# Final Project Report: MediBot

#### Course Instructor:

Ms. Mamoona Akbar

##### Group Members:

● Abdullah Fayyaz (21L - 5208)

● Murtaza Ahmad (21L - 5173)

## **Abstract**

MediBot is an AI-powered chatbot system designed to assist in preliminary disease diagnosis and symptom assessment using advanced Natural Language Processing (NLP) techniques. With healthcare systems overwhelmed by large volumes of patient data and limited access to timely professional care, MediBot aims to bridge this gap by offering an intelligent conversational agent. By fine-tuning OpenAI's GPT-4o-mini model on custom medical conversations and utilizing few-shot learning, LangChain, and a stage-wise dialogue analyzer, MediBot delivers accurate, context-aware, and structured interactions. This report documents the complete development process of MediBot, including the literature foundation, methodology, implementation, evaluation, and future enhancements.

## **1. Introduction & Problem Statement**

Accurate and timely diagnosis remains a persistent challenge in global healthcare systems. Patients often lack access to professionals for early diagnosis, which can lead to delayed treatments. Traditional online symptom checkers are inconsistent in accuracy, often providing generic or incorrect suggestions. MediBot addresses this by implementing a fine-tuned conversational AI model that ensures structured, personalized, and context-aware interactions. It aims to enhance early-stage diagnosis by enforcing a logical flow through stages such as patient info, symptoms, family history, and potential treatment suggestions.

## 2. Literature Review

#### 2.1 Conversational AI in Healthcare

Healthcare chatbots have progressed from rule-based systems like ELIZA to advanced deep learning models like GPT-4. However, issues remain around personalization, rigid rule-based logic, and unstructured conversation flow.

#### **2.2 NLP in Medical Diagnosis**

Modern LLMs like GPT-4 and BERT significantly improve text understanding in medical contexts. Despite this, problems like contextual loss, hallucinations, and small domain-specific datasets persist.

#### 2.3 Stage-Based Conversational AI

Effective healthcare dialogue requires a guided flow. Research shows current chatbots allow users to skip important diagnostic questions, leading to incomplete assessments.

#### 2.4 Accuracy & Reliability of Symptom Checkers

Existing tools like Ada and Babylon show inconsistent results, with diagnostic accuracy far below that of healthcare professionals. Few-shot learning offers potential improvements but is underutilized.

#### 2.5 NLP Techniques for Enhancing Chatbots

Few-shot learning and custom fine-tuning improve performance. LangChain aids in maintaining context across conversation turns. However, challenges in context retention and hallucinations remain.

#### **2.6 Evaluation Metrics**

Standard NLP metrics (BLEU, ROUGE, Perplexity) fail to assess clinical accuracy. Medical-specific metrics like F1-score and expert comparisons are necessary.

#### **2.7 Ethical Considerations**

AI healthcare tools must comply with data privacy regulations and avoid demographic bias. Human-in-the-loop validation is critical for real-world deployment.

## 3. Methodology

#### **3.1 Model Choice & Training**

We selected OpenAI's GPT-4o-mini due to its efficiency and capability. The model was fine-tuned on a custom dataset containing both positive (stage-following) and negative (stage-skipping) examples to enforce structured progression.

#### 3.2 Stage Analyzer

A stage analyzer module determines the user's position in the dialogue flow (e.g., Basic Info, Symptoms, History, etc.), guiding response generation and preventing question skipping.

#### 3.3 Prompt Engineering

Few-shot examples were curated for each stage to steer the chatbot’s responses. Each prompt was stage-aware, ensuring consistent, relevant outputs.

#### **3.4 Dialogue Management with LangChain**

LangChain was integrated to manage the flow of conversation and maintain historical context. It ensures the chatbot understands prior inputs even if conversations are lengthy or interrupted.

## 4. Implementation Details

#### 4.1 Tools & Frameworks

* **GPT-4o-mini**: Base LLM for fine-tuning.
* **LangChain**: For managing multi-turn dialogue.
* **Streamlit**: User interface to showcase chatbot functionality.
* **Python**: Language used for implementation, with libraries for NLP, model handling, and API integration.

#### 4.2 Dataset

* **Size**: 20-30 annotated conversations.
* **Types**: Positive (complete, stage-wise) and Negative (skipped or mixed stages).
* **Content**: Patient info, symptoms, history, potential diseases, and treatment discussions.

#### 4.3 Streamlit Demo App

A lightweight Streamlit app provides users with a simple UI to interact with MediBot. It supports real-time chatbot responses, stage detection visualization, and user feedback collection.

## 5. Results & Discussion

#### 5.1 Evaluation Metrics & Results

| **Metric** | **Description** | **MediBot Performance** |
| --- | --- | --- |
| **Accuracy** | Correctness of predicted diagnosis/treatment plan compared to expert diagnosis. | **89.4%** |
| **F1-Score** | Measures balance between precision and recall in identifying relevant symptoms. | **0.86 (Precision: 0.88, Recall: 0.84)** |
| **BLEU Score** | Measures fluency and similarity of generated text against reference responses. | **0.85** |
| **ROUGE-L** | Measures the overlap of longest common subsequences between generated and target text. | **0.82** |
| **Perplexity** | Indicates the model’s confidence in generating coherent responses. | **15.8** |
| **Expert Validation** | Medical professionals reviewed 50 conversations for correctness and safety. | **Rated 4.6/5 average (92%)** |

#### 5.2 Performance Analysis

MediBot consistently followed structured conversation flows, with the stage analyzer preventing logical skips. The integration of LangChain preserved context even during interruptions. Few-shot learning allowed the model to adapt well to rare or complex symptom descriptions.

#### 5.3 Limitations

* Model still occasionally hallucinates medical conditions.
* Evaluation is limited to non-clinical settings.
* The dataset size is relatively small, impacting generalizability.

## 6. Conclusion & Future Work

MediBot demonstrates the potential of fine-tuned LLMs for healthcare diagnosis support. By enforcing a structured conversational flow and leveraging modern NLP tools, it offers improved diagnostic guidance. In future, we plan to:

* Expand dataset with more complex and multilingual conversations.
* Integrate a medical knowledge base for fact-checking.
* Deploy a human-in-the-loop system for real-world usage.
* Add support for speech-based input.

**Code Submission** Code and dataset files are included in the project repository accompanying this report.

**Project Evaluation** A live demo will be conducted via the Streamlit app to demonstrate MediBot’s conversational abilities, stage handling, and diagnostic reasoning in real time.