When Raw Data Prevails: Are Large Language Model Embeddings Effective in Numerical Data Representation for Medical Machine Learning Applications?

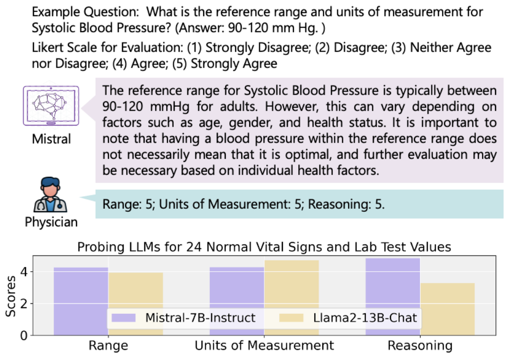
Anonymous ACL submission

Abstract

The introduction of Large Language Models (LLMs) has advanced data representation and analysis, bringing significant progress in their use for medical questions and answering. Despite these advancements, integrating tabular data, especially numerical data pivotal in clinical contexts, into LLM paradigms has not been thoroughly explored. In this study, we examine the effectiveness of vector representations from last hidden states of LLMs for medical diagnostics and prognostics using electronic health record (EHR) data. We compare the performance of these embeddings with that of raw numerical EHR data when used as feature inputs to traditional machine learning (ML) algorithms that excel at tabular data learning, such as eXtreme Gradient Boosting. We focus on instruction-tuned LLMs in a zero-shot setting to represent abnormal physiological data and evaluating their utilities as feature extractors to enhance ML classifiers for predicting diagnoses, length of stay, and mortality. Furthermore, we examine prompt engineering techniques on zero-shot and few-shot LLM embeddings to measure their impact comprehensively. Although findings suggest the raw data features still prevails in medical ML tasks, zero-shot LLM embeddings demonstrate competitive results, suggesting a promising avenue for future research in medical applications.

1 Introduction

Numerical data plays a pivotal role across various domains. For instance, much of the data used for analytics from electronic health records (EHRs) are numerical values in tabular formats, documenting patient demographics (e.g., age), vital signs, laboratory tests, and nurse assessments. Utilizing numerical data for predictive modeling has been instrumental in facilitating accurate diagnoses (Pang et al., 2021), risk stratifying (Zeiberg et al., 2019; Green et al., 2018), and outcome predictions (Akel et al., 2021; Chang et al., 2019) in healthcare. Machine learning (ML) classifiers like gradient boosted (Chen and Guestrin, 2016) have excelled in these tasks for making accurate clinical predictions (Churpek et al., 2024; Lolak et al., 2023; Moore and Bell, 2022).



Example Question: What is the reference range and units of measurement for Systolic Blood Pressure? (Answer: 90-120 mm Hg.)

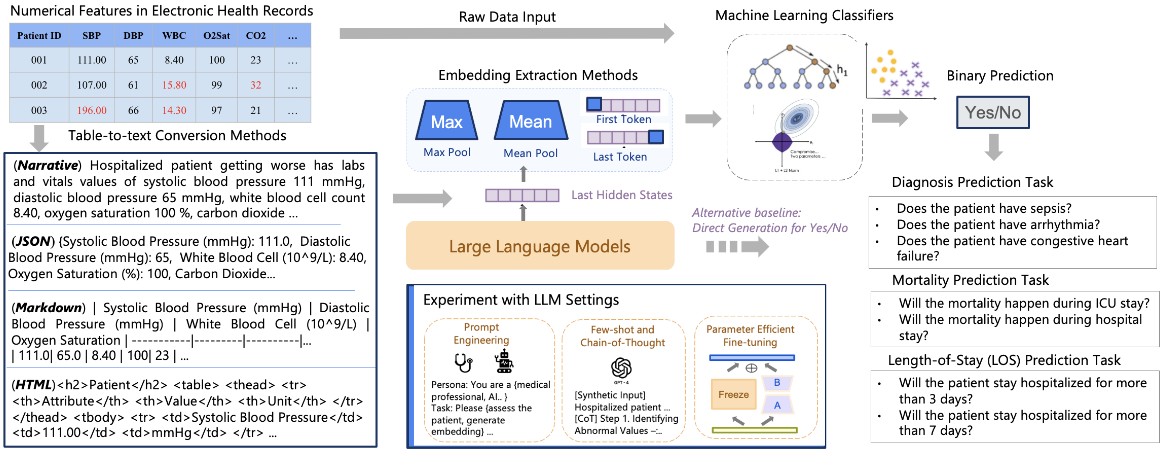
Likert Scale for Evaluation: (1) Strongly Disagree; (2) Disagree; (3) Neither Agree nor Disagree; (4) Agree; (5) Strongly Agree

Mistral: The reference range for Systolic Blood Pressure is typically between 90-120 mmHg for adults. However, this can vary depending on factors such as age, gender, and health status. It is important to note that having a blood pressure within the reference range does not necessarily mean that it is optimal, and further evaluation may be necessary based on individual health factors.

Physician: Range: 5; Units of Measurement: 5; Reasoning: 5.

Figure 1: Physician Evaluation of LLMs' Knowledge on Normal Vital Sign and Lab Test Values. This experiment probes Mistral-7B-Instruct and Llama2-13B-Chat on reference ranges for 24 vital signs and lab tests. Results show these models have a strong understanding of normal medical values, crucial for clinical applications. Table 1 listed all 24 feature names, and more output examples are in Appendix B.

Recent work shows Large Language Models (LLMs)' vast potential on text generation over structured data input, including Chain-of-Thought (CoT) reasoning over tabular data (Zheng et al., 2023), classification on diseases (Hegselmann et al., 2023). LLMs have also exhibited exceptional promise in medical NLP tasks, evident in their stellar performance in the United States Medical Licensing Examination (MedQA) (Nori et al., 2023). However, the use of embedding representations, particularly for medical diagnostics and outcome predictions using standard EHR numerical data, remains largely unexplored. In these areas, raw data inputs have traditionally dominated feature representation for ML algorithms before the era of LLMs. This is exemplified by their use in critical applications such as mortality prediction and early sepsis warnings (Deng et al., 2022; Hou et al., 2020), and patient infection (Bashiri et al.,


Numerical Features in Electronic Health Records
Table-to-text Conversion Methods
Raw Data Input
Embedding Extraction Methods
Max
Max Pool
Mean
Mean Pool
First Token
Last Token
Last Hidden States
Large Language Models
Machine Learning Classifiers
h_1
Compromise... Two parameters ...
L1 + L2 Norm
Binary Prediction
Yes/No
Alternative baseline:
Direct Generation for Yes/No
Experiment with LLM Settings
Prompt Engineering
Few-shot and Chain-of-Thought
GPT - 4
Parameter Efficient Fine-tuning
Freeze
B
A
Diagnosis Prediction Task
Mortality Prediction Task
Length-of-Stay (LOS) Prediction Task

Patient ID	SBP	DBP	WBC	O2Sat	CO2	…
001	111.00	65	8.40	100	23	…
002	107.00	61	15.80	99	32	…
003	196.00	66	14.30	97	21	…

(***Narrative***) Hospitalized patient getting worse has labs and vitals values of systolic blood pressure 111 mmHg, diastolic blood pressure 65 mmHg, white blood cell count 8.40, oxygen saturation 100 %, carbon dioxide ...

(***JSON***) {Systolic Blood Pressure (mmHg): 111.0, Diastolic Blood Pressure (mmHg): 65, White Blood Cell (10^9/L): 8.40, Oxygen Saturation (%): 100, Carbon Dioxide...

(***Markdown***) | Systolic Blood Pressure (mmHg) | Diastolic Blood Pressure (mmHg) | White Blood Cell (10^9/L) | Oxygen Saturation | -----------|---------|----------|...
| 111.0| 65.0 | 8.40 | 100| 23 | ...

(***HTML***)<h2>Patient</h2> <table> <thead> <tr> <th>Attribute</th> <th>Value</th> <th>Unit</th> </tr> </thead> <tbody> <tr> <td>Systolic Blood Pressure</td> <td>111.00</td> <td>mmHg</td> </tr> ...

Persona: You are a {medical professional, AI.. }
Task: Please {assess the patient, generate embedding} ...

[Synthetic Input]
Hospitalized patient ...
[CoT] Step 1. Identifying Abnormal Values –:...

- Does the patient have sepsis?
- Does the patient have arrhythmia?
- Does the patient have congestive heart failure?

- Will the mortality happen during ICU stay?
- Will the mortality happen during hospital stay?

- Will the patient stay hospitalized for more than 3 days?
- Will the patient stay hospitalized for more than 7 days?

Figure 2: This study investigates the feasibility of using LLM embeddings for numerical EHR data features representation in medical machine learning applications. To use LLMs, raw features are transformed into queries via templates. Under a zero-shot setting, these queries are encoded into embeddings for ML classification. We explore the effects of prompt engineering, few-shot learning using synthetic data generation, and parameter efficient tuning on LLM embeddings.

2022; Bhavani et al., 2020). The potential of LLM-derived features as a viable alternative to raw data features in ML applications is still unclear.

This study aims to address this knowledge gap by examining the use of LLM embeddings for EHR numerical data representation in ML algorithms. Although LLMs are renowned for text generation, their embeddings may offer multiple advantages, such as leveraging LLMs' pre-trained knowledge and sophisticated text understanding to enhance domain-specific tasks. Moreover, using LLMs to represent tabular data allows for a unified model that encodes both structured and unstructured text in EHRs, seamlessly integrating and contextualizing information across modalities, such as embedded tables in clinical notes (Soenksen et al., 2022; Kline et al., 2022).

Our work presents novel examination of the impact of different formats and embedding methods on LLM last layers and ML classifiers. We focus on open-source, zero-shot LLMs suitable for single-GPU systems, considering the resource limitations prevalent in many hospitals and academic research settings. To establish a foundation for this work, we probed Mistral-7B-Instruct and Llama2-13B, two open-source, general-domain LLMs, for their knowledge of reference ranges for vital signs and lab test values. We directly asked about the standard physiological values and units of measurement for 24 EHR features identified as critical predictor variables for detecting clinical deterioration (Akel et al., 2021). As in Figure 1, physician judgment indicates that LLMs possess this knowledge, providing initial evidence for further investigation.

Our study utilizes three clinical prediction tasks derived from two independent EHRs and four ML classifier input settings. We investigate the impact of table-to-text conversion formats, embedding extraction methods, prompt engineering, and few-shot techniques, along with early results from parameter-efficient fine-tuning, on the quality of LLM embeddings. Our main contributions are threefold:

- We present a comprehensive study exploring various factors that influence the performance of numerical EHR feature embeddings generated by LLMs for medical ML applications.
- Our findings show that while LLM embeddings paired with XGB classifiers can achieve performance comparable to traditional raw data features on some tasks, performance gaps persist, necessitating further improvements to maximize their effectiveness.
- We discuss the efficiency and robustness of LLM feature representation for numerical data versus raw data in training ML classifiers.

Results show that, despite external evidence indicating that LLMs possess extensive knowledge of medical facts, extracting usable representations of this knowledge for downstream tasks will require significant additional methodological progress.

2 Related Work

Recent studies highlight LLMs in tabular data analysis: Hegselmann et al. (2023) introduces TableLLM, which converts tables to text using a manual template. Zheng et al. (2023) studies CoT reasoning over tables. Akhtar et al. (2023) examines the abilities of LLMs on numerical data understanding. Zhu et al. (2024), closest to our work, explores zero-shot LLM for structured longitudinal EHR data and finds that GPT-4 can outperform XGB on clinical prediction tasks. Our study, however, uniquely focuses on open-box LLM embeddings for enhancing ML algorithms.

Raw EHR data are commonly used in medical ML applications, as found by a survey on medical ML research (Si et al., 2021). They noted that labs and vital signs as frequent data types for patient representation learning. Churpek et al. (2024) introduces an XGB algorithm predicting clinical deteriorations using EHR features like demographics and lab values. Wang et al. (2020) used 104 clinical EHR features across various ML algorithms to establish baselines for clinical tasks such as mortality predictions. Our work uses the same dataset and tasks as (Wang et al., 2020) to compare LLM embeddings against traditional ML classifier outcomes on the same raw data feature baseline.

3 Datasets and Tasks

3.1 Diagnosis prediction for clinical deterioration

Early warning systems often use rule-based and ML algorithms to identify patients at risk of deterioration or death without providing diagnoses (Churpek et al., 2014; Kipnis et al., 2016). To address this, experts from multiple hospitals created a dataset that labels the diagnoses for patients who had a clinical deterioration event during their hospitalization. These expert-annotated diagnoses were performed with a full review of the EHR and served as the labels for our training data. Twenty-four tabular data features including demographics, vital signs, labs, interventions, and nursing assessments were extracted from the structured EHR (eg. tabular data). They were previously identified as critical variables for clinical deterioration (Akel et al., 2021). The final datasets encompassed EHR data from 660 adult patients in medical-surgical ward within a U.S. health system. The primary diagnoses were Sepsis, Arrhythmia (Arrhy.), and Congestive Heart Failure (CHF) volume overload, with prevalence rates of 43.18% for Sepsis, 15.30% for Arrhythmia, and 11.82% for CHF, respectively. We used 5-fold validation on all 660 samples to generate five distinct test sets. [1]

Input features	Age, Systolic Blood Pressure, Diastolic Blood Pressure, Oxygen Saturation, Temperature in Celsius, Proton Pump Inhibitor, Alert, Voice, Pain, Unresponsive Scale (AVPU), Albumin, Alkaline Phosphatase, Anion Gap, Total Bilirubin, Blood Urea Nitrogen, Blood Urea Nitrogen to Creatinine Ratio, Calcium, Chloride, Carbon Dioxide, Creatinine , Serum Glucose, Hemoglobin, Platelet Count, Potassium, Serum Glutamic-Oxaloacetic Transaminase, Sodium, Total Protein, White Blood Cell Count
Target prediction	Sepsis, Arrhythmia, Congestive Heart Failure (CHF) Volume Overload

Table 1: Raw clinical data features from the EHR for diagnosis prediction task.

Table 1 outlines the structured input features from the cohort EHR dataset and target diagnoses utilized in our analysis. The input features comprised a comprehensive set of clinical data points including demographic information like age, vital signs such as Systolic and Diastolic Blood Pressure, and body Temperature, as well as a range of serum laboratory tests including electrolytes, liver function panel, renal function, red blood counts, etc. These inputs served as predictors and are relevant findings in making diagnoses like Sepsis, Arrhythmia, and CHF. Despite its smaller sample size, this EHR dataset includes physicians' manual chart reviews and carefully curated data, providing accurate annotations for patient diagnoses.

3.2 Mortality and length-of-stay prediction

The MIMIC-III dataset, derived from the EHR of the Critical Care Units (ICU) at Beth Israel Deaconess Medical Center, has been utilized extensively in research (Johnson et al., 2016). Wang et al. (2020) further developed an open-source pipeline for extracting, preprocessing, and representing data from the MIMIC-III database, namely MIMIC-Extract. This pipeline aggregates various data types, such as tabular demographic data available at admission, vital signs with repeated measures, laboratory test results, time-varying intervention signals, and prediction labels needed for clinical tasks. MIMIC-Extract introduces two clinical prediction tasks: mortality and length-of-stay (LOS) predictions. The mortality prediction task uses tabular data from the first 24-hour window of a patient's ICU stay to predict mortality as a binary classification task. The LOS prediction task, in contrast, determines whether a patient's stay will exceed three (LOS 3) or seven days (LOS 7) based on the same 24-hour data period. Importantly, to avoid competing risk outcomes between death and

---

[1] The dataset used in this study has been detailed in a clinical journal article currently under review, with a preprint also available. To maintain the anonymity of this paper, references to the journal preprint were omitted. Details about the demographic characteristics of the patients, including gender, age, and race, are included in Appendix.

Diagnosis dataset Hospitalized patient of age *[value]* getting worse has labs and vitals values of systolic blood pressure *[value]* mmHg, diastolic blood pressure *[value]* mmHg, oxygen saturation*[value]* %, body temperature *[value]* celsius degree, ... total protein *[value]*, white blood cell *[value]*. What are the diagnoses for this patient?

MIMIC-Extract Hospitalized patient with lab and vital signs available: in the past 24 hours, the observed alanine aminotransferas values are [*list of unique values sorted by temporal order*], albumin values are [*list of unique values sorted by temporal order*], anion gap values are [*list of unique values sorted by temporal order*]...Predict if the patient mortality will occur in-hospital.

Table 2: The template for NARRATIVE serialization method for diagnosis prediction dataset (top) and MIMIC-Extract dataset (bottom).

Dataset	Size	Average input tokens
Diagnosis	660	346.97 ±2.21
MIMIC-Extract	23,884	1829.57 ±497.02

Table 3: Dataset description

LOS, patients who died within the 3- or 7-day LOS window were excluded from the LOS prediction.

We adopted the same data partitioning used in (Wang et al., 2020), comprising 16,700, 2,394, and 4,790 patient records for the training, development, and testing sets. Each patient record includes 104 time-varying tabular data features. More detailed demographic information can be found in the MIMIC-Extract study (Wang et al., 2020). The labels in the MIMIC-Extract dataset are highly skewed, with positive label distributions of 42.82% for LOS 3, 7.66% for LOS 7, 10.27% for Mort Hosp, and 7.10% for Mort ICU.

4 Methods and Experiment Setup

Figure 2 illustrates the study overview and experiment setup. We began with a patient's tabular data input, represented using the Pandas DataFrame data structure (*raw data*). This raw data was converted to text using four distinct conversion methods, detailed in §4.1, and LLM encoded the converted text, with the last hidden states extracted to generate embedding features (§4.2). These embeddings were subsequently used to train various ML classifiers on two datasets for binary prediction tasks.

We started with zero-shot, off-the-shelf LLMs for experiments (§4.3). We then investigated the impact of prompt-engineering techniques and few-shot learning configurations on the embeddings and subsequent predictions (§4.4). An initial investigation was also conducted to assess the effects of parameter-efficient fine-tuning on LLM embeddings for ML tasks, focusing on two of the models (§4.5).

As baselines, we included traditional ML classifiers trained directly on raw tabular data inputs. To benchmark the effectiveness of LLM embeddings, we used randomly initialized embeddings of the same size as the LLM-generated embeddings.

4.1 Table-to-text conversion

We employed four different methods to convert EHR tables into input formats for LLMs: NARRATIVES, JSON, HTML, and MARKDOWN. NARRATIVES provide a continuous text description of patient data, offering context and readability similar to clinical notes (Yu et al., 2023). JSON structures the data hierarchically, making it easy to parse and interpret programmatically (Zhao et al., 2023). HTML format leverages web-based structures to present the data with tags (García-Ferrero et al., 2024). MARKDOWN offers a lightweight markup language that provides formatting while remaining readable in plain text (Zhao et al., 2023).

Table 2 includes two NARRATIVES templates used to format these varied clinical measurements into a standardized query. These templates detail the format in which data from the EHR dataset are presented, integrating both laboratory results and vital signs into a single descriptive snapshot of a patient's current state. Each placeholder in the template is populated with actual data points from patient records, facilitating the transformation of tabular EHR data into a format suitable for LLM input, from which we then generate embeddings.

The primary distinction between the templates for the diagnosis prediction dataset and the MIMIC-Extract dataset lies in the types of values incorporated. For diagnosis prediction, data are values collected immediately before the early warning system triggers for clinical deterioration. In contrast, MIMIC-Extract tasks include laboratory and vital signs data from the 24 hours prior to the event. We extracted all unique values observed during the first 24 hours of ICU admission in chronological order, compiling these into a list format. If a feature has no observations, it is omitted, resulting in variable length sequences.

4.2 Embedding extraction methods

This section introduces the methods used to convert input text to fixed-size vector for ML input. We focused on the *last hidden states* of LLMs (as in (Lu et al., 2021)), and employed three different embedding extraction methods: **Max Pooling** captures the most salient features by taking the maximum value across all token embeddings for each dimension (Bao et al., 2023); **Mean Pooling** computes the average value of the token embeddings, providing a balanced representation reflecting the overall content (Ram et al., 2023); **Last Token** uses the embedding of the last token as the representation, capturing the concluding context or final summary (Shani et al., 2023; Fu et al., 2023). We included embeddings extracted from **first token** as a reference point despite it is not ideal due to the nature of decoder-only models.

Our choice of ML classifiers comprised two tree-based methods and a linear model to provide a comprehensive assessment of various predictive approaches. Specifically, we utilized eXtreme Gradient Boosting (XGB)(Chen and Guestrin, 2016) and Random Forest (RF)(Breiman, 2001) as our tree-based classifiers due to their robustness and efficiency in handling diverse datasets with accuracy. Additionally, Logistic Regression with regularization (LR) as our linear model was chosen for its effectiveness in preventing overfitting via Ridge and Least Absolute Shrinkage and Selection Operator regularization(Zou and Hastie, 2005). Together, these classifiers form a balanced baseline setup that caters to both non-linear and linear decision boundaries in our data.

4.3 Selection of LLMs

We assessed a mix of general-domain models and models trained on medical text. Three widely-used, general-domain LLMs that have been instruction-finetuned are Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), Llama2-13B-chat-hf, Llama2-70B-chat-hf (Touvron et al., 2023), and Llama3-8B-instruct (LLaMa, 2024). These models are compatible with one Nvidia 80GB A100 GPU, making them popular choices among available LLMs. For the domain specific LLM, We selected Meditron-7B (Chen et al., 2023), a Llama2-7B based model continuously pretrained on medical text. We also included ClinicalBERT (Alsentzer et al., 2019), pre-trained on MIMIC EHR text, representing encoders pre-trained on clinical text baseline compared to decoder-only LLMs.

4.4 Prompt design and few-shot learning

Because the majority of LLMs we tested are instruction-tuned and require varying input formats, we utilized the chat templates to ensure proper integration of input data (Zheng, 2024). In our study, the default setting involves including only the task-relevant question (shown on the right side of Fig 2) in the system message and the converted EHR data in the user input, without additional system instructions, predefined personas, or other context. Given that instruction-tuned LLMs are known to be sensitive to system instructions, we designed four system instructions that vary by persona (medical professional, AI system), tasks (assess patients, generate embeddings for ML classifiers), thinking style (chain-of-thoughts), and question type (general assessment, binary question), enabling us to explore the influence of prompt characteristics on the embeddings. All prompts were paraphrased for better perplexity scores, following prompt optimization strategies (Gonen et al., 2023; Lu et al., 2023).

Two few-shot settings were explored besides zero-shot prompt engineering. We generated synthetic data for diagnosis prediction, by prompting GPT-4 to generate values based on the attribute names in Table 1. For each target diagnosis, GPT-4 generated one example confirming the diagnosis (positive) and one example negating it (negative). Moreover, GPT-4 was asked to generate CoT explanations identifying abnormal values and their clinical significance. An expert physician and clinical informaticist reviewed these synthetic data pairs for quality assurance. The complete set of prompts are presented in Table 8.

4.5 Parameter efficient fine-tuning

While our paper primarily focuses on evaluating zero-shot LLMs for numerical feature representation, we included parameter-efficient fine-tuning experiment to suggest future directions for improvement. We employed QLoRA (Dettmers et al., 2024) on Mistral-7B-Instruct and Llama3-8B-Instruct, using the MIMIC-Extract dataset due to its larger training set compared to the diagnosis dataset. We trained Mistral with a sequence classification head on top, saving checkpoints with the lowest validation loss. Based on validation performance, we optimized the (q, k, v, o) layers with $r = 16$, a learning rate of 3e-5, and a LoRA dropout of 0.1. Each model was trained for 3 epochs with early stopping to prevent overfitting.

4.6 Experiment setup

We used a 5-fold cross-validation on the diagnosis dataset (660 patient records), resulting in 528 patients for training and 132 for testing per fold. For mortality and LOS prediction tasks from MIMIC-Extract data, we followed the data split from (Wang et al., 2020). We evaluated performance using Area

Model	**Sepsis AUROC** (95% CI)	**Arrhythmia AUROC** (95% CI)	**CHF AUROC** (95% CI)	Average (95% CI)
Raw Data Features Baseline				
LogisticRegression	71.10 (67.01, 75.18)	74.40 (69.35, 79.56)	54.79 (47.74, 61.79)	66.76 (61.37, 72.18)
RandomForest	65.26 (61.79, 68.48)	53.07 (50.58, 55.80)	50.89 (49.01, 53.43)	56.41 (53.79, 59.24)
XGB	**71.17 (67.06, 75.11)**	**76.49 (71.32, 84.13)**	58.47 (51.36, 65.15)	**68.71 (63.25, 74.80)**
LLM embedding + XGB classifier				
Random	54.01 (49.89,58.44)	49.65(44.02,54.62)	50.02 (47.13, 52.29)	51.22 (47.01, 55.19)
Mistral-7b-Instruct$_{best}$	71.12 (67.54, 74.92)	68.00 (61.52, 73.93)	51.80 (44.48, 58.65)	63.40 (57.73, 68.77)
Llama3-8b-Instruct$_{best}$	63.84 (57.31, 69.87)	71.08 (65.69, 75.87)	**63.84 (56.77, 70.37)**	66.25 (60.15,72.35)
Llama2-13b$_{best}$	66.02 (61.64, 70.32)	58.62 (52.62, 64.46)	49.69 (48.83, 62.58)	58.11 (54.36, 65.79)
Llama2-70b-chat$_{best}$	68.57 (63.88, 71.53)	69.15 (67.08, 71.17)	53.87 (49.83, 58.52)	63.86 (60.93, 67.07)
Meditron$_{best}$	66.74 (62.30, 66.15)	72.26 (65.28, 77.43)	58.11 (50.64, 64.48)	63.90 (58.28, 65.45)
ClinicalBERT	58.80 (54.44, 63.04)	64.91 (61.84, 70.27)	49.67 (41.94, 57.51)	57.79 (52.74, 63.11)
LLM embedding + Logistic Regression classifier				
Random	49.58 (47.68, 51.12)	49.22 (48.09, 50.43)	49.36 (47.12 51.06)	49.39 (47.63, 50.87)
Mistral-7b-Instruct$_{best}$	62.61 (58.17, 66.95)	69.59 (64.67, 74.71)	48.98 (42.96,55.62)	60.39 (55.27, 65.76)
Llama3-8b-Instruct$_{best}$	66.54 (62.32, 70.62)	70.22 (64.82, 74.11)	63.52 (55.91,69.20)	66.76 (61.50, 72.02)
Llama2-13b-chat-hf$_{best}$	66.95 (62.82, 70.88)	66.04 (60.04, 71.22)	58.54 (52.09, 65.09)	63.84 (58.32, 69.06)
Llama2-70b-chat-hf$_{best}$	69.50 (65.37, 73.43)	68.11 (61.75, 70.57)	62.72 (56.47, 68.39)	66.78 (61.20, 70.80)
Meditron$_{best}$	66.91 (62.83, 71.09)	68.61 (63.49, 73.72)	57.60 (51.02, 63.89)	64.37 (59.11, 69.90)
ClinicalBERT	47.28 (43.07, 51.63)	44.62 (38.79, 50.29)	46.98 (42.96, 55.62)	46.29 (41.61, 52.51)

Table 4: Comparing raw data features and LLM embeddings features for ML classifiers on Diagnosis dataset. We report the best AUROC scores from LLM embedding across various embedding extraction and table-to-text conversion methods. The "Random" row indicates the randomly initialized embedding input. For ClinialBERT, we used [CLS] token embedding as the final representation. We use green color to highlight the LLM+ML results where it has CI overlapping with the best results (in bold fonted text).

Under the Receiver Operating Characteristic (AUROC) with 95% confidence intervals (CI).

For all ML classifiers, we determined the best parameters through grid search on the validation set. Specifically, we tuned the number of estimators, maximum depth, learning rate, and minimum child weight for XGB classifiers, and alpha and L1 ratio for LR classifiers (see Appendix E). For LLMs under 13B, the maximum input length was 1042 for the diagnosis dataset and 3076 for the MIMIC dataset, resulting in a 4096-dimensional embedding. For 70B LLM, the max input length was 1500 and 4-bit quantization was set to avoid GPU memory errors, producing an 8192-dimensional embedding. All experiments ran on an Ubuntu server with an Nvidia 80GB A100 GPU.

5 Results

5.1 Main results for diagnosis prediction

Table 4 presents AUROC scores for predicting Sepsis, Arrhythmia, and CHF with different ML models, demonstrating the effects of using LLM embeddings compared to raw data features. XGB with raw data features stood out in the baseline, demonstrating the highest AUROC for Sepsis and Arrhythmia, and the highest average AUROC across all diagnoses. LR and RF, while using raw data input, showed moderately lower effectiveness.

For LLM embeddings with zero-shot setting, we observed performance gain over a randomly initialized embedding approach into XGB with substantial gains in all decoder LLMs. ClinicalBERT was the only model, as an older pre-trained encoder, that did not show notable performance gains over the randomly initialized embedding model. Further, Mistral embedding with XGB classifiers achieved a competitive AUROC of 71.12 (vs. 71.16 of raw data with XGB). Llama2-13B scored an AUROC of 58.54 on CHF prediction, the best among all models. This demonstrates that LLM embeddings can match or nearly match the performance of models trained with raw data inputs.

5.2 Main results for mortality prediction and length-of-stay

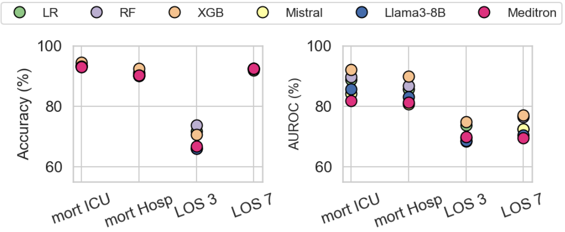


Figure 3: Accuracy (left) and AUROC (right) for in-ICU mortality (mort ICU), in-Hospital morality (mort Hosp), hospital LOS exceeding 3 days (LOS 3) and 7 days (LOS 7). The Logistic Regression (LR) and Random Forest (RF) baselines are reported from (Wang et al., 2020). The LLM results are from LLM embeddings + XGB settings. The CIs mostly overlap; for clarity in presentation, they were omitted from this figure.

Figure 3 displays performance for various models on tasks of in-ICU mortality (mort ICU), in-hospital mortality (mort Hosp), and hospital length-of-stay for more than 3 (LOS 3) and 7 days (LOS 7). The raw data features with XGB model consistently outperforms others with an AUROC of 92.02 in mort ICU and 89.83 in mort Hosp. LLM embeddings from Mistral, Llama3-8b, and Meditron, while slightly lagging behind the raw data features with ML classifiers in the mortality tasks, performed comparably in the LOS 7 tasks. Mistral with XGB achieved accuracy of 92.34 and AUROC of 72.36 on LOS 7 task, showing competitive performance to XGB with raw data features, with accuracy of 92.32 and an AUROC of 76.93. The gap between LLM embeddings and raw data for mortality and LOS tasks suggests a need to improve time-varying feature representation.

5.3 Comparisons across different embedding methods and data conversion methods

Figure 4 presents AUROC values for different embedding methods and data conversion formats across three models: Mistral, Meditron, and Llama3-8b. Max pooling achieves the highest performance for Mistral (64.62) and Meditron (62.54), while mean pooling is most effective for Llama3-8b (64.69). The last token method yields moderate performance across all models, with AUROCs around 57, while first token embeddings result in the lowest AUROC values, indicating a less effective representation for these models.

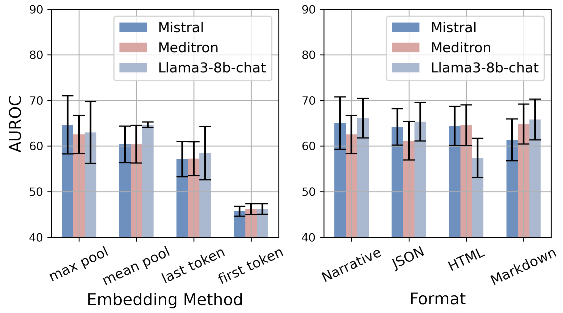


Figure 4: Comparison across different embedding methods and different format on the Diagnosis dataset. For simplicity, we used NARRATIVE and max pooling for the other analysis after this section.

When encoding data with different formats, Mistral shows preference for NARRATIVES, JSON, and HTML. The MARKDOWN format generally yielded the lowest performance across the models, particularly for Mistral. JSON and HTML formats showed competitive performance, with JSON being slightly more effective for Meditron and Llama3-8b. Notably, Llama3-8b exhibited the highest variability across formats, with AUROCs ranging from 57.40 (HTML) to 66.13 (NARRATIVES).

5.4 Impact of prompt engineering and few-shot learning

We compared performance of Mistral and Llama3 using different system instructions under zero-shot and few-shot settings, as well as CoT examples. Mistral, under 0-shot with a system instruction with persona of medical professional and the task of assessing patient condition (prompt 1 in Table 8), achieved an AUROC of 71.35 on Sepsis prediction, the highest of all models. Llama3 with zero-shot prompting using prompt 1 in Table 8 showed reported AUROC of 73.51 on Arrhythmia, surpassing its counterpart at 71.08 but still below raw data XGB baseline (76.49). CoT and few-shot exhibited various performance and often resulted in lower AUROC scores compared to Table 4. Full results are provided in Appendix C.

5.5 Parameter efficient fine-tuning results

Setting	LOS 3	LOS 7	Mort ICU	Mort Hosp
Mistral	67.84	72.36	84.16	80.71
Mistral_{QLoRA}	65.26	67.66	75.69	73.66
Performance Drop	2.58	4.70	8.47	7.05
Llama3-8b	68.54	70.38	85.61	83.06
Llama3-8b_{QLoRA}	66.69	68.56	75.14	71.15
Performance Drop	1.85	1.82	10.47	11.91

Table 5: AUROC comparison before and after training LLM with QLoRA on MIMIC tasks.

Table 5 presents results of Mistral and Llama3-8b under the QLoRA across all four tasks from MIMIC-Extract. The performance drops are noticeable, especially in the two mortality predictions. To further understand the reason behind the performance drops, we plotted the confusion matrices for LOS 3 and Mort ICU, comparing Mistral's predictions before and after QLoRA in Figure 5. For LOS 3 prediction, the Mistral model with QLoRA shows an increase in true negatives and a decrease in false positives. However, the false negatives rises from 1133 to 1473, and true positive drops from 918 to 578. On the Mort ICU task, the Mistral model with QLoRA correctly predicts no false positives, but it fails to predict any positive cases (0 true positives). The performance drop can be attributed to the imbalanced class distribution in the dataset, as the models show a tendency to favor the majority class (negative cases). During QLoRA, the LLM might learn the class prevalence, biasing its representation and making it challenging to correctly

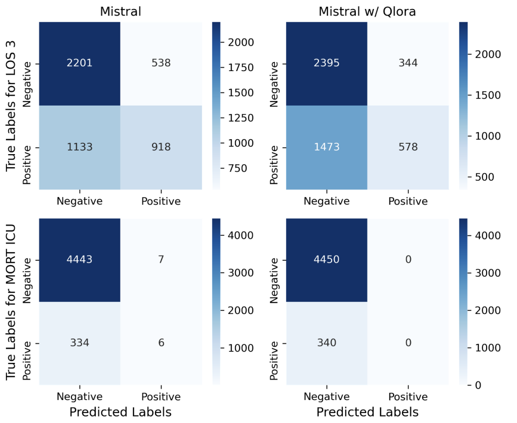


Figure 5: Confusion matrices for Mistral prediction on LOS 3 and Mort ICU tasks. Right: Mistral without QLoRA; left: Mistral after QLoRA.

identify the minority class.

5.6 LLM Embedding vs LLM Generation

The final experiments compare the performance of LLM embeddings combined with ML classifiers against direct outputs from LLMs. This comparison shows that, although LLM embeddings generally do not outperform raw data features, they offer a more robust and reliable solution than relying on LLMs to directly answer Yes or No questions. Our exploration revealed significant limitations in LLM generation for binary prediction tasks. For instance, Mistral frequently predicted 'Yes' for sepsis, arrhythmia, and CHF AORC, resulting in AUROC scores being 50, whereas LLM embeddings achieved AUROCs of 71.12 for sepsis, 72.26 for arrhythmia, and 63.54 for CHF AUROC. Similar patterns were observed from Llama3-8b results (Table 12). On MIMIC-Extract tasks with highly skewed class distributions, Mistral and Llama3-8b, when generating direct Yes/No answers, again showed reduced ability to discriminate between positive and negative cases (Table 13). These findings underscore the need for embeddings, which provide a more nuanced and effective approach for clinical predictions. We refer readers to Appendix D for more details.

6 Discussion

To understand the discrepancy between the two data representations, we examined the training effectiveness of raw data features and LLM embeddings by controlling the training set size. Figure 6 compares the performance of the raw data XGB baseline model with the Mistral and Meditron embeddings across different training set sizes for two tasks in the MIMIC dataset. The raw data XGB baseline model shows a significant increase in AUROC scores with larger training sets, achieving high performance. In contrast, both the Mistral and Meditron embeddings paired with XGB models exhibit much smaller improvements, consistently performing lower than the raw data XGB baseline. This highlights the greater effectiveness of XGB when learning from raw data features compared to LLM embeddings for these prediction tasks.

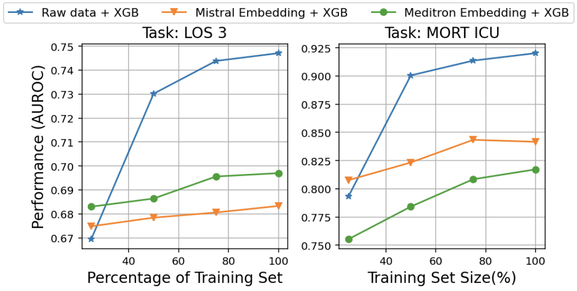


Figure 6: AUROC comparison between Raw data, embeddings from Mistral and Meditron with XGB classifiers, by controlling the training set size on two MIMIC tasks.

Our findings suggest that raw data features provide more informative input for ML models compared to LLM-generated embeddings. While LLM embeddings capture complex representations, they may not be as tailored for binary medical prediction tasks. Additionally, computing efficiency is an important consideration, as LLMs require significantly more GPU memory than raw data features.

However, zero-shot LLM embeddings achieve comparable performance in certain scenarios, highlighting their potential for rapid deployment without extensive training. A promising direction is distilling these embeddings into a smaller space while retaining their extensive knowledge (Lee et al., 2024). BehnamGhader et al. (2024) recently proposes LLM2Vec, a method to train decoder-only LLMs as text encoders with unsupervised training, which merits further investigation.

7 Conclusion

We present the first analysis of LLM embeddings for numerical EHR data features in medical ML applications, showing the opportunity and challenges of using LLM embeddings as a substitute of raw data features. We hope to encourage future research on improving LLM embeddings, particularly for imbalanced label prediction, and advancing health predictions with multi-modal data, while addressing interpretability and bias.

8 Limitation

In our study, we focused on investigating some of the most common LLMs, including Meditron, Mistral, Llama2, and Llama3. Due to GPU constraints, some experiments, such as Qlora, were conducted on only one or two models, limiting the comprehensiveness of our analysis. We did not include black-box LLMs via API because, despite using fully de-identified data, both EHR datasets are protected under Data Use Agreement, restricting us sharing with third parties. Additionally, we acknowledge that we did not explore all possible methods of prompting LLMs, which may have influenced our results. Furthermore, our examination was restricted to the last layers of the LLMs, potentially overlooking valuable information encoded in other layers.

Regardless of these limitations, our findings are consistent across models: zero-shot LLM embeddings paired with machine learning classifiers generally underperform compared to raw data features, though they sometimes achieve comparable performance.

9 Ethical Statement

Following the ACL's ethical review guidelines, our study on leveraging LLMs for medical diagnosis within EHR emphasizes ethical integrity by prioritizing harm avoidance, privacy protection, fairness, transparency, and respect for intellectual property. While our research aims to advance medical diagnostics through LLMs, there is a potential risk that misinterpretations of model predictions could inadvertently lead to diagnostic errors or bias in clinical decision-making. Therefore, rigorous validation protocols, including expert medical review and bias detection mechanisms are needed to ensure that model predictions are both accurate and equitable across diverse patient populations.

We have rigorously ensured data de-identification, obtained ethical approvals, actively mitigated biases, and maintained openness in our methodologies and findings to uphold honesty and reproducibility. Our commitment extends to respecting intellectual property through proper attribution and license adherence, with the overarching goal of contributing positively to healthcare outcomes and societal well-being. This approach underscores the importance of robust, secure research practices in developing computational tools for healthcare, aligning with our ethical responsibility to advance the field for the public good.

References

MA Akel, KA Carey, CJ Winslow, MM Churpek, and DP Edelson. 2021. Less is more: Detecting clinical deterioration in the hospital with machine learning using only age, heart rate, and respiratory rate. *Resuscitation*, 168:6–10.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Forrest Bao, Ruixuan Tu, Ge Luo, Yinfei Yang, Hebi Li, Minghui Qiu, Youbiao He, and Cen Chen. 2023. Docasref: An empirical study on repurposing reference-based summary quality metrics as reference-free metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1226–1235.

Fereshteh S Bashiri, John R Caskey, Anoop Mayampurath, Nicole Dussault, Jay Dumanian, Sivasubramanium V Bhavani, Kyle A Carey, Emily R Gilbert, Christopher J Winslow, Nirav S Shah, et al. 2022. Identifying infected patients using semi-supervised and transfer learning. *Journal of the American Medical Informatics Association*, 29(10):1696–1704.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Sivasubramanium V Bhavani, Zachary Lonjers, Kyle A Carey, Majid Afshar, Emily R Gilbert, Nirav S Shah, Elbert S Huang, and Matthew M Churpek. 2020. The development and validation of a machine learning model to predict bacteremia and fungemia in hospitalized patients using electronic health record data. *Critical care medicine*, 48(11):e1020–e1028.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Wenbing Chang, Yinglai Liu, Yiyong Xiao, Xinglong Yuan, Xingxing Xu, Siyue Zhang, and Shenghan Zhou. 2019. A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, 9(4):178.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

M. M. Churpek, K. A. Carey, A. Snyder, C. J. Winslow, E. Gilbert, N. S. Shah, B. W. Patterson, M. Afshar, A. Weiss, D. N. Amin, D. J. Rhodes, and D. P. Edelson. 2024. Multicenter development and prospective validation of ecartv5: A gradient boosted machine learning early warning score. *medRxiv*.

Matthew M Churpek, Trevor C Yuen, Christopher Winslow, Ari A Robicsek, David O Meltzer, Robert D Gibbons, and Dana P Edelson. 2014. Multicenter development and validation of a risk stratification tool for ward patients. *American journal of respiratory and critical care medicine*, 190(6):649–655.

Hong-Fei Deng, Ming-Wei Sun, Yu Wang, Jun Zeng, Ting Yuan, Ting Li, Di-Huan Li, Wei Chen, Ping Zhou, Qi Wang, et al. 2022. Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. *Iscience*, 25(1).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Harvey Fu, Qinyuan Ye, Albert Xu, Xiang Ren, and Robin Jia. 2023. Estimating large language model capabilities without labeled test data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9530–9546.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, et al. 2024. Medmt5: An open-source multilingual text-to-text llm for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Malcolm Green, Harvey Lander, Ashley Snyder, Paul Hudson, Matthew Churpek, and Dana Edelson. 2018. Comparison of the between the flags calling criteria to the mews, news and the electronic cardiac arrest risk triage (ecart) score for the identification of deteriorating ward patients. *Resuscitation*, 123:86–91.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. 2020. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of translational medicine*, 18:1–14.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Patricia Kipnis, Benjamin J Turk, David A Wulf, Juan Carlos LaGuardia, Vincent Liu, Matthew M Churpek, Santiago Romero-Brufau, and Gabriel J Escobar. 2016. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the icu. *Journal of biomedical informatics*, 64:10–19.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Meta LLaMa. 2024. Llama3. `https://github.com/meta-llama/llama3`. Accessed: 2024-06-08.

Sermkiat Lolak, John Attia, Gareth J McKay, and Ammarin Thakkinstian. 2023. Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: Retrospective cohort study. *JMIR cardio*, 7:e47736.

Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. 2023. What makes pre-trained language models better zero-shot learners? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2303.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3855–3865.

Alexander Moore and Max Bell. 2022. Xgboost, a novel explainable ai technique, in the prediction of myocardial infarction: A uk biobank cohort study. *Clinical Medicine Insights: Cardiology*, 16:11795468221133611.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. 2021. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What are you token about? dense retrieval as distributions over the vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2481–2498.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. 2023. Towards concept-aware large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. 2021. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics*, 115:103671.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. 2022. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Unified language representation for question answering over text, tables, and images. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4756–4765.

Daniel Zeiberg, Tejas Prahlad, Brahmajee K Nallamothu, Theodore J Iwashyna, Jenna Wiens, and Michael W Sjoding. 2019. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PloS one*, 14(3):e0214465.

Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. 2023. Large language models are complex table parsers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14786–14802.

Chujie Zheng. 2024. Chat templates for huggingface large language models. `https://github.com/chujiezheng/chat_templates`.

Mingyu Zheng, Hao Yang, Wenbin Jiang, Zheng Lin, Yajuan Lyu, Qiaoqiao She, and Weiping Wang. 2023. Chain-of-thought reasoning in tabular language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11006–11019, Singapore. Association for Computational Linguistics.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

A Diagnosis Prediction Dataset Patient Demographics

Group	Distribution
Total	660 Patients
Gender	Male (52%), Female (48%)
Ages	Adults (36%), Geriatric (64%)
Race	White\Caucasian (89%), Black\African American (6%), Asian\Mideast Indian (2%), American Indian\Alaska Native (1%), Pacific Islander\Hawaiian Native (<1%) , Declined\Unknown (<1%)

Table 6: Patient demographic description of diagnosis prediction dataset. Note that in this work, we exclude the demographic information from ML input.

B Probing LLMs for Inherent Knowledge of Normal Ranges

A foundational question for using LLM embeddings for numerical data representation is whether they possess inherent knowledge about the normal range of values for clinical data. To assess this, we first asked the LLMs about standard physiological ranges, measurement units, and reasoning from the feature set of tabular data outlined in Table 1. A board-certified physician assessed the LLM generations using a 1 to 5 Likert scale across three dimensions: correctness of the range, accuracy of measurement units, and quality of explanations ("Reasoning"). Our probing experiments were conducted on Mistral and Llama2-13b, as these are general-domain LLMs that have been fine-tuned to follow instructions. This experiment was done prior to the release of Llama3 models, therefore we did not include them. We did not employ automated metrics because there is no single set of reference ranges for some features, as they are highly dependent on patients' ages, sex, and other demographic factors.

In our exploration of language model capabilities within the healthcare domain, we specifically probe the model's understanding of standard ranges and measurement units for EHR data features. The following table illustrates Mistral's performance on tasks designed to evaluate its ability to accurately identify and articulate the normal physiological ranges and corresponding units for 24 distinct EHR features. This assessment not only tests the model's grasp of medical terminology and concepts but also its potential application in clinical settings where precise knowledge of such parameters is crucial.

Question: What is the reference range for the Pulse Pressure Index?
Reference: 0.1, 1
LLM (Mistral): The reference range for Pulse Pressure Index (PPI) is not a commonly used vital sign or lab result label. It measures the ratio of the pulse pressure to the systolic blood pressure, used to assess cardiovascular risk. The PPI is typically reported as a percentage, with a normal range being less than 10%. However, the specific reference range may vary by population and clinical context. It is important to consult healthcare professionals for personalized PPI interpretations.
Physician Evaluation (Range, Unit, Reasoning): 3, 5, 5

Question: What is the reference range for Serum Carbon Dioxide Level?
Reference: 22, 29 mmol/L
LLM (Mistral): The reference range for serum carbon dioxide (CO2) level is typically between 21 to 29 millimoles per liter (mmol/L) for serum measurements, or 2.5 to 3.5 mmol/L for arterial blood gas (ABG) measurements. Note that CO2 level units are mmol/L or ppm, and the range may vary by laboratory or tested population.
Physician Evaluation (Range, Unit, Reasoning): 5, 4, 2

Table 7: Mistral's probing output examples for Normal Ranges and Units of EHR data features

C Prompt Design

To test the impact of prompt engineering on LLM embeddings, we designed four distinct prompts, varying by persona, thinking style, and question types. Our prompt engineering strategy builds on prior work that utilizes perplexity to select optimal prompts (Gonen et al., 2023). Initially, we crafted meta-language prompts delineating clinical scenarios and tasks. We then employed GPT-4, a state-of-the-art LLM, to generate five paraphrases for each prompt to capture a range of expressions. We evaluated each paraphrase's naturalness and fluency by calculating its perplexity score, selecting the most coherent paraphrase for each scenario for our experiments. The optimized prompts are detailed in Table 8.

We also used these prompts in few-shot experiments. Given the complexity of the MIMIC-Extract dataset, which includes 104 features with data collected over 24 hours, creating high-quality synthetic data is challenging. Thus, we focused our efforts on the diagnosis prediction use case, which involves simpler inputs of twenty-four clinical features, facilitating quality control over the synthetic data generation.

As illustrated in Figure 2, we used GPT-4 to create synthetic data depicting patient cases of clinical deterioration with features in Table 1. This data set includes both positive and negative diagnosis cases, which were reviewed by an expert physician and clinical informaticist for quality assurance. Our few-shot experiments varied in complexity: the first modified Prompt 1 to include example input-output pairs ("Simple" few-shot setting), while the second added a CoT explanation detailing the diagnostic reasoning into Prompt 3. The CoT was structured to identify and reason over abnormal values to conclude diagnoses, enhancing the data's interpretability and educational value.

D Results of LLM Direct Generation

We tested the ability of Mistral and Llama3 to directly predict Yes or No answers to questions from the Diagnosis and MIMIC-Extract datasets. To achieve this, we added specific instructions directing the LLMs to respond only with "Yes" or "No," then parsed the outputs to 1 or 0 labels and computed AUROC and Accuracy. For this experiment, we set the maximum token limit to 25 and the top

Prompt Description
Prompt 1 -**Persona: Medical Professional** As a healthcare provider, please assess the patient's condition provided below and outline the likely causes or diagnoses for their clinical worsening. List only the diagnoses and keep your response brief.
Prompt 2 -**Persona: AI System** You are an AI with medical expertise. Create an embedding for the probable problems or diagnoses that are causing clinical deterioration, based on the patient's condition detailed below, to aid in training a diagnostic prediction machine learning classifier. Be brief in your description.
Prompt 3 -**Persona: Medical Professional (Chain-of-Thought)** As a medical expert, please examine the patient's condition by first identifying any abnormal values. Next, critically analyze these values to assess their impact, and clearly state your final diagnosis regarding what might be causing the clinical deterioration. Keep your summary brief.
Prompt 4 -**Persona: Medical Professional (Binary Question)** You are a medical doctor. Based on the patient's condition, determine the likelihood that diagnosis X is causing their clinical deterioration. Be aware that diagnosis X occurs in Y% of similar cases.

Table 8: System prompts for medical diagnosis assistance with different persona settings.

Model	Setting	AUROC CI (%)
Mistral-7b-instruct	sys1	54.85 [48.18, 62.11]
	sys2	53.88 [47.37, 60.67]
	sys3	51.16 [44.34, 57.67]
	sys4	54.04 [46.84, 61.56]
	Fewshot	54.43 [46.62, 61.05]
	CoT	57.96 [60.72, 69.24]
Llama2-13b-chat	sys1	56.49 [49.90, 63.09]
	sys2	55.61 [48.43, 62.31]
	sys3	50.41 [43.19, 57.33]
	sys4	60.24 [53.28, 67.09]
	Fewshot	53.12 [46.38, 59.84]
	CoT	54.10 [51.84, 60.59]
Llama3-8b-instruct	sys1	52.81 [46.47, 59.21]
	sys2	51.11 [44.64, 57.27]
	sys3	49.03 [42.19, 55.74]
	sys4	55.23 [48.28, 61.79]
	Fewshot	53.24 [46.60, 59.99]
	CoT	51.44 [48.58, 53.77]

Table 9: AUROCs for various models and settings on CHF Volume Overload prediction.

Model	Setting	AUROC CI (%)
Mistral-7b-instruct	sys1	62.27 [56.47, 67.67]
	sys2	63.84 [58.06, 69.31]
	sys3	64.92 [58.98, 70.05]
	sys4	66.11 [60.20, 71.52]
	Fewshot	68.43 [62.82, 73.98]
Llama2-13b-chat	sys1	69.24 [63.52, 74.84]
	sys2	61.90 [56.04, 67.63]
	sys3	61.44 [56.35, 66.68]
	sys4	64.43 [58.63, 69.95]
	sys5	67.74 [62.35, 73.22]
Llama3-8b-instruct	sys1	71.12 [65.91, 76.05]
	sys2	72.13 [66.12, 77.88]
	sys3	70.24 [64.94, 75.58]
	sys4	**73.51 [68.09, 78.54]**
	sys5	73.10 [67.29, 78.18]

Table 10: One Time AUROC and Confidence Intervals for various models and settings on Arrhythmia prediction. Scores are multiplied by 100.

k to 50.

Table 12 presents results of Mistral directly generating "Yes/No" answers for the Diagnosis dataset. For all tasks (Sepsis, Arrhythmia, CHF), Mistral achieved an AUROC of 50.00, indicating no discriminatory ability. Accuracy varied across tasks, with Sepsis at 43.18%, Arrhythmia at 15.30%, and CHF at 11.82%, corresponding to the positive class distribution, demonstrating poor performance in direct prediction. Llama3 exihibited similar performance: it reported AUROC scores between 47.12 (Arrhythmia) to 50.28 (Sepsis), underperforming its embedding counterparts reported in Table 4. Admittedly, extra effort in prompt engineering and parameter searching could improve direct generation results. However, compared to their embedding + ML classifier counterparts in the same zero-shot setting and input format (NARRATIVES, without additional system instructions such as personas), their performance is significantly lower.

On the MIMIC-Extract tasks, table 13 shows the results of Mistral-7b-Instruct and Llama3-8b-Instruct in directly generating "Yes/No" answers for various tasks in the MIMIC-Extract dataset. Both models demonstrated no discriminatory ability, with AUROC scores close to 50 for all tasks. Accuracy varied, with notable high accuracy for MORT ICU and MORT HOSP tasks, particularly for Llama3-8b-Instruct (92.88% and 89.71%, respectively). However, these high accuracy scores likely reflect class imbalance rather than model performance. The contrast between the LLM direct prediction performance and LLM embedding + classifier performance further suggests that LLM embeddings provide a more robust method.

E Parameter Grids for ML Classifiers

We conducted a comprehensive grid search for hyperparameter optimization on two classifiers: XGBoost (XGB) and Logistic Regression. For the XGB classifier, the parameter grid in-

Model	Setting	AUROC CI (%)
Mistral-7b-instruct	sys1	**71.35 [67.39, 75.73]**
	sys2	67.63 [63.53, 71.80]
	sys3	65.67 [61.50, 69.78]
	sys4	67.87 [63.29, 71.88]
	Fewshot	67.32 [63.26, 71.20]
	CoT	64.29 [60.72, 69.24]
Llama2-13b-chat	sys1	68.79 [64.87, 72.59]
	sys2	69.82 [65.66, 73.92]
	sys3	68.92 [64.64, 73.09]
	sys4	64.62 [60.40, 68.73]
	Fewshot	66.49 [62.21, 70.59]
	CoT	65.13 [62.30, 69.25]
Llama3-8b-instruct	sys1	67.05 [62.96, 71.11]
	sys2	66.07 [61.94, 70.28]
	sys3	64.80 [60.74, 69.11]
	sys4	66.81 [63.08, 70.90]
	Fewshot	66.87 [62.90, 70.84]
	CoT	62.12 [58.37, 66.96]

Table 11: AUROC Confidence Intervals for various models and settings on Sepsis prediction. Scores are multiplied by 100.

Model	Task	AUROC	Accuracy
Mistral-7b-instruct	Sepsis	50.00	43.18
	Arrythmia	50.00	15.30
	CHF	50.00	11.82
Llama3-8b-instruct	Sepsis	50.28	54.69
	Arrythmia	47.12	73.63
	CHF	47.61	77.12

Table 12: Results of Mistral and Llama3-8B directly generating "Yes/No" to the Diagnosis dataset. To align with the results reported for emebdding+ML classifiers settings, the LLMs are zero-shot, and no additional system instructions are included in the chat template.

Model	Task	AUROC	Accuracy
Mistral-7b-Inst	LOS 3	49.72	56.05
	LOS 7	49.86	87.93
	MORT ICU	50.04	91.98
	MORT HOSP	49.79	86.93
Llama3-8b-Inst	LOS 3	50.73	57.24
	LOS 7	50.10	92.29
	Mort ICU	49.99	92.88
	Mort Hosp	49.99	89.71

Table 13: Results of Mistral and Meditron direct generation of "Yes/No" to the MIMIC-Extract dataset

cluded $nestimators$ set to [50, 100, 250, 500], $maxdepth$ ranging from [2, 5, 10, 15, 20], $learningrate$ values of [0.005, 0.01, 0.05, 0.1], and $minchildweight$ values of [1, 2, 3]. This extensive search aimed to identify the best combination of hyperparameters to enhance model performance.

Parameter	Values
`n_estimators`	50, 100, 250, 500
`max_depth`	2, 5, 10, 15, 20
`learning_rate`	0.005, 0.01, 0.05, 0.1
`min_child_weight`	1, 2, 3

Table 14: Parameter grid for XGBoost (XGB) classifier.

Parameter	Values
`alpha`	0.1, 0.5, 1.0
`l1_ratio`	0.1, 0.5, 0.9

Table 15: Parameter grid for Logistic Regression (LR).

For the Logistic Regression classifier, we varied alpha with values of [0.1, 0.5, 1.0] and $l1ratio$ with values of [0.1, 0.5, 0.9]. This grid search was designed to fine-tune the regularization parameters to achieve optimal balance between model complexity and performance.

Grid-searching on XGB parameters took 25-40 minutes on GPU. On LR, it took about 25 minutes to search for the best parameters. Training both classifiers took less than 5 minutes, even on the MIMIC-Extract dataset where there are more than 16000 samples.