## **Internship Documentation**

**Duration:** 3 Months  
 **Program:** AI Medical Data Research & Model Development  
 **Focus Areas:** Oncology, Hematology, Allergy/Immunology, Cardiology, Pediatric Pulmonology  
 **Role:** Data Researcher & Model Developer  
 **Track:** Artificial Intelligence in Healthcare

**Overview**

During this 3-month internship, I participated in a multidisciplinary project aimed at developing a conversational AI model capable of answering medical queries across various specialties. The work focused on **collecting, organizing, and modeling medical data** using state-of-the-art **NLP models** and clinical guidelines.

**Specialties Covered**

#### **1. Oncology**

* Curated datasets covering cancer types, treatments, side effects, staging, and patient FAQs.
* Extracted structured Q&A pairs from reputable sources such as National Cancer Institute, [Mayo Clinic](https://www.mayoclinic.org), and PubMed.
* Created symptom-treatment-answer mappings for common cancers (e.g., breast, lung, leukemia).

#### **2. Hematology**

* Focused on disorders like anemia, sickle cell disease, hemophilia, and thrombocytopenia.
* Built a structured database of diagnostic criteria, symptoms, lab findings, and treatments.
* Applied entity recognition for terms like “platelet count,” “RBC,” “hemoglobin,” etc.

#### **3. Allergy & Immunology**

* Gathered Q&A on food allergies, asthma, eczema, autoimmune conditions, and immune deficiencies.
* Integrated AAAAI guidelines and CDC vaccine safety data.
* Developed multi-turn dialogues for chatbot simulation (e.g., seasonal allergy assessments).

#### **4. Cardiology Q&A**

* Extracted clinical insights on hypertension, heart failure, congenital defects, and arrhythmias.
* Sourced content from AHA, [Cleveland Clinic](https://my.clevelandclinic.org/health), and WHO.
* Integrated guideline-based management (e.g., ACC/AHA hypertension classification).

#### **5. Pediatric Pulmonology**

* Collected detailed data for diseases like asthma, bronchiolitis, pneumonia, and croup.
* Used sources such as GINA, AAP, and [CDC](https://www.cdc.gov).

**Tools & Technologies Used**

| **Category** | **Tools** |
| --- | --- |
| **Data Gathering** | BeautifulSoup, Requests, PubMed API and manually |
| **Preprocessing** | Pandas, Regex, spaCy, NLTK |
| **Modeling** | HuggingFace Transformers, BioGPT, Sentence-BERT, TextCat |
| **Deployment** | Gradio UI, Google Colab |
| **Storage Format** | JSONL, CSV, Markdown |

**Models Built**

**Chatbot Simulation**

* Designed an **interactive chatbot** capable of:  
  + Entity recognition (SciSpaCy)
  + Intent classification (Zero-shot with BART)
  + Follow-up question generation
  + Realistic doctor-patient dialogue simulation

**Outcome**

* 📊 Created a scalable Q&A pipeline across 5 major specialties.
* 📁 Developed a reusable dataset with **medical Q&A pairs**.
* 🤖 Enabled real-time conversational AI interaction for medical training and triage.

CHARLHAM EL

INTERN