# CareConnect

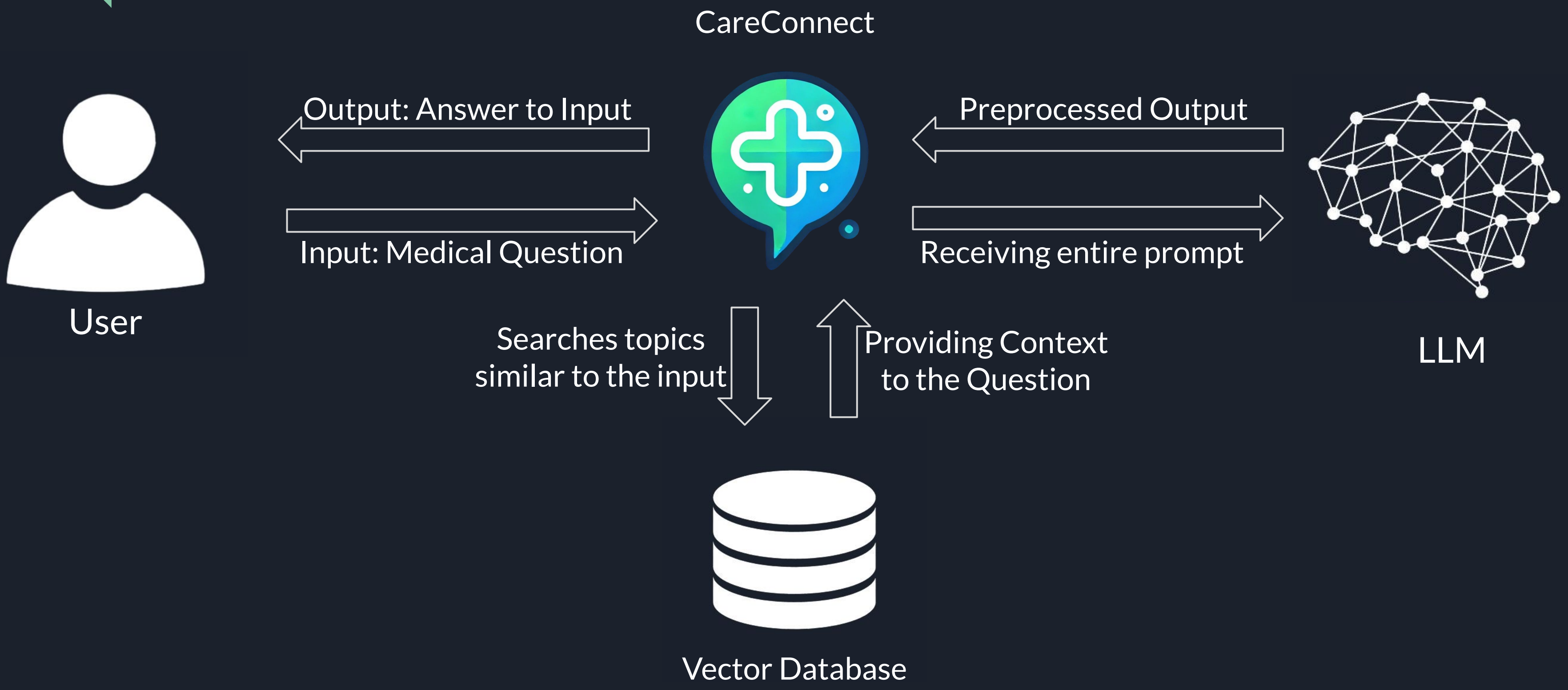
## A Personal Health Assistant

Atreyo Das & Atyab Hakeem

# Introduction

- Access to proper authentic medical information is extremely important for people to make informed decisions about their health.
- The rising cost of healthcare and increasing wait time to visit physicians can make it very difficult to get these information.
- Furthermore, it can be intimidating to discuss personal issues.
- To help this situation, we built a personal healthcare assistant to answer questions for your medical needs.

# System Design



# Project Workflow

### 1 Data Collection and Preprocessing

Data was collected and compiled from various sources and processed to be suitable for the model.

### 2 Model Selection and Training

A large language model was selected and then subsequently trained on the processed data.

### 3 Hyperparameter Tuning and RAG Implementation

RAG was implemented to enhance the accuracy and relevancy of the generated text. Hyperparameters were tuned to optimize inference time and performance.

### 4 Model Evaluation and Deployment

The model was then evaluated and deployed on a web application for a user-friendly interface.

# Dataset

MedAlpaca Dataset

- The dataset was obtained from MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data.
- It contains almost 160,000 medical question-answer pairs, taking data from Stack Exchange forums, WikiDoc and flashcards meant for medical students.
- We obtained a cleaned and processed subset of 30,000 entries from the original.

HealthcareMagic Dataset

- This dataset compiles about a 100,000 patient-doctor interactions from the site Ask A Doctor - 24×7
- Used for the model to provide more conversational answers to the user queries.

Alignment Datasets

- We further refined the model by tuning it on datasets to enhance social niceties and to filter out irrelevant non-medical questions.
- The dataset was synthetically generated through OpenAI’s GPT-4 and GPT-4o.

# Data Preprocessing

- Prior to training, we had to remove private information like names of doctors from the data.
- For training the model, we experimented with various different templates to see which one worked the best.
- Due to limited resources and to reduce their consumption we provided both the actual question and answer as the context.
- To simulate deployment environment, we provided the correct context only 50% of the time.
- Furthermore, we summarized the interactions between the doctors and patients using a BERT model to provide only the vital information as the context.

# Model Selection

- For the purpose of this project, we wanted to choose a good open-sourced model that could efficiently manage a large quantity of data while minimizing loading, training, and inference times.
- We considered and tried out models like Vicuna, Bloom and Llama 3 since their extensive training on large amounts of data, ensures high-quality responses.
- Finally, we went ahead with the 8B version of the Llama 3 model.

**Why Llama 3?**

Besides the fact that Llama 3 had SOTA performance on many NLP task, it can efficiently scale and handle large volumes of data. It also answered basic medical questions with ease, even before fine tuning, which made it perfect for our project.

# Model Training

## Training Specs:

- 1 Tesla V100 32G GPU
- Batch Size=4
- Scheduled learning rate of 1e6 to 1e4
- Maximum Sequence Length =1024
- Early stopping activated
- WandDB used for monitoring training

| Phase I | Phase II | Phase III |
| --- | --- | --- |
| **Data:** MedAlpaca Dataset | **Data:** MedAlpaca & HealthcareMagic Dataset | **Data:** MedAlpaca, HealthcareMagic Dataset & the Alignment Dataset |
| **Training Time:** 8 hours | **Training Time:** 8 hours | **Training Time:** 4 hours |
| **Number of Epochs:** 1 | **Number of Epochs:** ≈1 | **Number of Epochs:** 0.5 |

# RAG Implementation

- RAG was implemented to enhance quality and relevance of responses.
- We used the Pinecone to create the vector database.
- Similarity metrics, like cosine and dot product, were tested and evaluated.
- Although they produced similar results, we chose cosine similarity since it is normalized.
- Additionally, thresholding was also applied to prevent the retrieval of unnecessary data.

# Hyperparameter Tuning

- Hyperparameter tuning was performed on the decoding parameters to optimize inference time and performance.
- We employed Random Search to assess the impact of varying different parameters on the results.
- Since quantitative metrics can be unreliable sometimes, we performed qualitative analysis to obtain the best configuration.

# Model Evaluation- Quantitative

| Evaluation | Base Llama-3 model | Base Llama-3 model with RAG | CareConnect |
|---|---|---|---|
| BERT Precision | 0.83289 | 0.83381 | **0.85413** |
| BERT Recall | 0.85143 | 0.85178 | **0.86649** |
| BERT F1 | 0.84149 | 0.84215 | **0.85965** |
| Semantic Similarity | 0.69809 | 0.69939 | **0.77653** |

# Model Evaluation- Qualitative

## Query

Hi..I am having irritation on my both palms and sometimes cracks do occur (it was there for almost a year)..This started after i used VIM liquid. THk i got allergy. I have used prednisone 10 mg for 6 days recently and it got cured..but even now i feel irritation now and then and some very minor crack do occur...suggest wat i hv to do

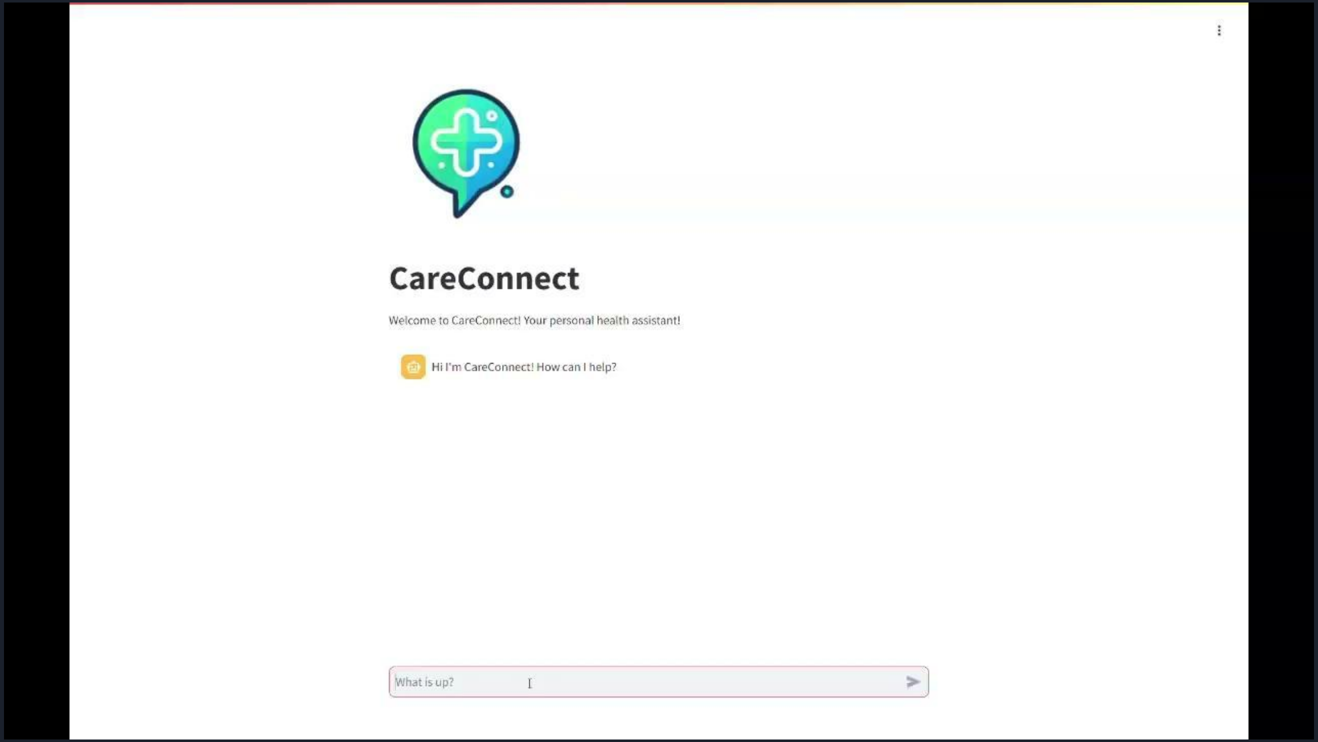
## Llama-3

Hi..I am having irritation on my both palms and sometimes cracks do occur (it was there for almost a year)..This started after i used VIM liquid. THk i got allergy. I have used prednisone 10 mg for

## CareConnect

Hi, Welcome to CareConnect, I understand your concern. The symptoms that you have described are suggestive of allergic contact dermatitis. Allergic contact dermatitis is a type of allergic reaction that occurs when the skin comes into contact with an allergen.

# Demo



CareConnect
Welcome to CareConnect! Your personal health assistant!
Hi I'm CareConnect! How can I help?
What is up?

# Challenges Faced

- Optimizing the usage of resources.
- Determination of the optimal prompt format.
- Preventing the model from learning the typos and noise.
- Tuning the model such that it ignores the context if it isn't relevant to the query.
- Optimizing the inference time and performance of the model.
- Optimizing the parameters for dialogue performance at deployment.
- Misleading evaluation metrics.

# Conclusion & Future Directions

- Meta Llama-3 was tuned on medical dialogue and alignment datasets and integrated into a web application for user-friendly interface.
- Although the model achieved high metrics both qualitatively and quantitatively, its performance during deployment did not meet expectations. The decoding parameters have to be further optimized.
- Furthermore, the inference time also has to be reduced significantly to make it feasible for usage.
- Although the design changes made in terms of providing context, summarizing the data, and threshold seemed to improve the models functionality, further work is to be done to address the remaining deployment performance issues.

# Thank you!

Any Questions?