**Documentation of my internship journey**

**Program Title:** 3-Month Internship in Artificial Intelligence – Pediatric Pulmonology Chatbot Project  
 **Domain:** AI for Healthcare  
 **Focus Area:** Data Gathering, NLP, and Conversational AI

### **Duration:**

**April 2025 – July 2025** (3 months)

### **Host Organization:**

*Omeife technology*  
 Supervised under [Mrs. Peace Okafor]

### **Overview:**

This internship focused on applying **artificial intelligence and natural language processing (NLP)** techniques to build a **clinical decision-support chatbot**. The goal was to assist patients, caregivers, and healthcare professionals by providing symptom-based information, disease insights, and early triage advice.

### **Key Responsibilities:**

1. **Medical Data Collection & Curation**
   1. Researched many diseases under valvular heart disease, dermatology, orthopedics, nephrology and pediatric pulmonology.
   2. Gathered text from reputable sources (Mayo Clinic, Cleveland Clinic, Medscape, PubMed).

## **Official & Reputable Medical Sources**

### **General Medical Institutions**

|  |  |  |
| --- | --- | --- |
| **Source** | **Description** | **Official Link** |
| **Mayo Clinic** | Trusted source for patient education and disease information. | [https://www.mayoclinic.org](https://www.mayoclinic.org/) |
| **Cleveland Clinic** | Comprehensive health library & clinical insights. | <https://my.clevelandclinic.org/health> |
| **Medscape** | Clinical news, drug info, and disease reference for professionals. | [https://www.medscape.com](https://www.medscape.com/) |
| **WebMD** | Popular consumer health information site. | <https://www.webmd.com> |

### **Scientific Research & Literature**

|  |  |  |
| --- | --- | --- |
| **Source** | **Description** | **Official Link** |
| **PubMed (NIH/NLM)** | Massive database of peer-reviewed biomedical literature. | <https://pubmed.ncbi.nlm.nih.gov> |
| **NCBI (National Center for Biotechnology Information)** | Research articles and datasets. | [https://www.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov/) |
| **ScienceDirect** | Peer-reviewed articles on clinical medicine and health science. | [https://www.sciencedirect.com](https://www.sciencedirect.com/) |
| **Springer Nature** | Academic medical and life sciences journals. | <https://link.springer.com> |

### **Public Health & Epidemiology**

|  |  |  |
| --- | --- | --- |
| **Source** | **Description** | **Official Link** |
| **CDC (Centers for Disease Control and Prevention)** | U.S. disease surveillance, prevention, and guidelines. | [https://www.cdc.gov](https://www.cdc.gov/) |
| **WHO (World Health Organization)** | Global health topics, diseases, and data reports. | [https://www.who.int](https://www.who.int/) |
| **ECDC (European Centre for Disease Prevention and Control)** | Europe’s health monitoring and infectious disease database. | <https://www.ecdc.europa.eu> |

### **Pediatric & Specialty Societies**

|  |  |  |
| --- | --- | --- |
| **Organization** | **Description** | **Official Link** |
| **AAP (American Academy of Pediatrics)** | Pediatric guidelines, diseases, and best practices. | <https://www.aap.org> |
| **GINA (Global Initiative for Asthma)** | Asthma care guidelines including pediatric focus. | <https://ginasthma.org> |
| **ERS (European Respiratory Society)** | Respiratory care, pediatric pulmonology included. | <https://www.ersnet.org> |
| **ATS (American Thoracic Society)** | Resources for respiratory, critical care, sleep medicine. | <https://www.thoracic.org> |

### **Orthopedic & Related**

|  |  |  |
| --- | --- | --- |
| **Source** | **Description** | **Official Link** |
| **American Orthopedic Society for Sports Medicine (AOSSM)** | Sports-related orthopedic injury and recovery research. | <https://www.sportsmed.org> |
| **AAOS (American Academy of Orthopedic Surgeons)** | Musculoskeletal conditions and clinical guidelines. | <https://www.aaos.org> |

* 1. Parsed and structured datasets.

1. **Knowledge Extraction**
   1. Extracted and labeled entities such as definition/ description, symptoms, causes, risk factors, diagnosis and test, treatments, prevention, possible complication, epidemiology, diagnostic consideration, differential diagnosis, and drug information and their side effects using regular expressions and NLP models.
   2. Developed a comprehensive question-answering (Q&A) dataset with de identified doctor patient conversation.

The final project was for the interns in each group to build a medical chatbot using the pediatric pulmonology dataset.

After the dataset gathering, we converted to txt, then tokenize and cleaned it to avoid noise and hallucination.

1. **Model Building**

The dataset was used to build and evaluate multiple models for the pediatric pulmonology chatbot, with a focus on both **traditional classification** and **advanced generative reasoning**:

#### **1. TextCatBoW (Text Categorizer with Bag-of-Words)**

* Implemented a lightweight text classification model using spaCy’s textcat\_bow architecture.
* Used for **intent recognition** and categorizing user queries into:
  + Symptoms
  + Causes
  + Treatment
  + Prevention
  + Emergency
* Achieved fast inference and high accuracy on structured queries.

#### **2. BioGPT (Biomedical Generative Transformer)**

* Integrated **Microsoft/BioGPT**, a domain-specific large language model trained on biomedical literature.
* Used BioGPT for:
  + Generating **reasoned, natural responses** to medical questions.
  + Answering **free-form clinical queries** with high fluency and medical relevance.
* Applied in tandem with Q&A dataset matching to improve answer precision and diversity.

This hybrid setup allowed the chatbot to both classify and understand queries contextually, while generating human-like answers with clinical depth.

1. **NER & Medical Term Extraction**
   1. Applied spaCy and SciSpaCy for medical entity recognition (e.g., medications, diseases, anatomy).
   2. Enhanced the chatbot’s ability to understand user queries and extract relevant clinical concepts.
2. **Gradio Chatbot Interface**
   1. Developed an interactive chatbot UI using **Gradio**.
   2. Simulated **doctor-patient multi-turn conversations** with follow-up logic.
   3. Incorporated reasoning-based generation using transformer pipelines.
3. **Documentation & Deployment**
   1. Wrote detailed technical documentation in Markdown for GitHub.
   2. Converted .docx internship logs into structured reports and README files.
   3. Ensured reproducibility of the project via Google Colab notebooks and proper versioning.

### **Tools & Technologies:**

* **Python**, **Pandas**, **Regex**, **BeautifulSoup**
* **Transformers (Hugging Face)**, **spaCy**, **SciSpaCy**
* **Gradio**, **scikit-learn**, **Sentence Transformers**
* **Git**, **GitHub**, **Google Colab.**

### **Outcome:**

Successfully built a scalable, data-driven, medical chatbot prototype capable of answering pediatric respiratory health questions with clinical accuracy. The solution supports further fine-tuning for production-grade deployment in hospitals or telehealth platforms.

### **Reflection:**

This internship gave me hands-on experience in merging **AI**, **healthcare**, and **NLP**. It deepened my understanding of clinical datasets, medical terminologies, ethical AI design, and building end-to-end AI systems that could improve healthcare accessibility.

LESLIE EL

INTERN