Student 1 – Lorena Gomez – 21734359

Student 2 – Darragh Manning - 21506373

Supervisor - Tomas Ward

Project Name – TimelineXtract

Functional Specification - 11th November 2024

Functional Specification Contents

1. Introduction

Overview

Clinical trials are crucial for testing new treatments, but they are often slowed down by the manual work involved in managing complex protocols, particularly those involving patient and clinician-reported outcomes. Our project seeks to address this by developing a machine learning-powered application to extract and streamline PRO and ClinRO schedules and assist patients with procedural guidance through AI-generated instructional videos/images.

Business Context

This system could be deployed within organizations that conduct clinical trials, such as pharmaceutical companies, clinical research organizations (CROs), and healthcare institutions involved in the development and testing of new medical treatments. These organizations would benefit from the automation of manual tasks, allowing for faster and more accurate trials, potentially accelerating the approval of new drugs and treatments.

Our project also leverages publicly available clinical trial documents from ClinicalTrials.gov, which provides access to protocols, patient information sheets, and other relevant data. This allows us to test and refine our machine learning models on a diverse range of clinical trial formats.

Glossary

- **PRO (Patient-Reported Outcomes):** Feedback provided directly by the patient on how they feel during the trial.
- **ClinRO (Clinician-Reported Outcomes):** Observations and assessments recorded by clinicians during the trial.
- **NLP (Natural Language Processing):** A branch of AI focused on understanding and processing human language.
- **Generative AI:** AI technology that can create new content, such as images or videos, from textual descriptions.
- **GDPR (General Data Protection Regulation):** A regulation in the European Union designed to protect individuals' personal data and privacy. It applies to any organization processing personal data of EU residents and imposes strict guidelines on data collection, storage, and sharing.

2. General Description

2.1 Product/System Functions

The application will:
1. Extract timelines for PROs and ClinROs from clinical trial protocols using NLP.
2. Generate instructional videos or images to help patients follow procedures at home, based on provided trial documents.

The system will also utilize publicly available databases, such as ClinicalTrials.gov, to access clinical trial protocols and patient information documents, which serve as input for the NLP and AI-based media generation features.

2.2 User Characteristics and Objectives

Vendors: The vendors (clinical trial managers) are expected to have intermediate-level knowledge of clinical trial procedures and protocols. Their primary objective is to reduce the time and effort spent on extracting and organizing trial schedules.

Patients: Patients participating in trials may have varying levels of comfort with technology, so the AI-generated videos and images aim to simplify instructions, ensuring they follow trial procedures correctly.

2.3 Operational Scenarios

Scenario 1: A vendor uploads a clinical trial protocol into the application. The system analyzes the document and automatically generates a calendar for the trial's PRO and ClinRO schedules.

Scenario 2: A patient receives an instructional video generated by the system, showing them how to perform a specific medical test at home, following the steps described in the "Patient Information Sheet."

2.4 Constraints

The system must handle diverse protocol formats and non-standardized document structures.

It should ensure patient privacy and compliance with healthcare regulations such as GDPR.

3. Functional Requirements

Requirement 1: NLP-driven Questionnaire Extraction (PROs and ClinROs)

Description: The system must automatically extract patient-reported outcomes (PROs) and clinician-reported outcomes (ClinROs) questionnaires from clinical trial protocols. This will involve identifying and parsing sections of the protocol that contain relevant questionnaires and associated details.

Criticality: High, as accurate extraction of PRO and ClinRO questionnaires directly impacts the ability to provide meaningful data for clinical trial management and patient monitoring.

Technical Issues: Variability in how questionnaires are formatted or represented in trial documents may require highly adaptive NLP techniques for accurate extraction. The system must handle various document structures, including tables, bullet points, or free-text descriptions.

Dependencies: Requires access to well-structured clinical trial protocols that contain relevant sections with questionnaires, including text or table formats, and the use of advanced NLP models for pattern recognition and data extraction.

Requirement 2: NLP-driven Timetable Extraction

Description: Automatically extract timelines for PROs and ClinROs from clinical trial protocols.
Criticality: High, as it directly impacts the efficiency of trial management.
Technical Issues: Variability in document formats may require a highly adaptive NLP algorithm.
Dependencies: Requires access to well-structured protocol documents and metadata.

Requirement 3: AI-based Media Generation for Patients

Description: Generate instructional videos/images based on text instructions from the Patient Information Document.
Criticality: Medium, important for improving patient compliance but secondary to timetable extraction.
Technical Issues: Ensuring accurate representation of the procedure in the generated media.
Dependencies: Dependent on the quality of the text input from the trial documents.

Requirement 4: User Feedback System

Description: The system should allow both vendors and patients to provide feedback on the accuracy and usefulness of extracted schedules and generated media, which can be used to improve the system.
Criticality: Medium, useful for continuous improvement of the system.
Technical Issues: Creating a feedback loop mechanism that collects and analyses user input.
Dependencies: Integration with a feedback management system (e.g., a simple database or a third-party tool).

Requirement 5: System Scalability

Description: The system should be scalable to handle increasing amounts of clinical trial data and users without significant degradation in performance.
Criticality: High, as large-scale clinical trials could require handling significant volumes of data.
Technical Issues: Efficient load balancing, database partitioning, and horizontal scaling.
Dependencies: Use of scalable cloud infrastructure (e.g., AWS, Google Cloud) and databases that can grow with the application.

Requirement 6: User Authentication

Description: The system must implement a secure user authentication mechanism to ensure that only authorized individuals can access the system and its data. Users will need to authenticate via a secure login process before interacting with the system.
Criticality: High, as unauthorized access to clinical trial data or patient information must be prevented to ensure data privacy and regulatory compliance.
Technical Issues: The system must handle secure authentication methods, including password management and two-factor authentication (2FA) if necessary.
Dependencies: Use of a reliable authentication service (e.g., OAuth, Google Authentication, or custom authentication mechanisms) to ensure secure login.

Requirement 7: Data Security and Compliance

Description: The system must implement strong data encryption and security measures to protect sensitive patient information and clinical trial data.
Criticality: High, as the system deals with sensitive personal and clinical data.
Technical Issues: Compliance with GDPR, HIPAA, and other healthcare regulations must be ensured. Data encryption during both storage and transmission will be required.
Dependencies: Integration with secure cloud storage services (e.g., AWS S3 with encryption) and use of secure communication protocols (e.g., HTTPS).

Requirement 8: Data Integrity and Validation

Description: The system must implement validation mechanisms to ensure the integrity and accuracy of the extracted data, including timelines, schedules, and patient instructions.
Criticality: High, as incorrect or missing data could lead to significant issues with clinical trial execution and patient safety.
Technical Issues: Requires the ability to cross-check extracted data against predefined templates or expected formats to ensure consistency.
Dependencies: Integration with data validation tools or algorithms to verify the correctness of extracted schedules and media content.

Requirement 9: Media Quality Assurance

Description: The system must include mechanisms to review and verify the quality of the generated instructional media, ensuring the media is clear, accurate, and adheres to clinical guidelines before being delivered to patients.
Criticality: High, as inaccurate or unclear instructional media could lead to patient confusion and errors in following trial protocols.
Technical Issues: Implementing a quality assurance process that ensures the generated videos/images are of sufficient quality, both in terms of content accuracy and visual clarity. This may require a review or approval process before media is sent to patients.
Dependencies: Integration with manual or automated review systems for quality checking, possibly involving clinical experts or AI-based content validation.

Requirement 10: Multi-format Media Support (Video, Images, Animations)

Description: The system should support generating media in multiple formats, including videos, images, and animations, based on the complexity and nature of the patient instructions. Different formats should be generated depending on the task or procedure described in the Patient Information Document.
Criticality: Medium, as some tasks may be better suited for images, others for videos, and some may benefit from animations. Providing flexibility in media format enhances patient understanding.
Technical Issues: Determining the optimal format for different types of instructions and ensuring the generative AI system can handle various output types effectively.
Dependencies: Use of multi-modal generative AI models capable of creating video, image, and animated content from textual descriptions (e.g., text-to-video, text-to-image technologies).

Requirement 11: Adaptive Feedback for Media Refinement

Description: The system must be capable of refining the generated media based on patient feedback. If patients report that they are having trouble understanding the instructions, the system should be able to adjust the content (e.g., provide clearer visual cues or add additional guidance).
Criticality: Medium, as this feedback loop would ensure that patients receive the best possible support throughout the trial, improving adherence.
Technical Issues: Building a mechanism to capture patient feedback on media clarity and integrating it into the media refinement process. This will require natural language understanding for feedback interpretation and automated content adjustment.
Dependencies: Integration with the user feedback system, as well as AI-driven content modification capabilities.

4. System Architecture

The system is divided into **frontend** and **backend** components, with additional support from external tools and services. The architecture is designed to streamline the extraction of patient and clinician-reported outcomes (PROs and ClinROs) from clinical trial protocols, as well as to generate instructional media (videos or images) for patients.

Frontend

- **User Interface (UI):**
  - Allows users (vendors and patients) to interact with the system.
  - Vendors upload clinical trial protocols, while patients view or receive instructional content.
- **Document Upload Module:**
  - Provides an interface for users to upload PDF documents containing clinical trial protocols.
  - Directs the uploaded files to the backend for processing.

Backend

- **NLP Processing Module:**

    - Extracts schedules and timelines for PROs and ClinROs from clinical trial protocols.
    - Powered by natural language processing (NLP) techniques, this module handles the parsing of text and table extraction from non-standardized clinical documents.
- **Data Validation Steps:**
    - Validates the extracted information to ensure accuracy and consistency.
    - Ensures that the data extracted by the NLP module is reliable before further processing.
- **Generative AI Module:**
    - Converts textual descriptions (e.g., from a patient information document) into instructional videos or images.
    - Uses third-party generative AI models to create media content that patients can easily understand and follow.
    - This module is critical for assisting patients in conducting medical procedures at home.

Data Storage

- **MongoDB:**
    - A third-party NoSQL database used to store the structured data extracted from clinical trial protocols.
    - Stores both the schedules generated by the NLP module and any media content produced by the generative AI module.
    - Provides scalability and flexibility to handle diverse data types.

External Services/Third-Party Tools

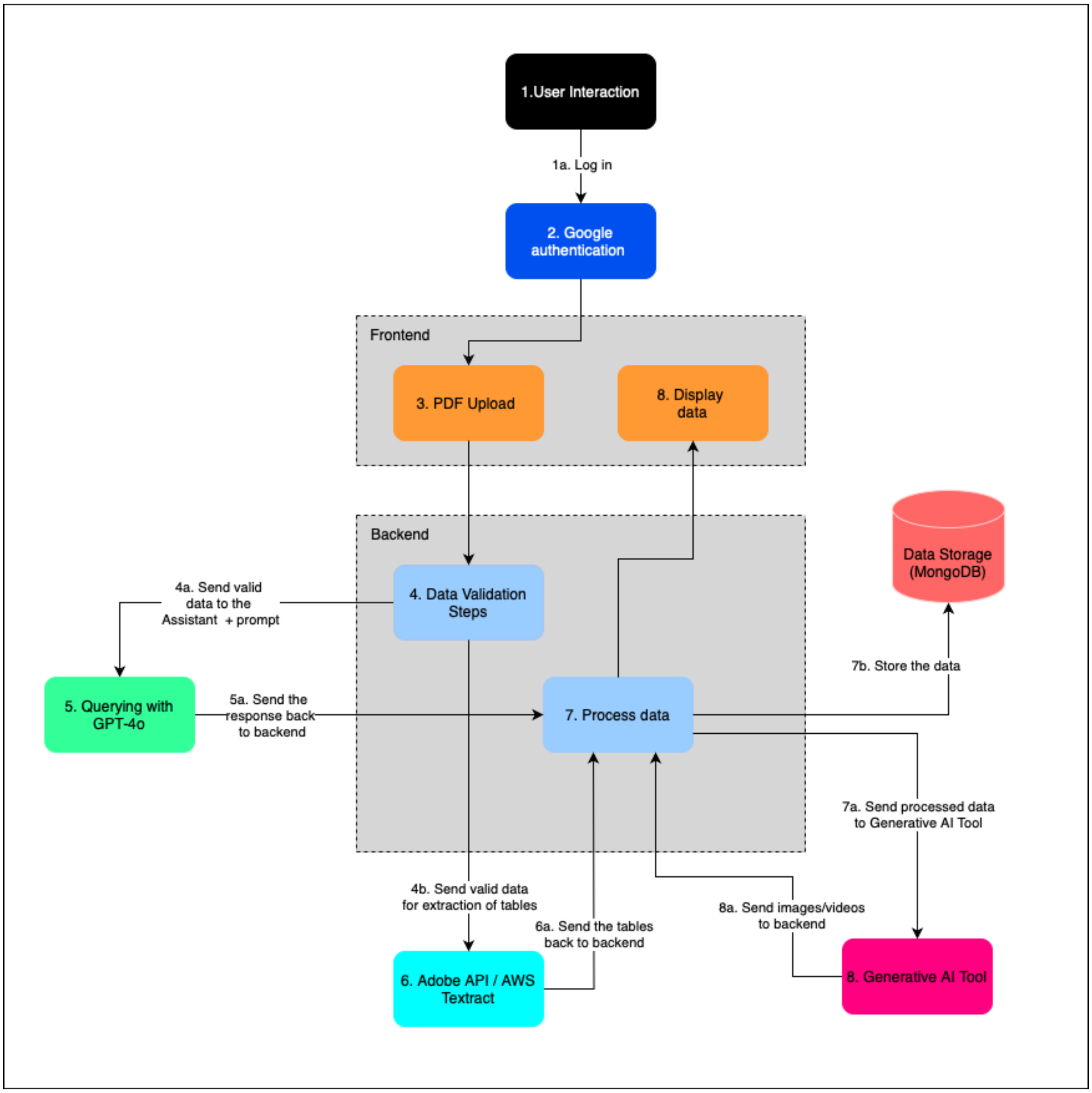
- **Adobe API / AWS Textract (for Document Parsing):**
    - Used to process PDFs and extract relevant text and tables from clinical trial protocols before sending the content to the NLP module.
    - These APIs provide efficient and reliable document parsing, allowing the system to handle different formats and structures.
- **GPT-4 API (for Querying and NLP Processing):**
    - Integrated for advanced natural language understanding and to enhance the performance of the NLP processing module.
    - Handles complex queries related to clinical data and outcomes.
- **Generative AI Tool (for Creating Images/Videos):**
    - A third-party generative AI tool is integrated into the system to convert text-based instructions (such as those found in patient information sheets) into images or videos.
    - This tool is crucial for creating step-by-step instructional content for patients to follow medical procedures at home, such as administering a COVID test.
    - The generated media ensures that patients receive personalized, easy-to-understand visual guidance, increasing adherence to clinical protocols and reducing errors.
    - This generative AI tool supports both image generation and video creation, leveraging cutting-edge AI models to produce high-quality content that accurately reflects the input text.

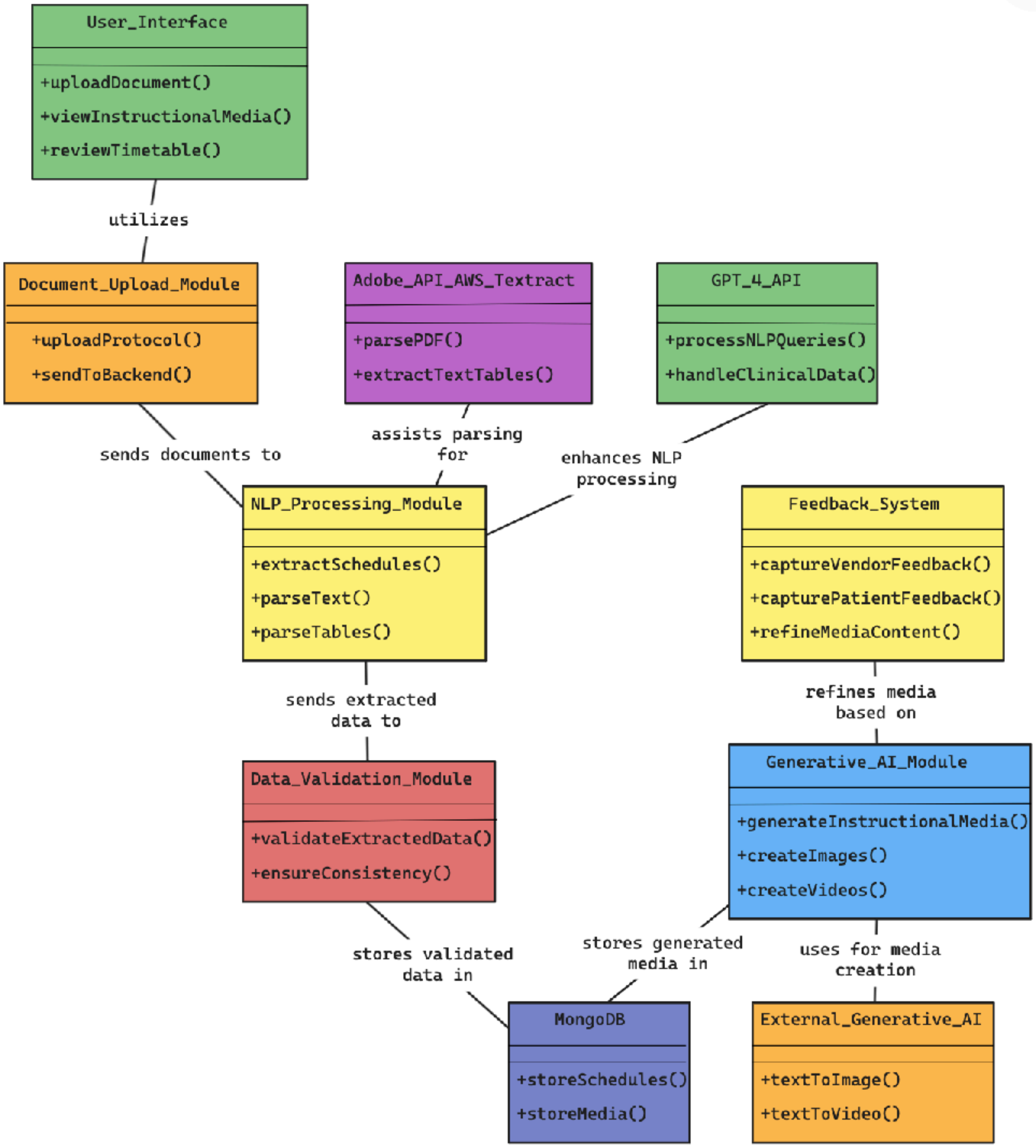
5. High-Level Design

NLP Module: Responsible for parsing and extracting relevant timetable data from unstructured clinical trial protocols.

Generative AI Module: Utilizes text-to-image/video techniques to create instructional media.

User Interface: Separate views for vendors (document upload and timetable review) and patients (viewing instructional media).


1.User Interaction
1a. Log in
2. Google authentication
Frontend
3. PDF Upload
8. Display data
Backend
4. Data Validation Steps
4a. Send valid data to the Assistant + prompt
5. Querying with GPT-4o
5a. Send the response back to backend
7. Process data
Data Storage (MongoDB)
7b. Store the data
7a. Send processed data to Generative AI Tool
4b. Send valid data for extraction of tables
6a. Send the tables back to backend
8a. Send images/videos to backend
6. Adobe API / AWS Textract
8. Generative AI Tool



6. Preliminary Schedule

***October** – Initial Phase (Research and Requirements Gathering)*

- **Week 1-2 (Mid-October to End of October):**
  - Finalize project requirements.
  - Gather clinical trial protocols and patient information documents from ClinicalTrials.gov for initial testing.
  - Research and select third-party tools (e.g., Adobe API, AWS Textract, GPT-4, Generative AI tools).

November *– Model Development and System Setup*

- **Week 3-4 (Early November to Mid-November):**
  - Develop the Natural Language Processing (NLP) module for extracting timelines from clinical trial protocols.
  - Integrate document parsing tools (Adobe API / AWS Textract) into the backend.
  - Initial database setup (MongoDB) for storing extracted data.
- **Week 5-6 (Mid-November to End of November):**
  - Develop and test the data validation module to ensure extracted information is accurate.
  - Begin integrating the GPT-4 API for advanced querying and NLP capabilities.

December *– AI Integration and Prototyping*

- **Week 7-8 (Early December to Mid-December):**
  - Start developing the Generative AI module for generating images and videos.
  - Integrate the text-to-image and text-to-video generation tools.
  - Prototype user interface for vendors to upload protocols and review generated schedules.
- **Week 9 (Mid-December to End of December):**
  - Internal testing of NLP and data validation modules.
  - Initial prototyping of the media generation process.

January *– Frontend and Backend Integration*

- **Week 10-11 (Early January to Mid-January):**
  - Complete the frontend development (UI for uploading documents and viewing results).
  - Finalize integration of the backend (NLP and Generative AI modules).
- **Week 12 (Mid-January to End of January):**
  - Perform end-to-end testing: document upload, extraction, data validation, media generation, and data storage.

February *– Testing and Optimization*

- **Week 13-14 (Early February to Mid-February):**
  - Optimize NLP performance for handling diverse clinical trial document formats.
  - Test and fine-tune the generative AI output for accuracy and relevance to the text input.
- **Week 15-16 (Mid-February to End of February):**
  - Conduct user acceptance testing with mock clinical trial documents and real-world scenarios.

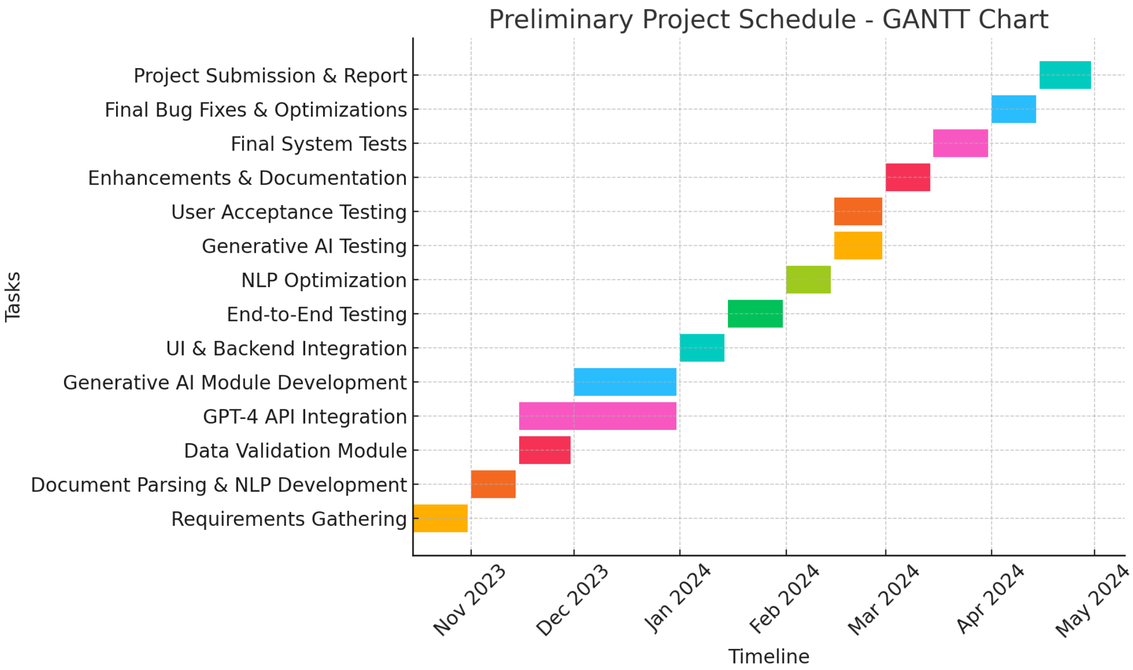
March *– Final Enhancements and Documentation*

- **Week 17-18 (Early March to Mid-March):**

  - Address any feedback from user testing.
  - Enhance the user interface and user experience.
  - Finalize integration of external APIs and third-party tools.
- **Week 19 (Mid-March to End of March):**
  - Prepare technical documentation, including system architecture and user guides.
  - Conduct final system tests.

April *– Final Delivery Preparation*

- **Week 20-21 (Early April to Mid-April):**
  - Final bug fixes and optimizations.
  - Perform any last-minute testing with new clinical trial documents.
- **Week 22 (Mid-April to End of April):**
  - Prepare for project submission, including a project report, final presentation, and any demonstration materials.



6. Appendices

ClinicalTrials.gov Database:
ClinicalTrials.gov is a free and publicly available database of clinical trial documents, including protocols, patient information sheets, and results. It provides detailed information about clinical studies conducted globally, making it an invaluable resource for extracting and testing PRO and ClinRO schedules.

Website: https://clinicaltrials.gov