Cover Page

LINK FOR CODE AND PRESENTATION :
HTTPS://GITHUB.COM/FABLIHA-BD/LLM-COMPARATIVE-STUDY

VIDEO LINK:
HTTPS://UTEXAS.HOSTED.PANOPTO.COM/PANOPTO/PAGES/VIEWER.ASPX?ID=AFD3BB72-A0B5-4340-85A0-B2CE002EFEC7

Post-COVID Symptom Analysis: A Comparative Study of OpenAI, DeepSeek, and Traditional Machine Learning Models

Fabliha Fairooz

The University of Texas, Austin

1 INTRODUCTION

The long-term effects of COVID-19, often called long COVID or post-acute sequelae of SARS-CoV-2 infection (PASC), have become a serious issue, affecting many people's lives even after they recover from the initial infection. Patients can experience ongoing or new symptoms in their lungs, heart, brain, and other parts of the body. I wanted to do this project because long COVID has been like a villain, silently impacting people's health, and I wanted to explore which type of AI model would be better at analyzing and understanding these symptoms. My goal was to help people better understand whether the symptoms they are experiencing are truly linked to their past COVID-19 infection, or if they might be caused by something else that needs different treatment.

In this study, three different methods are compared for classifying post-COVID symptoms:

- Using an OpenAI large language model (LLM) with custom embedding-based retrieval to provide clinical context before querying
- Using an open-source DeepSeek-R1 LLM, and
- Using a traditional machine learning method trained on patient data.

By comparing these approaches on the same dataset, we aim to see which method works best and discuss how AI models can support medical decision-making for long COVID patients.

research used statistical methods, clustering techniques, and traditional machine learning models like logistic regression and tree-based approaches to identify symptom patterns and predict long COVID risk factors from electronic health records (EHRs). Recently, large language models (LLMs) trained on huge amounts of medical text have been tested for clinical reasoning tasks. OpenAI's GPT-3 and GPT-4 models, for example, have done well at answering medical questions, but mostly without any special task-specific training. DeepSeek-R1 is a newer open-source LLM that offers similar reasoning abilities compared to expensive proprietary models. However, applying LLMs to classify EHR data, such as deciding whether a symptom is related to COVID-19, is still new and being explored.

2 RELATED WORK

Early research used statistical methods, clustering techniques, and traditional machine learning models like logistic regression and tree-based approaches to identify symptom patterns and predict long COVID risk factors from electronic health records (EHRs) [1]. Recently, large language models (LLMs) trained on huge amounts of medical text have been tested for clinical reasoning tasks. OpenAI's GPT-3 and GPT-4 models, for example, have done well at answering medical questions, but mostly without any special task-specific training. DeepSeek-R1 is a newer open-source LLM that offers similar reasoning abilities compared to expensive proprietary models [2]. However, applying LLMs to classify EHR data, such as deciding whether a symptom is related to COVID-19, is still new and being explored.

3 METHODOLOGY

Figure 1 illustrates the main workflows used in this project. The details are described in brief below. At a high level, the workflows include:

(1) Process 10K Covid-related dataset (EHR-like) to filter patient symptom narratives, map labels, and split into train and test sets.

(2) Generate embeddings for patient narratives:

- Use OpenAI's text-embedding-ada-002 model for logistic regression and OpenAI querying.
- Use Nomic-embed-text-v1 model for context retrieval with DeepSeek.

(3) Invoke models with patient narratives and/or retrieved context:

- OpenAI GPT 3.5- turbo model was queried with context retrieved using FAISS and OpenAI embeddings.
- DeepSeek-R1 model was queried locally with context retrieved using FAISS and Nomic embeddings.
- Logistic regression model was trained using OpenAI narrative embeddings.

(4) Given ground-truth labels and model predictions, calculate evaluation metrics:

- Area Under the ROC Curve (AUROC)
- Area Under the Precision-Recall Curve (AUPRC)

3.1 DATA PREPARATION:

This project, 10K covid-related publicly available dataset has been used [3] . From the dataset, only the patients, conditions, observations, medications, immunizations and encounters were loaded. At first, the data was filtered for the patients who had confirmed Covid-19. Then the recovery date was determined with covid stop datetime. If some covid stopped time is missing, then it filled with +14 days from covid start time (this can be modified).

Then the observation data is selected for only post covid recovery time. I kept only some specific symptoms here for simplifications. Calculated days after recovery, then labeled the data by 'likely related' or 'unlikely related'. By looking at the data distribution I prefered 60 as cutoff days to covid recovery. For now, the symptoms which occurred after 60 days are taken as unlikely related and otherwise likely related to covid.

Added some other attributes to the dataset as necessary. The final data has these attributes- descriptions (they are basically symptoms), days_after_recovery, value (such as scores of severity of symptom) , gender, age, post_covid_symptom (label), comorbidities, medications, immunizations, vaccination_before_covid. To create inputs for the large language models (LLMs), a method was developed to generate patient narratives from these attributes. Finally, the dataset was split into training and testing sets using an 80/20 ratio.

3.2 MODELLING APPROACHES:

I implemented three modeling approaches on the above data. The dataset was first processed by converting patient data into text narratives and splitting them into training and testing sets. These prepared narratives were used across all modeling approaches: an OpenAI LLM approach, a DeepSeek LLM approach, and a traditional machine learning model. To support

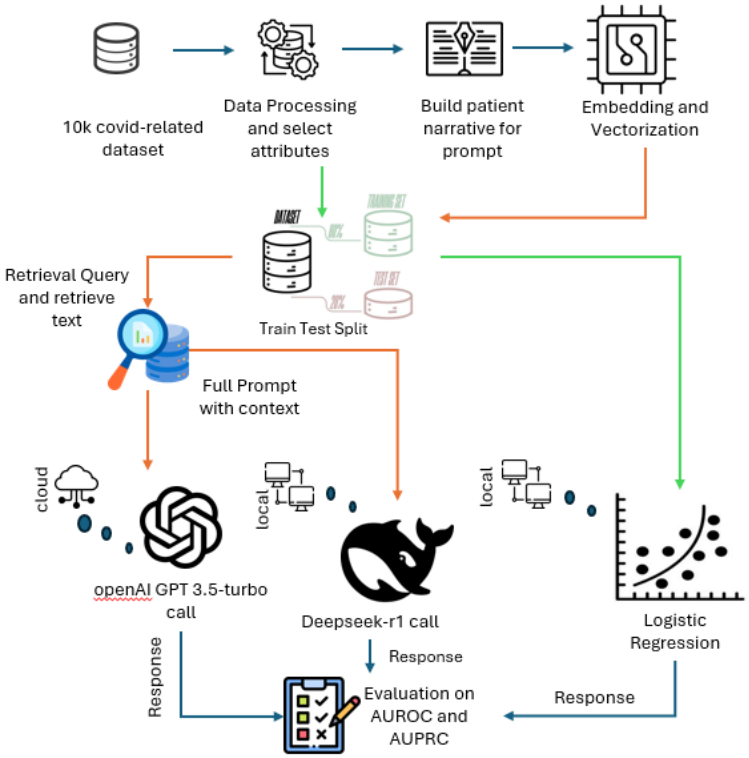

Figure 1: Workflow of the project , data preparations, build patient narrative, embedding and vectorization, RAG based query with full context, openAIGPT call and Deepseek call, prediction with linear regression.

similarity search, all patient narratives were converted into dense vector embeddings, and a FAISS index was built for fast retrieval based on vector similarity.

3.2.1 OpenAI GPT -3.5 turbo testing:

Each patient's narrative (including age, gender, symptoms, days after recovery, etc.) was used to create a prompt asking whether the symptom was related to COVID. Figure 2 shows some sample prompts and their responses. To improve GPT-3.5's reasoning, the three most similar narratives from the training set were retrieved using OpenAI's Ada embeddings stored in a FAISS index [4]. These retrieved examples were added to the prompt, and GPT's yes/no answer was taken as the model's prediction. Since GPT-3.5 Turbo is only available via API and not open-source, all inference was performed remotely with openAI API key [5]. The response time was moderate.

3.2.2 DeepSeek-R1 testing:

The DeepSeek-R1 model, hosted locally, was used in a similar way. However, for retrieval of similar examples, the Nomic-embed-text-v1 model was used to generate embeddings. The FAISS index built on Nomic embeddings enabled fast similarity search. As with GPT, DeepSeek was asked the yes/no question with retrieved examples for context, and its response was treated as the prediction. Although DeepSeek-R1 was run locally on an NVIDIA GeForce RTX 3060 Laptop

Prompt: Suppose , you are a doctor. I am a female patient 30 years old. I am having fever. Previously 1 month ago I had covid. does my fever a post covid symptom? give me the answer in one word yes or not .

Response: Yes.

Prompt: I am a male patient 40 years old. I am having sore throat. Previously 1 year ago I had covid. I had taken vaccine. Does my issue a post covid symptom? give me the answer in one word yes or not .

Response: No.

Figure 2: Sample Prompt and Response

GPU, the model inference was relatively slow due to hardware limitations, leading to longer overall processing time for the test set compared to the GPT-3.5-turbo API.

3.2.3 Traditional Machine Learning (Logistic Regression) testing:

A logistic regression model was trained using attributes of the data (not on patient narrative) on the training set of 3745 Data and corresponding labels ("likely related" or "unlikely related"). Then it was tested against the heldout dataset of 937 patient data.

3.3 Evaluating LLM Responses:

The OpenAI GPT3.5-turbo model was evaluated on a held-out test set of 500 patient symptom narratives. Linear regression based model was evaluated on around 800 patient dataset. The Deepseek-R1 was evaluated on the same held-out test set, but on 50 patient symptom narratives (as it took longer time for resource unavailability). The main evaluation metrics were Area Under the ROC Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC), suitable for binary classification under class imbalance. For logistic regression, predicted probabilities were used to calculate the metrics. For the LLM approaches (OpenAI GPT and DeepSeek-R1), binary outputs ("yes" or "no") were interpreted as predictions, and evaluation was based on these outcomes. Both LLM methods were used by incorporating retrieved context examples through embedding-based retrieval to assist the model's reasoning. Logistic regression was trained fully supervised. All experiments were conducted on the same dataset to ensure a fair comparison, with OpenAI GPT accessed via API and DeepSeek-R1 run locally using the Ollama backend.

4 RESULTS

In this study, both OpenAI's GPT-3.5 Turbo and DeepSeek-R1 large language models (LLMs) were evaluated for classifying post-COVID symptoms using retrieval-augmented prompts. While neither model was fine-tuned for this task, both were provided similar examples through embedding-based retrieval. GPT-3.5-turbo struggled with discrimination,

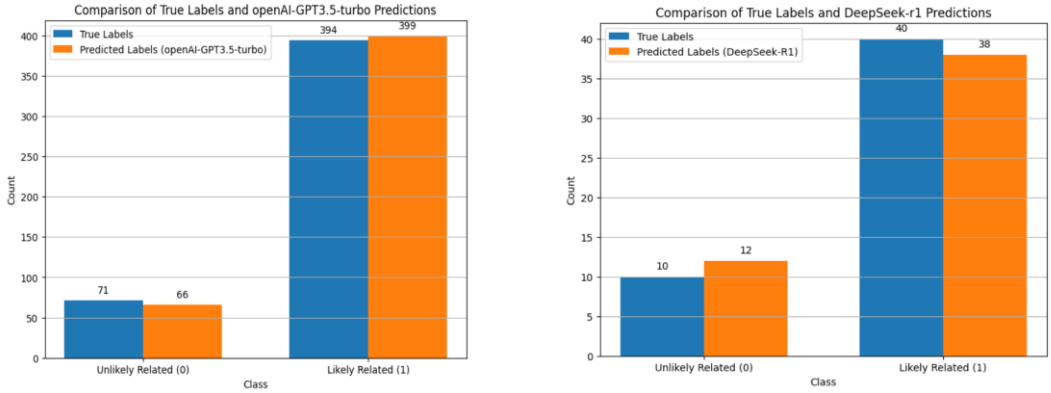

Figure 3: The prediction of OpenAI's model against ground truth values (Left), The prediction of Deepseek's model against ground truth values (Right)

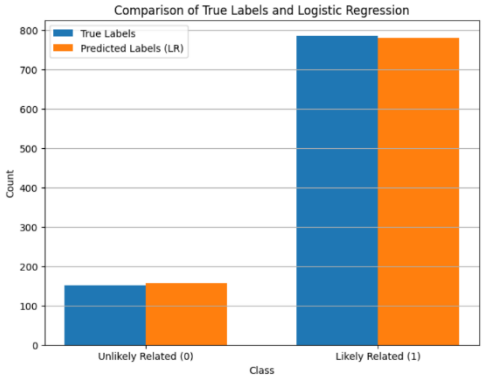

Figure 4: The prediction of Logistic Regression model against ground truth values

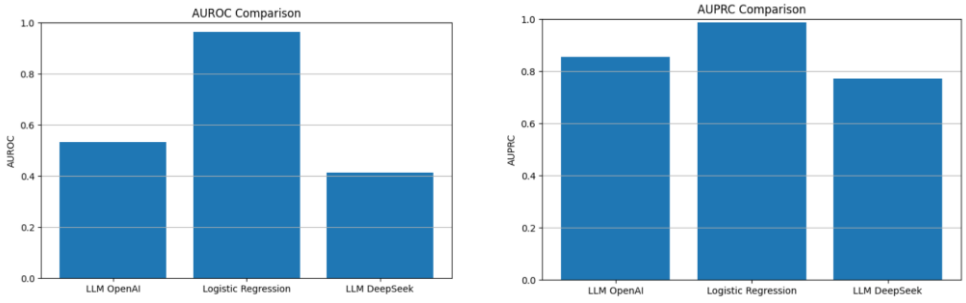

Figure 5: The AUROC (left) and AUPRC score (right) comparison on the 3 different models

frequently overpredicting symptoms as related to COVID, which led to a low AUROC close to random performance. Figure 3 (left) shows the prediction of OpenAI's model against ground truth values. This behavior likely reflects its broad exposure to medical texts where long COVID symptoms are emphasized, making it less sensitive to our specific 60-day labeling rule. DeepSeek-R1 showed slightly better ability to distinguish related and unrelated cases, achieving a higher

AUROC than GPT-3.5-turbo. Its smaller model size (8B parameters) and possibly different training dynamics may have made it more conservative in positive predictions. Figure 3 (right) shows the prediction of DeepSeek's model against ground truth values. However, both models still fell short in capturing the subtle timing and symptom patterns needed for high-precision classification. For comparison, a logistic regression model trained directly on embeddings achieved much stronger performance, highlighting that direct supervision still offers an advantage for structured clinical tasks. Figure 4 shows the prediction of LinearRegression model against ground truth values. Figure 5 shows the AUROC and AUPRC score comparison on the 3 different models . Overall, while LLMs demonstrated some reasoning ability, lack of task-specific fine-tuning limited their effectiveness for precise symptom classification.

5 CONCLUSION

In this study tested how large language models (LLMs) like OpenAI's GPT-3.5 Turbo and DeepSeek-R1 can classify post-COVID symptoms using patient narratives. Both models were used with similar case retrieval to help their reasoning.While they showed some understanding, they often struggled with the specific task. GPT-3.5-turbo tended to say symptoms were related to COVID even when they were not, and DeepSeek, although a little better, still missed some real cases. A simple trained model still performed better in this case. Some limitations of this study include the small dataset, the simplified labeling rule, and practical issues like the time and resources needed to run the models. Future Improvements:

- Use all the symptoms from the dataset instead of specific symptoms
- Instead of using 60 days as cutoff recovery days, it can be changed for better data availability and actual medical preference.
- For the experiments, more attributes can be added or removed depending on the resource's availability
- Experiment with more prompt engineering techniques.
- Try to include external medical knowledge to assist the large language model in generating the correct answer
- Experiments with other open sourced LLMs.
- Fine-tuning LLMs on labeled data or using LLMs to explain or highlight tricky cases could help make them more useful in clinical tasks.

REFERENCES

[1] A. I. Su et al., "Machine learning identifies long COVID patterns from electronic health records," *Nature Medicine*, vol. 29, pp. 47–48, 2023. Nature

[2] E. Glover, "What Is DeepSeek-R1? This high-profile AI model from the Chinese startup DeepSeek achieves comparable results to its American counterparts — at a fraction of the cost," *Built In*, Feb. 2025. builtin

[3] 54MB of 10k COVID related data. https://mitre.box.com/shared/static/9iglv8kbs1pfi7z8phjl9sbpjk08spze.zip

[4] Karen Ka Yan Ng, PharmD. https://orcid.org/0009-0002-2848-0571, Izuki Matsuba, B.A. https://orcid.org/0009-0002-2848-0571, and Peter Chengming Zhang, PharmD - "RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations" nejm AI

[5] *OpenAI*, "GPT-3.5 Technical Specifications." (Online). Available: OpenAI API documentation, 2023. (Mock reference for OpenAI model details.)