**Literature Review**

**MediBot**

**Focused on Checking Symptoms and Diagnosing Diseases**

**Course Instructor:**

Ms. Mamoona Akbar

**Group Members:**

* Abdullah Fayyaz (21L - 5208)
* Murtaza Ahmad (21L - 5173)

## **1. Introduction**

The integration of artificial intelligence (AI) in healthcare has seen rapid advancements, particularly in the development of conversational agents for preliminary medical diagnosis. These AI-driven chatbots assist in symptom assessment, disease prediction, and patient engagement. However, challenges persist in ensuring diagnostic accuracy, conversation flow management, and contextual relevance. This literature review examines prior research in conversational AI for healthcare and highlights gaps that the proposed **MediBot** system aims to address.

## **2. Conversational AI in Healthcare**

### **2.1 Evolution of Healthcare Chatbots**

Healthcare chatbots have evolved from simple rule-based systems to sophisticated AI-powered assistants. Early implementations, such as ELIZA (Weizenbaum, 1966), relied on pattern-matching techniques but lacked contextual understanding. Recent chatbots utilize deep learning and large language models (LLMs) like GPT-4, enhancing their ability to process and generate human-like responses (Brown et al., 2020).

**Limitations in Existing Systems:**

* **Limited personalization**: Most chatbots provide generic responses rather than personalized medical advice (Laranjo et al., 2018).
* **Rule-based constraints**: Traditional chatbots rely on predefined rules, restricting their ability to handle diverse user inputs dynamically.
* **Lack of structured interaction**: Few systems enforce a stage-wise flow for collecting patient data systematically (Young et al., 2013).

### **2.2 Application of NLP in Medical Diagnosis**

Natural Language Processing (NLP) enables chatbots to extract meaningful information from user inputs. Pre-trained language models such as BERT (Devlin et al., 2019) and GPT-4o (OpenAI, 2024) have significantly improved medical text interpretation. Research highlights that fine-tuned models on medical corpora enhance chatbot performance in clinical applications (Peng et al., 2020).

**Challenges in NLP-based Medical Chatbots:**

* **Contextual relevance**: Maintaining conversation context across multiple exchanges is difficult.
* **Medical hallucinations**: AI models sometimes generate incorrect yet plausible-sounding medical information.
* **Data scarcity**: Medical datasets are often small, requiring domain-specific fine-tuning to improve performance.

### **2.3 Stage-Based Conversational AI**

Dialogue systems in healthcare must ensure structured and logical progression. Research in dialogue management (Serban et al., 2016) emphasizes the need for **stage-aware models** that enforce a structured flow. Existing symptom checkers fail in this regard, often allowing users to skip crucial questions, leading to incomplete or inaccurate assessments (Semigran et al., 2015).

**Gaps Identified:**

* **Lack of stage-aware conversation modeling** in medical chatbots.
* **Inadequate mechanisms to prevent skipping of necessary diagnostic questions.**

## **3. Accuracy and Reliability of Symptom Checkers**

Studies on existing symptom checkers (e.g., WebMD, Ada, Babylon Health) reveal inconsistencies in diagnostic accuracy. A comparative analysis by Semigran et al. (2015) found that accuracy rates range between **34% and 58%**, significantly lower than those of general practitioners. Misdiagnoses primarily arise from:

* **Bias toward common diseases** at the expense of rare conditions.
* **Failure to interpret multi-symptom inputs effectively.**
* **Lack of contextual patient history consideration.**

To address these limitations, recent research recommends **few-shot learning techniques** to improve chatbot adaptability to rare and complex conditions (Peng et al., 2020). However, most symptom checkers do not implement such methodologies effectively.

### **MediBot’s Contribution:**

* Uses **fine-tuned GPT-4o-mini** for improved **contextual understanding**.
* Implements **few-shot learning** to enhance diagnostic flexibility.
* Incorporates **stage-wise progression**, preventing users from bypassing crucial questions.

## **4. NLP Techniques for Improving Healthcare Chatbots**

### **4.1 Few-Shot Learning and Fine-Tuning**

Pre-trained LLMs require fine-tuning for domain-specific applications. Few-shot learning, where models learn from minimal examples, has gained popularity in healthcare AI (Brown et al., 2020). Peng et al. (2020) demonstrated that fine-tuning GPT-3 on medical texts improved diagnostic accuracy by **20%**.

However, challenges persist:

* **Risk of AI hallucinations** due to lack of proper control mechanisms.
* **Difficulties in maintaining consistency** in multi-turn conversations.

MediBot **addresses these issues** by:

* Fine-tuning with a **custom dataset containing both positive and negative examples**.
* Implementing a **stage analyzer** to ensure consistency in conversations.

### **4.2 Handling Context Retention and Interruptions**

Conversational AI in healthcare must effectively manage user inputs across multiple exchanges. Razzak et al. (2019) highlight that **context retention** is critical for reliable medical advice. Many existing chatbots fail when users provide fragmented or incomplete inputs.

Solutions proposed in the literature:

* **LangChain-based dialogue management** to maintain context effectively (Serban et al., 2016).
* **Reinforcement learning** for optimizing dialogue progression based on user behavior (Jurafsky & Martin, 2020).

**MediBot’s Approach:**

* Uses **LangChain** to **retain context** across interactions.
* Employs a **stage analyzer** to ensure that interruptions do not lead to information gaps.

## **5. Evaluation Metrics for Healthcare Chatbots**

### **5.1 Standard NLP Metrics**

Most AI chatbots are evaluated using text-generation metrics such as:

* **BLEU & ROUGE**: Assess fluency and coherence (Razzak et al., 2019).
* **Perplexity**: Measures the naturalness of chatbot responses.

However, these metrics do not adequately assess **medical accuracy**.

### **5.2 Medical-Specific Evaluation Metrics**

Recent studies propose specialized evaluation frameworks:

* **F1-score for symptom extraction** (Peng et al., 2020).
* **Clinical accuracy comparison against expert diagnoses** (Semigran et al., 2015).

**MediBot’s Evaluation Approach:**

* Combines **BLEU, ROUGE, and Perplexity** for NLP assessment.
* Uses **F1-score and expert validation** to measure medical relevance.

## **6. Ethical Considerations in Medical Chatbots**

The deployment of AI in healthcare raises ethical concerns, particularly regarding:

* **Data privacy** (GDPR compliance).
* **Bias in AI models** that may disproportionately affect different demographic groups.
* **Liability for incorrect diagnoses** and potential harm to users.

To mitigate these risks, researchers recommend **human-in-the-loop approaches** where medical professionals validate AI outputs before clinical implementation (Laranjo et al., 2018). MediBot aligns with these best practices by ensuring that **AI-generated diagnoses are reviewed by medical experts before deployment**.

## **7. Conclusion**

The literature highlights significant challenges in existing medical chatbots, including poor accuracy, lack of structured dialogue management, and weak contextual retention. The proposed **MediBot** system addresses these gaps by:

* **Fine-tuning GPT-4o-mini** for improved diagnostic accuracy.
* **Implementing stage-based dialogue progression** to ensure structured interactions.
* **Using LangChain** for enhanced context retention.
* **Incorporating specialized evaluation metrics** tailored for medical chatbot performance.

By integrating advanced NLP techniques and structured conversation flows, MediBot aims to set a new benchmark for AI-driven healthcare diagnosis