Generative LLM Powered Conversational AI Application for Personalized Risk Assessment: A Case Study in COVID-19

Mohammad Amin Roshani, Xiangyu Zhou, Yao Qiang, Srinivasan Suresh, Steve Hicks, Usha Sethuraman, Dongxiao Zhu

Submitted to: JMIR AI
on: October 09, 2024

Table of Contents

Generative LLM Powered Conversational AI Application for Personalized Risk Assessment: A Case Study in COVID-19

Mohammad Amin Roshani[1] BSc; Xiangyu Zhou[1] BE, MCS; Yao Qiang[2] BS, MCS, PhD; Srinivasan Suresh[3] MD; Steve Hicks[4] MD; Usha Sethuraman[5] MD; Dongxiao Zhu[1] BS, MS, PhD

[1]Wayne State University Detroit US
[2]Oakland University Rochester US
[3]UPMC Children's Hospital of Pittsburgh Pittsburgh US
[4]Penn State College of Medicine Hershey US
[5]Children's Hospital of Michigan Detroit US

Corresponding Author:
Dongxiao Zhu BS, MS, PhD
Wayne State University
5057 Woodward Ave
Suite 14101.3
Detroit
US

Abstract

Background: Large Language Models (LLMs) have demonstrated powerful capabilities in natural language tasks and are increasingly being integrated into healthcare for tasks like disease risk assessment. Traditional machine learning methods rely on structured data and coding, limiting their flexibility in dynamic clinical environments. This work presents a novel approach to disease risk assessment using generative LLMs via conversational AI, eliminating the need for programming.

Objective: This study explores the use of pre-trained generative LLMs, including LLaMA2-7b and Flan-T5-xl, to assess COVID-19 severity in real time. The goal is to compare their performance with traditional classifiers, such as Logistic Regression, XGBoost, and Random Forest, which are trained on structured tabular data.

Methods: We fine-tuned LLMs using few-shot natural language examples from a dataset of 393 pediatric patients, developing a mobile application that integrates these models to provide real-time, no-code COVID-19 severity risk assessment through clinician-patient interaction. The LLMs were compared with traditional classifiers across different experimental settings, using Area Under the Curve (AUC) as the primary evaluation metric. Feature importance derived from LLM attention layers was also analyzed to enhance interpretability.

Results: Generative LLMs consistently outperformed traditional machine learning models, particularly in low-data settings. In zero-shot scenarios, the T0-3b model achieved an AUC of 0.75, whereas traditional classifiers like Logistic Regression and XGBoost lagged behind, with AUCs of 0.57 and 0.50, respectively. LLMs maintained their lead even as the number of training examples increased, outperforming traditional models up to 32-shot settings. For instance, the Flan-T5-xl model achieved an AUC of 0.70 in 32-shot experiments, further highlighting the LLMs' effectiveness in few-shot learning scenarios. Moreover, the mobile application provided real-time COVID-19 severity assessments and personalized insights through attention-based feature importance, adding value to the clinical interpretation of the results.

Conclusions: Generative LLMs provide a robust alternative to traditional classifiers, particularly in scenarios with limited labeled data. Their ability to handle unstructured inputs and deliver personalized, real-time assessments without coding makes them highly adaptable to clinical settings. This study underscores the potential of LLM-powered conversational AI in healthcare and encourages further exploration of its use for real-time disease risk assessment and decision-making support.

(JMIR Preprints 09/10/2024:67363)
DOI: https://doi.org/10.2196/preprints.67363

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Generative LLM Powered Conversational AI Application for Personalized Risk Assessment: A Case Study in COVID-19

Mohammad Amin Roshani, BS[1]; Xiangyu Zhou, MS[1]; Yao Qiang, PhD[2]; Srinivasan Suresh, MD[3]; Steve Hicks, MD[4]; Usha Sethuraman, MD[5]; Dongxiao Zhu, PhD[1]

[1]Department of Computer Science, Wayne State University, Detroit, Michigan, USA
[2]Department of Computer Science, Oakland University, Rochester, Michigan, USA
[3]Department of Pediatrics, UPMC Children's Hospital of Pittsburgh, Pittsburgh, Pennsylvania, USA
[4]Department of Pediatrics, Penn State College of Medicine, Hershey, Pennsylvania, USA
[5]Division of Emergency Medicine, Department of Pediatrics, Children's Hospital of Michigan, Detroit, Michigan, USA

Abstract

Background: Large Language Models (LLMs) have demonstrated powerful capabilities in natural language tasks and are increasingly being integrated into healthcare for tasks like disease risk assessment. Traditional machine learning methods rely on structured data and coding, limiting their flexibility in dynamic clinical environments. This work presents a novel approach to disease risk assessment using generative LLMs via conversational AI, eliminating the need for programming.

Objective: This study explores the use of pre-trained generative LLMs, including LLaMA2-7b and Flan-T5-xl, to assess COVID-19 severity in real time. The goal is to compare their performance with traditional classifiers, such as Logistic Regression, XGBoost, and Random Forest, which are trained on structured tabular data.

Methods: We fine-tuned LLMs using few-shot natural language examples from a dataset of 393 pediatric patients, developing a mobile application that integrates these models to provide real-time, no-code COVID-19 severity risk assessment through clinician-patient interaction. The LLMs were compared with traditional classifiers across different experimental settings, using Area Under the Curve (AUC) as the primary evaluation metric. Feature importance derived from LLM attention layers was also analyzed to enhance interpretability.

Results: Generative LLMs consistently outperformed traditional machine learning models, particularly in low-data settings. In zero-shot scenarios, the T0-3b model achieved an AUC of 0.75, whereas traditional classifiers like Logistic Regression and XGBoost lagged behind, with AUCs of 0.57 and 0.50, respectively. LLMs maintained their lead even as the number of training examples increased, outperforming traditional models up to 32-shot settings. For instance, the Flan-T5-xl model achieved an AUC of 0.70 in 32-shot experiments, further highlighting the LLMs' effectiveness in few-shot learning scenarios. Moreover, the mobile application provided real-time COVID-19 severity assessments and personalized insights through attention-based feature importance, adding value to the clinical interpretation of the results.

Conclusions: Generative LLMs provide a robust alternative to traditional classifiers, particularly in scenarios with limited labeled data. Their ability to handle unstructured inputs and deliver personalized, real-time assessments without coding makes them highly adaptable to clinical settings. This study underscores the potential of LLM-powered conversational AI in healthcare and encourages further exploration of its use for real-time disease risk assessment

and decision-making support.

Keywords: Personalized Risk Assessment; Large Language Model; Conversational AI; COVID-19

Introduction

Background

Disease risk assessment is a critical tool in public health surveillance, where demographic variables and social determinants are often utilized to assess a patient's susceptibility to disease, predict treatment response, and forecast severity outcomes. These predictions have been carried out using traditional classification models that are trained *de novo* for each disease or condition using curated tabular data [1-3]. For example, Wang et al. [2] developed a linear model-based multi-task learning approach to predict the risk of childhood obesity according to their geolocations. Li et al. [3] developed a mixture neural network approach to stratify patients and predict heart failure risk within each group.

The advent of transformers has marked a significant shift, allowing researchers to deploy these advanced models for various tasks, thereby improving prediction accuracy and handling complex data structures more effectively. Researchers have extensively used BERT-style models [4] in various healthcare tasks. Notable examples include ClinicalBERT [5] and BioClinicalBERT [6], both trained on clinical notes in the MIMIC-III database. Additionally, MedBERT [7] was further trained on electronic health records (EHRs), resulting in high Area Under the Curve (AUC) scores for disease risk prediction. However, BERT-based models, primarily used for *discriminative* tasks, are limited in their ability to process streaming question and answer (QA) pairs, such as in conversational data science tasks, due to their architecture.

Generative LLMs for Healthcare

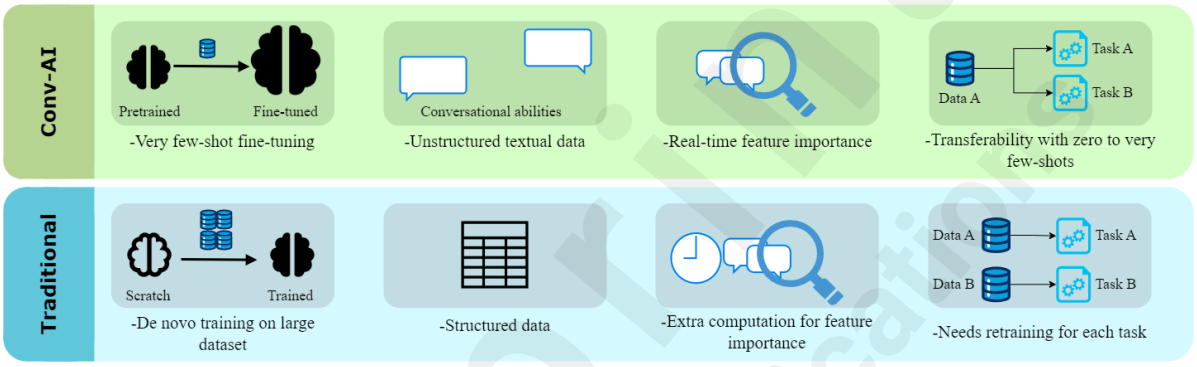
Generative large language models (LLMs), such as OpenAI's GPT-3 [8], have introduced significant advancements in Natural Language Processing (NLP) for healthcare by transcending the limitations of discriminative models like BERT. Unlike BERT-style models, which often require extensive preprocessing and are primarily tailored for specific tasks with structured inputs, generative LLMs excel at handling diverse data formats, including both structured clinical data and unstructured text such as patient narratives and medical histories. This versatility allows them to integrate and synthesize information from multiple sources, making them highly effective for complex tasks such as predicting disease severity.

With increasingly longer context windows, up to 8,192 tokens in OpenAI's GPT-4 [9], generative LLMs can efficiently manage extensive patient records and interaction histories. This capability to process long, streaming, and varied inputs, coupled with their extensive pre-training on diverse datasets, allows generative LLMs to generalize effectively even with limited labeled domain-specific data. Furthermore, their ability to handle multi-hop questions and answers positions them uniquely for real-time conversational applications, facilitating no-code disease assessment via interactive patient engagements. These strengths make generative LLMs particularly suitable for tasks such as disease severity risk assessment, where leveraging pre-trained world knowledge and user-provided natural language inputs allows for accurate predictions without the need for coding.

Despite the remarkable performance of proprietary black-box LLMs, such as GPT-4 [10] and MedPaLM-2 [11], researchers are increasingly interested in deploying white-box models in healthcare and other high-stakes domains since these models can mitigate risks related to data privacy breaches and hallucination. Their transparency allows for task-specific and domain-specific fine-tuning at a reduced cost, providing researchers with complete control over the process. This shift towards encoder-decoder and decoder-only models is exemplified by PMC-LLaMA [12], a general-purpose LLM adapted from LLaMA and fine-tuned using instruction tuning on health and medical

corpora, which has outperformed LLaMA-2-70B and ChatGPT-175B in several health/medical Question-and-Answer (QA) benchmarks.

Despite these advancements, there remains a notable gap in research regarding the use of generative LLMs for disease diagnosis and risk assessment tasks. Addressing this gap is crucial for fully leveraging the potential of LLMs in healthcare applications, as they offer advanced capabilities in handling complex medical data and providing accurate predictions. One of the few studies in this area is CPLLM [13], which fine-tunes Llama2 [14] as a general LLM and BioMedLM [15], trained on biological and clinical text, for different prediction tasks. Our work, however, opens a new avenue of research in conversational data science to enable no-code personalized risk assessment via a conversational interface *anytime and anywhere*. We experiment with a broader range of white-box LLMs, including LLaMA2, Flan-T5, and T0 models, integrating them into a conversational agent mobile application with a natural language interface for no-code personalized risk assessment and patient-clinician communication. A comparison of our work to traditional methods is shown in Figure 1.



Contributions

Our contributions to the field of LLM-based disease risk assessment are multifaceted. First and foremost, we propose a paradigm shift from traditional machine learning-based health outcome prediction, which typically relies on structured tabular data, to conversational agent-based no-code prediction using streaming QAs. This is realized through the development of a GenAI-powered mobile application that integrates fine-tuned LLMs as the core for personalized risk assessment and patient-clinician communication. The application not only assesses disease risk for patients but also provides contextual insights related to risk surveillance and mitigation through natural language conversation.

Secondly, we demonstrate that generative LLMs can outperform traditional machine learning methods (Table 1), such as Logistic Regression [16], Random Forest [17], and XGBoost [18], in *low-data regimes*, which is critical for medical applications where labeled data is scarce. For instance, our results show that LLMs like the T0-3b model achieve an AUC of 0.75 in zero-shot settings, underscoring the ability of pre-trained LLMs to achieve high accuracy without task-specific training. Additionally, we provide a comprehensive comparison of both decoder-only and encoder-decoder models, fine-tuned using the widely adopted parameter-efficient LoRA (Low-Rank Adaptation) method [19].

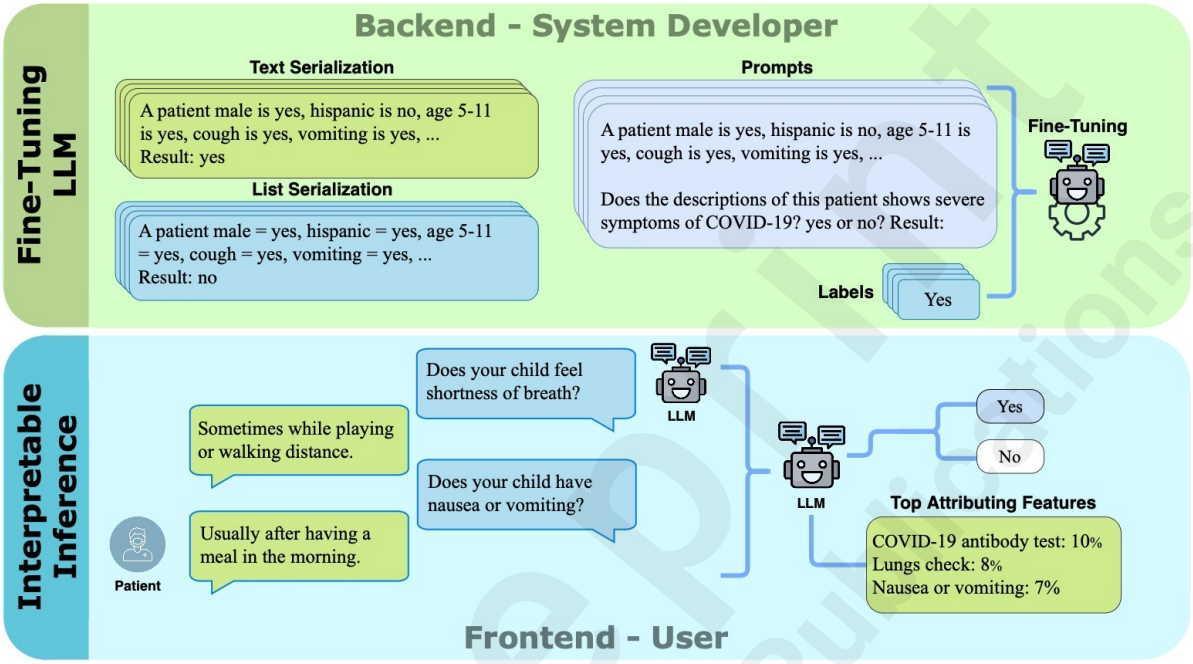
Thirdly, we introduce a feature importance analysis derived from the LLM's attention layers, providing personalized insights into the most influential factors driving the model's predictions. This enhances the interpretability and utility of the risk assessment for both patients and clinicians, offering real-time, instance-specific explanations during inference.

Methods

Our Research Objective

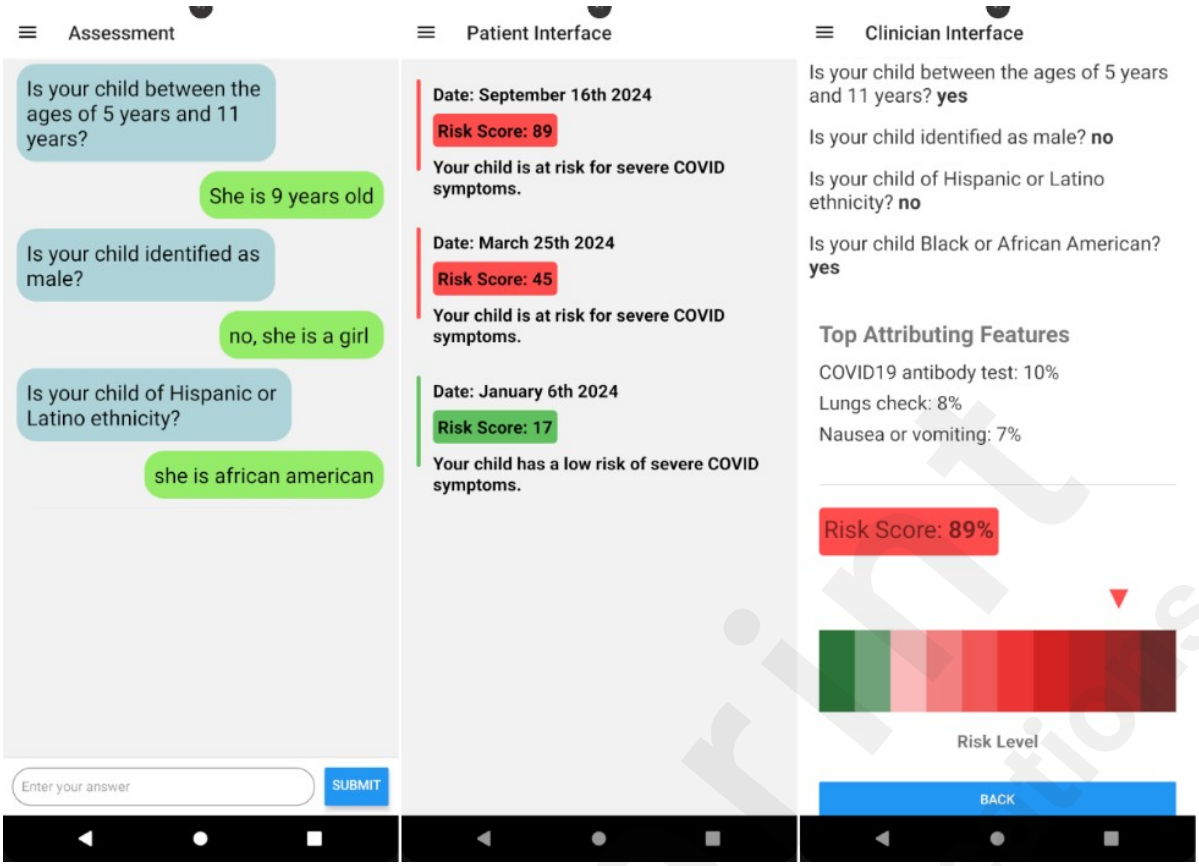
The primary objective of this research is to explore the effectiveness of pre-trained generative LLMs

in no-code risk assessment of disease severity using few-shot multi-hop QAs. We aim to evaluate how these generative LLM-powered conversational agents can utilize streaming QAs to accurately classify patient outcomes as severe or non-severe, which is crucial for early risk assessment and optimizing healthcare resource allocation. Through a case study of COVID-19 severity risk assessment, we develop an application that employs open-source generative LLMs to determine the severity of COVID-19 outcomes. This involves leveraging the models' capabilities in zero-shot and few-shot settings, with a focus on the use of serialization techniques to enhance their effectiveness and generalizability. We also integrate real-time feature importance to provide interpretable risk assessments. The workflow of our approach, from fine-tuning generative LLMs using serialized QA pairs to real-time risk assessment via a conversational interface, is illustrated in Figure 2.



Data Collection

A dataset was collected from the emergency departments (EDs) of Children's Hospital of Michigan and UPMC Children's Hospital of Pittsburgh between March 2021 and February 2022. The dataset includes $n=393$ participant records, each characterized by responses to a series of carefully designed questions. See Figure 3 for sample QAs. The severity of outcomes was defined as the need for supplemental oxygen ($\geq 50\%$ FiO2), non-invasive positive pressure or mechanical ventilation, extracorporeal membrane oxygenation, vasopressors or inotropes, cardiopulmonary resuscitation, or death from a related cause during hospitalization or within one month after discharge. These outcomes, categorized as severe or non-severe, were determined through chart reviews and parent surveys conducted thirty days post-discharge [20].



Tabular Data for Traditional Models

As traditional machine learning methods require tabular data as input, we formalize the questionnaire QA pairs as $D=\{(x_i, y_i)\}_{i=1}^{n}$, where $n=393$. $x_i \in \{0,1\}^d$ represents the binary feature vector of the i-th instance where $d=15$, and $y_i \in \{0,1\}$ denotes the binary class label indicating the presence or absence of severe COVID-19 symptoms determined by clinicians.

Each feature vector x_i consists of binary indicators representing social determinants, clinical, and demographic factors that may influence the severity of COVID-19, such as age, pre-existing conditions, vital signs, and laboratory test results. The feature names are denoted as $F=\{f_1, f_2, \ldots, f_d\}$, where each f_j is a natural-language string describing the corresponding attribute.

The task is to predict the binary outcome y_i based on the information provided in x_i. This constitutes a supervised learning problem where the objective is to train a model to minimize prediction error on unseen data.

Serialization for New Conversational AI

At the time of data collection during 2021-2022, we did not yet have a conversational agent (chatbot) for automated data donation from users, so we used a questionnaire to collect answers from each patient based on a set of questions designed for this study. As a result, the native format of the dataset consists of QA pairs, which were subsequently serialized to fine-tune the generative LLMs for the risk assessment task. It is important to note that the fine-tuned model is capable of assessing risk using streaming QAs in real time (Figures 2, 3).

To achieve serialization, the features in our dataset are denoted as $f_1, f_2, \ldots, f_d$, and their associated values as $v_1, v_2, \ldots, v_d$. This notation provides a structure that is transformed into natural language prompts for the LLM.

We used two main serialization methods, the *List Template* and the *Text Template*, to create natural language representations of the data. As shown in Figure 2, the List Template links each feature with

its value using an equal sign ('='), while the Text Template uses a narrative structure with the word "is" to connect each feature with its value. These templates enable us to evaluate which serialization approach better translates the data into actionable insights by the LLM.

Generative LLMs

We explore the capabilities of three white-box LLMs---LLaMA2 [14], T0 [21], and Flan-T5 [22]---focusing on their application in risk prediction for COVID-19 using both the native QA pairs and the formatted tabular dataset.

To our knowledge, this is *one of the the first attempts* leveraging generative LLMs and conversational data science for disease risk assessment across various LLMs and few-shot settings. Our selection includes both decoder-only (LLaMA2) and encoder-decoder architectures (T0 and Flan-T5), allowing for a comprehensive assessment and comparison of their performance. The white-box nature of these models is particularly advantageous as it enables setup on local hosts with private datasets, ensuring precise risk assessment by allowing direct access to model weights and logits.

The input to the LLMs is a serialized string generated from the tabular data using the previously explained serialization strategies. Given a feature vector $x_i=[f_1,f_2,\ldots,f_d]$ and their associated values $v_1,v_2,\ldots,v_d$, the serialized input string S_i can be represented using either the List Template or Text Template serialization methods (Figure 2).

The LLM processes the serialized input string S_i and outputs logits for the next token in the sequence. We focus on the logits corresponding to the tokens 'yes' and 'no', which indicate severe or non-severe symptoms respectively. The probabilities for these tokens are obtained by applying the softmax function to the logits:

$$p(yes|S_i)=\frac{e^{logit\,s_{yes}}}{e^{logit\,s_{yes}}+e^{logit\,s_{no}}}$$

The probability $p(yes\vee S_i)$ indicates the likelihood of severe symptoms based on the input data S_i. This probability is directly used as the severity risk score for evaluation purposes.

To determine the binary predicted label $\hat{y}_i$ from this probability:

$$\hat{y}_i=\begin{cases}1, if\ p(yes|S_i)>0.5\\ 0,\wedge otherwise\end{cases}$$

The probability score $p(yes\vee S_i)$, reflecting the severity risk, is used to compute the AUC for evaluation (Figure 2).

Evaluation Setting

Zero-Shot Setting

In the zero-shot setting, our approach leverages the intrinsic capabilities of LLMs. These models, unlike traditional classifiers such as Logistic Regression and XGBoost, have been extensively pre-trained on diverse datasets. This extensive pre-training enables them to apply their accumulated world knowledge directly to specific classification tasks without additional training, demonstrating exceptional generalizability.

We assess the zero-shot prediction effectiveness of these LLMs by presenting them with tasks aligned with our study's objectives that they have not been specifically trained on. The models interpret and classify new, unseen data solely based on their pre-trained knowledge. This approach not only highlights the potential of LLMs in real-world applications but also evaluates their ability to generalize from their training to novel scenarios in healthcare.

This zero-shot methodology allows us to evaluate how well these LLMs can recognize and classify complex, previously unseen patterns in healthcare data, providing valuable insights into their practical applicability and limitations in clinical settings.

Few-Shot Fine-Tuning

In the few-shot setting, we utilize sample sizes of 2, 4, 8, 16, and 32 to fine-tune the LLMs, aiming to examine the effect of training sample size on model performance compared to traditional classifiers. To ensure fairness and reduce bias in the fine-tuning process, we maintain a balanced ratio of positive ($y_i=1$) and negative ($y_i=0$) samples, with an equal number of examples from each class in each sample size.

To enhance computational efficiency in adapting the LLMs to our specific tasks, we employ a parameter-efficient fine-tuning approach using LoRA (Low-Rank Adaptation) [19]. Instead of adjusting all parameters within the model, LoRA involves training a small proportion of parameters by integrating trainable low-rank matrices into each layer of the pre-trained model. This method allows the model to quickly adapt to new tasks by optimizing only a subset of parameters, thereby preserving the general capabilities of the LLM while enhancing its performance on task-specific features.

Feature Importance Analysis

In disease risk assessment, interpretability is as critical as accuracy, particularly when both are provided to the user in real-time. Here, we introduce a novel approach for analyzing feature importance by leveraging the attention mechanisms inherent in the output layers of generative LLMs. This method provides additional insights into the risk assessment process of the model, which is valuable for both clinicians and patients in understanding the factors contributing to the model's ouput.

Our approach involves extracting attention scores from the model's output layer, where the attention assigned to each input token is interpreted as an indicator of feature importance. We compute the attention for each feature-value pair and associate the average attention score with the corresponding feature. This provides a holistic view of which features, along with their associated values, influence the model’s output.

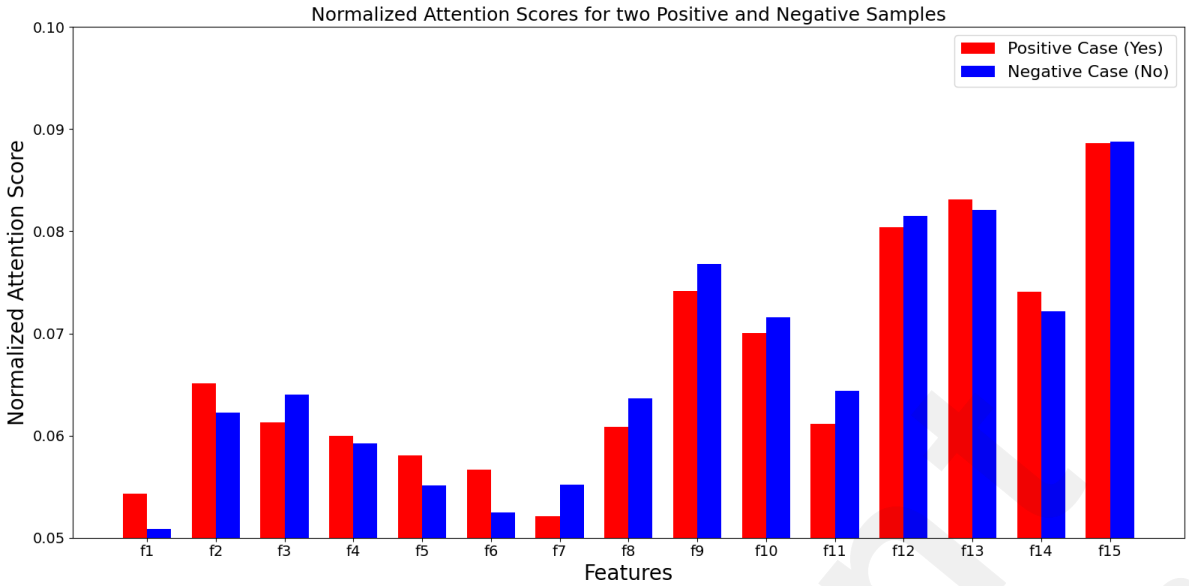
For an input sequence such as:

A patient with $f_1=v_1, f_2=v_2, \ldots, f_{15}=v_{15}$.

Does this patient have COVID-19, yes or no?

We calculate attention scores for each feature-value pair in the original sequence. The average attention score for each feature-value pair is then computed, and the score is associated with the feature itself, offering a representation of feature importance in the context of disease severity risk.

This normalized attention score serves as a proxy for feature importance, offering clinicians and patients a clearer understanding of which features (e.g., age, pre-existing conditions, vital signs, etc.) are most influential in the model's assessment of COVID-19 severity risk. As illustrated in Figure 4, the plot shows the normalized attention scores from the LLaMA2-7b model in the 32-shot setting for two test cases: one positive (yes) and one negative (no).



For the positive case, the top five features with the highest attention scores, as shown in this figure, are:

1. f15: COVID-19 antibody test
2. f13: Lungs check
3. f12: Nausea or vomiting
4. f9: Cough
5. f14: Eye redness

By integrating this analysis into our mobile application, we enhance the interpretability of LLM-based risk assessments, empowering users with deeper insights into the model's reasoning process.

Mobile Application

To provide users with code-free disease severity risk assessment and enhance user experience, we developed a mobile conversational agent powered by the aforementioned generative LLMs. This application is designed to facilitate the assessment and management of COVID-19 in children, with potential applicability to other diseases and conditions. It offers two versions: one for patients to donate their health information via answering the questions and receive real-time severity risk assessments, and another for clinicians to manage, review, and interpret the sessions donated by patients. The primary goals are to enhance early detection of severe outcomes, improve patient-clinician communication, and streamline the overall risk assessment process.

The application targets patients, clinicians, and other healthcare providers involved in managing pre-clinical cases. It leverages the capabilities of generative LLMs to analyze patient responses and provide immediate feedback on the risk of severe symptoms. Developed using React Native and JavaScript for the front end, Firebase for database management, and various frontend technologies, the application provides a user-friendly, efficient, and effective solution for managing disease risks. It aims to improve patient outcomes by facilitating timely and informed decision-making.

Database Structure

Our mobile application utilizes Firebase for database management, structured into three primary collections: Users, Questions, and Answers.

- Users: This collection includes essential user information such as ID, Email, and isAdmin. The ID uniquely identifies each user, the Email serves as contact information, and the isAdmin field (a boolean) indicates whether the user has administrative privileges (clinicians) or not (patients).

- Questions: Each document in this collection has a unique ID and a Description field. The ID is used to reference questions in the Answers collection, and the Description contains the text of the question posed to the user, ensuring clarity and specificity in data mapping.
- Answers: This collection records user responses during their sessions. Each document includes a session ID, an array of Answers where each entry links to the relevant QuestionID from the Questions collection. Additionally, it contains a Text field for the user's detailed response, an Answer field for the LLM-generated response (e.g., Yes or No), a Date field marking the session's completion time, a Risk Score field, which is derived from the user's responses and utilized for subsequent risk prediction by the LLM, and an Important Features field, which stores the key features identified by the LLM's attention scores that contributed to the risk assessment.

User Interface - Assessment

As shown in Figure 3, on the *Assessment* page, we leverage the power of LLMs to engage in a conversation with the patient. This interaction allows us to ask questions and gather contextual information for each response. By doing so, we retrieve a binary answer (Yes/No) using the LLM, which is then provided to the primary care physician along with the patient's context to aid in decision-making.
After the user responds to each question, we use our LLM to generate a binary answer. This involves providing the LLM with an instruction that includes the question and the user's response, asking the LLM to interpret the response into a binary answer (Yes or No). This sequential process is performed for all questions. Currently, the input for the final LLM-based risk assessment, which predicts the COVID-19 severity risk, is based solely on the set of binary answers generated by the LLM. Future enhancements could incorporate the original user responses to improve context understanding.

We currently utilize the Llama2-7b API for answer retrieval. Our long-term goal is to integrate a fine-tuned LLM hosted on our servers to ensure better optimization and accuracy specific to our dataset, as evidenced by the improved performance results discussed in this paper.

User Interface - Patient and Clinician Results

Patients can submit a session at any time, receiving an immediate risk assessment in the *Patient Results* section (Figure 3). This section displays all sessions submitted by the current user, along with their respective risk assessments.
In the *Clinician Results* section, clinicians can access all sessions from their patients, organized by patient ID for efficient review. Each session includes a comprehensive report featuring the predicted risk score, ensuring transparency and aiding in clinical decision-making.
Upon submission, a patient's session is instantly available in both the patient's and clinician's panels. While patients can only view their own sessions, clinicians can review all sessions from their assigned patients. This setup supports real-time updates through Firebase, facilitating seamless communication and follow-up between patients and their healthcare providers. Moreover, the application provides personalized feature importance analysis based on the LLM's attention layers, giving both patients and clinicians additional insights into the most critical factors influencing the risk assessment.

Results

Training and Fine-Tuning Settings

In our experiments, we employed a rigorous setup using five specific random seeds—0, 1, 32, 42, and 1024—to ensure diverse dataset initialization and mitigate potential biases in data allocation.

For traditional machine learning methods, the dataset of 393 samples was divided into 65% training, 15% validation, and 20% testing segments. Although the full training set is available, we focus specifically on training the models with up to 32 shots to examine performance in the few-shot regime. For LLMs, we similarly fine-tune the models using up to 32 shots, highlighting their capability to generalize in low-data settings with minimal task-specific examples.
When fine-tuning LLMs using LoRA, we monitored the validation loss to select the best model checkpoint, aiming to minimize overfitting and enhance generalization to the test set.

Effects of Serialization

Table 1 shows the performance of different serialization methods for the LLMs across various few-shot settings. We evaluated two primary serialization methods: List Template and Text Template, across models tested with 0, 2, 4, 8, 16 and 32 training shots to observe performance variations with the number of training examples.

Model	Number of Shots					
	0	**2**	**4**	**8**	**16**	**32**
Llama2-7b-L	$0.54_{.05}$	$0.69_{.07}$	$0.69_{.06}$	$0.68_{.04}$	$0.63_{.04}$	$0.66_{.07}$
Flan-t5-xl-L	$0.62_{.03}$	$0.64_{.02}$	$0.63_{.02}$	$0.68_{.06}$	$0.66_{.05}$	$0.69_{.06}$
Flan-t5-xxl-L	$0.60_{.03}$	$0.61_{.03}$	$0.61_{.05}$	$0.62_{.06}$	$0.59_{.10}$	$0.65_{.11}$
T0pp(8bit)-L	$0.69_{.04}$	$0.70_{.07}$	$0.70_{.05}$	$0.70_{.05}$	$0.68_{.06}$	$0.70_{.10}$
T0-3b-L	$0.68_{.04}$	$0.67_{.04}$	$0.68_{.05}$	$0.70_{.04}$	$0.67_{.04}$	$0.67_{.07}$
Llama2-7b-T	$0.59_{.05}$	$0.69_{.03}$	$0.69_{.01}$	$0.64_{.07}$	$0.63_{.05}$	$0.67_{.06}$
Flan-t5-xl-T	$0.69_{.03}$	$0.69_{.02}$	$0.69_{.03}$	$0.71_{.05}$	$0.69_{.04}$	$0.70_{.05}$
Flan-t5-xxl-T	$0.61_{.04}$	$0.58_{.03}$	$0.63_{.08}$	$0.59_{.10}$	$0.62_{.09}$	$0.63_{.10}$
T0pp(8bit)-T	$0.67_{.02}$	$0.65_{.05}$	$0.66_{.05}$	$0.68_{.04}$	$0.65_{.08}$	$0.67_{.08}$
T0-3b-T	$0.75_{.04}$	$0.65_{.06}$	$0.65_{.05}$	$0.68_{.03}$	$0.67_{.04}$	$0.65_{.08}$
Logistic Regression	-	$0.57_{.07}$	$0.55_{.10}$	$0.64_{.06}$	$0.61_{.11}$	$0.69_{.08}$
Random Forest	-	$0.57_{.07}$	$0.57_{.06}$	$0.62_{.08}$	$0.66_{.07}$	$0.68_{.07}$
XGBoost	-	$0.50_{.00}$	$0.50_{.00}$	$0.50_{.00}$	$0.54_{.06}$	$0.65_{.03}$

Table 1. Performance of models across different shot settings. All values represent the AUC rounded to two decimal places. Standard deviations given across five random seeds are shown as subscripts. The suffixes -L and -T represent List Serialization and Text Serialization, respectively.

The List Template often exhibited better performance at lower shot counts, while the Text Template typically outperformed the List Template as the number of training examples increased. The following summarizes the performance trends for each model:

- Llama2-7b:

In the zero-shot setting, the Text Template achieved an AUC of 0.59 compared to 0.54 for the List Template. At 2 training shots, both templates achieved an AUC of 0.69, but the Text Template began to outperform, reaching an AUC of 0.67 at 32 training shots compared to 0.66 for the List Template.

- Flan-t5-xl:

The Text Template consistently outperformed the List Template across most shot settings. At 2 training shots, the Text Template achieved an AUC of 0.69 compared to 0.64 for the List Template, and this lead continued up to 32 shots, where the Text Template achieved an AUC of 0.70 compared to 0.69 for the List Template.

- Flan-t5-xxl:

Both templates showed similar performance in the early few-shot settings. At 2 training shots, the List Template achieved an AUC of 0.61, slightly outperforming the Text Template, which achieved an AUC of 0.58. By 32 training shots, the List Template achieved an AUC of 0.65, slightly outperforming the Text Template, which achieved an AUC of 0.63.

- T0pp (8bit):

In the zero-shot setting, the List Template led with an AUC of 0.69 compared to 0.67 for the Text Template. This lead was maintained through most shot settings, with both templates achieving around 0.70 AUC by 32 shots.

- T0-3b:

The Text Template outperformed the List Template in the zero-shot setting, achieving an AUC of 0.75 compared to 0.68 for the List Template. In the 2-shot setting, the List Template performed slightly better with an AUC of 0.67 compared to 0.65 for the Text Template. At 32 shots, the Text Template closed the gap with an AUC of 0.65 compared to 0.67 for the List Template.

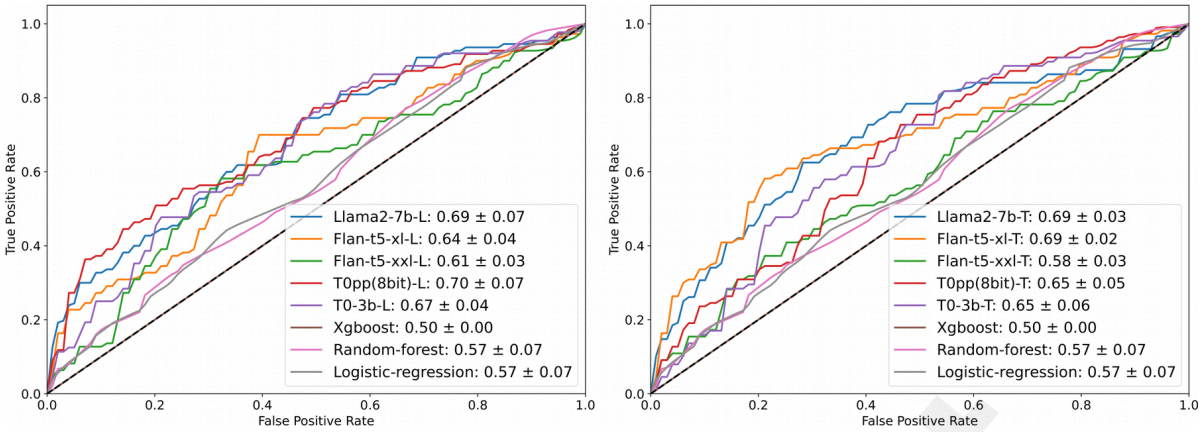
Overall, while the List Template often provides an initial advantage in early few-shot settings, the Text Template shows competitive performance as the number of training examples increases. This suggests that serialization choice can be important in low-data regimes. The Text Template's strong performance in the zero-shot setting, particularly for the T0-3b model, highlights its potential when no training data is available.

LLMs vs Traditional Machine Learning Methods

Our study highlights the versatility of LLMs for various healthcare applications, particularly in scenarios with limited data. To benchmark their performance against traditional machine learning methods, we compared LLMs with Logistic Regression, Random Forest, and XGBoost.

LLMs benefit from extensive pre-training, allowing them to generalize well to “unseen” data, unlike traditional methods that require substantial amounts of training data. As shown in Table 1, LLMs like T0-3b-T achieved an AUC of 0.75 in the zero-shot setting, outperforming traditional methods even without task-specific fine-tuning. This demonstrates the effectiveness of LLM-powered risk assessment without the need for additional labeled data.

In the 2-shot setting, LLMs continue to show strong performance relative to traditional methods. For instance, Figure 5 compares the average AUC across five different seeds in this scenario. The left panel shows results using the List Serialization (-L) approach, while the right panel shows results using the Text Serialization (-T) approach. In this 2-shot scenario, LLMs such as T0pp(8bit)-L and Flan-t5-xl-T achieve AUCs of 0.70 and 0.69, respectively, clearly outperforming traditional methods, including Logistic Regression, Random Forest, and XGBoost, which achieved AUCs of 0.57, 0.57, and 0.50, respectively.



LLMs' ability to perform well with minimal data highlights their advantage in low-data regimes. This makes them particularly suitable for real-time, no-code healthcare applications where rapid decision-making is required, even in scenarios where labeled data is scarce.

Furthermore, LLMs' capacity to handle streaming data formats, such as multi-hop question-answering (QA), enhances their integration into conversational interfaces, supporting real-time patient-clinician interactions. This flexibility offers significant utility in clinical settings where personalized and immediate risk assessments are needed (Figure 1).

Overall, while traditional methods may improve with larger datasets, LLMs demonstrate a clear advantage in dynamic, low-data healthcare environments. Their ability to handle incomplete data and streaming input formats makes them robust for real-world applications requiring adaptability and speed.

Discussion

Key Findings

Our research demonstrates that generative LLMs provide a robust and no-code approach for predicting COVID-19 severity, particularly effective in low-data regimes. These models excel in zero-shot and few-shot settings, showcasing their ability to perform well without extensive domain-specific training. This is crucial for real-time applications requiring immediate and reliable predictions, highlighting their exceptional generalizability compared to traditional classifiers like Logistic Regression, Random Forest, and XGBoost, which typically require more labeled data to achieve comparable performance.

Generative LLMs effectively handle diverse input formats, integrating both structured clinical data and unstructured natural language inputs from patient interactions. This flexibility enables them to synthesize information from various sources, such as patient medical histories and symptom descriptions, enhancing their utility in dynamic healthcare settings. In our study, we incorporated these models into a conversational interface, which facilitates real-time patient-clinician interactions and immediate risk assessments. This setup supports continuous data collection and leverages the conversational capabilities of LLMs to optimize clinical decision-making and resource allocation.

Future Directions and Limitations

Future work should focus on integrating continuous clinician-patient conversational data for fine-tuning or in-context learning (ICL), extending the application of LLMs beyond static disease prediction models. Techniques like Chain of Thought (CoT) and Chain of Interaction (CoI), which align with the interactive nature of medical consultations, show promise for enhancing model performance in interpreting and responding to patient data in real-time settings [23,24].

While our study utilized models like T0pp with parameter-efficient fine-tuning using LoRA, future research could explore newer and more advanced small language models such as LLaMA3-8b and Mistral-7b-Instruct, which have demonstrated exceptional performance in low-data regimes. These

models could offer greater efficiency and accuracy as computational resources and methodologies advance, supporting more sophisticated and scalable applications in healthcare.
However, as these models continue to evolve, addressing their vulnerabilities remains critical. Studies have demonstrated that adversarial attacks can hijack LLMs during in-context learning, undermining their performance in sensitive tasks such as disease risk assessment [25]. In adversarial in-context learning (ICL) scenarios, an attacker can manipulate inputs, influencing the model to produce inaccurate or harmful predictions. This poses significant risks in high-stakes settings like healthcare, where incorrect assessments could lead to adverse patient outcomes. As LLMs gain wider adoption in healthcare, enhancing their resilience against such adversarial techniques is essential to ensure safe and reliable patient outcomes.

Conclusions

In conclusion, generative LLMs offer a valuable tool for no-code risk assessment in low-data regimes. Their ability to perform zero-shot or few-shot transferability to new diseases or conditions and handle complex, varied inputs positions them as key assets for enhancing healthcare interventions and resource management. Furthermore, the incorporation of feature importance analysis derived from the LLM's attention layers provides an additional layer of interpretability, offering personalized insights into the decision-making process for both patients and clinicians.

Acknowledgements

MR conducted the experiments, designed the application, and wrote the manuscript. XZ contributed to the application design, assisted with the experiments, and provided revisions. DZ, as the supervisor, supported the project through funding and manuscript revisions. YQ offered suggestions on the experiments and revisions. SS, SH, and US assisted with dataset collection and provided feedback on the manuscript draft.

Conflicts of Interest

None declared.

Abbreviations

LLM: Large Language Model
COVID-19: Coronavirus Disease 2019
AUC: Area Under the Curve
LoRA: Low Rank Adaptation
QA: Question and Answer
GPT: Generative Pre-trained Transformer

References

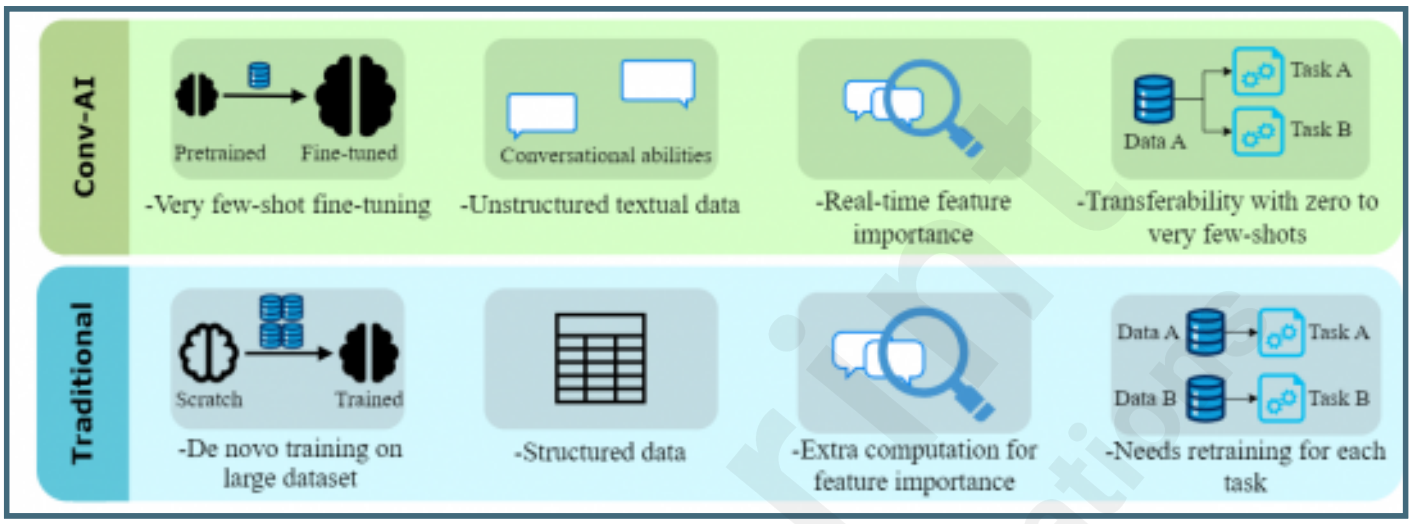
1. Li X, Zhu D, Levy P. Leveraging auxiliary measures: a deep multi-task neural network for predictive modeling in clinical research. BMC Med Inform Decis Mak 2018;18:45-53. doi:10.1186/s12911-018-0626-9
2. Wang L, Dong M, Towner E, Zhu D. Prioritization of multi-level risk factors for obesity. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019:1065-1072.
3. Li X, Zhu D, Levy P. Predicting clinical outcomes with patient stratification via deep mixture neural networks. AMIA Summits Transl Sci Proc 2020:367.
4. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
5. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting

hospital readmission. arXiv preprint arXiv:1904.05342, 2019.
6. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323, 2019.
7. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4:86. doi:10.1038/s41746-021-00427-8
8. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901.
9. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
10. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375, 2023.
11. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617, 2023.
12. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLAMA: towards building open-source language models for medicine. arXiv preprint arXiv:2304.14454, 2023.
13. Shoham OB, Rappoport N. CPLLM: clinical prediction with large language models. arXiv preprint arXiv:2309.11295, 2023.
14. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. LLAMA 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
15. Venigalla A, Frankle J, Carbin M. BioMedLM: a domain-specific large language model for biomedical text. MosaicML 2022;23(3):2.
16. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. John Wiley & Sons; 2013.
17. Breiman L. Random forests. Mach Learn 2001;45:5-32. doi:10.1023/A:1010933404324
18. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.
19. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
20. Hicks SD, Zhu D, Sullivan R, Kannikeswaran N, Meert K, Chen W, et al. Saliva microRNA profile in children with and without severe SARS-CoV-2 infection. Int J Mol Sci 2023;24(9):8175. doi:10.3390/ijms24098175
21. Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207, 2021.
22. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res 2024;25(70):1-53.
23. Han G, Liu W, Huang X, Borsari B. Chain-of-interaction: enhancing large language models for psychiatric behavior understanding by dyadic contexts. arXiv preprint arXiv:2403.13786, 2024.
24. Gramopadhye O, Nachane SS, Chanda P, Ramakrishnan G, Jadhav KS, Nandwani Y, et al. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. arXiv preprint arXiv:2403.04890, 2024.
25. Qiang Y, Zhou X, Zhu D. Hijacking large language models via adversarial in-context learning. arXiv preprint arXiv:2311.09948, 2023.

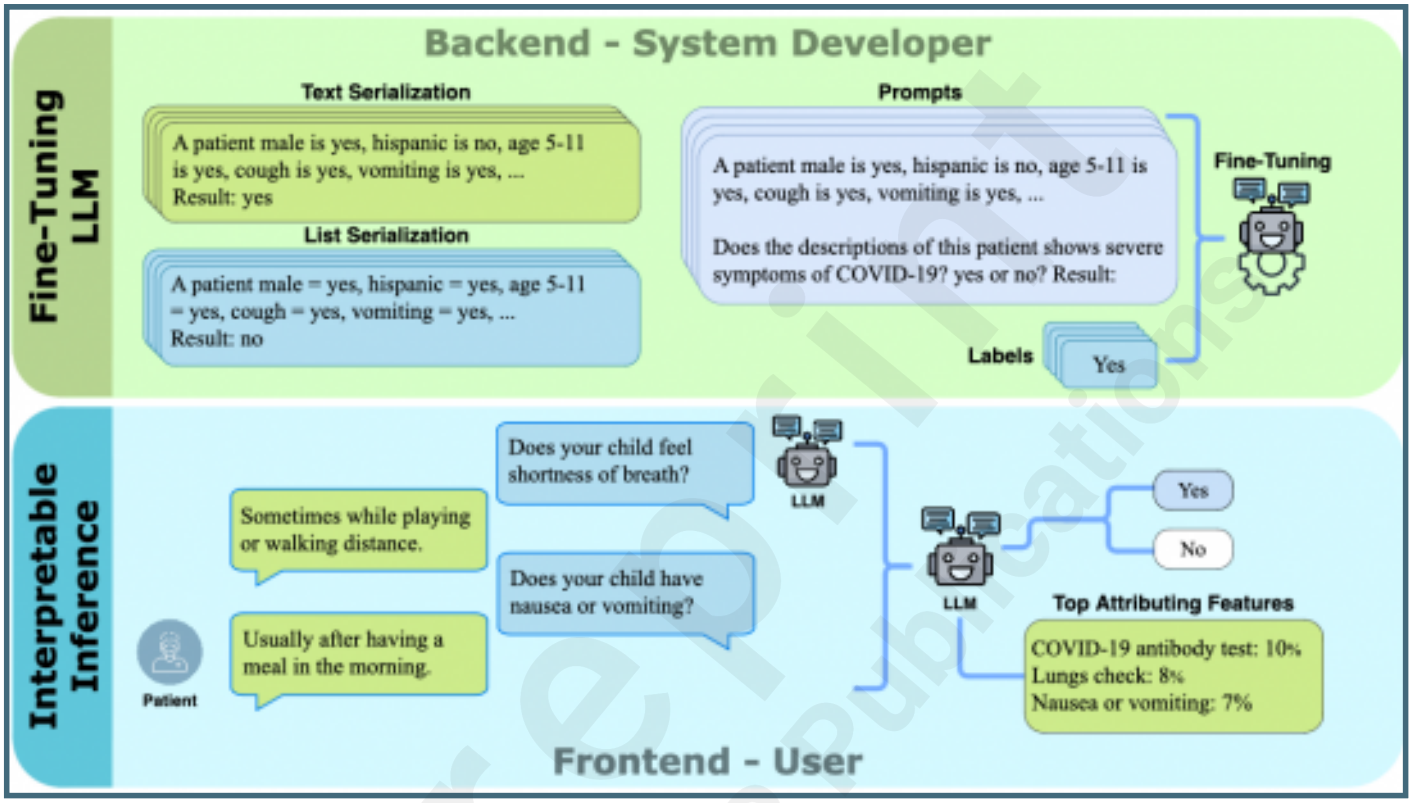
Supplementary Files

Figures

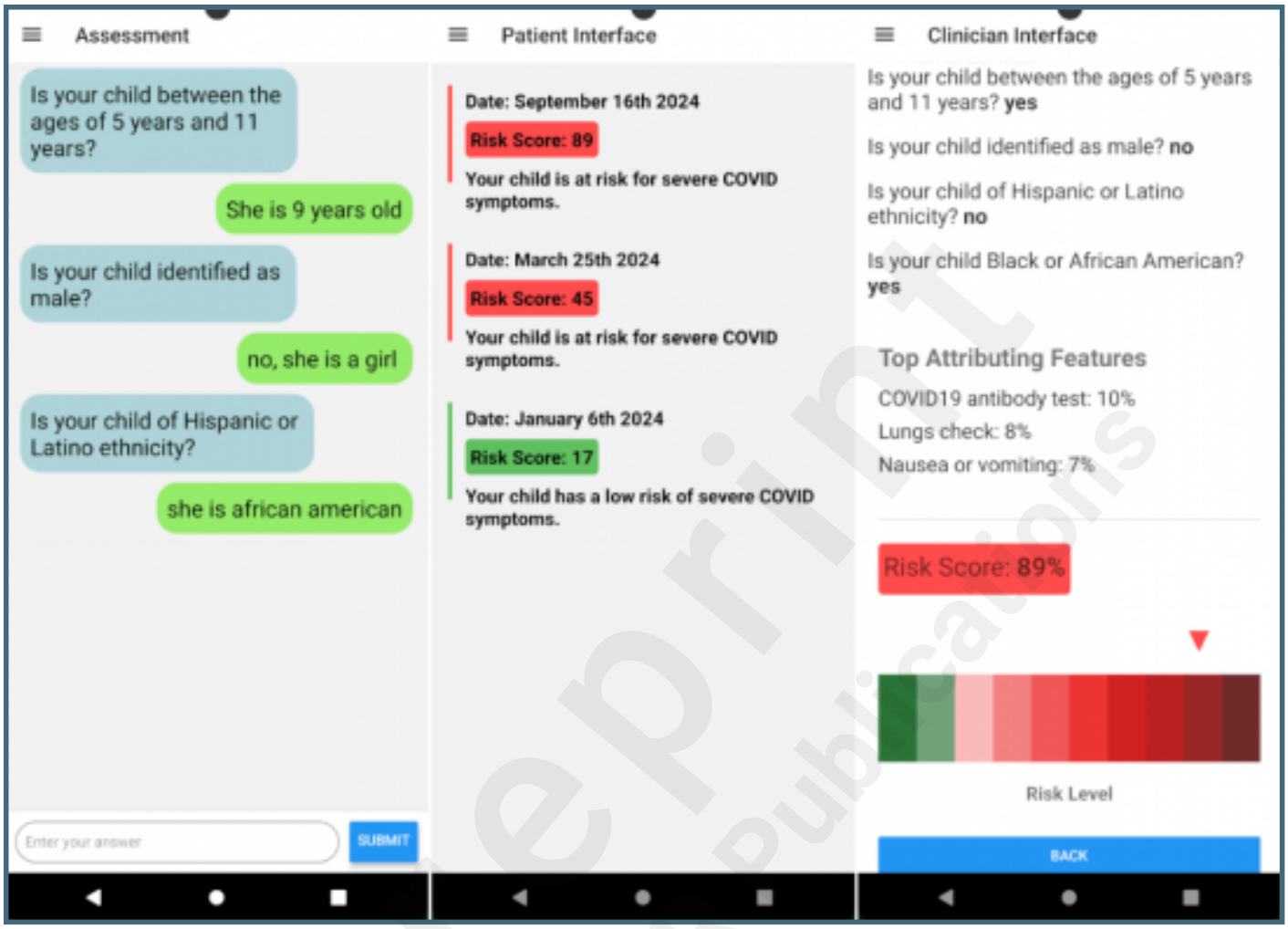
comparison between LLM-based conversational AI (Conv-AI) and traditional machine learning methods for disease risk assessment. The Conv-AI leverages pretrained models that require only very few-shot fine-tuning, can handle unstructured textual data, provide real-time feature importance for each risk assessment it provides, and offer transferability with zero to very few-shots for new risk assessment tasks. In contrast, traditional machine learning methods require large datasets for de novo training, process structured data, rely on extra computational steps for instance-specific post-hoc feature importance (e.g., SHAP), and need retraining for each new task.



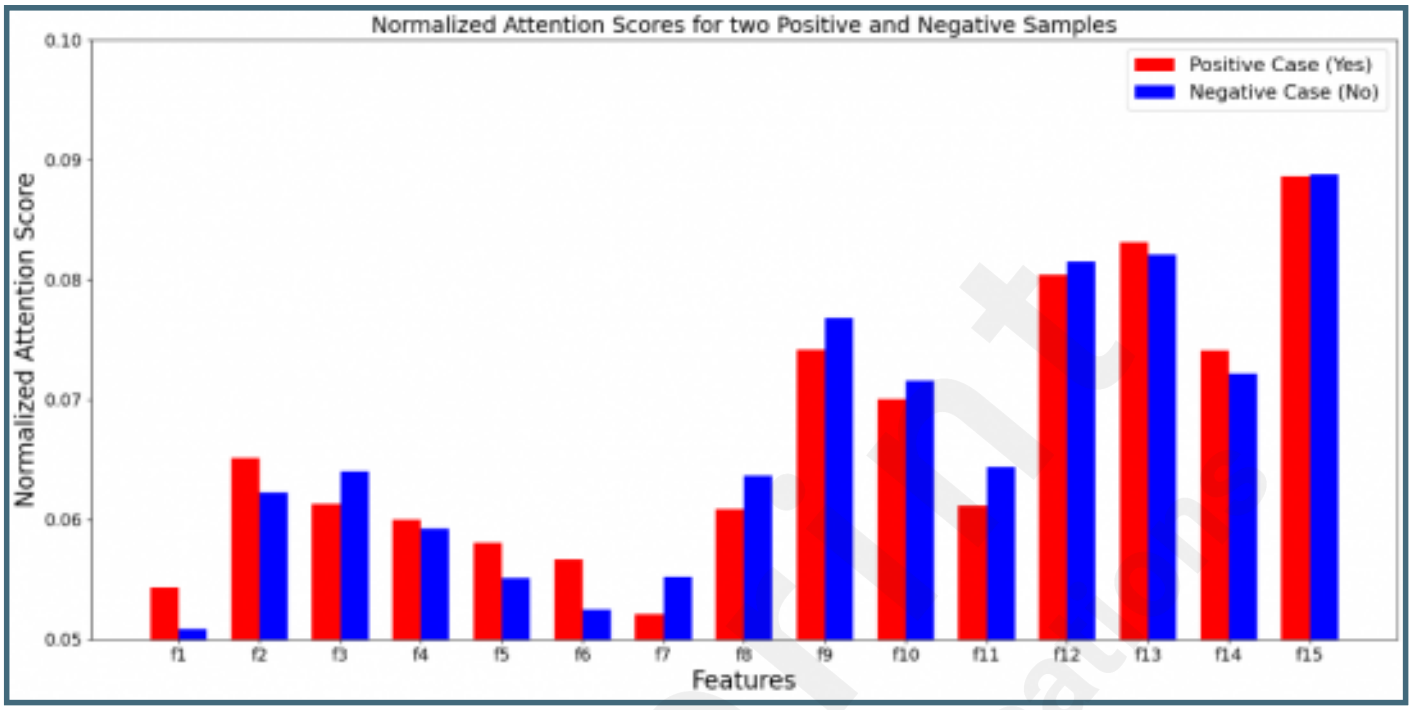
Workflow for few-shot COVID-19 severity risk assessment using generative LLMs with different serialization techniques. The top section, labeled Backend - System Developer, shows the fine-tuning phase where a few-shot sample of patient data, serialized via List and Text Templates, is used to fine-tune the LLMs. This backend process includes the creation of prompts and corresponding labels for model fine-tuning. The bottom section, labeled Frontend - User, illustrates how a conversational chatbot interacts with users through our application to gather responses via streaming QAs. These responses are analyzed by the fine-tuned LLM in real-time, providing risk assessments and highlighting the top attributing features that explain the model's risk assessment.



Overview of our mobile application design, showcasing patient data collection, real-time risk assessment using LLMs, and clinician review interface.



Normalized attention scores from LLaMA2-7b in the 32-shot setting, showing feature importance for two test cases, one positive (yes) and one negative (no), simultaneously with the risk assessment.



Average AUC in 2-shot setting over five different seeds. The left panel shows results using the List Serialization (-L) approach, while the right panel shows results using the Text Serialization (-T) approach.