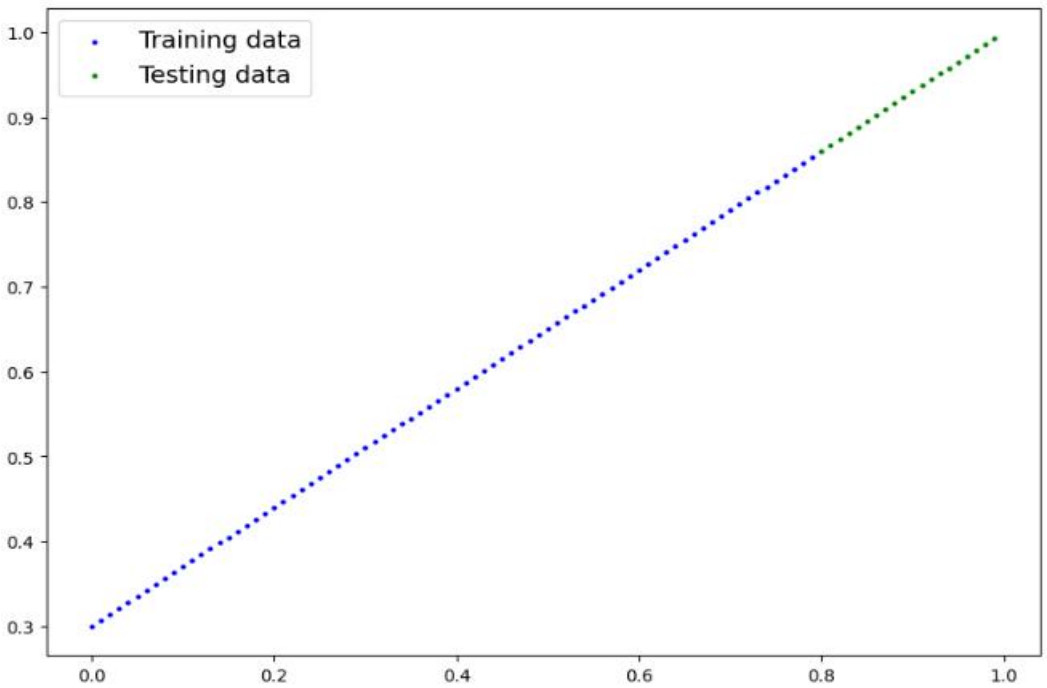


Figure 2.1 Linear regression model

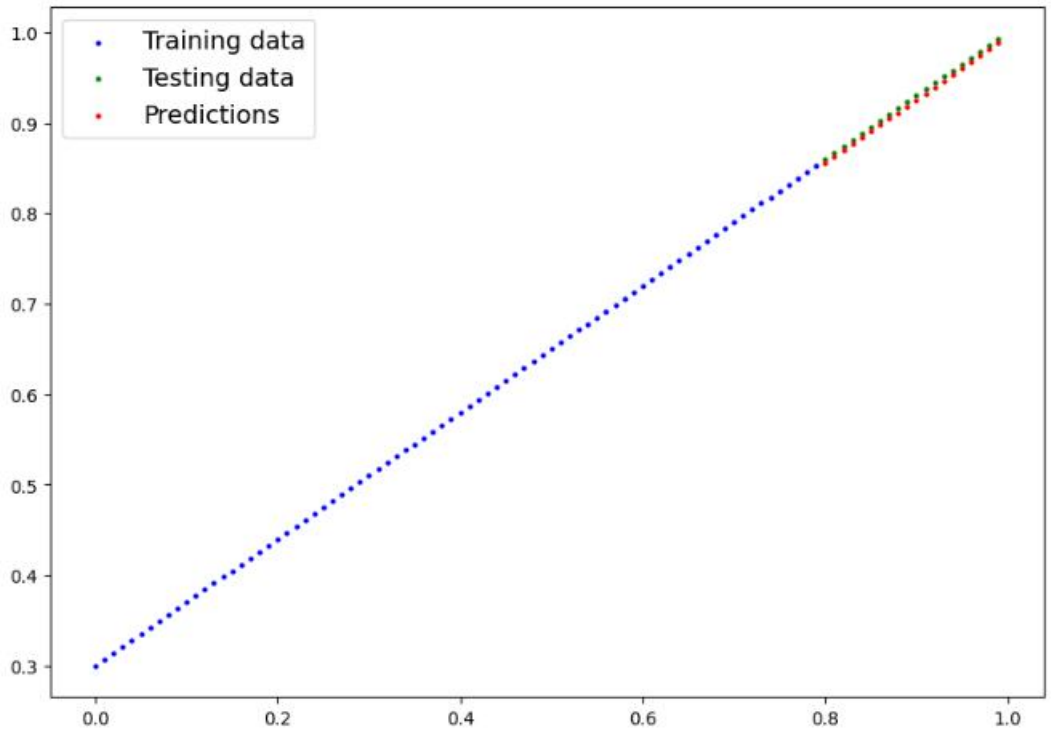


Figure 2.2 Prediction done by using linear regression model

- ## Learning Activities of Week 6 & 7 :

  In the next two weeks, I gained knowledge related to the fundamentals of machine learning classification. I studied how to classify inputs and outputs and understood the architecture of a classification neural network. A key part of this learning involved turning our data into tensors, the fundamental data structures in PyTorch, to facilitate efficient computation.I coded a neural network specifically designed for classification tasks using the torch.nn.Sequential module. This module allows for easy and clear stacking of neural network layers. I also explored essential components like loss functions, such as Cross-Entropy Loss, and optimizers, including Stochastic Gradient Descent (SGD) and Adam, which are critical for minimizing the loss and improving model performance.I learned how to handle model logits, convert them into prediction probabilities, and ultimately into prediction labels, which are essential steps for making accurate classifications. To evaluate the model, I implemented training and testing loops to iteratively train the neural network on the training data and assess its performance on the testing data.Additionally, I created a simple straight-line dataset to understand how the classification model predicts outputs and evaluated the model's predictions against

actual values. Finally, I delved into the concept of non-linearity by using non-linear activation functions, which enable neural networks to model complex relationships within the data. This comprehensive approach provided me with a solid foundation in building, training, and evaluating classification models using PyTorch.

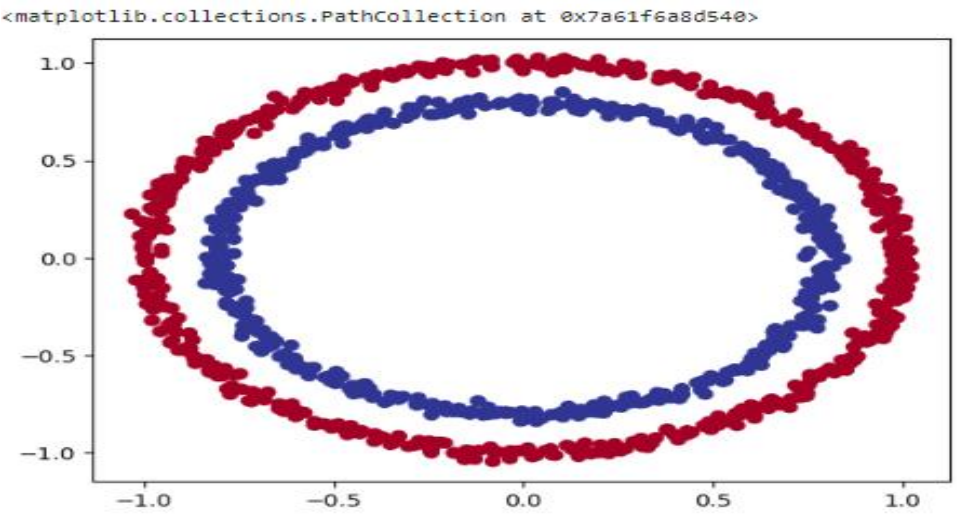


Figure 2.3 Binary classification model

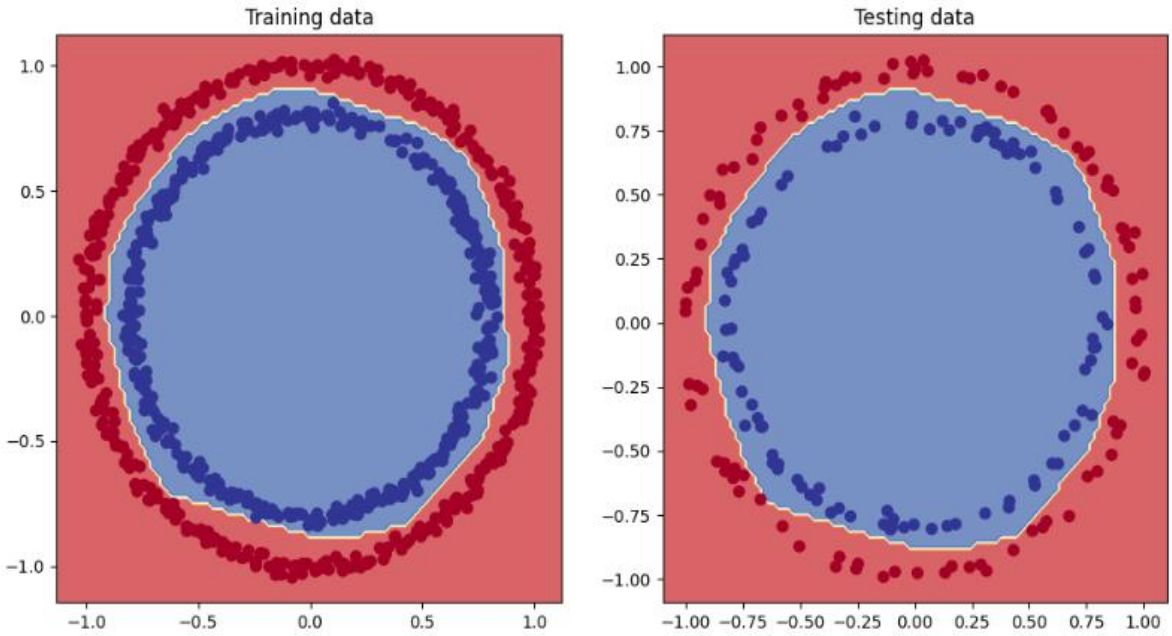


Figure 2.4 Model output of train and test data

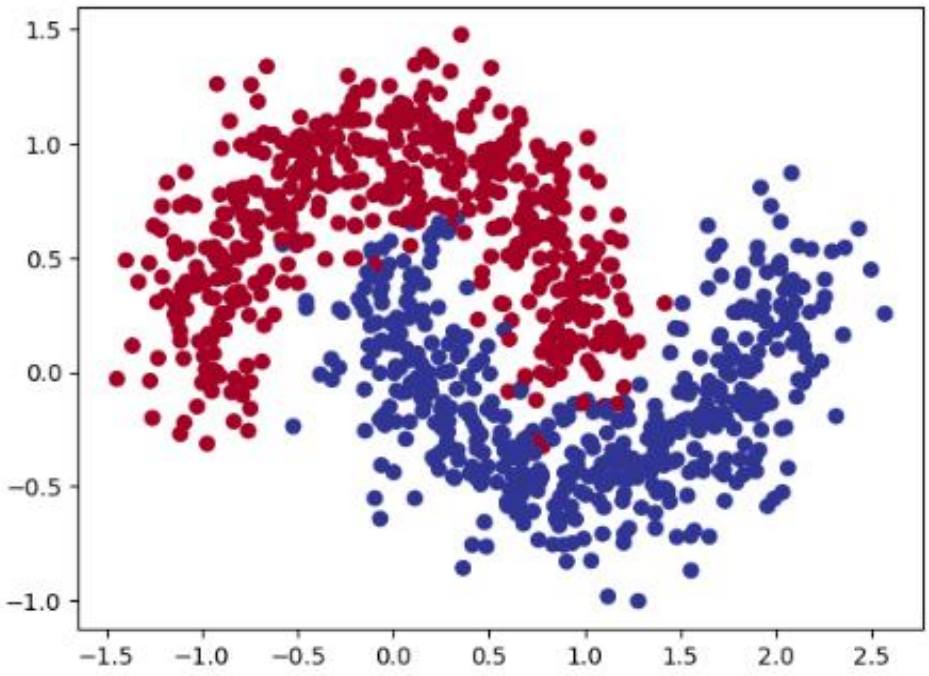


Figure 2.5 Binary classification model dataset

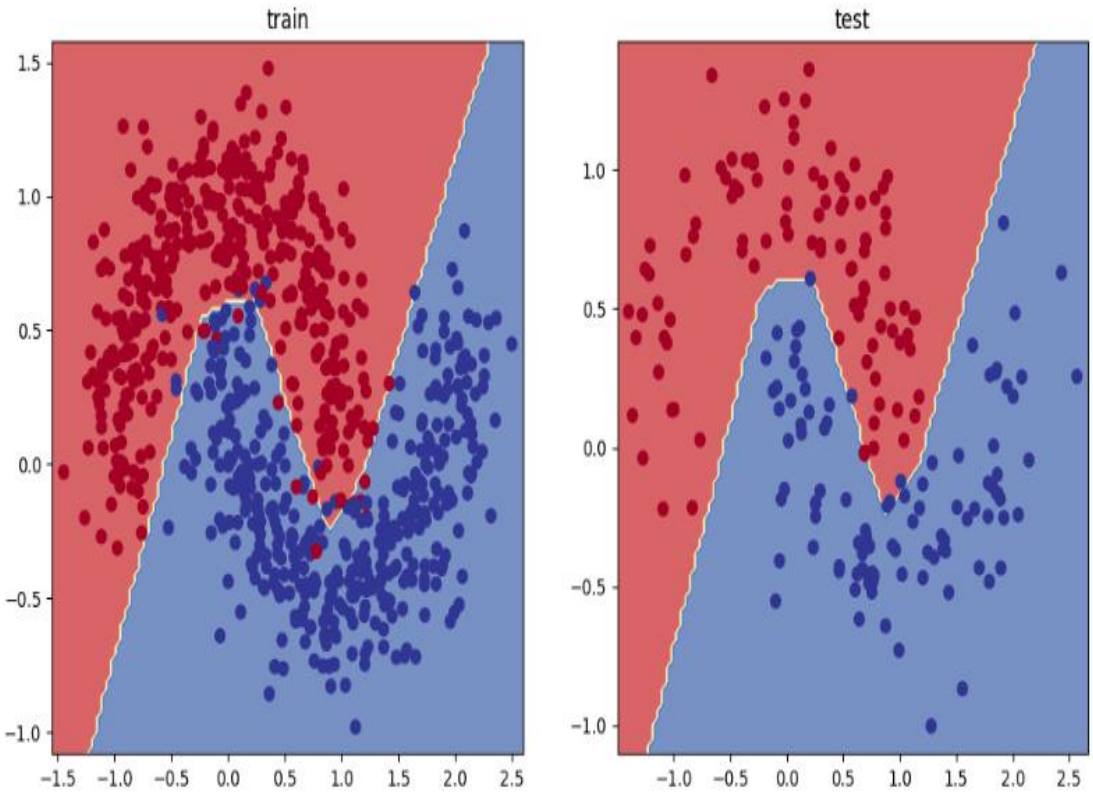


Figure 2.6 Output of train and test data of binary classification model

- **Learning Activities of Week 8 & 9 :**

  I applied the techniques I learned to build a multiclass classification model. This involved several key steps, starting with the classification of inputs and outputs to ensure the data was properly labeled for multiple categories. I designed the architecture of a classification neural network, which included determining the appropriate number of layers and nodes for the task.Next, I converted the data into tensors, the fundamental data structures in PyTorch, to facilitate efficient computation. I then coded the neural network for classification using the torch.nn.Sequential module, which allowed me to stack layers in a clear and straightforward manner. I selected appropriate loss functions, such as Cross-Entropy Loss, and optimizers like Stochastic Gradient Descent (SGD) and Adam to minimize the loss and improve the model's accuracy.To move from model logits to prediction probabilities and then to prediction labels, I implemented the necessary transformations and activation functions. I developed training and testing loops to iteratively train the neural network on the training data and evaluate its performance on the testing data

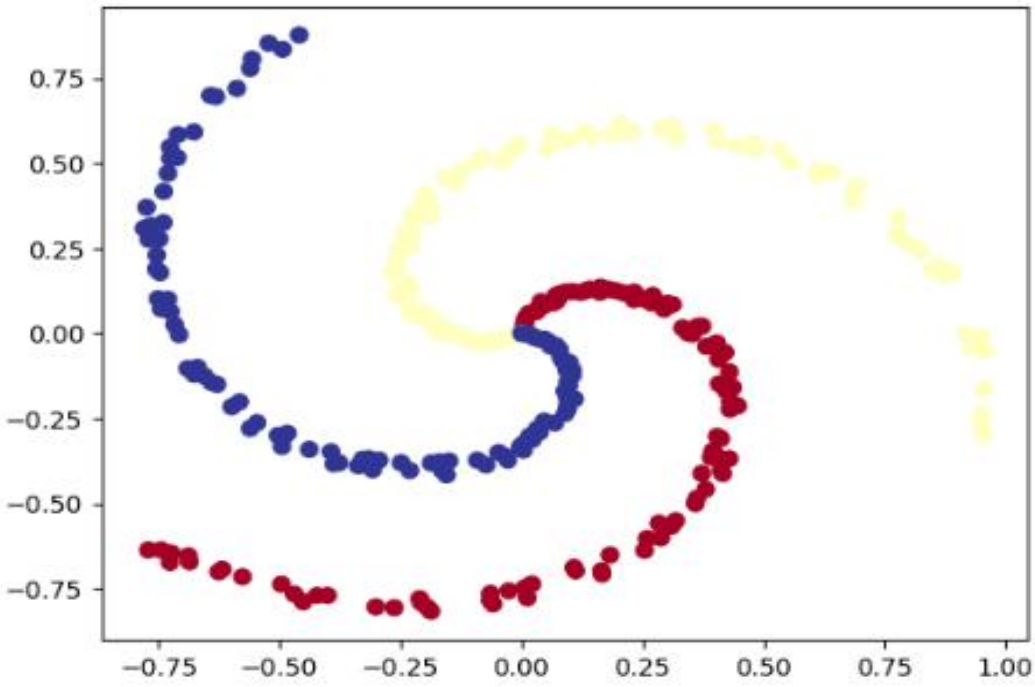


Figure 2.7 Spiral dataset for multiclass classification

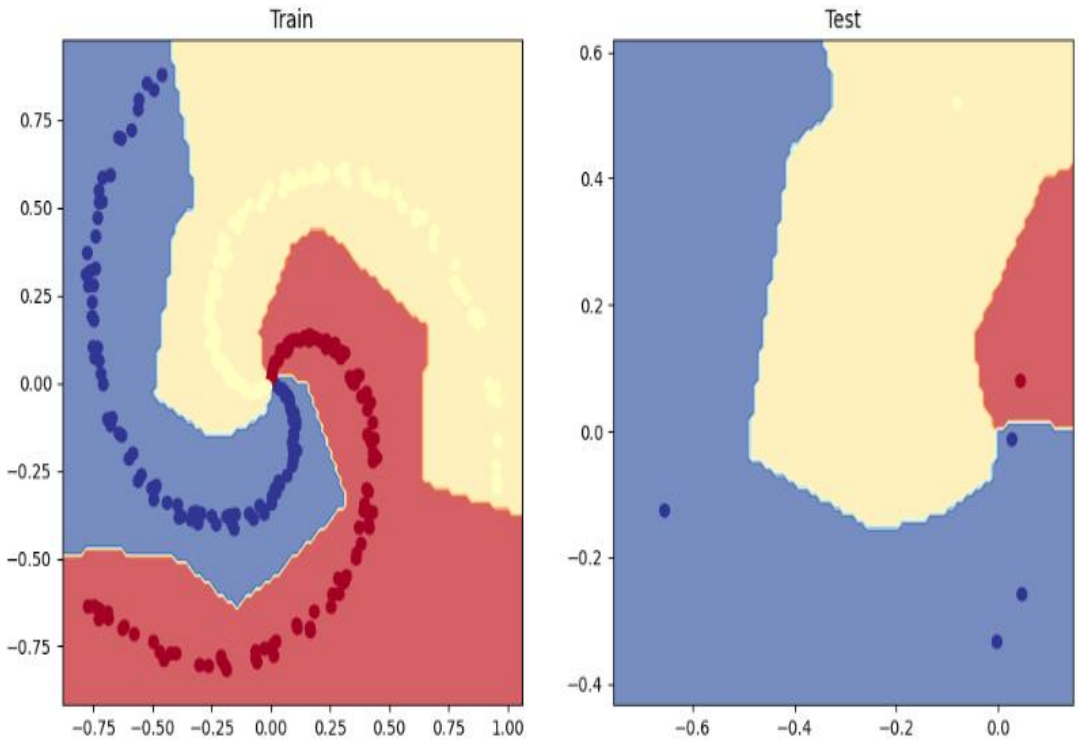


Figure 2.8 Output of train and test data of multiclass classification

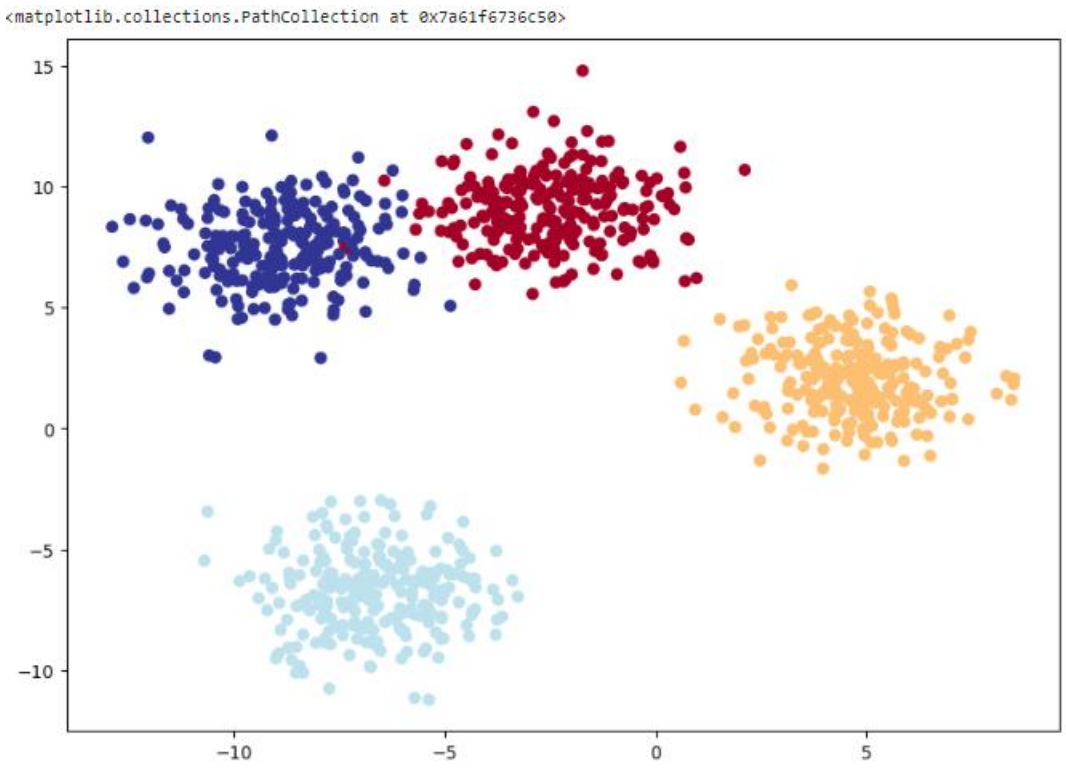


Figure 2.9 dataser for multiclass classification

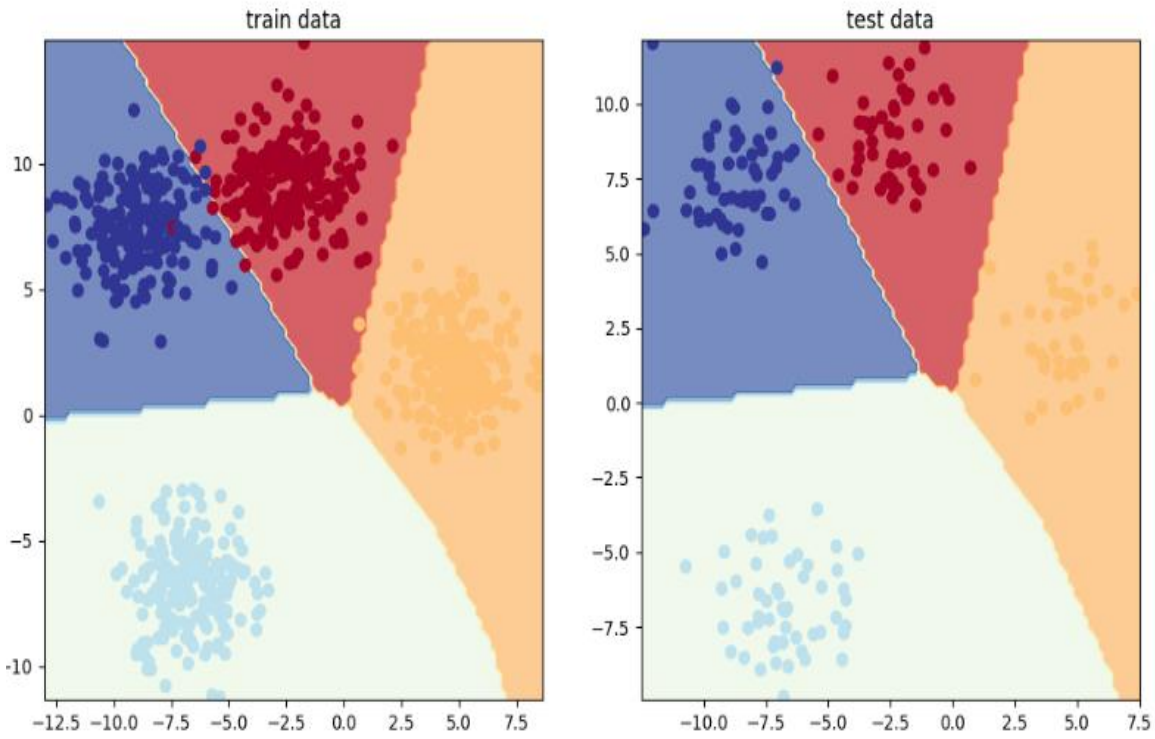


Figure 2.10 Output of train and test data of multiclass classification

- **Learning Activities of Week 10 & 11 :**

  In my upcoming project, I'm leveraging a Large Language Model to create an AI chatbot, facilitating seamless communication between AI systems and humans. The foundation of this application lies in LangChain, renowned for its capabilities in constructing powerful generative AI solutions. Through diligent study, I've gained a comprehensive understanding of the LangChain library and successfully integrated it into my project.One of the standout features I've developed is the "Chat with PDF Documents" functionality. Here, users can upload PDF files and inquire about their contents directly to the chatbot. Utilizing LangChain's capabilities, the chatbot processes these queries and provides accurate responses based on the information extracted from the PDF documents. This innovative feature not only showcases the versatility of AI-driven communication tools but also demonstrates the practical application of natural language interaction in extracting insights from PDF documents.

- **Learning Activities of Week 12 & 13 :**

  In preparation for future projects involving data scraping, I dedicated time to mastering the BeautifulSoup library, a powerful tool widely employed for extracting information from websites. This newfound skill will play a crucial role in my upcoming work, enabling me to efficiently gather data from various online sources. By harnessing BeautifulSoup's capabilities, I aim to enhance user experiences by providing access to relevant and valuable information. With its intuitive interface and robust functionality, the library empowers developers to navigate through web pages, locate specific content, and extract it seamlessly. Incorporating BeautifulSoup into my toolkit not only expands my repertoire of technical skills but also opens doors to innovative possibilities in data-driven applications.

- **Learning Activities of Week 14 & 15 :**

  In the development of an application titled "Chat with Website from URL," I integrated LangChain's chatbot functionality with a Streamlit GUI interface. This innovative application facilitates seamless conversations between AI and humans, enhancing user engagement and accessibility. Users can effortlessly upload URL links in the sidebar, initiating the data scraping process from the specified website. Once the data is extracted, users can interact with the chatbot through prompts, posing questions and receiving relevant query results in real-time. This feature-rich application not only streamlines the data retrieval process but also provides an intuitive platform for dynamic interactions, catering to diverse user needs with ease.

- **Learning Activities of Week 16 & 17 :**

  In preparation for upcoming machine learning projects, I've been tasked with mastering the sklearn library, a versatile toolkit for building classification and regression models. This essential learning endeavor equips me with the necessary skills to implement a wide range of machine learning algorithms effectively. By delving into sklearn's functionalities, I gain insight into its vast array of tools and techniques for model development and evaluation. From classification tasks, such as predicting categories or labels, to regression analysis, aimed at predicting continuous outcomes, sklearn offers comprehensive solutions for various machine

learning challenges. Armed with this knowledge, I am well-prepared to tackle future projects with confidence, leveraging sklearn's capabilities to create robust and efficient machine learning solutions.

- **Learning Activities of Week 18 :**

  In reinitiating the previous project, I aim to integrate two key features into the chatbot: interaction with both PDFs and website URLs. The application's functionality revolves around seamlessly incorporating these features, allowing users to upload either a PDF file or a website URL. Once uploaded, users can pose questions related to the content of the PDF or the website and receive accurate responses based on their queries. This integration enhances the versatility and usability of the chatbot, providing users with a comprehensive platform to access information from various sources. By combining these capabilities, the application offers a seamless and efficient solution for users to interact with both PDF documents and online content, catering to diverse information needs effectively.

CHAPTER 3 SYSTEM PLANNING

3.1 PROJECT DEVELOPMENT APPROACH

For a project involving the development of an application that enables chattingwith a website or multiple PDFs using LangChain, OpenAI, and Streamlit, the **Agile Project Development** is a suitable fit. Here's why Agile is appropriate for this project:

- **Iterative and Incremental:** Agile is an iterative approach to software development where the project is divided into small, manageable units called sprints or iterations.
- **Flexible and Adaptive:** Agile is designed to be adaptable to changing requirements and feedback throughout the development process.
- **Collaborative:** Emphasizes close collaboration between cross-functional teams and stakeholders.
- **Customer-Centric:** Focuses on delivering value to the customer through continuous feedback and iteration.
- **Continuous Improvement**: Regular reviews and retrospectives to identify and implement improvements.
- **Rapid Deployment and Feedback Loops:** Agile allows for the deployment of working software at the end of each sprint, enabling quick feedback from users and stakeholders. This is particularly useful for applications that integrate emerging technologies like LangChain and OpenAI, where user experience and functional requirements can evolve rapidly.
- **Risk Management:** Agile's iterative approach helps in identifying and mitigating risks early in the development process. Regular testing and integration mean potential issues are addressed promptly, reducing the risk of significant project delays or failures.
- **User Stories and Prioritization:** Agile uses user stories to capture functional requirements from the user's perspective. This ensures that development focuses on features that provide the most value to users first, enhancing the overall user experience.

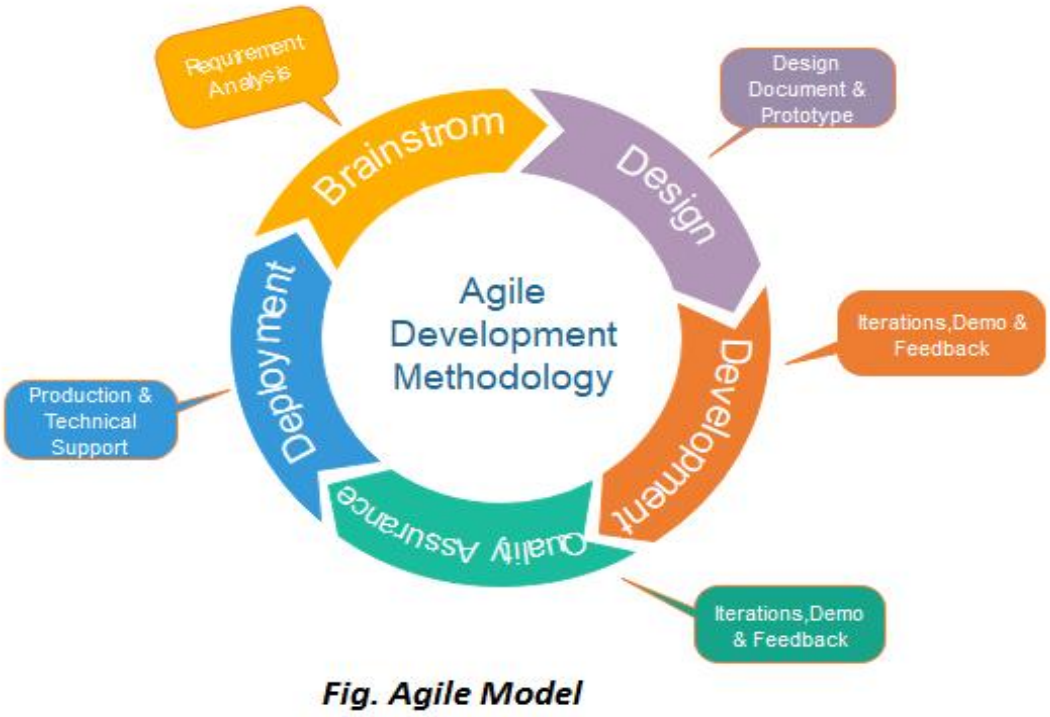


Figure 3.1 Agile Project Development Model

✧ **Justify: Why you have selected this model for your project?**

I opted for the Agile Project Development model for my project because it offers unparalleled flexibility, iterative progress, and a strong emphasis on collaboration and customer feedback. Agile's iterative cycles allow for frequent reassessment and adjustment of project goals and deliverables, ensuring the final product meets user needs effectively. This adaptability is crucial for projects where requirements can evolve rapidly, as Agile facilitates continuous improvement and responsiveness to change.Moreover, Agile promotes a collaborative environment where cross-functional teams work closely together, fostering innovation and problem-solving. Regular feedback loops with stakeholders ensure that any issues are identified and addressed promptly, enhancing the overall quality of the project. This model also supports incremental delivery, providing tangible results at the end of each iteration, which can be valuable for maintaining stakeholder engagement and satisfaction. Agile's focus on delivering a functional product at every stage reduces the risk of project failure and ensures that the project remains aligned with user expectations. By breaking down the project into manageable chunks and

prioritizing tasks, Agile helps in maintaining a clear vision and achieving better time management. Overall, Agile's dynamic and user-centric approach makes it an ideal choice for ensuring the success and adaptability of my project.

✧ **Advantages of your software model**

- High amount of risk analysis hence, avoidance of Risk is enhanced
- Strong approval and documentation control.
- Additional functionality can be added at a later date.
- Software is produced early in the Software Life Cycle.

3.2 SYSTEM MODULES

An image enhancement website typically consists of various modules, each serving a specific function. Here are some key modules :

- **PDF Loading:** The app reads multiple PDF documents and extracts their text content.
- **Text Chunking**: The extracted text is divided into smaller chunks that can be processed effectively.
- **Language Model:** The application utilizes a language model to generate vector representations (embeddings) of the text chunks.
- **Similarity Matching:** When you ask a question, the app compares it with the text chunks and identifies the most semantically similar ones.
- **Response Generation:** The selected chunks are passed to the language model, which generates a response based on the relevant content of the PDFs.
- **Large Language Model Integration:** Compatibility with models like GPT-4, Llama2 etc. In this code I am using GPT-4, but you can change it to any other model.
- **Streamlit GUI:** A clean and intuitive user interface built with Streamlit, making it accessible for users with varying levels of technical expertise.

3.3 FUNCTIONAL REQUIREMENTS

Table 3.1 Functional Requirements

ID	Title & Description
FR1	Title : Upload Multiple PDF Desc : User can upload more than one pdf in the given section.
FR2	Title : Prompt section Desc : User can ask question with the prompt bar.
FR3	Title : Splitting text into chuks Desc : The entire data of the document will be split into small chunks and ready for vectorazation.
FR4	Title : Data embedding Desc :The Uploaded data will be transform single time in vector database and user can ask queries with single time data embeddings.
FR5	Title : create a vector store for data Desc : All the data embeddings will be store in vector database.
FR6	Title : Ranked search Desc : The user queries will be use for semantic search with the data store in vector database.

3.4 NON FUNCTIONAL REQUIREMENTS

Nonfunctional Requirements (NFRs) define system attributes such as **security, reliability, performance, maintainability, scalability, usability, Portability, and Compliance.**

- **Security:** The system provides username and password to prevent the system from unauthorized access.
- **Reliability:** Reliability describes the ability of a solution or component to perform its required functions under stated conditions for a specified period of time.
- **Performance:** Performance Efficiency describes how quickly and predictably the system responds to user input or other events. In other words, this is about ensuring that your website is not slow.
- **Maintainability:** Maintainability describes ease with which a solution or component can be modified to correct faults, improve performance or other attributes, or adapt to a changed environment.
- **Scalability:** Scalability describes how easily the system can grow to handle more users, data, transactions, servers, or other extensions without compromising performance or correctness.
- **Usability:** The system provides a help and support menu in all interfaces for the user to interact with the system. The user can use the system by reading help and support
- **Portability:** This includes requirements related to the ability of the system to be easily transferred to different hardware or software environments.
- **Compliance:** This includes requirements related to adherence to laws, regulations, industry standards, or company policies.

3.5 TIMELINE CHART

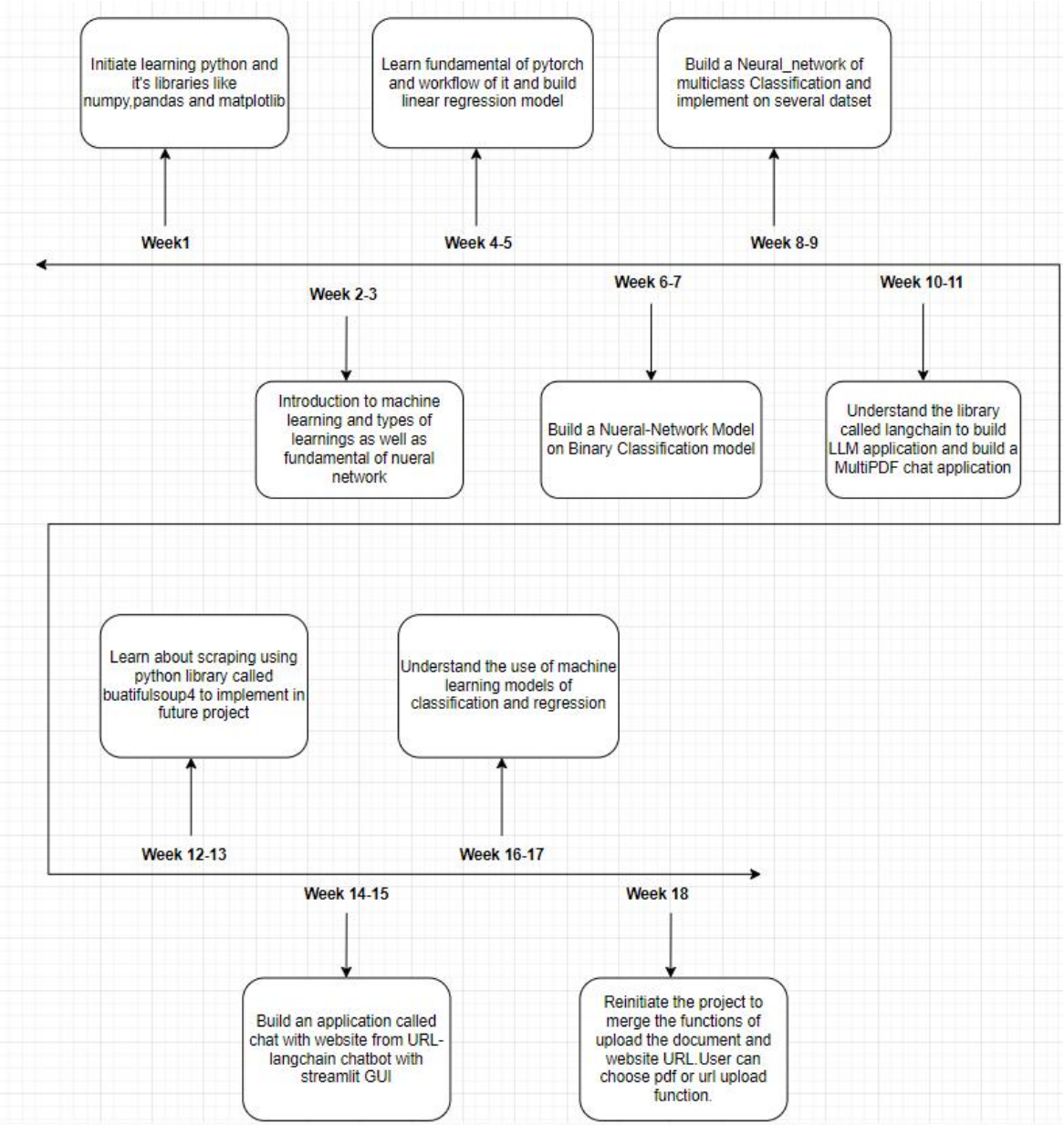
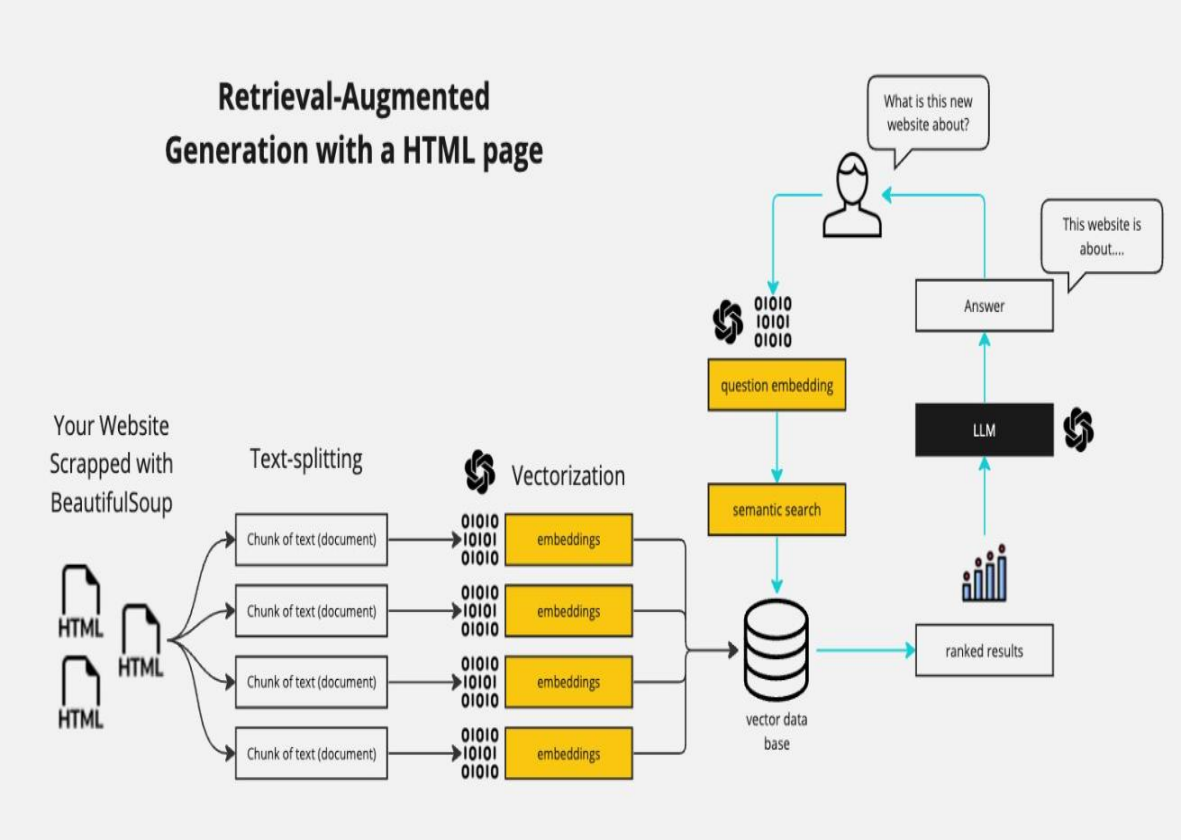


Figure 3.2 Timeline chart of entire internship period

CHAPTER 4 SYSTEM DESIGN

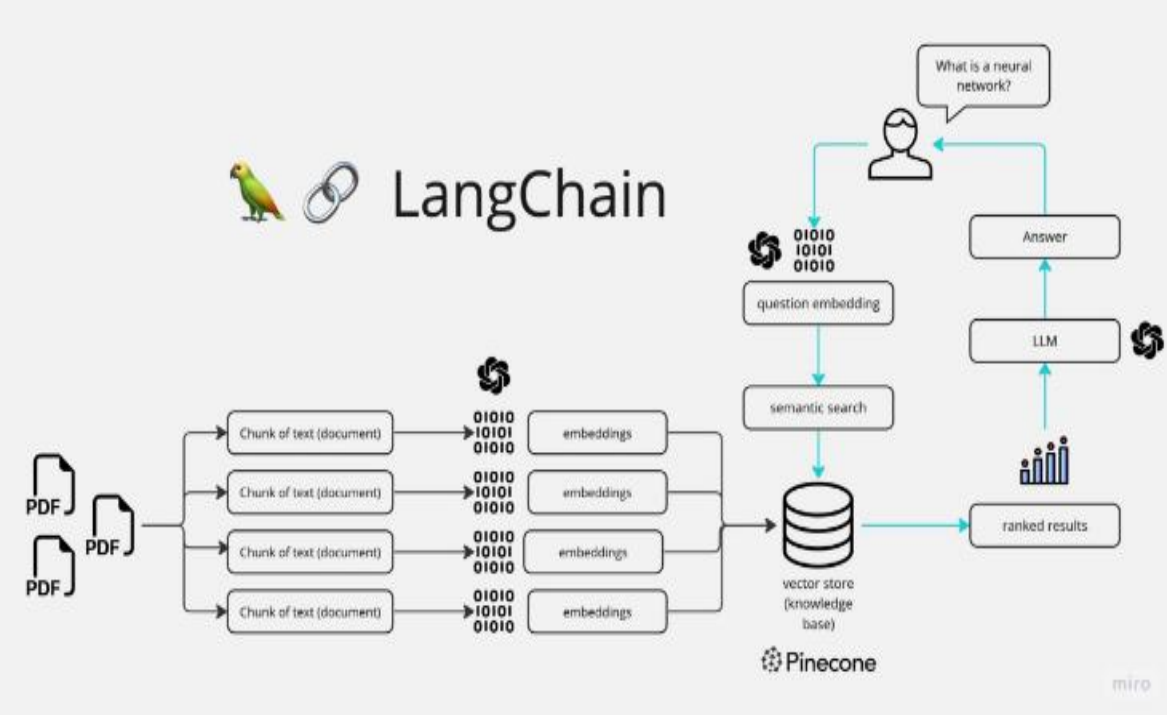
4.1 Diagram of chat with web application



4.1 Workflow of Chat With Website application

The diagram illustrates a workflow for retrieval-augmented generation using a website's HTML content. First, the website is scraped using BeautifulSoup, and the text is split into manageable chunks. These chunks are then vectorized to create embeddings, which are stored in a vector database. When a user queries about the website's content, the question is embedded and a semantic search is conducted against the vector database to find relevant text chunks. The most relevant results are ranked and passed to a large language model (LLM), which generates a coherent answer for the user.

4.2 Diagram of chat MultiPDF application



4.2 Workflow of Chat With MultiPDF application

The diagram depicts a workflow for using LangChain to process and retrieve information from PDF documents. Initially, the PDFs are divided into text chunks, which are then transformed into embeddings. These embeddings are stored in a vector store managed by Pinecone. When a user asks a question, it is converted into an embedding, and a semantic search is conducted against the vector store to find the most relevant text chunks. The ranked results are then provided to a large language model (LLM), which generates a suitable answer for the user.

CHAPTER 5 IMPLEMENTATION AND TESTING

5.1 HARDWARE AND SOFTWARE REQUIREMENTS

- **Software Requirements:**
  - **Operating System:** windows 7+ requirements
  - **Programming Language:** Python
  - **Framework and Libraries:** Langchain, Streamlit
  - **Database**: ChromaDB,Pinecone(Vector store and similarity search)
  - **Development tool:** VS Code
- **Hardware Requirements:**
  - **Processor:** Ryzen 3
  - **Hard Disk:** 2 GB RAM.
  - **Memory:** 512 GB

5.2 SNAPSHOTS

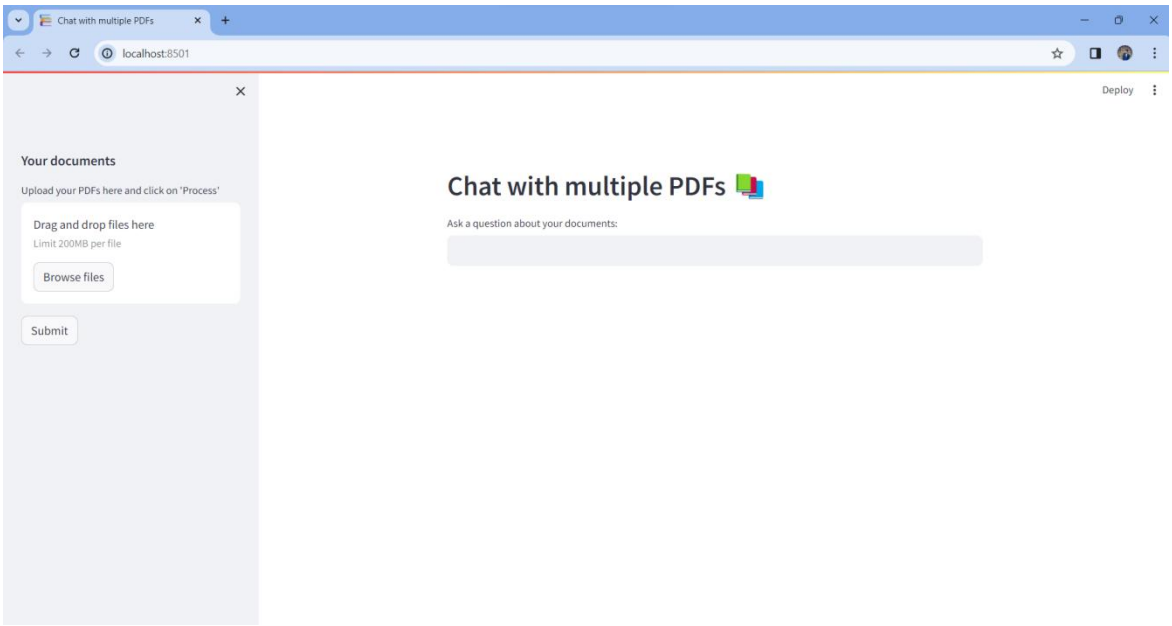


Figure 5.1 Home page of Chatbot Application

User need to upload PDF document shown in the left sidebar.

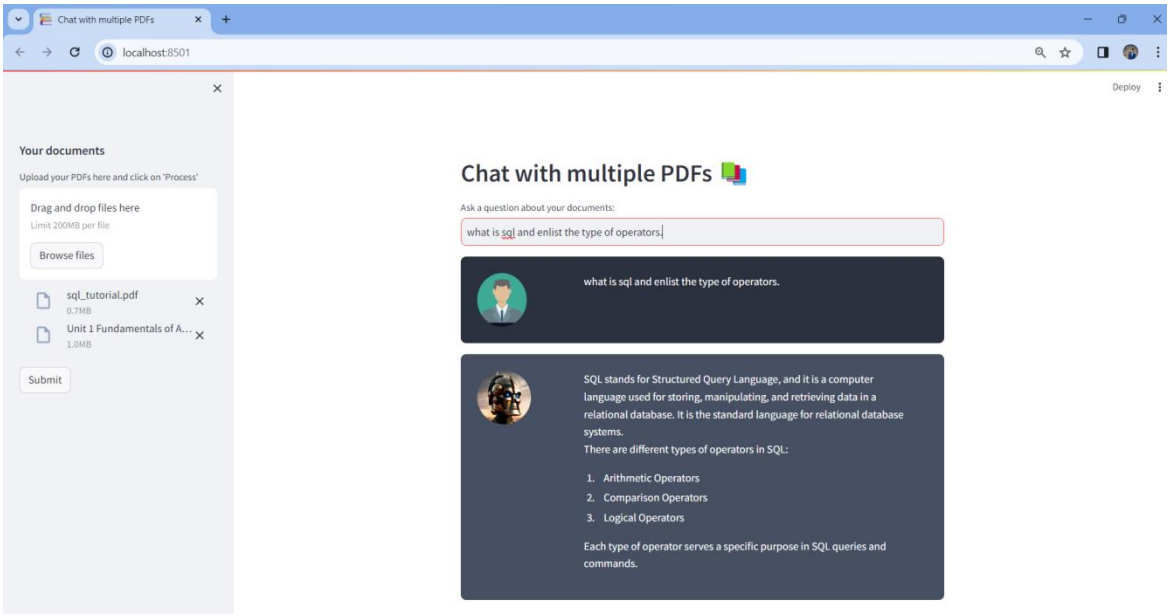


Figure 5.2 Interaction with Chatbot

User can ask question based on the uploaded pdf and get answer from chatbot

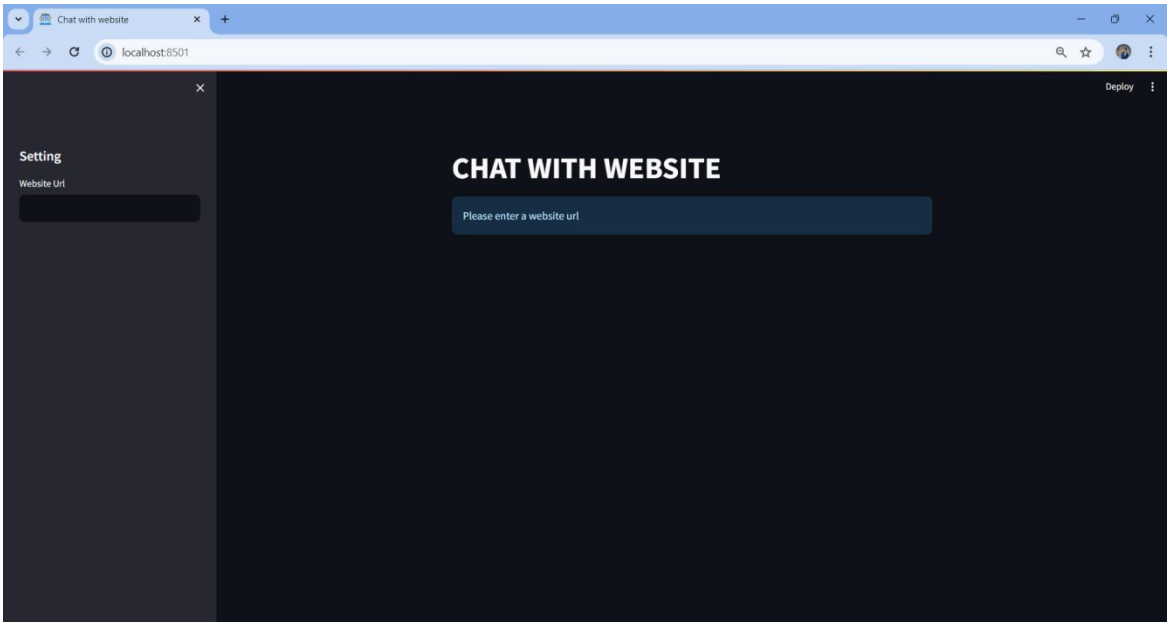


Figure 5.3 Home page of Chatbot Application through website URLS

User need to upload website URL shown in the left sidebar.

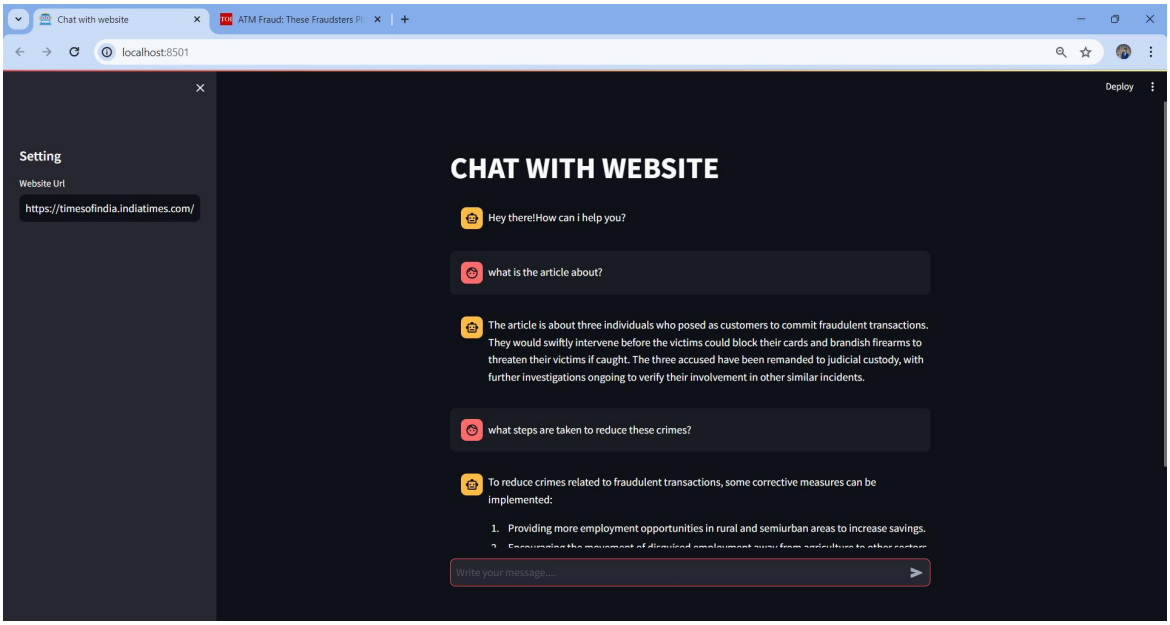


Figure 5.4 Interaction with Chatbot

User can ask question based on the uploaded pdf and get answer from chatbot

5.3 TEST CASES

Sample test cases are given as below in

Test Table 5.1

TestID	Case	Test Data	Expected Result	Actual Result	Pass/Fail
001	User asks a question about the content of a specific PDF	"What is the main argument in Chapter 3 of PDF1?"	The response accurately summarizes the main argument in Chapter 3 of PDF1	As expected system has provide accurate answr of the question.	Pass
002	User asks a general question about the website content	"What are the services offered by the website?"	The response lists the services offered by the website	Response correctly summarizes or answers the query based on	Pass

				the website content	
003	User inputs multiple PDFs and asks a question that spans content in all PDFs	"Compare the main arguments of PDF1 and PDF2."	The response accurately compares the main arguments of PDF1 and PDF2	In some complex pdf files, System rarely producing bias data.	Pass
004	User asks a question not relevant to any provided content	"What is the capital of France?" with no relevant content in PDFs or website	The response indicates that the information is not available in the provided content	The response was expected with no such content or data is available.	Pass
005	User requests a summary of a specific section of a PDF	"Summarize the introduction of PDF2."	The response provides an accurate summary of the introduction of PDF2	The summary of the asked queries was absolutely accurate and easy to comprehend.	Pass
006	User asks a follow-up question related to the previous query	Initial question: "What are the key findings of PDF1?" Follow-up: "Can you explain the first finding in more	The response maintains context and provides a detailed explanation of the first finding of	The response maintains context and provides a detailed explanation of the first finding of PDF1	Pass

		detail?"	PDF1		
007	User requests the application to display a specific PDF	"Show me PDF3."	The application displays or provides access to PDF3	The application displays or provides access to PDF3	Pass

CHAPTER 6 CONCLUSION AND FUTURE SCOPE

In conclusion, the LangChain Chatbot with Streamlit GUI offers a seamless and user-friendly platform for engaging with a website through natural language conversation. By integrating advanced language processing capabilities with an intuitive interface, it provides an efficient means for users to interact with web content, seek information, and accomplish tasks more effectively.

In future work, Integrate Chatbot in Website to smooth the interaction with user requirements. Finetune the model for enhance the user experience.Combine the feature for upload option like website URL or PDF documents, For Instance, User can choose either pdf or website URL.

REFERENCE

Web references

[1] [Online] **Pytorch Documentation:** https://pytorch.org/

[Date Accessed: 10 JAN 2024].

[2] [Online] **Langchain Documentation:** https://www.langchain.com/

[Date Accessed: 25 FEB 2024].

[3] **[**Online**] Streamlit Documentation:** https://streamlit.io/

[Date Accessed: 25 FEB 2024].

[4] [Online] **Beutifulsoup4_Documentation:** https://beautiful-soup-4.readthedocs.io/

[Date Accessed: 15 APRIL 2024].