CAREPAL: AI-Driven Health & Activity Partner

Project Report - Fall 2024

I. Header

Project Title: CAREPAL: AI-Driven Health & Activity Partner

Team Members:

1. Dhyey Joshi (dhyjoshi@iu.edu)
2. Gowthami Gokul (ggokul@iu.edu)
3. Matthew Yeseta (matyeset@iu.edu)
4. Sahithi Vangala (svangal@iu.edu)

Sponsors: Crisis Technologies Innovation Lab, Kyle Stirling

II. Overview

The Crisis Technologies Innovation Lab launched this project to explore the transformative potential of large language models (LLMs), such as ChatGPT, in domains like business, art, music, and humanitarian efforts. These tools redefine human-technology interaction by leveraging vast datasets to generate human-like responses and innovative solutions. This project aims to understand and evaluate the societal impact of LLMs and identify meaningful applications, such as personalized education, crisis response, and enhanced professional productivity. By summarizing and interacting with digitized information, LLMs enable novel approaches to solving real-world challenges.

Among these domains, the team chose healthcare to demonstrate the practical applications of LLMs and developed CAREPAL, a Personal Health Assistant chatbot. This decision was based on the need for accessible, personalized, and reliable health information, which can be difficult to navigate due to the complexity of medical guidelines. By focusing on healthcare, the team showed how LLMs can bridge this gap, turning static health documents into interactive resources that provide actionable, tailored advice. This highlights the unique potential of LLMs to improve healthcare communication and accessibility.

CAREPAL is an AI-powered health and wellness chatbot designed to provide personalized, evidence-based health advice by leveraging comprehensive guidelines from authoritative sources like WHO and the U.S. Department of Health Services. It simplifies complex medical information into actionable recommendations, addressing the growing need for accessible, real-time health and fitness support. Powered by cutting-edge Large Language Models (LLMs), CAREPAL utilizes Llama 2, ensuring conversational fluency and accuracy in delivering personalized health-related insights. CAREPAL personalizes responses by considering user-specific variables like age, health history, and lifestyle. By automating health guidance delivery, it reduces the workload on medical professionals while empowering individuals to proactively manage their health.

III. Purpose

The CAREPAL project aims to address the pressing need for accessible, reliable, and personalized health and wellness guidance. Many individuals struggle to navigate the overwhelming and often conflicting health information available online, compounded by busy schedules, aging concerns, and unique health conditions. CAREPAL seeks to bridge this gap by providing actionable, evidence-based advice tailored to individual needs, empowering users to make informed lifestyle decisions.

By leveraging the conversational fluency and accuracy of Llama 2, CAREPAL offers non-diagnostic, personalized recommendations in areas such as fitness, nutrition, and mental well-being. The chatbot simplifies complex health guidelines into practical advice, ensuring accessibility for diverse user groups while respecting data privacy and compliance with public health standards. This project is important because it democratizes access to reliable health information, promotes healthy habits, and supports better lifestyle choices.

Stakeholders include:

- Individuals and Families: Access to age-appropriate and condition-specific guidance for maintaining a healthy lifestyle.
- Healthcare Professionals: A supplemental tool to enhance patient understanding of wellness practices.
- Fitness Enthusiasts and Trainers: Clear advice on fitness routines and dietary practices.
- Public Health Organizations: A scalable resource to disseminate health guidelines and promote awareness.

By addressing common challenges such as a lack of time, age-related concerns, and misinformation, CAREPAL has the potential to positively impact communities by fostering healthier, more informed living.

IV. Methodology

To address the challenge of delivering personalized, evidence-based health and wellness advice, CAREPAL was developed using a systematic approach that integrates advanced AI technologies and domain-specific methodologies. The primary objective was to create a scalable, interactive platform that could provide accurate responses while adhering to token limits and maintaining contextual relevance.

Data Sources:

The foundational knowledge for CAREPAL was derived from reputable health guidelines and datasets, including:

- WHO Guidelines on Physical Activity and Sedentary Behavior. https://3.basecamp.com/3947469/buckets/31399169/uploads/8055641038
- Dietary Guidelines for Americans, 2020-2025. https://3.basecamp.com/3947469/buckets/31399169/uploads/8055641338
- Curated datasets focusing on health, fitness, and wellness. https://3.basecamp.com/3947469/buckets/31399169/uploads/8055641475

These sources ensured the chatbot provided accurate and reliable advice based on trusted health information.

Technologies:

Large Language Models (LLMs):

Llama 2: This state-of-the-art large language model was chosen for its advanced conversational fluency, scalability, and adaptability. Llama 2 ensures that the chatbot generates coherent, contextually relevant, and user-friendly responses across diverse health and wellness queries. It provides the necessary balance between flexibility for general queries and specificity for health-related information, making it the cornerstone of CAREPAL's interactive capabilities.

(Previously Considered) BioBERT and ClinicalBERT: Initially explored for their domain-specific healthcare precision, these models were ultimately set aside to

streamline the implementation process. While highly effective for clinical tasks, their smaller token context (512 tokens) and added complexity made Llama 2 a more practical choice for the project's broader goals.

Vector Databases:

Pinecone: Pinecone was employed to manage semantic indexing and similarity searches, crucial for retrieving contextually relevant information from the chatbot's knowledge base. By converting textual data into embeddings (numerical vector representations), Pinecone enables fast and accurate matching between user queries and stored knowledge, significantly enhancing response relevance. Its scalability ensures that CAREPAL can handle large datasets and evolving content efficiently.

Backend Framework:

Flask: This lightweight Python framework serves as the backbone of the application's API layer, managing interactions between the user interface, semantic search engine, and LLM. Flask was chosen for its simplicity and flexibility, allowing seamless integration with other components and ensuring efficient processing of user requests and responses.

Frontend Framework:

React: CAREPAL's user interface was built using React, a powerful JavaScript library known for its responsiveness and interactive capabilities. React provides a dynamic, user-friendly chatbot experience, allowing users to engage intuitively with the application. Features like real-time query submission and visually appealing response displays enhance the overall usability of the platform.

Preprocessing Tools:

Python Libraries (SpaCy, NLTK): These tools were instrumental in cleaning and preparing the textual data from knowledge sources.

Development:

1. Data Preprocessing:

    - Extracted and cleaned text from static documents by removing stop words, unwanted characters, and extra spaces.

- Split the text into manageable chunks (e.g., 200 tokens) with overlaps (20 tokens) to preserve context while staying within Llama 2's 4,096-token limit.

2. Semantic Indexing:

    - Converted text chunks into embeddings using OpenAI's embedding models.
    - Indexed embeddings into Pinecone to enable fast similarity searches for retrieving contextually relevant chunks.

3. Question Answering Workflow:

    - User queries were embedded and matched against stored embeddings to retrieve the most relevant text chunks.
    - The retrieved chunks were passed to Llama 2, which generated personalized, non-diagnostic responses.

4. Integration and Testing:

    - Combined preprocessing, semantic search, and LLM response generation into a cohesive pipeline.
    - Developed a React-based chatbot interface to provide an intuitive user experience.
    - Iterative testing and refinement were conducted based on stakeholder feedback to ensure accuracy and conversational flow.

Challenges:

1. Token Limitations: Initial model outputs exceeded token limits. This was resolved by chunking large texts into smaller segments with overlaps, maintaining context while ensuring compatibility with the LLM's token window.

2. Contextual Relevance: Challenges in aligning outputs with user queries were addressed by refining embedding models and improving the query-matching process.

3. Response Quality: Early iterations occasionally generated responses that lacked fluency or relevance. Iterative tuning and evaluation helped align the outputs with user expectations.

Why This Approach Was Correct:

This methodology was chosen to balance the trade-offs between accuracy, scalability, and user engagement. Semantic Indexing allowed for efficient retrieval of relevant information, ensuring that responses were both contextually appropriate and timely. Llama 2 Integration provided conversational fluency and ensured the chatbot could interact naturally with users. Preprocessing and Chunking enabled the application to handle large datasets while preserving the context necessary for meaningful responses.

The use of widely adopted tools and techniques, such as embedding-based semantic search and advanced LLMs, ensures the reproducibility of this project. Additionally, drawing from domain-specific datasets and iterative feedback loops demonstrates our commitment to delivering a robust, user-centered solution. This approach not only addressed the immediate challenges but also established a scalable framework for future enhancements and domain expansions.

V. Impact / Outcomes

CarePal, an AI-driven chatbot focused on providing evidence-based health and wellness guidance, has demonstrated significant potential in addressing key challenges in healthcare accessibility and personalized wellness advice.

Key Outcomes:

1) **Improved Access to Health Information:** CarePal has successfully integrated guidelines from trusted authorities like the CDC and WHO, providing users with reliable, actionable health advice on topics such as physical activity, nutrition, and mental well-being.
2) **Personalization Without Compromise on Privacy**: Utilizing Meta's Llama 2 and ClinicalBERT ensures tailored recommendations while adhering to stringent privacy norms.
3) **Positive Social Impact**: The chatbot's ability to assist healthcare professionals, fitness trainers, and the general public fosters a collaborative and informed community.

VI. Results

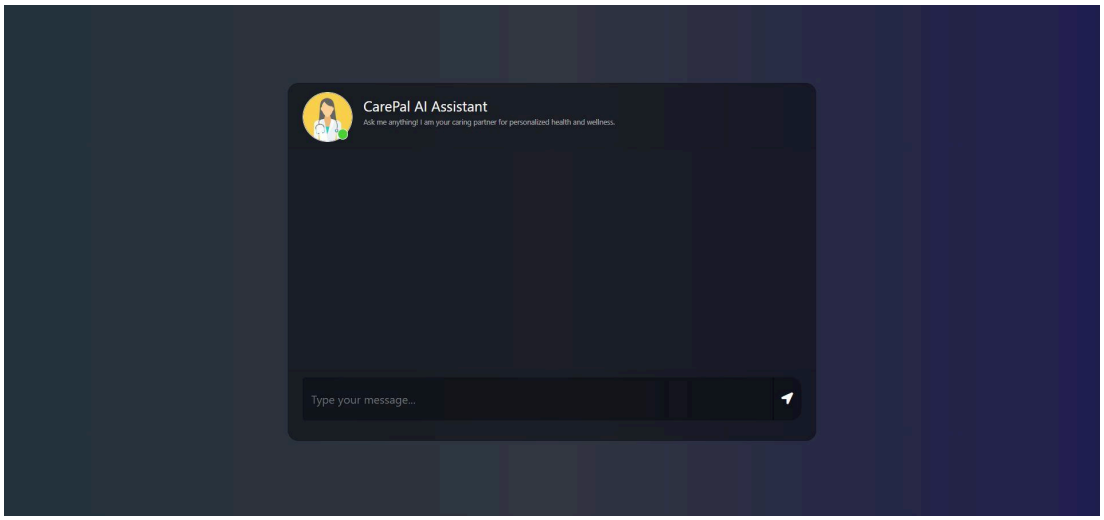

Figure 1: UI of CarePal

The above image showcases the user-friendly interface of CarePal, making it accessible for personalized health and wellness advice.

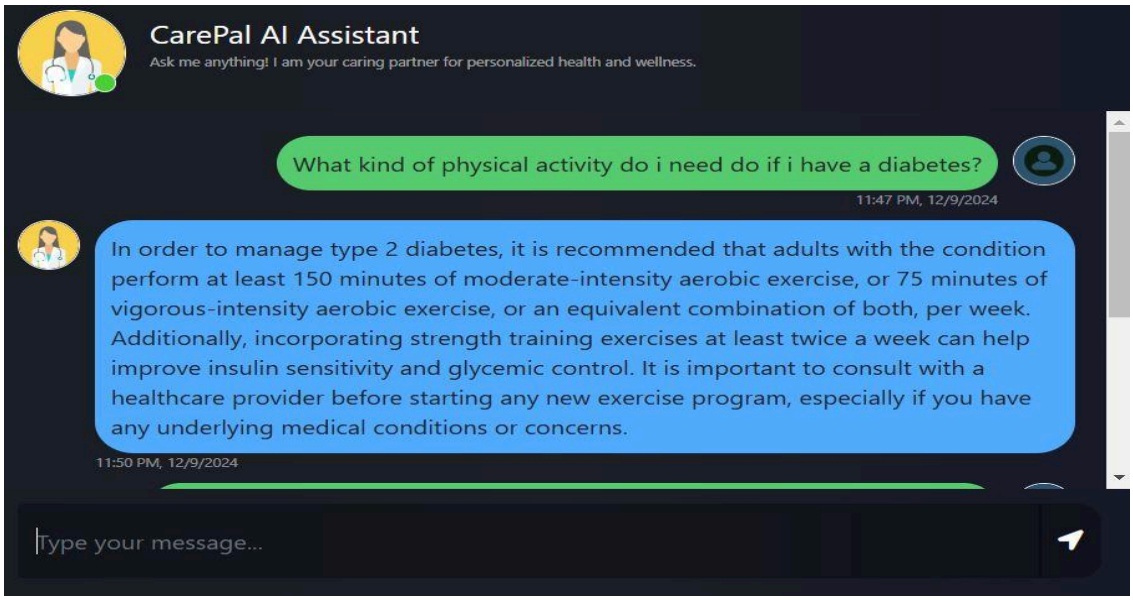

Figure 2: Chat response about physical activity for diabetes

The above image demonstrates CarePal's ability to provide specific, evidence-based advice for managing diabetes through physical activity. It shows the AI's capability to deliver tailored health guidance in real-time.

- The recommendation for **150 minutes of moderate-intensity aerobic exercise** or **75 minutes of vigorous-intensity exercise** per week aligns with **CDC and WHO guidelines**.
- CarePal emphasizes **strength training twice a week** to improve insulin sensitivity, showing its focus on holistic diabetes management.
- This personalized guidance can lead to better long-term health outcomes for users managing diabetes.

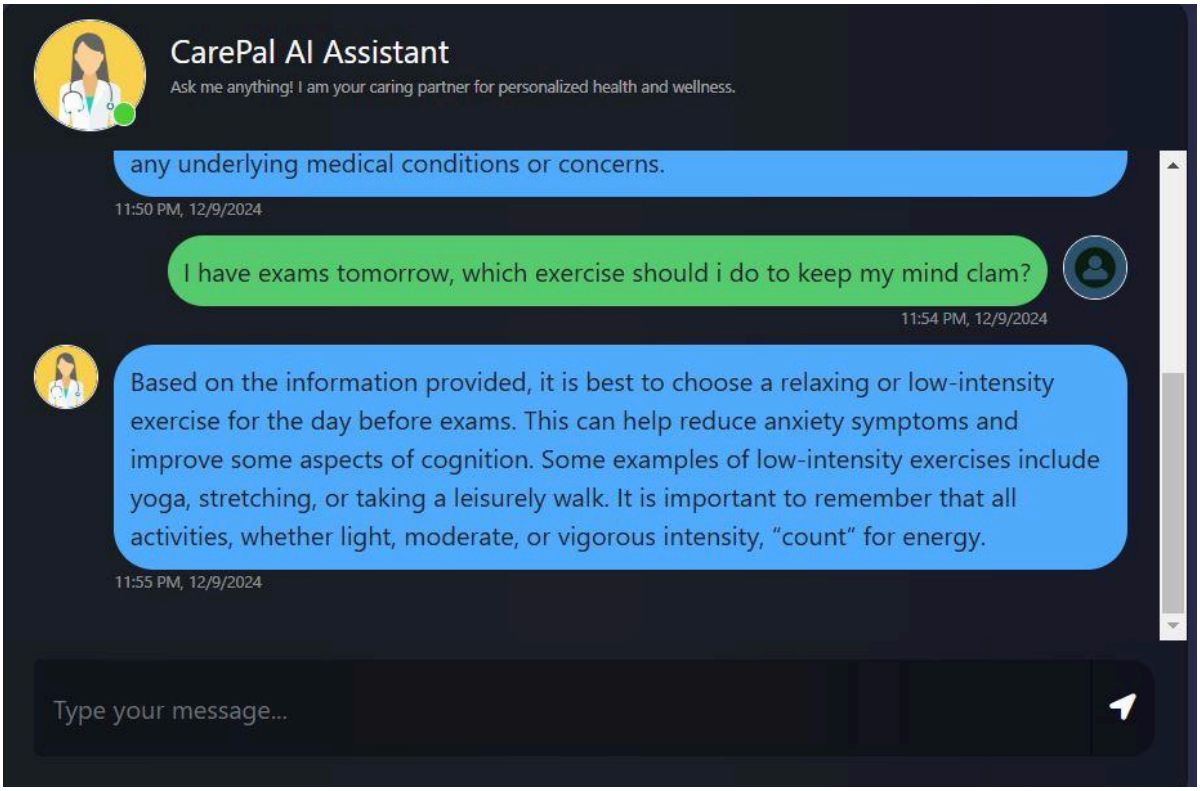

Figure 3: Exercise recommendation for exams

The above image shows CarePal's response to stress management and mental well-being, suggesting low-intensity exercises to calm anxiety before an exam.

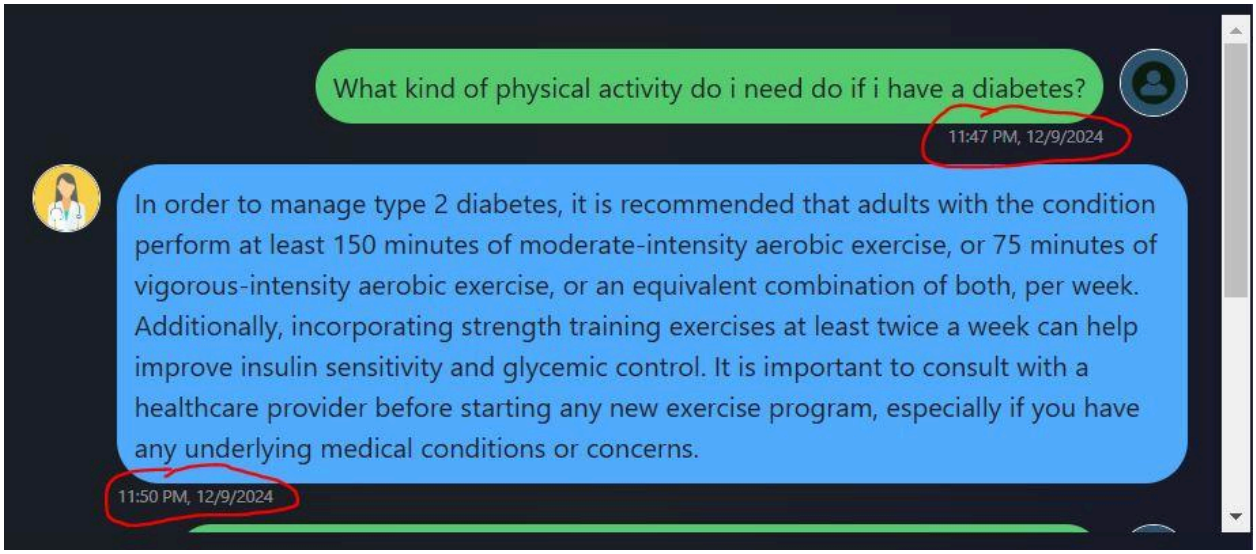

Figure 4: Real-Time Response with Timestamp

This image highlights the chatbot's ability to provide responses in real-time, demonstrating system efficiency and responsiveness.

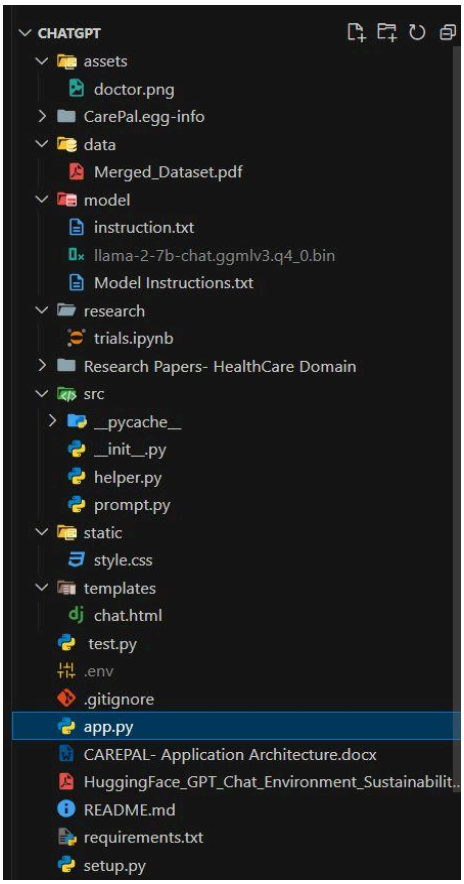

Figure 5: CarePal Project Folder Structure

The above image shows the organized folder structure of the CarePal project, highlighting key components for chatbot development.

The highlighted file **app.py**, is the central application script for running the CarePal chatbot. This file typically initializes the web server, sets up routes, and manages the main backend logic.

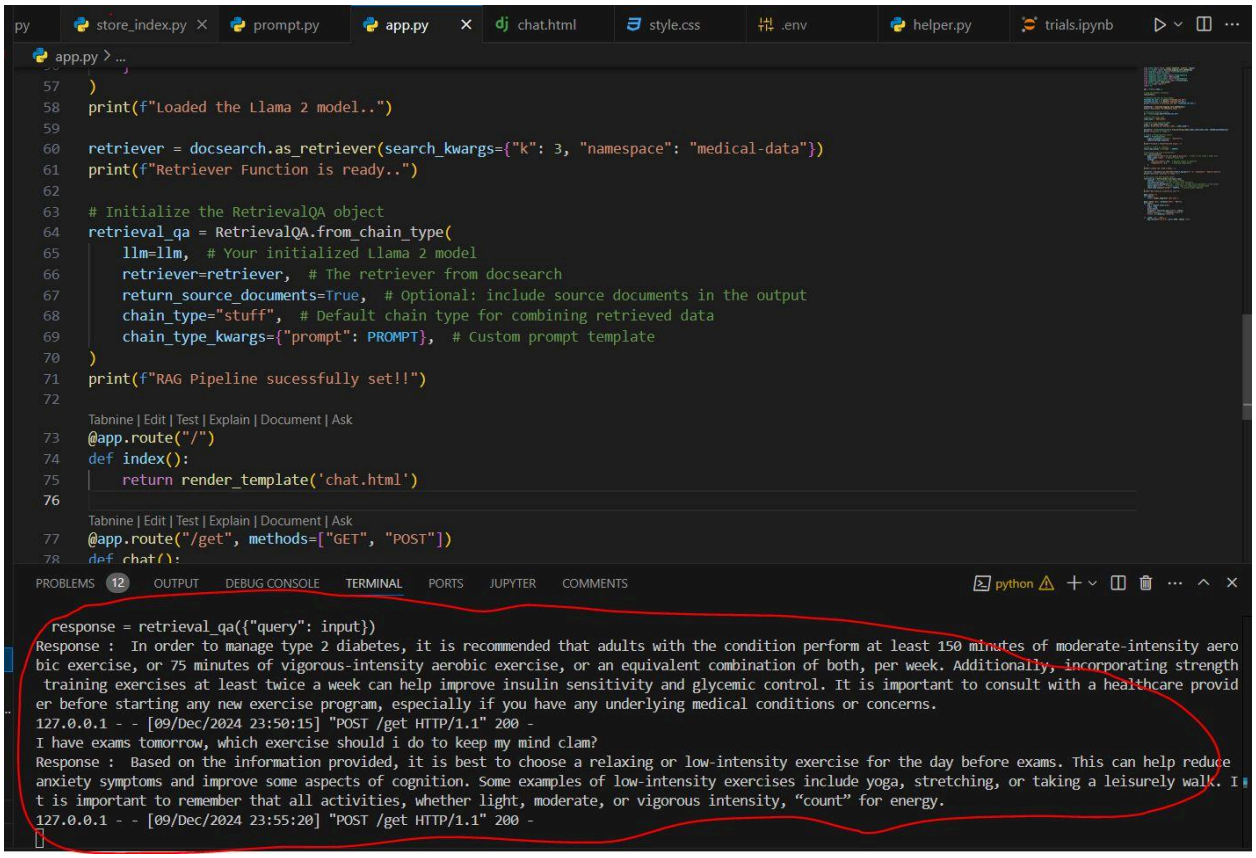

Figure 5: System Logs and Responses

CarePal processes user queries and delivers detailed, accurate responses through a streamlined **Retrieval-Augmented Generation (RAG)** pipeline.

- This image shows the **backend functionality** of CarePal, specifically how it handles queries through the Retrieval-Augmented Generation (RAG) pipeline.
- The code uses a **Llama 2 model** to process the queries and retrieve accurate medical advice from the knowledge base.

VII. Conclusion

The "CarePal" application integrates generative AI, semantic search, and healthcare domain expertise to provide a robust solution for personalized health and wellness guidance. By leveraging models like BioBERT and ClinicalBERT, it ensures domain-specific accuracy, while tools like Pinecone and Flask provide efficient data retrieval and a scalable API backend.

CarePal has successfully addressed the need for accessible and reliable health advice by leveraging Generative AI while adhering to strict data privacy protocols. The chatbot offers an effective solution for promoting healthier lifestyles and empowering users with knowledge. By providing personalized, non-diagnostic recommendations based on reputable public health guidelines, CarePal has demonstrated its potential to positively impact public health outcomes and support healthcare professionals.

VIII. Recommendations for Future Work

1) Expansion of Knowledge Base: We included data mainly from CDC and WHO. In the future, incorporate additional reputable sources to broaden the scope of health topics covered, including emerging health concerns and specialized areas of wellness.

2) Integration with Wearable Devices: Explore the possibility of integrating CarePal with popular fitness trackers and health monitoring devices to offer more precise, data-driven recommendations.

3) Multilingual Support: We have used English. In the future, develop versions of CarePal in multiple languages to extend its reach to diverse global communities.

4) Continuous Evaluation: Implement a robust system for ongoing evaluation of CarePal's performance, including regular updates to its knowledge base and fine-tuning of its language model.

5) Ethical AI Development: Continue to prioritize ethical considerations in AI development, addressing potential biases and ensuring transparency in the chatbot's decision-making processes.

By implementing these recommendations, CarePal can continue to evolve, expanding its positive impact on global health and wellness while addressing the complex challenges of AI in healthcare.

IX. REFERENCES

- **CDC:** Physical Activity Guidelines for Americans, 2nd Edition
    - https://health.gov/sites/default/files/2019-09/Physical_Activity_Guidelines_2nd_edition.pdf
- **WHO:** WHO Guidelines on Physical Activity and Sedentary Behaviour
    - https://apps.who.int/iris/bitstream/handle/10665/336656/9789240015128-eng.pdf
- **Mayo Clinic:** Health and Wellness Information
    - https://www.mayoclinic.org
- **Meta Llama 2 Model:**
    - https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML/tree/main