Companion Agentic AI for COVID-19 Patient Support

Author: Sujay Gopinathan (SG59258), Email: sujay.gopinathan@utexas.edu

Abstract

This paper presents the design, development, and evaluation of a Companion Agentic Artificial Intelligence (AI) system tailored for supporting COVID-19 patients. Using fine-tuned large language models (LLMs) and embeddings derived from a synthetic COVID-19 dataset, we developed a retrieval-augmented generation (RAG) system that enhances human-AI interaction through explainable outputs. Traditional zero-shot Open-API approaches were found to be insufficiently accurate; thus, we constructed a custom dataset, fine-tuned embeddings, and applied logistic regression, achieving significant performance improvements (AUROC 92%, AUPRC 93%). The proposed system demonstrates the power of integrating fine-tuned models and embeddings to create trustworthy, transparent, and explainable AI companions for healthcare.

1 Introduction

The COVID-19 pandemic imposed unprecedented challenges on global healthcare systems, highlighting the need for innovative solutions to augment clinical decision-making. Companion AI systems, particularly those leveraging advanced language models, offer a promising avenue to support overburdened healthcare professionals. However, reliability, interpretability, and domain-specific accuracy remain critical barriers to real-world adoption.

This project explores the creation of an Agentic AI companion trained on COVID-19 synthetic datasets. By fine-tuning embeddings and employing a retrieval-based interaction system, we aim to address the shortcomings of zero-shot general-purpose models and provide explainable, clinically relevant outputs thus increasing the trust between humans and AI.

2 Related Work

Large language models (LLMs) such as GPT-3 and GPT-4 have demonstrated remarkable capabilities across diverse tasks. However, their application in specialized domains like healthcare often suffers due to a lack of domain-specific fine-tuning, mainly because of lack of data available for the models due to privacy reasons. Fine-tuning on domain data significantly enhances performance by aligning the model's latent representations with task-specific distributions.

Retrieval-augmented generation (RAG) has emerged as a powerful paradigm to combine pre-trained models with knowledge retrieval components, providing updated and contextually relevant information during generation. Recent works have explored explainable AI (XAI) approaches to foster transparency, trust, and regulatory compliance in healthcare AI applications.

Our work integrates these advances, demonstrating the effectiveness of fine-tuned embeddings, RAG architecture, and explainable conversational agents in a healthcare setting.

3 Methodology

3.1 Dataset

The project utilizes the synthetic COVID-19 dataset published by MITRE, containing anonymized patient records. The following tables were analyzed:

- patients.csv.gz

- observations.csv.gz
- conditions.csv.gz,
- procedures.csv.gz.

3.2 Data Preprocessing

COVID-19 cases were first filtered using SNOMED code 840539006. Following key input features relevant to treating COVID were extracted from the observations table.

- Oxygen saturation (LOINC 2708-6)
- Respiratory rate (LOINC 9279-1)
- Ferritin levels (LOINC 2276-4)

Output Label for each patient were created from the procedures table. Ventilator use was identified by SNOMED code 26763009.

3.3 Model Training

The data set was split into training and test dataset.

Zero-shot Open-API approaches were initially tested but found lacking in consistency. Accuracy however improved to around 50 % by giving context from the training dataset.

To further improve the accuracy, embeddings from the last layer of the model were generated and a logistic regression classifier was training on the embeddings. This resulted in the accuracy improving to above 90% on the test data.

This proved general purpose LLM models can be finetuned with domain specific knowledge at a very low cost in terms of GPU usage to provide very high accuracy.

3.4 Retrieval-Augmented Generation (RAG) Agent

Since we now have a fine-tuned domain specific COVID model, a virtual agent was developed to demonstrate how an agent can be used to generate responses to user queries.

The embeddings were stored in a vector database and a retrieval mechanism was developed to assist an agent in generating properly curated responses.

4 Experiments

4.1 Experimental Setup

The dataset was split into 80% training and 20% testing with stratified sampling. Metrics evaluated included AUROC and AUPRC.

Zero-shot Open-API responses served as baseline data.

5 Results

Embedding-based logistic regression significantly outperformed baselines:

Zero-shot Open-API: AUROC 49%, AUPRC 55%
Logistic Regression (embeddings): AUROC 92%, AUPRC 93%

6 Discussion

This study illustrates that zero-shot approaches are inadequate for critical clinical tasks. Fine-tuning on domain-specific datasets and using embeddings greatly improved both AUROC and AUPRC.

Retrieval mechanisms facilitated contextualized, explainable responses. Although the synthetic dataset provides a strong start, real-world validation remains necessary.

Future work involves expansion into multimodal inputs and real-world clinical validation.

7 Conclusion

Fine-tuned embeddings and retrieval-augmented generation enable the creation of clinically meaningful, explainable AI agents. These agents have the potential to assist healthcare professionals during public health emergencies and improve patient outcomes.

References

1. Brown, T.B., Mann, B., Ryder, N., et al. (2020). "Language Models are Few-Shot Learners." NeurIPS 2020.
2. Gururangan, S., Marasović, A., Swayamdipta, S., et al. (2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." ACL 2020.
3. AutoGen framework: https://microsoft.github.io/autogen/stable/
4. Project code: https://github.com/sujaycloud/aih/blob/main/project.ipynb