Large Language Models are Powerful Electronic Health Record Encoders

Stefan Hegselmann[1,2†], Georg von Arnim[1†], Tillmann Rheude[1], Noel Kronenberg[1], David Sontag[3,4], Gerhard Hindricks[2], Roland Eils[1,5], Benjamin Wild[1]

[1]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health, Berlin, Germany.
[2]Deutsches Herzzentrum der Charité – Medical Heart Center of Charité and German Heart Institute Berlin, Berlin, Germany.
[3]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.
[4]Layer Health, Inc., MA, USA.
[5]Intelligent Medicine Institute, Fudan University, Shanghai, China.

[†]These authors contributed equally to this work.

Abstract

Electronic Health Records (EHRs) offer considerable potential for clinical prediction, but their complexity and heterogeneity present significant challenges for traditional machine learning methods. Recently, domain-specific EHR foundation models trained on large volumes of unlabeled EHR data have shown improved predictive accuracy and generalization. However, their development is constrained by limited access to diverse, high-quality datasets, and by inconsistencies in coding standards and clinical practices. In this study, we explore the use of general-purpose Large Language Models (LLMs) to encode EHR into high-dimensional representations for downstream clinical prediction tasks. We convert structured EHR data into markdown-formatted plain text documents by replacing medical codes with natural language descriptions. This enables the use of LLMs and their extensive semantic understanding and generalization capabilities as effective encoders of EHRs without requiring access to private medical training data. We show that LLM-based embeddings can often match or even surpass the performance of a specialized EHR foundation model, CLMBR-T-Base, across 15 diverse clinical tasks from the EHRSHOT benchmark. To demonstrate generalizability, we further evaluate the approach on the UK Biobank (UKB) cohort, a population distinct from that used to train CLMBR-T-Base. Notably, one of the

tested LLM-based models achieves superior performance for disease onset, hospitalization, and mortality prediction, highlighting robustness to shifts in patient populations. Our findings suggest that repurposed general-purpose LLMs for EHR encoding provide a scalable and generalizable alternative to domain-specific models for clinical prediction.

Keywords: electronic health records, clinical prediction, machine learning, large language models, foundation models

1 Introduction

EHRs are now widely used in modern healthcare, providing comprehensive, longitudinal views of a patient's health status [1]. Machine learning methods can use this rich data for risk stratification and to support clinical decision-making [2–4]. In recent years, researchers have explored a variety of prediction tasks based on EHRs, including hospital readmission [5, 6], length of hospital stay [6], sepsis onset detection [7, 8], mortality prediction [6, 9], discharge diagnoses [6], and heart failure outcomes [10]. The overarching goal is to harness EHR data using machine learning to improve clinical outcomes and reduce healthcare costs.

However, machine learning on EHR data poses significant challenges due to its inherent complexity. EHR data is characterized by variable-length sequences of patient visits, irregular sampling intervals, missing entries, heterogeneous and noisy information, and a wide range of hierarchical medical concepts [11]. As a result, deep learning models often achieve only modest improvements over traditional methods such as logistic regression or tree-based methods [6, 12, 13]. To mitigate these issues, recent approaches have employed large-scale foundation models that are pre-trained on unlabeled EHR data using unsupervised learning [14]. Many of these models adopt strategies from natural language processing, such as masked-word prediction as in BERT [15] or autoregressive next-word prediction as in GPT [16]. Treating EHR data as sequences of medical codes enables analogous methods such as masked code prediction [12, 17–19] or next code prediction [13]. However, these techniques also face significant limitations: coding standards and healthcare practices differ strongly across sites, and interoperable EHR foundation models would likely need to be trained on a wide variety of EHR datasets, which is difficult to achieve due to the sensitivity of healthcare data. Therefore, the development of EHR-specific foundation models remains constrained by the limited size and restricted availability of EHR data.

In contrast, LLMs benefit from training on vast general-purpose text corpora and a broad range of natural language tasks [20, 21]. This extensive pre-training enables their language comprehension and allows them to capture domain-agnostic patterns that can be adapted for healthcare applications. Consequently, LLMs have demonstrated strong performance in extracting medical concepts [22], summarizing medical texts [23], and predicting medical outcomes [24] even in low-resource settings. However, most modern LLMs, such as GPT [25] or Llama [26], use a decoder-only transformer

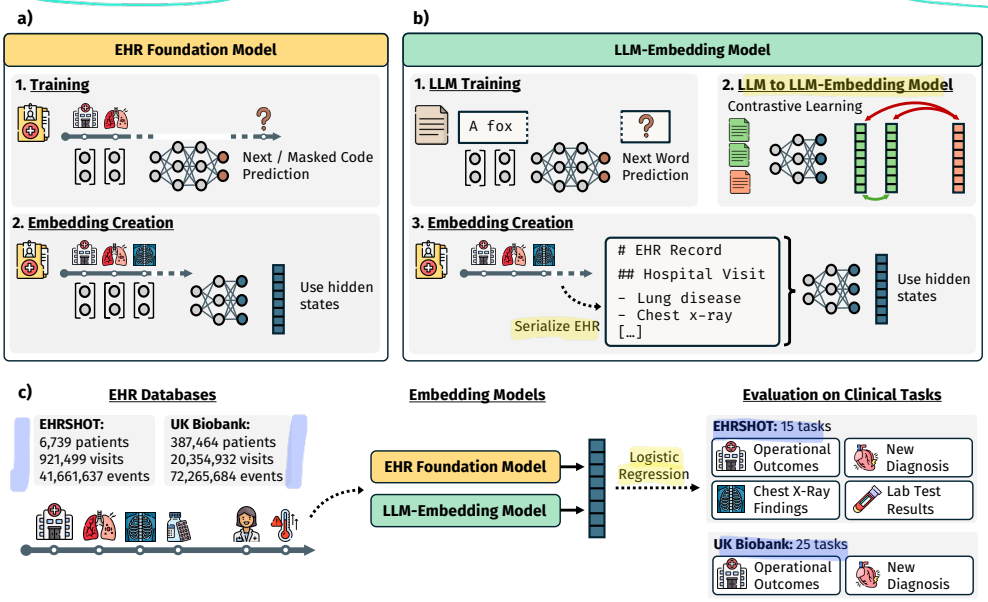

Fig. 1 Study Overview. (a) EHR foundation models are pre-trained on unlabeled EHR data. Common unsupervised learning tasks are masked code or next code prediction. To obtain a representation for an EHR, we use the hidden states of the pre-trained model. (b) LLMs are pre-trained on vast amounts of text data. To obtain an LLM embedding model, architectural changes are applied, and contrastive learning is used to improve representational performance. To obtain an EHR embedding, the data is first serialized as text and then processed by the LLM embedding model. Again, we use the hidden states for the embedding. (c) We use the EHRSHOT database and the UK Biobank for our experiments. Medical events of each patient are converted into numerical embeddings using an EHR foundation model and an LLM embedding model, respectively. A logistic regression classification head is trained, validated, and tested for each clinical prediction task. Icons from flaticon.com.

architecture, which complicates the generation of robust text representations. To overcome this limitation, recent work has introduced methods to convert decoder-only LLMs into effective embedding models for downstream prediction tasks [27–30]. Additionally, these state-of-the-art models offer an increased context window, making them well-suited for handling long inputs such as serialized EHR data.

In this study, we evaluate whether general-purpose LLM embedding models can effectively encode EHR data for clinical prediction [31] (see Fig. 1). To this end, we convert structured EHR records into plain text summaries that highlight key patient information at prediction time. Using two state-of-the-art LLM embedding models, GTE-Qwen2-7B-Instruct (GTE-Qwen2-7B) [30, 32] and LLM2Vec-Llama-3.1-8B-Instruct (LLM2Vec-Llama-3.1-8B) [27, 33], we generate high-dimensional EHR embeddings that serve as inputs to logistic regression classifiers across 15 clinical tasks from the EHRSHOT benchmark. We focus on the performance in few-shot settings to evaluate the generalization ability of this approach and conduct extensive ablation studies to identify the specific components that drive the LLM's effectiveness. Finally, we validate our approach on the UKB for predicting mortality, hospitalization and onset of 23 diseases [34] to assess its robustness and compare the generalizability with EHR-specific foundation models.

Table 1 Cohort Overview. Summary statistics for EHRSHOT and UK Biobank, including the number of patients, visits, events, and patient characteristics.

Attribute	EHRSHOT	UK Biobank
Num Patients	6,739	387,464
Num Visits	921,499	19,484,777
Num Events	41,661,637	72,265,684
Num Female	3,441	214,565
Num Male	3,298	172,899
Age, mean ± SD	59.3 ± 17.9	56.78 ± 8.11
American Indian	25	0
Asian	1,043	8,659
Black	298	5,751
Pacific Islander	74	0
Unknown	1,563	7,202
White	3,736	365,888
Hispanic	1,038	-
Non-Hispanic	5,701	-

2 Results

2.1 Experimental Setup

Our primary analysis was conducted using the EHRSHOT database, which contains EHRs from 6,739 adult patients treated at Stanford Health Care and Lucile Packard Children's Hospital between 1990 and 2023. This dataset includes 921,499 visits and over 41.6 million clinical events [31], and serves as a standardized benchmark with 15 clinical prediction tasks organized into 4 task groups, along with predefined splits and publicly available code. Table 1 summarizes cohort statistics, and task details are shown in Table 2. Further information on task definitions and preprocessing is provided in Section 4.1.

For external validation, we used the UKB, a population-based cohort comprising 502,489 UK participants [35, 36]. This allowed us to evaluate the generalizability of our approach across healthcare systems, especially since the CLMBR-T-Base model was trained on data from the same hospital system as EHRSHOT. Due to structural differences between the datasets, not all EHRSHOT tasks were transferable to the UKB (see Section 4.8). We focused on predicting one-year risk of hospitalization, mortality, and the onset of 23 chronic diseases [34]. The processed UKB subset used in our study consisted of 387,464 patients, approximately 19.5 million visits, and over 72 million clinical events. Table 6 provides task descriptions and label distributions.

To prepare EHR data for use with LLMs, we serialized structured patient records into plain text formatted in Markdown, with a maximum length of 4,096 tokens, prioritizing more recent data. For EHRSHOT, serialization began with a reference prediction date to which all event dates were normalized, followed by demographics, recent values for 24 key LOINC concepts across Body Metrics, Vital Signs, and Lab Results, including units and classification (low/normal/high). These were followed by a summary of patient visits and reverse-chronological listings of visit-level events such as conditions, medications, and procedures. An example EHR text serialization is shown in Fig. 2, and a detailed explanation appears in Section 4.2. For the UKB, a similar serialization approach was used, with event dates normalized to each patient's recruitment date and exclusion of detailed measurements due to data limitations.

```
# Electronic Healthcare Record

Current time: [2024-01-01](2024-01-01)

## Patient Demographics

- Patient age: 78
- Black
- FEMALE
- Hispanic or Latino

## Recent Body Metrics

- Body weight (oz): 1801
- Body height (inch): 62.0, 61.0
- Body mass index / BMI (kg/m2): 18.7 (normal)
- Body surface area (m2): 1.47

## Recent Vital Signs

- Heart rate (bpm): 121 (high), 85 (normal)
- Respiratory rate (breaths/min): 16 (normal)
- Systolic blood pressure (mmHg): 148 (high), 117 (normal)
- Diastolic blood pressure (mmHg): 81 (normal), 57 (low)
- Body temperature (°F): 97.4 (normal), 98.8 (normal)
- Oxygen saturation (%): 97 (normal), 97 (normal), 99 (normal)

## Recent Lab Results

- Hemoglobin (g/dL): 8.2 (low), 8.6 (low), 8.8 (low)
- Hematocrit (%): 24 (low), 26 (low), 26 (low)
- Erythrocytes: No recent data
- Leukocytes (10^3/uL): 2.7 (low), 8.8 (normal), 6.2 (normal)
- Platelets (10^3/uL): 215 (normal), 199 (normal)
- Sodium (mmol/L): 132 (low)
- Potassium (mmol/L): 4.2 (normal), 4.4 (normal)
- Chloride (mmol/L): 95 (low), 102 (normal)
- Carbon dioxide, total (mmol/L): No recent data
- Calcium (mg/dL): 8.9 (low), 8.5 (low)
- Glucose (mg/dL): 92 (normal), 112 (high)
- Urea nitrogen (mg/dL): 11 (normal), 8 (normal)
- Creatinine (mg/dL): 0.4 (low), 0.7 (normal)
- Anion gap: No recent data

## Past Medical Visits

- Inpatient Visit on [2023-12-17](2023-12-17) (14 days before prediction
  time, current visit)
- Office Visit on [2023-10-27](2023-10-27) (65 days before prediction time)

## General Medical Events

- Cigarette consumption: N, N, N
- Mitral valve disorder

## Detailed Past Medical Visits (most recent first)

### Inpatient Visit on [2023-12-17](2023-12-17) (14 days before prediction
time, current visit)

#### Conditions

- Acute posthemorrhagic anemia
- Partial thromboplastin time, activated
- pH measurement, venous: 7.25, 7.31, 7.31

#### Medications

- furosemide 20 MG Oral Tablet
- pantoprazole 20 MG Delayed Release Oral Tablet

#### Procedures

- Chest x-ray
- Electrocardiogram report

### Office Visit on [2023-10-27](2023-10-27) (65 days before prediction time)

[…]
```

Fig. 2 Example EHR Text Serialization. The EHR data is serialized into plain text to apply LLM embedding models. We use Markdown formatting and prioritize relevant medical information. All dates were normalized relative to a reference date. Next, the patient's demographics are listed. Time-series data coded via Logical Observation Identifiers Names and Codes (LOINC) was aggregated into 24 key concepts listed with the last three values, units, and classifications into low, normal, and high. Then, a list of all visits and all concepts not associated with a visit are given. Lastly, detailed visit entries beginning with the most recent are listed. Unique concepts are categorized into conditions, medications, and procedures. The last three values of a concept are given when present.

Table 2 EHRSHOT Prediction Tasks Overview. The EHRSHOT benchmark defines 15 clinical prediction tasks spanning four different task groups. The number of examples per task differs based on the prevalence and frequency of clinical events. Canonical splits for training, validation, and testing are defined to ensure reproducible experiments [31].

Attribute	Train Labels (Positive)	Valid Labels (Positive)	Test Labels (Positive)	Total Labels (Positive)
Operational Outcomes				
Long Length of Stay	2,569 (681)	2,231 (534)	2,195 (552)	6,995 (1,767)
30-day Readmission	2,609 (370)	2,207 (281)	2,189 (260)	7,005 (911)
ICU Transfer	2,402 (113)	2,052 (92)	2,037 (85)	6,491 (290)
Anticipating Lab Test Results				
Thrombocytopenia	68,776 (9,774)	54,504 (6,962)	56,338 (7,960)	179,618 (24,696)
Hyperkalemia	76,349 (1,215)	60,168 (886)	63,653 (948)	200,170 (3,049)
Hypoglycemia	122,108 (1,065)	95,488 (858)	100,568 (783)	318,164 (2,706)
Hyponatremia	81,336 (20,181)	64,473 (14,674)	67,028 (16,003)	212,837 (50,858)
Anemia	70,501 (9,544)	56,224 (7,445)	58,155 (7,636)	184,880 (24,625)
Assignment of New Diagnoses				
Hypertension	1,260 (184)	1,250 (177)	1,261 (160)	3,771 (521)
Hyperlipidemia	1,684 (205)	1,441 (189)	1,317 (172)	4,442 (566)
Pancreatic Cancer	2,576 (155)	2,215 (53)	2,220 (56)	7,011 (264)
Celiac	2,623 (62)	2,284 (11)	2,222 (21)	7,129 (94)
Lupus	2,570 (104)	2,226 (33)	2,243 (20)	7,039 (157)
Acute MI	2,534 (175)	2,177 (146)	2,127 (144)	6,838 (465)
Anticipating Chest X-ray Findings				
Chest X-Ray Findings	7,481 (4,771)	9,366 (6,032)	9,428 (6,400)	26,275 (17,203)

We evaluated two instruction-tuned LLM embedding models, GTE-Qwen2-7B [30, 32] and LLM2Vec-Llama-3.1-8B [27, 33], using task-specific prompts (e.g., "Given a patient's electronic healthcare record (EHR) in Markdown format, retrieve relevant passages that answer the query: has the patient anemia", see Table 8). Our primary objective was to assess how general-purpose LLM-based embedding models, trained on public text data, perform in encoding EHRs for clinical prediction tasks. For comparison, we included CLMBR-T-Base, a 141-million-parameter autoregressive foundation model trained on 2.57 million de-identified EHRs from Stanford Medicine [13, 31]. For each model, we computed patient-level embeddings and trained a logistic regression classifier on the training split, with hyperparameters selected using a validation set. As a baseline, we also included a Gradient Boosted Machine (GBM) classifier trained on count-based features, which has previously demonstrated strong performance on EHR tasks [31]. See Section 4.4 for additional details on model architectures.

2.2 General-Purpose LLM Embeddings Rival Domain-Specific EHR Models

Using all available training and validation examples, the GTE-Qwen2-7B model outperformed CLMBR-T-Base in two of the four task categories: lab test result prediction and assignment of new diagnoses (see Table 3). In contrast, CLMBR-T-Base achieved higher performance on tasks related to operational outcomes and chest X-ray findings. LLM2Vec-Llama-3.1-8B embeddings showed slightly lower performance than

Table 3 Performance on All Examples for EHRSHOT. Macro averaged area under receiver operating characteristic curve (AUROC) performance and bootstrapped 95% confidence intervals of included models for four task groups. The macro-averaged performance across all task groups is given in the right-most column. The LLM-embedding model GTE-Qwen2-7B with a logistic regression (LR) classification head performs slightly worse than the EHR foundation model CLMBR-T-Base, but outperforms the Counts-based baseline using a gradient boosted machine (GBM) head. LLM2Vec-Llama-3.1-8B only outperforms CLMBR-T-Base on the task group for assignment of new diagnosis. Combining the embeddings of the LLM-embedding models and CLMBR-T-Base by concatenation leads to a further increase in performance. Additional model variants with fewer parameters or using an encoder-only architecture show a lower overall performance.

Model	Operational Outcomes	Anticipating Lab Test Results	Assignment of New Diagnosis	Anticipating Chest X-ray Findings	Macro Avg. Across Task Groups
LLM-Embedding Models					
GTE-Qwen2-7B	0.771 .747-.795	0.865 .858-.873	0.716 .675-.757	0.666 .653-.680	0.755 .730-.780
GTE-Qwen2-1.5B	0.755 .729-.780	0.865 .859-.872	0.700 .657-.743	0.670 .658-.683	0.748 .722-.774
LLM2Vec-Llama-3.1-8B	0.753 .727-.778	0.778 .768-.789	0.728 .682-.774	0.678 .665-.692	0.734 .707-.762
LLM2Vec-Llama 2 1.3B	0.706 .680-.731	0.649 .638-.660	0.636 .598-.673	0.615 .603-.627	0.651 .627-.675
LLM-Embedding Model + EHR Foundation Model					
GTE-Qwen2-7B + CLMBR-T-Base	0.822 .802-.842	0.886 .880-.893	0.725 .682-.768	0.711 .699-.723	0.786 .762-.811
LLM2Vec-Llama-3.1-8B + CLMBR-T-Base	0.806 .785-.828	0.841 .833-.849	0.732 .688-.777	0.714 .702-.725	0.773 .747-.799
Baselines [31]					
CLMBR-T-Base	0.824 .803-.845	0.832 .824-.840	0.707 .667-.746	0.713 .702-.724	0.769 .746-.792
Counts-based + GBM	0.774 .752-.797	0.728 .716-.741	0.719 .669-.768	0.656 .641-.671	0.719 .691-.748
Encoder Language Models					
DeBERTaV3 large	0.720 .693-.747	0.709 .697-.721	0.670 .624-.715	0.624 .610-.638	0.681 .653-.709
DeBERTa V3 base	0.742 .717-.766	0.707 .694-.720	0.665 .614-.715	0.624 .609-.638	0.684 .655-.714
BERT large	0.727 .702-.752	0.720 .709-.732	0.667 .620-.714	0.641 .627-.654	0.689 .660-.717
BERT base	0.738 .712-.763	0.716 .704-.728	0.677 .630-.725	0.635 .622-.649	0.692 .663-.720
ClinicalBERT	0.741 .716-.766	0.735 .723-.746	0.703 .658-.747	0.650 .636-.663	0.707 .680-.734

Table 4 Performance on external validation. Macro averaged area under receiver operating characteristic curve (AUROC) performance. The LLM-embedding model GTE-Qwen2-7B with a logistic regression (LR) classification head outperforms the EHR foundation model CLMBR-T-Base and the Counts-based baseline using a gradient boosted machine (GBM) head. The *Assignment of New Diagnoses* prediction is based on the mean of all 23 provided diseases.

Model	Mortality prediction	Operational Outcomes (Hospitalization)	Assignment of New Diagnoses	Macro Avg. Across Task Groups
LLM-Embedding Models				
GTE-Qwen2-7B	0.826 .819-.833	0.655 .647-.663	0.727 .712-.742	0.736 .726-.746
LLM2Vec-Llama-3.1-8B	0.781 .776-.787	0.638 .630-.645	0.715 .699-.730	0.711 .702-.721
Baselines [31]				
CLMBR-T-Base	0.782 .766-.799	0.639 .629-.649	0.707 .691-.723	0.709 .695-.724
Counts-based + GBM	0.747 .728-.766	0.619 .611-.626	0.703 .681-.726	0.690 .673-.706

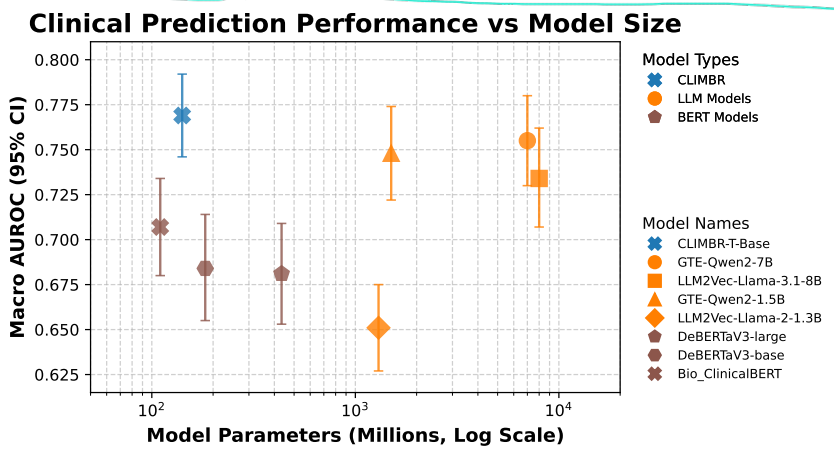

Fig. 3 Scaling Behavior of Models. Number of model parameters (x-axis) and macro-averaged area under receiver operating characteristic curve (AUROC) performance and 95% confidence intervals across all four task groups (y-axis). LLMs with more parameters show increased performance. The specialized EHR foundation model, CLMBR-T-Base, is the most efficient prediction model.

GTE-Qwen2-7B, exceeding CLMBR-T-Base only in the assignment of new diagnoses task group. Both LLM-based models were often on par with or outperformed the GBM-based baseline, which is considered a strong baseline for EHR prediction tasks [12, 13]. To assess whether the models encode orthogonal information, we concatenated CLMBR-T-Base embeddings with LLM-based embeddings and observed substantial performance improvements for both LLM models. The combined embeddings achieved an average area under the receiver operating characteristic curve (AUROC) of 0.786 (0.762–0.811) for GTE-Qwen2-7B, up from 0.755 (0.730–0.780), and 0.773 (0.747–0.799) for LLM2Vec-Llama-3.1-8B, up from 0.734 (0.707–0.762). The smaller GTE-Qwen2-1.5B model achieved an average performance of 0.748 (0.722–0.774), only marginally lower than its larger counterpart, suggesting it is a viable lightweight alternative for EHR embeddings. In contrast, LLM2Vec-Llama-2-1.3B performed notably worse than the larger version, potentially due to limitations of the older Llama 2 architecture. Figure 3 shows a general trend that larger models tend to yield better performance across tasks. Among the evaluated models, CLMBR-T-Base remained the most parameter-efficient relative to predictive performance. Nonetheless, it is important to note that the LLM-based embedding models were not specifically optimized for structured EHR data, which may explain their performance differences.

On the UKB dataset, GTE-Qwen2-7B consistently outperformed CLMBR-T-Base across all three prediction task groups (see Table 4). LLM2Vec-Llama-3.1-8B surpassed CLMBR-T-Base only in the task group related to the assignment of new diagnoses. The performance drop of CLMBR-T-Base was anticipated, as the UKB represents an out-of-domain dataset with numerous disease codes not included in the model's original vocabulary. Although we applied a careful mapping of UKB Observational Medical Outcomes Partnership (OMOP) concepts to those supported by CLMBR-T-Base to improve compatibility and predictive performance, many concepts remained

unmapped due to CLMBR-T-Base's limited vocabulary. This limitation underscores a key challenge in adapting domain-specific models to external datasets, even with considerable manual effort. In contrast, the LLM-based models required only a mapping from clinical concepts to their natural language descriptions, which were directly incorporated into the Markdown-based serialization. This simplified integration process substantially reduced implementation overhead and emphasized the strong generalization ability of LLM-based encodings across diverse coding systems and healthcare domains.

2.3 LLM Embeddings Achieve Strong Performance in Low-Data Regimes

To evaluate model performance with limited training data, we conducted experiments in a few-shot setting using small numbers of training examples. The experiments followed the EHRSHOT task definitions [31]. Fig. 4 illustrates the aggregated AUROC across all subtasks within the four task categories for varying numbers of training examples. LLM embedding models performed strongly in the few-shot setting, suggesting effective transfer from general text pretraining to serialized EHR data. Notably, GTE-Qwen2-7B consistently outperformed CLMBR-T-Base across all training sizes for lab test result prediction and assignment of new diagnoses. The strong results for lab test predictions indicate that the inclusion of the last three lab values in the EHR serialization provides highly predictive features, which GTE-Qwen2-7B effectively encodes. In contrast, LLM2Vec-Llama-3.1-8B showed comparatively weaker performance in the few-shot regime. It only surpassed CLMBR-T-Base in a limited subset of tasks, primarily those involving the assignment of new diagnoses. Nonetheless, both LLM-based models consistently outperformed the count-based baseline across most configurations, underscoring the benefits of extensive pretraining. Consistent with previous findings [31], the largest performance improvements over the baseline were observed at intermediate training set sizes. As the number of labeled data samples increased, the relative advantage of pre-trained LLM-based models diminished, highlighting the reduced marginal gains of foundation models in high-data regimes. The area under the precision-recall curve (AUPRC) results are shown in Fig. 7, with task-specific AUROC and AUPRC results provided in Fig. 8 and Fig. 9 in the Appendix.

In contrast to EHRSHOT, the LLM-based methods outperformed CLMBR-T-Base in most prediction tasks on the UKB dataset. For the assignment of new diagnoses, which includes 23 tasks, both GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B outperformed CLMBR-T-Base across all sample sizes (see Fig. 10 in the Appendix). For hospitalization and mortality prediction, CLMBR-T-Base showed stronger performance in the very low-data regime. However, with more samples the performance of the LLM embedding models increased and GTE-Qwen2-7B outperformed CLMBR-T-Base. Figure 5 presents the mean AUROC across all task groups for both EHRSHOT and UKB, highlighting the observed differences. On EHRSHOT, GTE-Qwen2-7B generally matched CLMBR-T-Base in performance, while LLM2Vec-Llama-3.1-8B fell between these models and the count-based baseline. On the UKB dataset, LLM-based models and CLMBR-T-Base performed comparably with very limited training data of

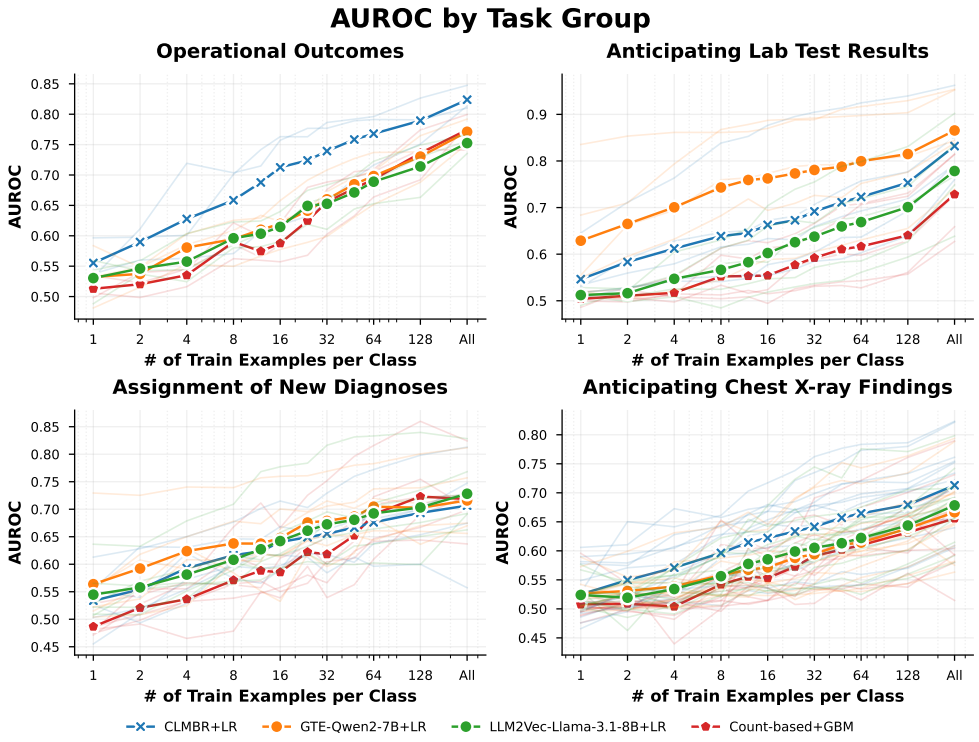

Fig. 4 Performance in Few-Shot Settings. Macro-averaged area under the receiver operating characteristic curve (AUROC) performance across subtasks for four task groups (bold). Blurred lines are averaged AUROC values across five bootstrapped runs using different seeds [31]. Similar to the EHR foundation model, CLMBR-T-Base, the LLM embedding models show the largest performance gains over the count-based model at intermediate numbers of training examples. With an increased number of training examples, the advantage of pre-trained LLM-based models decreases.

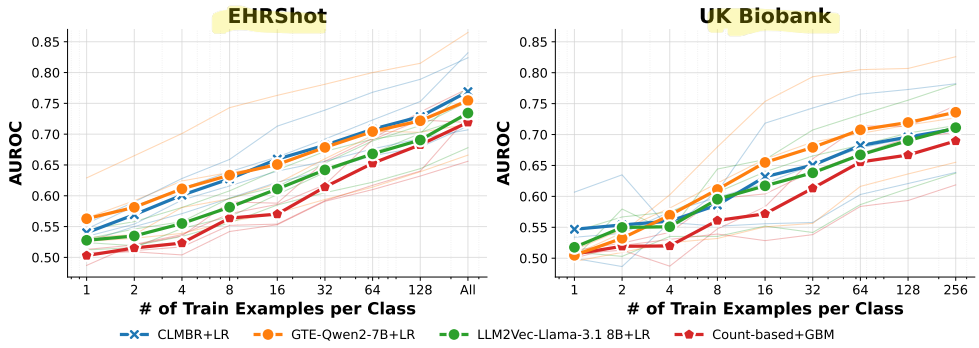

Fig. 5 Overall performance on EHRSHOT and UKB. Macro-averaged area under receiver operating characteristic curve (AUROC) performance across all subtasks of EHRSHOT (left) and the UKB (right) analysis. Blurred lines are averaged AUROC values of the different task groups.

Table 5 EHR Serialization Ablation Experiments for LLM-embedding Models. Macro averaged area under receiver operating characteristic curve (AUROC) performance and 95% confidence intervals for task groups and macro-averaged performance across all task groups for different EHR serialization ablations studies. Removing the task-specific instructions and aggregated information lead to the largest drop in performance for both LLM-embedding models. For LLM2Vec-Llama-3.1-8B, some ablations even show an increased overall performance.

Model	Operational Outcomes	Anticipating Lab Test Results	Assignment of New Diagnosis	Anticipating Chest X-ray Findings	Macro Avg. Across Task Groups
GTE-Qwen2-7B	0.771 .747-.795	0.865 .858-.873	0.716 .675-.757	0.666 .653-.680	0.755 .730-.780
generic instructions	0.761 .735-.786	0.807 .797-.817	0.717 .675-.759	0.673 .659-.686	0.739 .713-.765
no instructions	0.755 .728-.782	0.768 .757-.780	0.705 .665-.744	0.666 .653-.679	0.724 .698-.749
no demographics	0.762 .738-.786	0.864 .857-.872	0.702 .658-.747	0.675 .663-.686	0.751 .725-.777
no aggregated	0.769 .744-.793	0.714 .701-.726	0.741 .699-.784	0.669 .655-.683	0.723 .697-.749
no visits (both)	0.752 .728-.776	0.878 .872-.884	0.735 .690-.779	0.663 .649-.677	0.757 .731-.783
no conditions	0.761 .738-.785	0.867 .860-.874	0.731 .694-.769	0.653 .638-.667	0.753 .730-.777
no medications	0.766 .743-.789	0.866 .859-.873	0.709 .666-.751	0.672 .659-.686	0.753 .728-.778
no procedures	0.775 .752-.798	0.868 .861-.875	0.724 .684-.764	0.669 .655-.682	0.759 .735-.783
LLM2Vec-Llama-3.1-8B	0.753 .727-.778	0.778 .768-.789	0.728 .682-.774	0.678 .665-.692	0.734 .707-.762
generic instructions	0.753 .728-.778	0.748 .737-.760	0.726 .679-.773	0.669 .655-.683	0.724 .696-.752
no instructions	0.757 .732-.782	0.752 .741-.764	0.724 .678-.770	0.677 .663-.690	0.727 .700-.755
no demographics	0.755 .730-.780	0.779 .769-.790	0.703 .653-.752	0.680 .668-.693	0.729 .700-.758
no aggregated	0.753 .728-.779	0.706 .693-.719	0.715 .665-.766	0.676 .662-.690	0.713 .683-.743
no visits (both)	0.734 .709-.759	0.862 .856-.869	0.721 .668-.775	0.677 .663-.690	0.749 .718-.779
no conditions	0.750 .725-.774	0.786 .776-.796	0.705 .656-.755	0.662 .649-.675	0.726 .697-.755
no medications	0.744 .719-.769	0.788 .778-.799	0.714 .667-.762	0.677 .663-.690	0.731 .703-.759
no procedures	0.752 .728-.777	0.779 .769-.790	0.735 .692-.778	0.674 .661-.688	0.735 .709-.761

up to four examples. With additional training data, GTE-Qwen2-7B surpassed both CLMBR-T-Base and LLM2Vec-Llama-3.1-8B, while the latter closely approached the performance of CLMBR-T-Base. These findings underscore the capability of general-purpose LLM embeddings to rival or exceed the performance of domain-specific models, particularly in settings characterized by domain shift or limited labeled data. Detailed task-specific results are available in Fig. 12 and Fig. 13 in the Appendix.

2.4 Effect of Serialization Components on LLM-Based EHR Embedding Performance

To evaluate the contribution of individual serialization components, we conducted ablation studies for both LLM embedding models (see Table 5). Replacing task-specific instructions with a generic prompt led to a marked performance decline, reducing average AUROC from 0.755 (0.730-0.780) to 0.739 (0.713-0.765) for GTE-Qwen2-7B and from 0.734 (0.707-0.762) to 0.724 (0.696-0.752) for LLM2Vec-Llama-3.1-8B. Removing the instruction entirely caused a further drop for GTE-Qwen2-7B to 0.724 (0.698-0.749), whereas LLM2Vec-Llama-3.1-8B showed minimal change, suggesting greater reliance on instruction prompting by GTE-Qwen2-7B. The most significant drop was observed in lab result prediction tasks, highlighting the importance of guidance via task-specific instructions. Aggregated measurements (body metrics, vital signs, and lab values) were crucial, especially for lab test prediction, where their removal reduced AUROC from 0.865 (0.858-0.873) to 0.714 (0.701-0.726) for GTE-Qwen2-7B and from 0.778 (0.768-0.789) to 0.706 (0.693-0.719) for LLM2Vec-Llama-3.1-8B. This

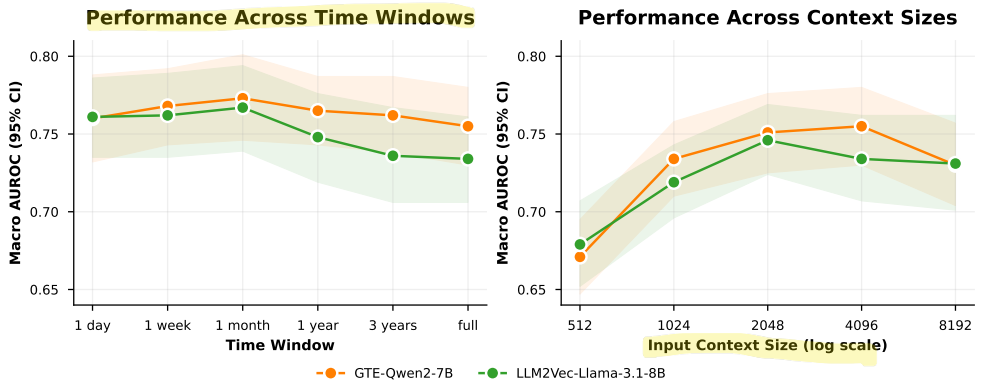

Fig. 6 Performance of LLM Embedding Models Across Time Windows and Context Sizes. Macro-averaged area under the receiver operating characteristic curve (AUROC) performance (y-axis) across all task groups for the LLM embedding models, shown for different time windows before prediction time (left) and different context sizes (right). Full results are presented in Table 9 and Table 10.

underscores the predictive value of recent lab results and their effective encoding by the models. For conditions like thrombocytopenia, hyperkalemia, and hyponatremia, the performance drop was especially pronounced without this information for GTE-Qwen2-7B (see Fig. 8). Interestingly, removing aggregated values slightly improved new diagnosis prediction for GTE-Qwen2-7B, suggesting a more focused input may benefit some tasks. Visit-level information had a strong influence on operational outcome tasks. Its removal led to AUROC reductions from 0.771 (0.747-0.795) to 0.752 (0.728-0.776) for GTE-Qwen2-7B and from 0.753 (0.727-0.778) to 0.734 (0.709-0.759) for LLM2Vec-Llama-3.1-8B. However, for LLM2Vec-Llama-3.1-8B, removing visit data unexpectedly improved lab result prediction, suggesting possible difficulty in filtering relevant signals from verbose input. Removing demographics, conditions, medications, and procedures had only minor effects, indicating these features were either redundant or contributed little predictive signal in isolation. While aggregated information and visit data were key drivers for lab and operational outcomes, prediction of new diagnoses and chest X-ray findings was less sensitive to individual serialization components. This suggests these tasks rely on broader contextual patterns rather than specific structured elements.

2.5 Effect of Temporal Scope and Context Length on Embedding Effectiveness

To assess model sensitivity to input length and recency, we performed ablations on temporal window, context size, and chunking using EHRSHOT data. Both LLM embedding models showed improved performance when restricted to more recent data, with the 1-month window outperforming longer histories. This suggests that recent clinical information may carry the most predictive signal, while older records may contribute less relevant or potentially distracting details (see Fig. 6 and Appendix A.2).

The models differed in optimal context size: GTE-Qwen2-7B achieved the best performance at the maximum 4,096-token context, whereas LLM2Vec-Llama-3.1-8B performed best at 2,048 tokens (Fig. 6 and Table 10). Both models suffered large drops at 512 tokens, a common constraint in older language models [37]. To test whether models processed the entire input in context, we evaluated chunked context representations, where fixed-size segments (e.g., 512–2,048 tokens) of a 4,096-token input were embedded separately and averaged (see Table 11). GTE-Qwen2-7B showed little degradation from chunking, indicating that simple averaging can be sufficient to preserve performance even under token limits. In contrast, LLM2Vec-Llama-3.1-8B benefited from chunking, suggesting it struggles with longer contiguous sequences and prefers shorter inputs. These results highlight differences in model-specific sequence processing and suggest practical approaches for adapting to context length constraints. Full tables and additional details are available in Appendix A.3 and Appendix A.4.

3 Discussion

Our study shows that general-purpose LLM embedding models, pre-trained on large-scale natural language corpora, can be effectively repurposed for clinical prediction using EHR data. Models such as GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B outperformed a strong count-based baseline and, in several clinical domains, matched or surpassed the dedicated EHR foundation model CLMBR-T-Base [13, 31]. This is notable, as CLMBR-T-Base was trained on data from the same health system as EHRSHOT. The strong performance of the LLM-based encoders on the UKB dataset highlights their generalization capabilities. GTE-Qwen2-7B consistently outperformed the domain-specific foundation model with at least 16 training examples, while LLM2Vec-Llama-3.1-8B achieved similar results. These results are particularly notable given the challenges posed by EHR data, including limited size, heterogeneity, and domain shift. Across a broad set of tasks, including few-shot settings, our findings indicate that knowledge acquired during large-scale text pre-training enables LLMs to extract clinically meaningful patterns from serialized EHRs. Model performance improved with larger model sizes and more advanced architectures, suggesting further gains are likely with future LLM development. In contrast, domain-specific EHR models remain constrained by the scarcity and fragmentation of large, diverse clinical datasets. Overall, our findings support general-purpose LLMs as scalable, adaptable, and high-performing alternatives for EHR representation and clinical prediction.

Our experiments suggest that several key factors contribute to the success of LLM embedding models for encoding EHR data. First, providing clear, task-specific instructions in the serialized EHR text was crucial, guiding models to focus on clinically relevant sections and improving tasks requiring targeted attention. This takes advantage of the instruction-tuned nature of LLM embedding models [27, 30], enabling them to extract meaningful patterns from heterogeneous EHR inputs. Second, aggregating structured data, particularly recent lab results and visit-level information, substantially improved performance, especially for lab test and operational outcome prediction. In contrast, components like demographics, medications, and condition codes contributed less consistently to model accuracy. Context size was also critical:

GTE-Qwen2-7B effectively handled up to 4,096 tokens, while both models performed poorly with very short or overly long inputs. Intermediate time windows, especially those covering one month prior to prediction, yielded optimal performance, suggesting that recent clinical data carries more predictive value than long-term history for many tasks. For instance, lab values reflect dynamic physiological states and may lose relevance outside of a narrow temporal window. This indicates that LLM embeddings are particularly effective at encoding temporally focused, task-relevant clinical signals. Notably, GTE-Qwen2-7B demonstrated greater robustness to extended inputs, highlighting its stronger capacity to extract salient information from longer contexts. Finally, combining LLM embeddings with CLMBR-T-Base representations improved overall performance, suggesting that these models capture complementary dimensions of patient history. This synergy may stem from CLMBR-T-Base's limited code-based input window of 496 medical codes [13], or from the broader world knowledge encoded in LLMs that is absent in domain-specific models.

Unlike count-based models and specialized EHR foundation models, LLM-based methods are agnostic to coding systems and data formats. Traditional EHR models rely on predefined vocabularies such as SNOMED, LOINC, and ICD, which can limit their applicability across healthcare settings. For instance, CLMBR-T-Base processes only 65,536 codes (26,249 unique), selected based on their entropy from the EHRSHOT dataset [31]. This restriction may lead to the exclusion of clinically important information when applied to external datasets with unseen codes. In contrast, LLMs operate on textual inputs, allowing them to interpret any clinical code mapped to a human-readable description. This flexibility is particularly valuable given the ongoing challenges in EHR interoperability due to varying coding practices, privacy constraints, and regulatory barriers, which often hinder the aggregation of large, standardized datasets for model pre-training [38]. Moreover, since LLMs are pre-trained on broad and diverse text corpora, including medical literature and case reports, they are well-equipped to capture the semantics of rare or underrepresented clinical concepts [34]. This mitigates the limitations of count-based models that often exclude infrequent events, ensuring that even rare but clinically significant phenomena are effectively encoded [39]. Finally, the ability to represent both structured and unstructured information makes LLMs well-suited for future multimodal EHR applications, including those that integrate clinical notes, diagnostic reports, and imaging metadata [40].

Our findings highlight that LLM embedding methods combine the strengths of count-based models, specialized EHR models, and the generalization capabilities of LLMs. While count-based and specialized EHR foundation models can be tuned to output well-calibrated probabilities, thereby enhancing trust in clinical applications, they are limited by predefined vocabularies and constrained training data. In contrast, LLMs can process any textual representation, enabling them to handle rare or unseen codes, though they often produce unstructured outputs that complicate clinical grounding and calibration [24]. LLM embedding methods bridge this gap by using the representational power of LLMs while maintaining compatibility with traditional predictive modeling frameworks. By converting EHR data into structured text and embedding it with general-purpose LLMs, we enable the use of standard classifiers

such as logistic regression to generate clinically meaningful predictions. This approach retains the calibration strengths of traditional models while benefiting from the rich contextual understanding of LLMs.

This study has several limitations. First, our approach relies on a manually designed EHR serialization designed to capture medically relevant information, which may introduce bias when comparing against models trained on raw data. Additionally, LLM embedding model performance is sensitive to the specific content and instructions in the serialized text, which may limit reproducibility and generalization. While the models achieve strong predictive accuracy across tasks, including few-shot settings, their large parameter counts lead to longer computation time and higher resource demands. Moreover, the use of LLM embeddings requires training a separate downstream classifier, foregoing the native zero-shot or few-shot capabilities of LLMs. We also limited EHR serialization to 4,096 tokens due to runtime constraints, which may exclude relevant long-term historical data. Finally, evaluation was conducted on only two datasets, which may limit the generalizability of our findings across diverse healthcare systems.

Future research should explore serialization-free approaches that allow LLMs to process raw EHR data directly, reducing biases introduced by manual text transformation. Integrating zero-shot and few-shot prompting into the LLM-based embedding framework could further enhance model flexibility and reduce dependency on downstream training. Extending the effective context window beyond 4,096 tokens will be important for capturing comprehensive patient histories. Additionally, developing techniques to distill large LLMs into smaller, more efficient models could improve their practical use in clinical settings. Expanding evaluations to real-world deployments and investigating how to combine the complementary strengths of domain-specific EHR models and general-purpose LLMs will be essential for building robust, scalable EHR foundation models.

4 Methods

4.1 EHRSHOT Database and Prediction Task

The EHR data used in our experiments is from the EHRSHOT benchmark for few-shot evaluation of EHR foundation models [31]. We obtained version 2.1 of the dataset, which is accessible via gated access under a research data use agreement. This dataset comprises longitudinal records for 6,739 patients, 921,499 visits, and 41,661,637 clinical events collected between 1990 and February 8, 2023. Each clinical event is linked to a specific patient and includes information such as start time, end time, a semantic code, a value, unit, visit ID, and the corresponding OMOP source table. We used the official ehrshot-benchmark repository[1] as a starting point to design our experiments, enabling us to build on existing functionalities and to facilitate comparisons with prior methods. The benchmark uses the Framework for Electronic Medical Records (FEMR)[2], which provides Python classes for efficient loading and processing of EHR data. All

[1] GitHub repository: https://github.com/som-shahlab/ehrshot-benchmark
[2] GitHub repository: https://github.com/som-shahlab/femr

extensions and experiments conducted for this paper have been made publicly available via our GitHub repository: https://github.com/stefanhgm/ehrshot-benchmark. The EHRSHOT benchmark defines a rigorous evaluation including 15 clinical prediction tasks categorized into four groups: operational outcomes, anticipating lab test results, assignment of new diagnoses, and anticipating chest X-ray findings [31]. Task labels are derived from clinical events; hence, a single patient could contribute multiple labels per task, resulting in significant variations in task-specific sample sizes. For instance, frequent events like lab tests provide considerably more examples compared to rarer events such as new diagnoses. The benchmark focuses on analyzing model performance in a few-shot setting, which is particularly relevant for large pre-trained foundation models [14] due to their ability to generalize from limited training data. To this end, the benchmark defines evaluation settings with a constrained number of training and validation examples. Specifically, for k in 1, 2, 4, 8, 12, 16, 24, 32, 48, 64, 128, the benchmark uses k positive and k negative training examples, along with k positive and k negative validation examples, to train and tune supervised classifiers. Testing is always performed on the full set of examples. The classifiers evaluated within the EHRSHOT framework include logistic regression, random forests, and GBMs [41]. Performance is reported using the AUROC and the AUPRC. For few-shot settings, we average the results over five runs with different seeds and compute bootstrapped 95% confidence intervals [31]. Macro averages are reported for each task group, and an overall macro average is provided across all groups.

4.2 EHR Text Serialization

To use LLM embedding models for representing EHR records, we serialized the records into textual formats. We limited the serialization length to 4,096 tokens (approximately 16,000 characters) due to computational constraints (see Section 4.6). Two experimental runs were conducted with serializations extending up to 8,192 tokens. The primary goal was to create a detailed and informative serialization requiring minimal preprocessing while ensuring that medically relevant information appeared early in the text. This approach mitigated truncation risks in lengthy records, preserving critical details even when older entries were omitted. To convert the visits and clinical events in the EHRSHOT dataset into text, we used the semantic information embedded in the dataset. Each clinical event was labeled using the format "ontology/code". EHRSHOT provided a set of prepared ontologies for resolving concept codes into their descriptions, which we incorporated. An analysis was performed to evaluate the use of these ontologies for all events of a subset of 200 patients across the task groups for operational outcomes and new diagnoses, covering 2,968 labels. We identified the following ontologies: Logical Observation Identifiers Names and Codes, SNOMED, RxNorm, CPT4, Domain, CARE_SITE, RxNorm Extension, Medicare Specialty, ICD10PCS, CMS Place of Service, Cancer Modifier, ICD9Proc, CVX, ICDO3, HCPCS, OMOP Extension, Condition Type. We excluded ontologies containing only a single value (Domain, Medicare Specialty, CMS Place of Service, OMOP Extension, and Condition Type). Codes of the ontologies CPT4, CARE_SITE, ICD10PCS, Cancer Modifier, CVX, and ICDO3 were not resolved with the provided ontologies. Cancer Modifier codes contained UICC cancer stages that we parsed manually. For CPT4, ICD10PCS,

and CVX we used custom mapping files that we manually added.[3] We excluded CARE_SITE and ICDO3 since we found no way to resolve these to useful descriptions.

Various approaches exist for serializing structured data, including row-by-row serialization [42], template-based methods [43], or structured data formats like JSON, HTML, and Markdown [44]. We used Markdown due to its minimal overhead and overall benefits of using a structured data input format for LLMs [45]. To harmonize dates, all timestamps were normalized relative to January 1, 2024, designated as the prediction reference time. Serialization began with patient demographics, typically the first event for each patient, where birthdates were converted into ages (in years) for simplicity. Since 65% of the dataset comprised time-series data encoded via LOINC, which we found imbalanced, we aggregated LOINC-coded events. Using the same patient subset as the ontology analysis, we identified the most frequent codes and categorized them into vital signs, body metrics, and lab values, selecting 24 key medical concepts. To avoid duplicates, we merged synonymous codes (see Table 7). The last three values of each concept are presented, and we filtered implausible values. To further enrich the text representation, we manually added default units and assessments (low, normal, high) based on standard ranges (see Table 7). Following the aggregated data, a summary of all visits was included to address the potential truncation of older visits. Events not associated with visits were then presented, using the same aggregation logic to display the last three values where applicable. Finally, a detailed chronological presentation of all visits was included, with events categorized into conditions (SNOMED, Visit, Cancer Modifier, CVX, HCPCS), medications (RxNorm, RxNorm Extension), and procedures (CPT4, ICD10PCS, ICD9Proc).

4.3 Potential Bias of Manually Defining an EHR Text Serialization

Defining an EHR serialization involved subjective decisions, which may have introduced bias. For instance, awareness of the prediction tasks could influence the prioritization of certain data elements, potentially favoring task-relevant information. To minimize this risk, we aimed to create an objectively defined serialization that encapsulates key aspects of the EHR records. Also, due to computational constraints, we evaluated only three alternatives to our final serialization for 4,096 tokens: (1) appending a list of all unique conditions at the beginning, (2) omitting the three values for the listed comments (e.g., for "Cigarette consumption" and "pH measurement, venous" in Fig. 2), and (3) combining both approaches. For GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B, the EHR serialization chosen for our experiments performed on par or better than the three alternatives. We also selected the current serialization for its simplicity and completeness of medical information. This approach avoided introducing additional entries of all unique conditions and preserved the most recent values for visit-level concepts.

[3] We downloaded CPT4 from https://gist.github.com/lieldulev/439793dc3c5a6613b661c33d71fdd185, ICD10PCS from https://hcup-us.ahrq.gov/toolssoftware/procedureicd10/procedure_icd10_archive.jsp, and CVX from https://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx.

4.4 LLM Embedding Models and Baselines

In this study, we evaluated two LLM embedding models, GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B, based on state-of-the-art decoder-only LLMs. We selected these models for their ability to handle the 4,096-token EHR serializations used in our experiments. For comparison, we also tested a smaller variant of both models. As additional baselines, we included commonly used encoder-only embedding models with smaller input sizes (512 tokens). To use them with 4,096 token inputs, the EHR serializations were split into up to eight 512-token chunks, and the resulting embeddings were averaged to generate a single representation. For all LLM embedding and smaller language models, we used mean pooling of the last layer for the final embedding [27, 46]. The LLM2Vec models used a slight variation that only incorporates the tokens that do not belong to the instruction. Below is an overview of all models:

GTE-Qwen2-7B

This LLM embedding model is based on the Qwen2-7B-Instruct LLM [32], which uses a decoder-only Transformer architecture with 28 layers, 28 attention heads, and a hidden size of 3,584[4]. It was trained with autoregressive next token prediction and converted into an embedding model using the General Text Embedding (GTE) method [30]. This conversion replaces causal attention with bidirectional attention, enabling the model to attend to both left and right contexts for token embedding, akin to BERT. Contrastive learning was applied using a mixture of private datasets to enhance embedding performance. The model also incorporates instructions tailored for embedding tasks and supports a context size of up to 32,000 tokens.

GTE-Qwen2-1.5B

A smaller variant of GTE-Qwen2-7B, this model is based on Qwen2-1.5B-Instruct, with 28 layers, 12 attention heads, and a hidden size of 1,536. It was also trained using the GTE method [30] and supports a context size of up to 32,000 tokens.

LLM2Vec-Llama-3.1-8B

This model is built upon the Llama-3.1-8B-Instruct LLM [33], which has a decoder-only Transformer architecture with 32 layers, 32 attention heads, and a hidden size of 4,096[5]. Initially trained for next-token prediction, it was converted to an embedding model using the LLM2Vec method [27]. This method adds bidirectional attention and fine-tunes the model with supervised contrastive learning on embedding tasks. The fine-tuning used curated data from the public E5 dataset [47, 48], containing approximately 1.5 million entries. The model supports task-specific instructions and supports a context size of up to 128,000 tokens.

LLM2Vec-Llama-2-1.3B

A smaller LLM2Vec variant, this model is derived from Sheared-LLaMA-1.3B [49], a pruned version of the Llama-2-7B-hf model [26]. It includes 24 layers, 16 attention

[4] Hugging Face identifier: Alibaba-NLP/gte-Qwen2-7B-instruct
[5] Hugging Face identifier: McGill-NLP/LLM2Vec-Meta-Llama-31-8B-Instruct-mntp-supervised

heads, and a hidden size of 2,048. The model follows the same LLM2Vec training methodology as the larger LLM2Vec-Llama-3.1-8B [27].

DeBERTa v3 base/large

DeBERTa v3 is an encoder-only Transformer model designed for token embeddings [50]. It improves upon its predecessor by replacing the masked language modeling objective with replaced token detection and using Gradient-Disentangled Embedding Sharing. We evaluated the base variant (12 layers, 12 attention heads, 768 hidden size) and the large variant (24 layers, 12 attention heads, 1,024 hidden size), with parameter counts of 183M and 434M, respectively[6].

BERT base/large

BERT is a well-established text embedding model using an encoder-only Transformer trained with the masked language modeling objective [15]. We included both the base (12 layers, 12 attention heads, 768 hidden size, 110M parameters) and large (24 layers, 16 attention heads, 1,024 hidden size, 340M parameters) variants as benchmarks[7]. While not state-of-the-art, BERT models remain widely used in embedding tasks.

Bio_ClinicalBERT

This model builds on BERT-Base, further fine-tuned on biomedical [51] and clinical data [52]. It is a widely adopted embedding model for medical text and was included as a baseline for comparison[8].

CLMBR-T-Base

CLMBR-T-Base is a specialized EHR foundation model trained on 2.57 million de-identified EHRs from Stanford Medicine with autoregressive next code prediction [13, 31]. It uses gated recurrent units and has 12 layers and a hidden dimension of 768[9]. The model has 141M parameters and allows for a context window of 496 codes. CLMBR-T-Base has demonstrated consistent improvements over count-based baselines for a variety of clinical prediction tasks [31]. It serves as a main baseline for our experiments to test specialized EHR models against general-purpose text embedding models for representing EHR records.

LLM Embedding Model and CLMBR-T-Base

To test whether the LLM embedding models and the EHR foundation model learn orthogonal information, we combined both models for the prediction. To this end, we simply appended both embeddings. The resulting embeddings have dimensions 4,352 (GTE-Qwen2-7B) and 4,864 (LLM2Vec-Llama-3.1-8B).

Count-based Model

Count models have proven to be strong baselines for EHR prediction tasks [6, 12, 13]. The basic idea is to encode all EHR events of a patient in a single vector where

[6] Hugging Face identifiers: microsoft/deberta-v3-base,large
[7] Hugging Face identifiers: google-bert/bert-base,large-uncased
[8] Hugging Face identifier: emilyalsentzer/Bio_ClinicalBERT
[9] Hugging Face identifier: StanfordShahLab/clmbr-t-base

each entry represents the number of occurrences of a medical concept. We used the counts baseline introduced in [31] that further extends this approach with ontology expansion, enriching the vectors with parent and child concepts.

Based on the embeddings or the counts vectors generated by the methods described above, a classification head was trained and validated for each prediction task. For the embedding models we used a logistic regression head. For the count-based model, we used a GBM [41], which proved superior [31]. We adopted the parameter tuning of the classification heads from the EHRSHOT benchmark to ensure comparability of results.

4.5 Instructions for LLM Embedding Models

The GTE and LLM2Vec models used instruction-tuned embeddings, requiring task-specific prompts. Hence, we added simple instructions for each prediction task based on their respective instruction templates. For instance, for prediction of anemia, we added "Given a patient's electronic healthcare record (EHR) in Markdown format, retrieve relevant passages that answer the query: has the patient anemia". The existing EHRSHOT benchmark encoded the EHRs of the same patient and the identical prediction times only once for efficiency reasons. However, to support task-specific instructions, we had to change this default behavior, leading to 1,161,412 instead of 406,379 EHRs to encode, resulting in longer processing times. The difference between 1,161,412 labels used in our experiments and the total number of labels of 1,178,665 (Table 2) is because some labels share the same task and prediction time and are therefore merged. For assignment of new diagnoses tasks in UKB, the same prompt was used for all diseases to reduce computational resources. We list all instructions in Table 8 and perform ablations to test the effect of the instructions.

4.6 Computational Setup and Running Times

All experiments were conducted on the Charité High-Performance Cluster using Nvidia A100 GPUs with 80 GB memory, configured with one, four, or eight GPUs (DGX systems). Running the GTE-Qwen2-7B model with our serialization truncated at 4,096 tokens on an 8-GPU DGX system required approximately 20 hours. For the LLM2Vec-Llama-3.1-8B model, runtime errors occurred during multi-GPU experiments with the full dataset. These issues were resolved by splitting the data into smaller batches, which introduced additional overhead. Additionally, we optimized the LLM2Vec code by removing an initial text pruning step that took approximately eight hours when using the full data. Using this modified setup, the LLM2Vec-Llama-3.1-8B model took approximately 30 hours to complete on eight A100 GPUs. The calculation of the embeddings for the 387,464 UKB patients took approximately 35 hours on a single GPU per task for the LLM-based methods. For CLMBR-T-Base, the calculation was faster, only requiring around eight hours.

4.7 Performance Results on EHRSHOT Prediction Tasks and Few-Shot Setting

Following the EHRSHOT benchmark, we evaluated all models across 15 prediction tasks under various few-shot settings. The benchmark includes a modular pipeline designed to execute key tasks, with the flexibility to optionally use a Slurm cluster for distributed execution. Running all steps within this pipeline ensures full reproducibility of results. Step four of the pipeline, which generates EHR representations with CLMBR-T-Base and the count-based model, was extended to incorporate our method for creating language model-based EHR representations. This adaptation allowed us to reuse significant portions of the existing code, including the task evaluation framework. Additionally, we implemented new functionality for EHR serialization and slightly modified other steps of the benchmark to accommodate our experimental setup. For instance, the label creation process was adjusted (step three) to enable task-specific instructions for the LLM embedding models. All modifications have been documented and can be tracked in our public GitHub repository.

4.8 External Validation on UK Biobank

External validation was performed using data from the UKB, a large-scale prospective cohort study comprising 502,489 UK participants recruited between 2006 and 2010, with a median follow-up of 13.8 years. We used linked EHR data from primary care (General Practitioner, GP) and secondary care (Hospital Episode Statistics, HES), providing information on diagnoses, procedures, and prescriptions.

Initial data preprocessing, including cleaning, feature extraction, missing value imputation, and endpoint selection followed the methodology described in [34]. All health records were mapped to the OMOP CDM using mapping tables provided by the UKB, SNOMED International, and the OHDSI community for mapping concepts from the provider and country-specific non-standard vocabularies to OMOP standard vocabularies. Participants lacking any recorded GP or HES events either before or after their recruitment date were excluded, resulting in a validation cohort of 387,464 individuals. Diagnostic codes were mapped to Phecodes X [53, 54] (derived directly from source ICD-10 or by mapping from SNOMED to ICD-10 codes and subsequently to Phecodes X) primarily for standardized endpoint definition and cohort selection. To avoid redundancy with source codes used as features, the Phecodes themselves were excluded during the creation of patient sequences for model input. Due to significant challenges in mapping and harmonizing UKB laboratory values [55], and to ensure comparability across models and tasks, laboratory data were excluded entirely, differing from the use of binary labels in [34]. The final feature set comprised conditions (SNOMED, CVX), medications (RxNorm), and procedures (SNOMED).

Given that UKB represents a general population cohort, we adjusted the proposed prediction tasks to define relevant longitudinal health trajectory tasks. These tasks included: (1) prediction of all-cause hospitalization within the next year (operational outcomes), (2) prediction of incident diagnoses for a set of selected conditions (assignment of new diagnoses), and (3) prediction of all-cause mortality (mortality prediction). The selection of incident diseases for the assignment of new diagnoses task

group largely followed [34], focusing on common conditions, diseases lacking established risk stratification tools, and specific cardiovascular conditions. From the initial 24 endpoints proposed in [34], we treated all-cause death as a separate task. For the assignment of new diagnoses and mortality tasks, patients with a diagnosis of the respective endpoint recorded prior to their UKB recruitment date were excluded to prevent data leakage; this exclusion was not applied to the hospitalization task due to high incidence. Following task-specific exclusions, the final sample size available for analysis ranged from 300,344 to 359,250 patients, depending on the prediction task (see Table 6 for details). For all three task groups, the prediction window was set to one year after the prediction year.

The pre-trained CLMBR-T-Base model operates with a fixed vocabulary (65,536 total, 26,249 unique codes). To use this model, we mapped the 50,702 unique medical codes (SNOMED CT, RxNorm, CVX) present in our processed UKB cohort to the CLMBR-T-Base vocabulary. This mapping followed the steps similar to those implemented in the FEMR package[10], involving direct code matching where possible, supplemented by indirect mapping via the OHDSI ATHENA vocabulary using 'Maps to' relationships and inclusion of ancestor concepts. This process successfully mapped 7,969 unique UKB codes to the CLMBR-T-Base vocabulary (in format ontology/-code), which means that the remaining codes were implicitly excluded from the input sequences for this model. After creating patient timelines, adding information about birth date, ethnicity, and visits, the data was converted into the MEDS standard [56], from which the embeddings were calculated. Additionally, UKB ethnicities were converted to the ethnicity groups used by CLMBR-T-Base. For transforming the patient information into embeddings, we mainly followed the code provided by FEMR (mainly the convert_patient function[11]) with minor modifications to enable batch processing.

Similar to the approach of Wornow et al. [31], we performed an ontology expansion to improve the counts baseline by adding direct parent and grandparent concepts. This increased the unique code count from 51,677 to 69,850, aiming to capture broader hierarchical relationships between medical concepts. However, the resulting high dimensionality posed memory challenges for a standard counts matrix. To mitigate this, we implemented feature selection by removing codes present in less than 50 individuals (<0.01% of the cohort), reducing the feature space to 24,509 unique codes. To ensure a more comparable baseline to the other models, which implicitly incorporate patient context, we explicitly added normalized age at prediction time and coded sex as additional features to the feature matrix used for the count-based baseline.

Similar to before, all approaches were evaluated using a five-fold cross-validation. Within each fold, hyperparameters were optimized on a dedicated validation set of similar sample size to that of the training partition in that fold. For each fold, patients were randomly assigned to train, validation, and test sets. The test set for each task within a fold was constructed by selecting all available positive cases (capped to 10,000), and an equal number of negative cases after allocating patients to the train/validation sets. To specifically assess performance in low-data scenarios, mimicking potential clinical applications with limited examples, we again evaluated models using incrementally

[10] https://github.com/som-shahlab/femr/tree/0ebf454303091fa83c9e4023e7db9ba9b68af09a
[11] https://huggingface.co/StanfordShahLab/clmbr-t-base

larger subsets of the training data, with the final evaluation point capped at a sample size of 256 patients from the train set, even when larger test sets were available.

4.9 Ablations of EHR Serialization

To better understand the contribution of various components in the EHR serialization process to the performance of the LLM embedding models, we conducted a series of ablation experiments. First, we assessed the impact of task-specific instructions by replacing them with a generic instruction (Table 8) or removing them entirely to determine their influence on the final embeddings. Subsequently, we performed additional ablations by systematically excluding specific components of the serialization. These included demographic data, aggregated LOINC codes, and both visit summaries and detailed entries. Furthermore, we examined the effect of removing specific fields from the detailed visit entries, such as conditions, medications, and procedures. Throughout these experiments, the rest of the pipeline was kept consistent to isolate the effects of the removed components. This approach allowed us to identify which parts of the serialization process were most critical for generating effective EHR representations, providing insights into how these models use structured medical data for prediction tasks.

4.10 Effect of Different Time Windows

To examine the influence of recency on predictive performance, we varied the time window preceding the prediction date used during EHR text serialization. We evaluated GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B across six intervals: one day, one week (7 days), one month (30 days), one year (365 days), three years (1,095 days), and full history. For each window, only events occurring within the specified interval before the prediction time were included. Hence, only data from the respective time window contributed to the aggregated information and the visit data in the EHR serialization. All other aspects of the serialization, including structure, formatting, and instruction prompts remained unchanged to isolate the effect of the temporal window.

4.11 Effect of Different Context Sizes

We investigated the impact of varying context sizes in the LLM embedding models. We wanted to determine whether encoding information from older visits enhances prediction performance and whether longer inputs might dilute critical details, such as laboratory values, in the final embeddings. Specifically, we evaluated GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B models with input token limits of 512, 1,024, 2,048, and 8,192 tokens. Input tokens exceeding these thresholds were discarded. Due to the design of our EHR serialization process, additional input tokens primarily consisted of medical concepts from past visits. By testing these varying context sizes, we aimed to assess the balance between capturing historical medical data and preserving the clarity of high-priority information within the embeddings.

4.12 Effect of Chunked Contexts

To further explore whether LLM embedding models effectively interpret their entire input, we compared the performance of models processing complete inputs versus segmented (chunked) inputs. For this, the 4,096-token inputs were divided into smaller chunks of sizes 512, 1,024, and 2,048 tokens[12]. Separate embeddings were generated for each chunk, and a final embedding was obtained by averaging the embeddings of all chunks. This approach aligns with the mean pooling applied to the last layer that we used to compute embeddings. In this setup, task-specific instructions, placed at the beginning of the inputs, were included only in the first chunk. This design allowed us to evaluate the models' ability to contextualize task instructions and whether chunking affects the overall performance.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 437531118 – SFB 1470 and German Federal Ministry of Education and Research (BMBF) Project-ID 01ZZ2317G.

Declaration of interest

The authors declare no competing interests.

Data availability

The EHRSHOT data is available through gated access via https://doi.org/10.57761/0gv9-nd83. UK Biobank data, including all linked routine health records, are publicly available to bona fide researchers upon application at http://www.ukbiobank.ac.uk/using-the-resource. In this study, only primary care data not subject to the Government's Control of Patient Information (COPI) notice was used (UK Biobank Category 3000).

Code availability

All extensions and experiments conducted for this paper have been made publicly available via our GitHub repository: https://github.com/stefanhgm/ehrshot-benchmark.

References

[1] Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S.: Big data in healthcare: management, analysis and future prospects. Journal of Big Data **6**(1), 54 (2019) https://doi.org/10.1186/s40537-019-0217-0 . Accessed 2025-02-15

[12] In our implementation, we used these chunk sizes decreased by eight tokens to cater for special tokens that might be added by the implementations of the language models.

[2] Ahsan, H., McInerney, D.J., Kim, J., Potter, C., Young, G., Amir, S., Wallace, B.C.: Retrieving Evidence from EHRs with LLMs: Possibilities and Challenges. Proceedings of machine learning research **248**, 489–505 (2024). Accessed 2025-02-15

[3] Rajkomar, A., Dean, J., Kohane, I.: Machine Learning in Medicine. New England Journal of Medicine **380**(14), 1347–1358 (2019) https://doi.org/10.1056/NEJMra1814259 . Accessed 2020-06-05

[4] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. arXiv. arXiv:2104.08836 [cs] (2021). https://doi.org/10.48550/arXiv.2104.08836 . http://arxiv.org/abs/2104.08836 Accessed 2024-07-23

[5] Golas, S.B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., Jethwani, K.: A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Medical Informatics and Decision Making **18**(1), 44 (2018) https://doi.org/10.1186/s12911-018-0620-z . Accessed 2024-12-14

[6] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J.: Scalable and accurate deep learning with electronic health records. npj Digital Medicine **1**(1), 1–10 (2018) https://doi.org/10.1038/s41746-018-0029-1 . Publisher: Nature Publishing Group. Accessed 2024-12-15

[7] Lauritsen, S.M., Kalør, M.E., Kongsgaard, E.L., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B.: Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artificial Intelligence in Medicine **104**, 101820 (2020) https://doi.org/10.1016/j.artmed.2020.101820 . Accessed 2024-12-15

[8] Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., Bühlmann, P., Borgwardt, K.: Predicting sepsis using deep learning across international sites: a retrospective development and validation study. eClinicalMedicine **62** (2023) https://doi.org/10.1016/j.eclinm.2023.102124 . Publisher: Elsevier. Accessed 2024-12-15

[9] Thorsen-Meyer, H.-C., Nielsen, A.B., Nielsen, A.P., Kaas-Hansen, B.S., Toft, P., Schierbeck, J., Strøm, T., Chmura, P.J., Heimann, M., Dybdahl, L., Spangsege, L., Hulsen, P., Belling, K., Brunak, S., Perner, A.: Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient

records. The Lancet. Digital Health **2**(4), 179–191 (2020) https://doi.org/10.1016/S2589-7500(20)30018-2

[10] Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T., Schneeweiss, S.: Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Network Open **3**(1), 1918962 (2020) https://doi.org/10.1001/jamanetworkopen.2019.18962 . Accessed 2024-12-15

[11] Kim, E., Rubinstein, S.M., Nead, K.T., Wojcieszynski, A.P., Gabriel, P.E., Warner, J.L.: The Evolving Use of Electronic Health Records (EHR) for Research. Seminars in Radiation Oncology **29**(4), 354–361 (2019) https://doi.org/10.1016/j.semradonc.2019.05.010 . Accessed 2025-02-15

[12] 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 Digital Medicine **4**(1), 1–13 (2021) https://doi.org/10.1038/s41746-021-00455-y . Publisher: Nature Publishing Group. Accessed 2024-12-15

[13] Steinberg, E., Jung, K., Fries, J.A., Corbin, C.K., Pfohl, S.R., Shah, N.H.: Language models are an effective representation learning technique for electronic health record data. Journal of Biomedical Informatics **113**, 103637 (2021) https://doi.org/10.1016/j.jbi.2020.103637 . Accessed 2024-06-12

[14] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S.v., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models. arXiv. arXiv:2108.07258 [cs] (2022). https://doi.org/10.48550/arXiv.2108.07258 . http://arxiv.org/abs/2108.07258 Accessed 2024-12-16

[15] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186 (2019) https://doi.org/10.18653/v1/N19-1423 . Accessed 2022-04-22

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. Technical report, OpenAi, 24 (2019)

[17] Odgaard, M., Klein, K.V., Thysen, S.M., Jimenez-Solem, E., Sillesen, M., Nielsen, M.: CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT. arXiv. arXiv:2404.15201 [cs] (2024). https://doi.org/10.48550/arXiv.2404.15201 . http://arxiv.org/abs/2404.15201 Accessed 2024-12-16

[18] Pang, C., Jiang, X., Kalluri, K.S., Spotnitz, M., Chen, R., Perotte, A., Natarajan, K.: CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. arXiv. arXiv:2111.08585 [cs] (2021). https://doi.org/10.48550/arXiv.2111.08585 . http://arxiv.org/abs/2111.08585 Accessed 2024-12-16

[19] Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: BEHRT: Transformer for Electronic Health Records. Scientific Reports **10**(1), 7155 (2020) https://doi.org/10.1038/s41598-020-62922-y . Publisher: Nature Publishing Group. Accessed 2024-12-16

[20] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., Launay, J.: The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. Advances in Neural Information Processing Systems **36**, 79155–79172 (2023). Accessed 2025-05-01

[21] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 140–54851405551 (2020)

[22] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large Language Models are Few-Shot Clinical Information Extractors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 1998–2022 (2022) https://doi.org/10.18653/v1/2022.emnlp-main.130 . Accessed 2024-04-12

[23] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C.P., Hom, J., Gatidis, S., Pauly, J., Chaudhari, A.S.: Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine, 1–9 (2024) https://doi.org/10.1038/s41591-024-02855-5 . Publisher: Nature Publishing Group. Accessed 2024-04-11

[24] Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., Sontag, D.: TabLLM: Few-shot Classification of Tabular Data with Large Language Models. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, 5549–5581 (2023). Accessed 2024-04-12

[25] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020). Accessed 2024-05-22

[26] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288 (2023). Accessed 2024-01-25

[27] BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., Reddy, S.: LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. (2024). https://openreview.net/forum?id=IW1PR7vEBf Accessed 2025-03-09

[28] Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv. arXiv:2405.17428 [cs] (2024). http://arxiv.org/abs/2405.17428 Accessed 2024-08-02

[29] Muennighoff, N., Su, H., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., Kiela, D.: Generative Representational Instruction Tuning. arXiv. arXiv:2402.09906 [cs] (2024). http://arxiv.org/abs/2402.09906 Accessed 2024-08-01

[30] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards General Text Embeddings with Multi-stage Contrastive Learning. arXiv. arXiv:2308.03281 [cs] (2023). https://doi.org/10.48550/arXiv.2308.03281 . http://arxiv.org/abs/2308.03281 Accessed 2024-08-02

[31] Wornow, M., Thapa, R., Steinberg, E., Fries, J.A., Shah, N.: EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. (2023).

https://openreview.net/forum?id=CsXC6IcdwI Accessed 2024-06-12

[32] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z.: Qwen2 Technical Report. arXiv. arXiv:2407.10671 [cs] (2024). https://doi.org/10.48550/arXiv.2407.10671 . http://arxiv.org/abs/2407.10671 Accessed 2025-01-08

[33] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., Linde, J.v.d., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., Maaten, L.v.d., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., Oliveira, L.d., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag,

Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V.,

Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The Llama 3 Herd of Models. arXiv. arXiv:2407.21783 [cs] (2024). https://doi.org/10.48550/arXiv.2407.21783 . http://arxiv.org/abs/2407.21783 Accessed 2025-01-08

[34] Steinfeldt, J., Wild, B., Buergel, T., Pietzner, M., Belzen, J., Vauvelle, A., Hegselmann, S., Denaxas, S., Hemingway, H., Langenberg, C., Landmesser, U., Deanfield, J., Eils, R.: Medical history predicts phenome-wide disease onset and enables the rapid response to emerging health threats. Nature Communications **16**(1), 585 (2025) https://doi.org/10.1038/s41467-025-55879-x . Publisher: Nature Publishing Group. Accessed 2025-02-15

[35] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine **12**(3), 1001779 (2015) https://doi.org/10.1371/journal.pmed.1001779 . Publisher: Public Library of Science. Accessed 2025-05-01

[36] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., Marchini, J.: The UK Biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018) https://doi.org/10.1038/s41586-018-0579-z . Number: 7726 Publisher: Nature Publishing Group. Accessed 2022-11-21

[37] Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., Miceli, M., Kim, N.C., Orillac, C., Schnurman, Z., Livia, C., Weiss, H., Kurland, D., Neifert, S., Dastagirzada, Y., Kondziolka, D., Cheung, A.T.M., Yang, G., Cao, M., Flores, M., Costa, A.B., Aphinyanaphongs, Y., Cho, K., Oermann, E.K.: Health system-scale language models are all-purpose prediction engines. Nature **619**(7969), 357–362 (2023) https://doi.org/10.1038/s41586-023-06160-y . Publisher: Nature Publishing Group. Accessed 2024-07-31

[38] Lehne, M., Sass, J., Essenwanger, A., Schepers, J., Thun, S.: Why digital medicine depends on interoperability. npj Digital Medicine **2**(1), 1–5 (2019) https://doi.org/10.1038/s41746-019-0158-1 . Publisher: Nature Publishing Group. Accessed 2025-02-15

[39] Kirchler, M., Ferro, M., Lorenzini, V., FinnGen, Lippert, C., Ganna, A.: Large

language models improve transferability of electronic health record-based predictions across countries and coding systems. medRxiv. Pages: 2025.02.03.25321597 (2025). https://doi.org/10.1101/2025.02.03.25321597 . https://www.medrxiv.org/content/10.1101/2025.02.03.25321597v1 Accessed 2025-02-15

[40] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023) https://doi.org/10.1038/s41586-023-05881-4 . Publisher: Nature Publishing Group. Accessed 2025-02-15

[41] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems **30** (2017). Accessed 2022-08-01

[42] Liu, Q., Chen, B., Guo, J., Ziyadi, M., Lin, Z., Chen, W., Lou, J.-G.: TAPEX: Table Pre-training via Learning a Neural SQL Executor. (2021). https://openreview.net/forum?id=O50443AsCP Accessed 2025-01-18

[43] Li, P., He, Y., Yashar, D., Cui, W., Ge, S., Zhang, H., Rifinski Fainman, D., Zhang, D., Chaudhuri, S.: Table-GPT: Table Fine-tuned GPT for Diverse Table Tasks. Proceedings of the ACM on Management of Data **2**(3), 1–28 (2024) https://doi.org/10.1145/3654979 . Accessed 2025-01-18

[44] Dong, H., Zhao, J., Tian, Y., Xiong, J., Zhou, M., Lin, Y., Cambronero, J., He, Y., Han, S., Zhang, D.: Encoding Spreadsheets for Large Language Models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 20728–20748. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1154 . https://aclanthology.org/2024.emnlp-main.1154/ Accessed 2025-01-18

[45] Sui, Y., Zhou, M., Zhou, M., Han, S., Zhang, D.: Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, pp. 645–654. ACM, Merida Mexico (2024). https://doi.org/10.1145/3616855.3635752 . https://dl.acm.org/doi/10.1145/3616855.3635752 Accessed 2025-01-18

[46] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410 . https://aclanthology.org/D19-1410 Accessed 2024-06-12

[47] Springer, J.M., Kotha, S., Fried, D., Neubig, G., Raghunathan, A.: Repetition Improves Language Model Embeddings. arXiv. arXiv:2402.15449 [cs] (2024). https://doi.org/10.48550/arXiv.2402.15449 . http://arxiv.org/abs/2402.15449 Accessed 2024-08-01

[48] Wang, A., Liu, C., Yang, J., Weng, C.: Fine-tuning Large Language Models for Rare Disease Concept Normalization. bioRxiv. Pages: 2023.12.28.573586 Section: New Results (2024). https://doi.org/10.1101/2023.12.28.573586 . https://www.biorxiv.org/content/10.1101/2023.12.28.573586v3 Accessed 2024-07-26

[49] Xia, M., Gao, T., Zeng, Z., Chen, D.: Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning. arXiv. arXiv:2310.06694 [cs] (2024). https://doi.org/10.48550/arXiv.2310.06694 . http://arxiv.org/abs/2310.06694 Accessed 2025-01-19

[50] He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. (2022). https://openreview.net/forum?id=sE7-XhLxHA Accessed 2025-01-19

[51] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020) https://doi.org/10.1093/bioinformatics/btz682 . Accessed 2022-04-26

[52] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., McDermott, M.: Publicly Available Clinical BERT Embeddings. In: Rumshisky, A., Roberts, K., Bethard, S., Naumann, T. (eds.) Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/W19-1909 . https://aclanthology.org/W19-1909/ Accessed 2025-01-19

[53] Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C., Theodoratou, E., Wei, W.-Q.: Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Medical Informatics **7**(4), 14325 (2019) https://doi.org/10.2196/14325 . Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada. Accessed 2025-05-01

[54] Wei, W.-Q., Bastarache, L.A., Carroll, R.J., Marlo, J.E., Osterman, T.J., Gamazon, E.R., Cox, N.J., Roden, D.M., Denny, J.C.: Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLOS ONE **12**(7), 0175508 (2017) https://doi.org/10.1371/journal.pone.0175508 . Publisher: Public Library of Science. Accessed 2025-05-01

[55] Denaxas, S., Shah, A.D., Mateen, B.A., Kuan, V., Quint, J.K., Fitzpatrick, N.,

Torralbo, A., Fatemifar, G., Hemingway, H.: A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. JAMIA Open **3**(4), 545–556 (2020) https://doi.org/10.1093/jamiaopen/ooaa047 . Accessed 2025-05-01

[56] Arnrich, B., Choi, E., Fries, J.A., McDermott, M.B.A., Oh, J., Pollard, T.J., Shah, N., Steinberg, E., Wornow, M., Water, R.: Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health. In: ICLR 2024 Workshop on Learning from Time Series For Health (2024). https://openreview.net/forum?id=IsHy2ebjIG

A Appendix

A.1 Additional Experimental Details

Table 6 UK Biobank Prediction Tasks Overview. Overview of clinical prediction tasks of UKB spanning three different task groups. Total Labels refers to the number of patients after filtering out patients with records prior to prediction. The prediction window for all tasks is 1 year (similar to *Assignment of New Diagnoses* Task as in [31]).

Attribute	**Total Labels (Positive)**
Mortality prediction	
Mortality (Death)	359245 (802)
Operational Outcomes	
Hospitalization	359250 (68505)
Assignment of New Diagnoses	
Cardiac arrest	358963 (87)
Abdominal aortic aneurysm	358926 (73)
Parkinson's disease (Primary)	358729 (86)
Aortic stenosis	358384 (114)
Rheumatic fever and chronic rheumatic heart diseases	358064 (154)
Mitral valve insufficiency	357949 (186)
Endocarditis	357876 (79)
Suicide ideation and attempt or self harm	357652 (175)
Cerebral infarction [Ischemic stroke]	357180 (310)
Pulmonary embolism	357158 (248)
Atrial fibrillation	357154 (306)
Heart failure	356657 (435)
Rheumatoid arthritis	356029 (316)
Chronic obstructive pulmonary disease [COPD]	354513 (923)
Psoriasis	353400 (410)
Chronic kidney disease	353396 (1001)
Pneumonia	353232 (802)
Myocardial infarction [Heart attack]	351970 (619)
Diabetes mellitus	344185 (1841)
Ischemic heart disease	341574 (1788)
Anemia	343235 (1814)
Back pain	300529 (3933)
Hypertension	300334 (7653)

Table 7 Semantic Codes for Aggregated Concepts in EHR Serialization. We aggregated time-series data encoded via Logical Observation Identifiers Names and Codes (LOINC) concepts and identified the most frequent concepts from which we selected 24 key medical concepts. To reduce duplicate information, we merged synonymous semantic codes. The primary LOINC codes is presented first in the column Semantic Codes followed by identified duplicates. We also defined a unit, minimum and maximum allowed values for filtering, a normal range to classify values in low, normal, and high, and a formatting strategy to create our EHR serialization.

Medical Concept	Semantic Codes	Unit	Min-Max Range	Normal Range	Formatting
Recent Body Metrics					
Body weight	LOINC/29463-7	oz	350-10000		One decimal
Body height	LOINC/8302-2	inch	5-100		One decimal
Body mass index / BMI	LOINC/39156-5	kg/m2	10-100	18.5-24.9	One decimal
Body surface area	LOINC/8277-6, SNOMED/301898006	m2	0.1-10		Two decimals
Recent Vital Signs					
Heart rate	LOINC/8867-4, SNOMED/364075005, SNOMED/78564009	bpm	5-300	60-100	Integer
Systolic blood pressure	LOINC/8480-6, SNOMED/271649006	mmHg	20-300	90-140	Integer
Diastolic blood pressure	LOINC/8462-4, SNOMED/271650006	mmHg	20-300	60-90	Integer
Body temperature	LOINC/8310-5	°F	80-120	95-100.4	One decimal
Respiratory rate	LOINC/9279-1	breaths/min	1-100	12-18	Integer
Oxygen saturation	LOINC/LP21258-6	%	1-100	95-100	Integer
Recent Lab Results					
Hemoglobin	LOINC/718-7, SNOMED/271026005, SNOMED/441689006	g/dL	1-20	12-17	One decimal
Hematocrit	LOINC/4544-3, LOINC/20570-8, LOINC/48703-3, SNOMED/28317006	%	10-100	36-51	Integer
Erythrocytes	LOINC/789-8, LOINC/26453-1	106/uL	1-10	4.2-5.9	Two decimals
Leukocytes	LOINC/20584-9, LOINC/6690-2	103/uL	1-100	4-10	One decimal
Platelets	LOINC/777-3, SNOMED/61928009	103/uL	10-1000	150-350	Integer
Sodium	LOINC/2951-2, LOINC/2947-0, SNOMED/25197003	mmol/L	100-200	136-145	Integer
Potassium	LOINC/2823-3, SNOMED/312468003, LOINC/6298-4, SNOMED/59573005	mmol/L	0.1-10	3.5-5.0	One decimal
Chloride	LOINC/2075-0, SNOMED/104589004, LOINC/2069-3	mmol/L	50-200	98-106	Integer
Carbon dioxide, total	LOINC/2028-9	mmol/L	10-100	23-28	Integer
Calcium	LOINC/17861-6, SNOMED/271240001	mg/dL	1-20	9-10.5	One decimal
Glucose	LOINC/2345-7, SNOMED/166900001, LOINC/2339-0, SNOMED/33747003, LOINC/14749-6	mg/dL	10-1000	70-100	Integer
Urea nitrogen	LOINC/3094-0, SNOMED/105011006	mg/dL	1-200	8-20	Integer
Creatinine	LOINC/2160-0, SNOMED/113075003	mg/dL	0.1-10	0.7-1.3	One decimal
Anion gap	LOINC/33037-3, LOINC/41276-7, SNOMED/25469001	mmol/L	-20-50	3-11	Integer

Table 8 Instructions for LLM-Embedding Models. The LLM-embedding models were trained using instructions; hence, we also defined simple task-specific prompts for each of the 15 clinical prediction tasks. Each prompt is prepended by the prefix given below, containing a general task description. The three tasks used for the external validation on UKB can be seen on the bottom.

Task	Prompt
Prefix (for all tasks)	Given a patient's electronic healthcare record (EHR) in Markdown format, retrieve relevant passages that answer the query:
EHRSHOT	
Long Length of Stay	will the patient stay in the hospital for more than 7 days
30-day Readmission	will the patient be readmitted to the hospital within 30 days
ICU Transfer	will the patient be transferred to the intensive care unit
Thrombocytopenia	has the patient thrombocytopenia
Hyperkalemia	has the patient hyperkalemia
Hypoglycemia	has the patient hypoglycemia
Hyponatremia	has the patient hyponatremia
Anemia	has the patient anemia
Hypertension	has the patient hypertension
Hyperlipidemia	has the patient hyperlipidemia
Pancreatic Cancer	has the patient pancreatic cancer
Celiac	has the patient celiac disease
Lupus	has the patient lupus
Acute MI	has the patient an acute myocardial infarction
Chest X-Ray Findings	what are the chest x-ray findings of the patient
Generic (ablation)	what are the key clinical features of the patient to predict future medical events
UK Biobank	
Mortality prediction	will the patient die in the next 1 year
Hospitalization	will the patient be hospitalized in the next 1 year
Assignment of New Diagnoses	has the patient a medical condition

A.2 Effect of Different Time Windows

To assess how the recency of medical history affects prediction performance, we varied the temporal window preceding the prediction time during EHR text serialization. The results are summarized in Fig. 6 (left) and detailed in Table 9. GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B exhibited similar overall trends, though GTE-Qwen2-7B performed better with larger time windows. The one-week and one-month windows consistently yielded the highest AUROC scores, suggesting that recent clinical data carries the most predictive value for downstream tasks. Notably, the 24-hour window produced better results than using the full historical record. This trend suggests that recent clinical data often holds the most predictive value, while older information may introduce irrelevant or outdated signals. Among the two models, GTE-Qwen2-7B demonstrated stronger performance with longer time windows, indicating a higher capacity to extract relevant features from extended clinical narratives.

A.3 Effect of Different Context Sizes

We further tested the impact of varying context sizes on the performance of the LLM embedding models. As summarized in Fig. 6 (right) and Table 10, the two models exhibited distinct behaviors. For GTE-Qwen2-7B, the best overall performance was achieved with a context of 4,096 tokens, with only a slight decline observed when using 2,048 or 1,024 tokens. This suggests that the most relevant information is contained within the first 1,024 tokens and that the model can accurately extract these details even when presented within a longer context. In contrast, LLM2Vec-Llama-3.1-8B showed a marked improvement with a context size of 2,048 tokens, indicating that the model struggles to effectively handle longer sequences up to 4,096 tokens. Both models experienced a significant drop in performance when limited to only 512 tokens, highlighting that such a short context omits crucial information; a noteworthy point given that many existing language models are constrained to 512 input tokens [37]. Furthermore, extending the context to 8,192 tokens resulted in decreased performance for both models, implying limitations in extracting relevant information from very long inputs. This decline may be partially attributed to the mean pooling of the last hidden layers used for generating embeddings, which can dilute the impact of important token-level information; an alternative strategy, such as incorporating an additional attention mechanism, might help mitigate this issue.

A.4 Effect of Chunked Contexts

To investigate whether the models can process the full 4,096-token context cohesively, we conducted an experiment in which the serialized EHR input was divided into chunks of 512, 1,024, and 2,048 tokens. For each chunk, separate embeddings were generated and then averaged to create a final representation (see Table 11). For GTE-Qwen2-7B, the performance decrease with smaller chunks was relatively modest, indicating that the information contained within the full 4,096-token input remains effectively used even when segmented. Notably, using 512-token chunks yielded an overall AUROC performance of 0.731 (0.706-0.756), compared to 0.671 (0.647-0.695) when processing a contiguous 512-token context, which suggests that chunking

can mitigate input constraints. In contrast, LLM2Vec-Llama-3.1-8B demonstrated improved performance with chunked inputs, consistent with its behavior on shorter context sizes. This enhancement was primarily driven by better lab value prediction, implying that LLM2Vec-Llama-3.1-8B is particularly effective when processing inputs of up to 2,048 tokens.

A.5 Detailed Ablation Results for EHRSHOT

Table 9 Performance of LLM-Embedding Models Across Time Windows. Macro averaged area under receiver operating characteristic curve (AUROC) performance and 95% confidence intervals for task groups and macro-averaged performance across all task groups for different time windows of the LLM-embedding models. Both GTE-Qwen2-7B and LLM2Vec-Llama-3.1-8B demonstrate improved overall performance with shorter time windows. The highest performance is observed when using data from up to one month prior to the prediction time.

Model	Operational Outcomes	Anticipating Lab Test Results	Assignment of New Diagnosis	Anticipating Chest X-ray Findings	Macro Avg. Across Task Groups
GTE-Qwen2-7B					
Full patient history	0.771 .747-.795	0.865 .858-.873	0.716 .675-.757	0.666 .653-.680	0.755 .730-.780
3 years (1,095 days)	0.774 .749-.798	0.867 .859-.874	0.724 .685-.763	0.684 .672-.697	0.762 .738-.787
1 year (365 days)	0.792 .769-.814	0.867 .860-.874	0.702 .668-.737	0.699 .686-.711	0.765 .743-.787
1 month (30 days)	0.805 .784-.826	0.862 .855-.869	0.722 .673-.770	0.704 .693-.715	0.773 .746-.801
1 week (7 days)	0.809 .787-.830	0.856 .849-.863	0.719 .676-.762	0.687 .675-.698	0.768 .743-.792
1 day	0.813 .793-.833	0.826 .819-.833	0.743 .693-.793	0.657 .646-.669	0.760 .732-.788
LLM2Vec-Llama-3.1-8B					
Full patient history	0.753 .727-.778	0.778 .768-.789	0.728 .682-.774	0.678 .665-.692	0.734 .707-.762
3 years (1,095 days)	0.767 .742-.792	0.780 .769-.790	0.708 .655-.761	0.690 .678-.703	0.736 .706-.767
1 year (365 days)	0.780 .757-.803	0.791 .781-.801	0.711 .660-.762	0.708 .696-.720	0.748 .719-.776
1 month (30 days)	0.798 .776-.820	0.821 .812-.829	0.733 .685-.781	0.715 .704-.725	0.767 .739-.794
1 week (7 days)	0.805 .783-.827	0.839 .832-.847	0.712 .664-.760	0.692 .681-.703	0.762 .735-.789
1 day	0.799 .776-.821	0.825 .818-.832	0.760 .717-.803	0.658 .646-.670	0.761 .735-.786

Table 10 Performance of LLM-Embedding Models Across Context Sizes. Macro averaged area under receiver operating characteristic curve (AUROC) performance and 95% confidence intervals for task groups and macro-averaged performance across all task groups for different context sizes of the LLM-embedding models. GTE-Qwen2-7B shows the best performance for 4,096-token context. LLM2Vec-Llama-3.1-8B shows the best performance for 2,048 tokens. Using a context size of 8,192 tokens does not show an improvement.

Model	Operational Outcomes	Anticipating Lab Test Results	Assignment of New Diagnosis	Anticipating Chest X-ray Findings	Macro Avg. Across Task Groups
GTE-Qwen2-7B					
8.192 context size	0.775 .752-.797	0.784 .774-.795	0.684 .638-.729	0.678 .666-.691	0.730 .704-.757
4.096 context size	0.771 .747-.795	0.865 .858-.873	0.716 .675-.757	0.666 .653-.680	0.755 .730-.780
2.048 context size	0.745 .720-.770	0.877 .871-.883	0.725 .683-.767	0.656 .642-.670	0.751 .725-.776
1.024 context size	0.731 .706-.756	0.885 .879-.891	0.677 .638-.716	0.643 .629-.656	0.734 .710-.758
512 context size	0.690 .664-.716	0.740 .729-.751	0.642 .606-.678	0.612 .597-.628	0.671 .647-.695
LLM2Vec-Llama-3.1-8B					
8.192 context size	0.767 .744-.791	0.760 .748-.771	0.708 .655-.762	0.689 .676-.701	0.731 .701-.762
4.096 context size	0.753 .727-.778	0.778 .768-.789	0.728 .682-.774	0.678 .665-.692	0.734 .707-.762
2.048 context size	0.729 .703-.755	0.880 .874-.886	0.722 .690-.754	0.654 .639-.669	0.746 .724-.769
1.024 context size	0.691 .666-.716	0.889 .883-.895	0.662 .626-.698	0.635 .620-.650	0.719 .696-.743
512 context size	0.658 .630-.686	0.820 .810-.829	0.631 .589-.674	0.609 .593-.625	0.679 .652-.707

Table 11 Performance of LLM-Embedding Models for Chunked Context. Macro averaged area under receiver operating characteristic curve (AUROC) performance and 95% confidence intervals for task groups and macro-averaged performance across all task groups for chunked inputs and averaged embeddings of the LLM-embedding models. The performance trends are similar to the context size experiments with less decrease in performance as all 4.096 input tokens are still incorporated.

Model	Operational Outcomes	Anticipating Lab Test Results	Assignment of New Diagnosis	Anticipating Chest X-ray Findings	Macro Avg. Across Task Groups
GTE-Qwen2-7B					
1 x 4.096 token chunks	0.771 .747-.795	0.865 .858-.873	0.716 .675-.757	0.666 .653-.680	0.755 .730-.780
2 x 2.048 tokens chunks	0.751 .726-.776	0.861 .854-.868	0.717 .677-.756	0.672 .659-.685	0.750 .725-.775
4 x 1.024 tokens chunks	0.753 .728-.778	0.854 .846-.861	0.708 .663-.753	0.669 .657-.682	0.746 .719-.773
8 x 512 tokens chunks	0.749 .725-.774	0.784 .774-.794	0.724 .685-.764	0.666 .654-.679	0.731 .706-.756
LLM2Vec-Llama-3.1-8B					
1 x 4.096 token chunks	0.753 .727-.778	0.778 .768-.789	0.728 .682-.774	0.678 .665-.692	0.734 .707-.762
2 x 2.048 tokens chunks	0.746 .719-.772	0.860 .853-.867	0.727 .691-.764	0.678 .665-.691	0.753 .729-.777
4 x 1.024 tokens chunks	0.755 .730-.780	0.855 .847-.863	0.732 .696-.769	0.680 .668-.692	0.755 .732-.779
8 x 512 tokens chunks	0.764 .740-.788	0.829 .820-.837	0.708 .665-.751	0.672 .659-.684	0.743 .717-.769

A.6 Additional Performance Results on EHRSHOT

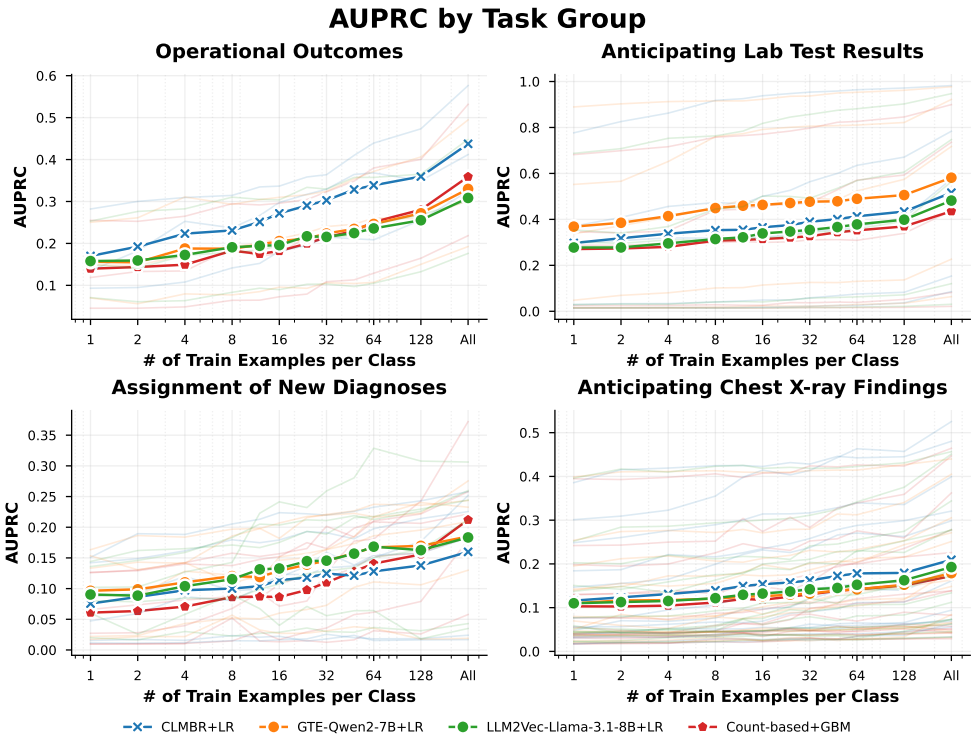

Fig. 7 AUPRC Performance in Few-Shot Settings for EHRSHOT. Macro-averaged area under the precision-recall curve (AUPRC) performance across subtasks for four task groups across (bold). Blurred lines are averaged AUPRC values across five bootstrapped runs using different seeds [31].

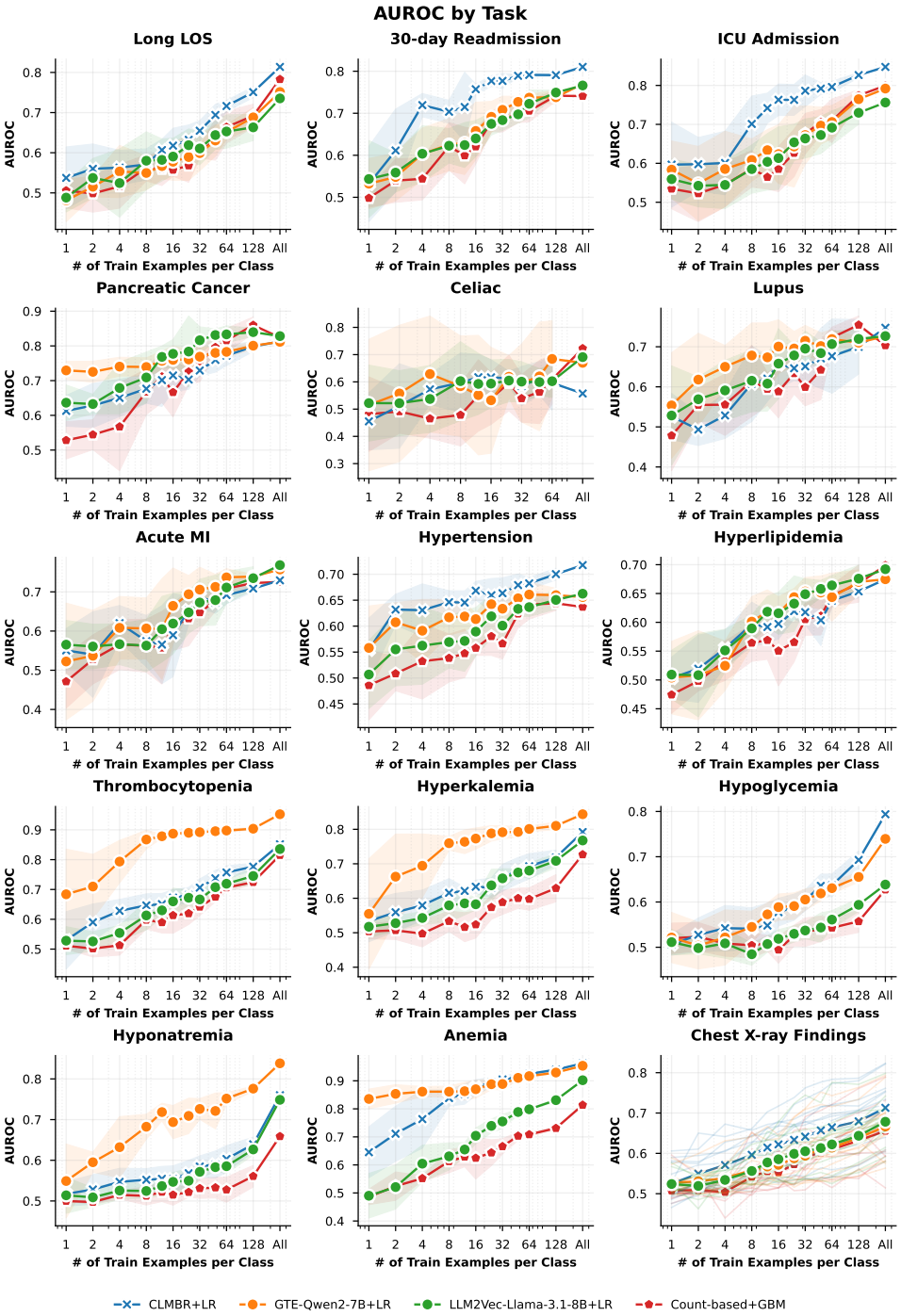

Fig. 8 Task-specific AUROC performance for EHRSHOT. Area under receiver operating characteristic curve (AUROC) performance with 95% confidence intervals across all 15 prediction tasks [31].

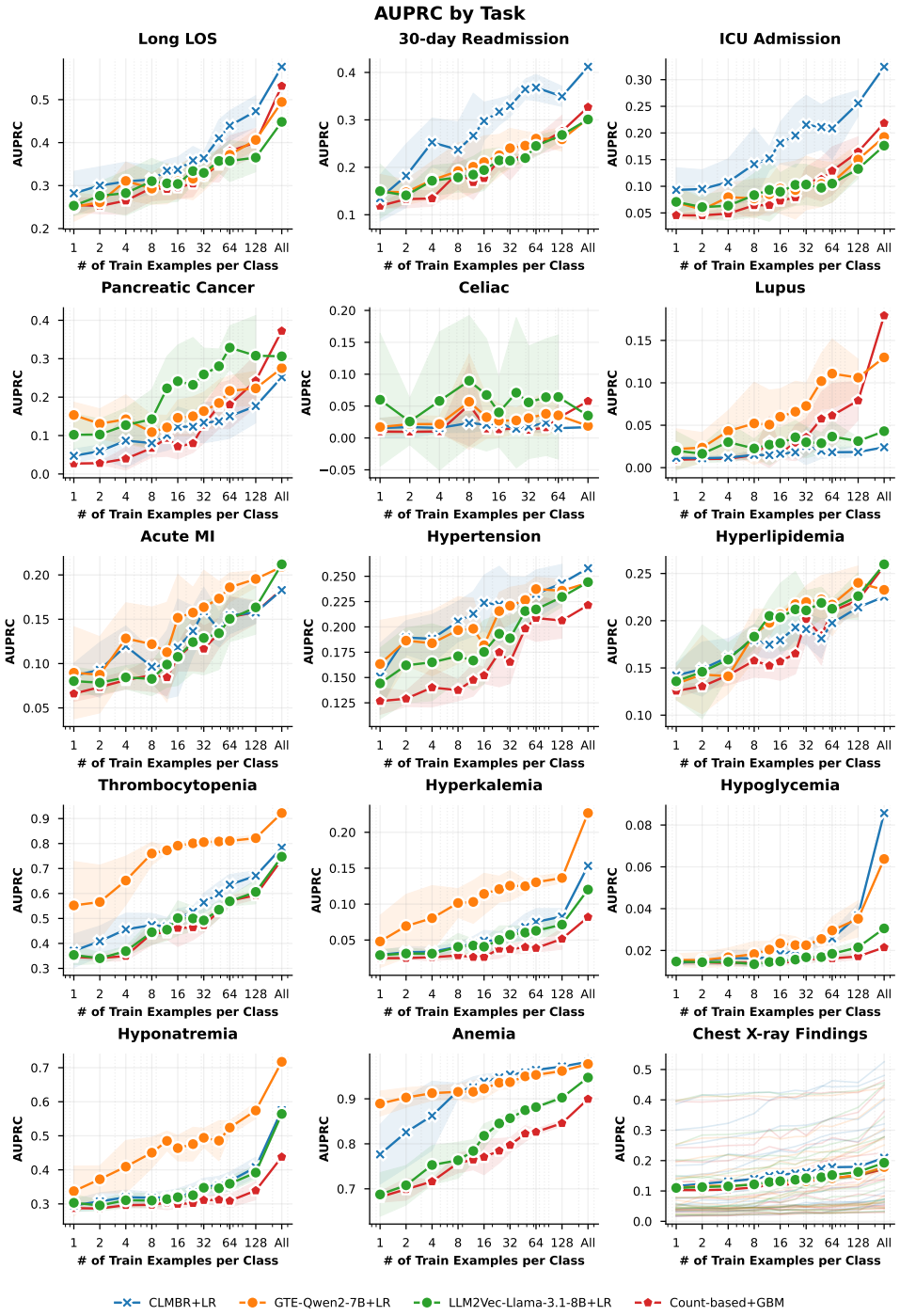

Fig. 9 Task-specific AUPRC performance for EHRSHOT. Area under the precision-recall curve (AUPRC) performance with 95% confidence intervals across all 15 prediction tasks [31].

A.7 Additional Performance Results on UK Biobank

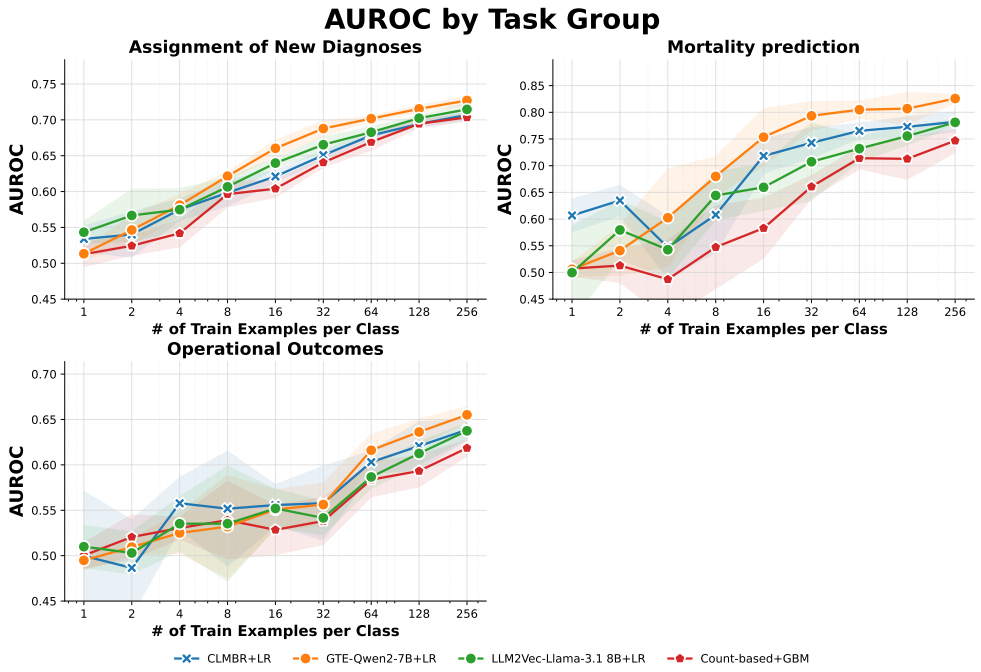

Fig. 10 AUROC Performance in Few-Shot Settings for UKB. Macro-averaged area under receiver operating characteristic curve (AUROC) performance across five CV rounds for three task groups across (bold).

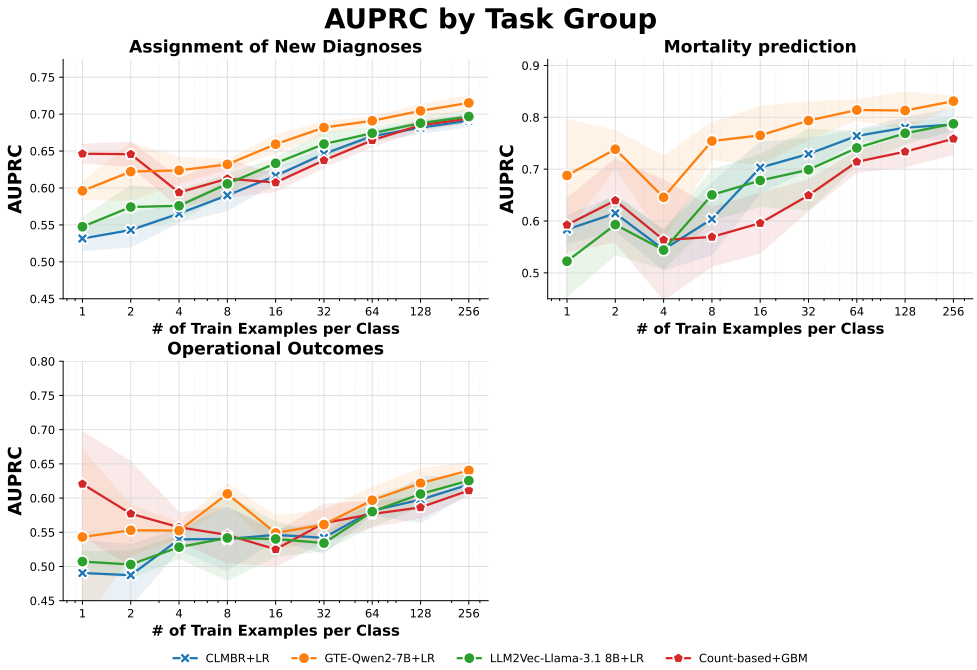

Fig. 11 AUPRC Performance in Few-Shot Settings for UKB. Macro-averaged area under the precision-recall curve (AUPRC) performance across five CV rounds for three task groups across (bold).

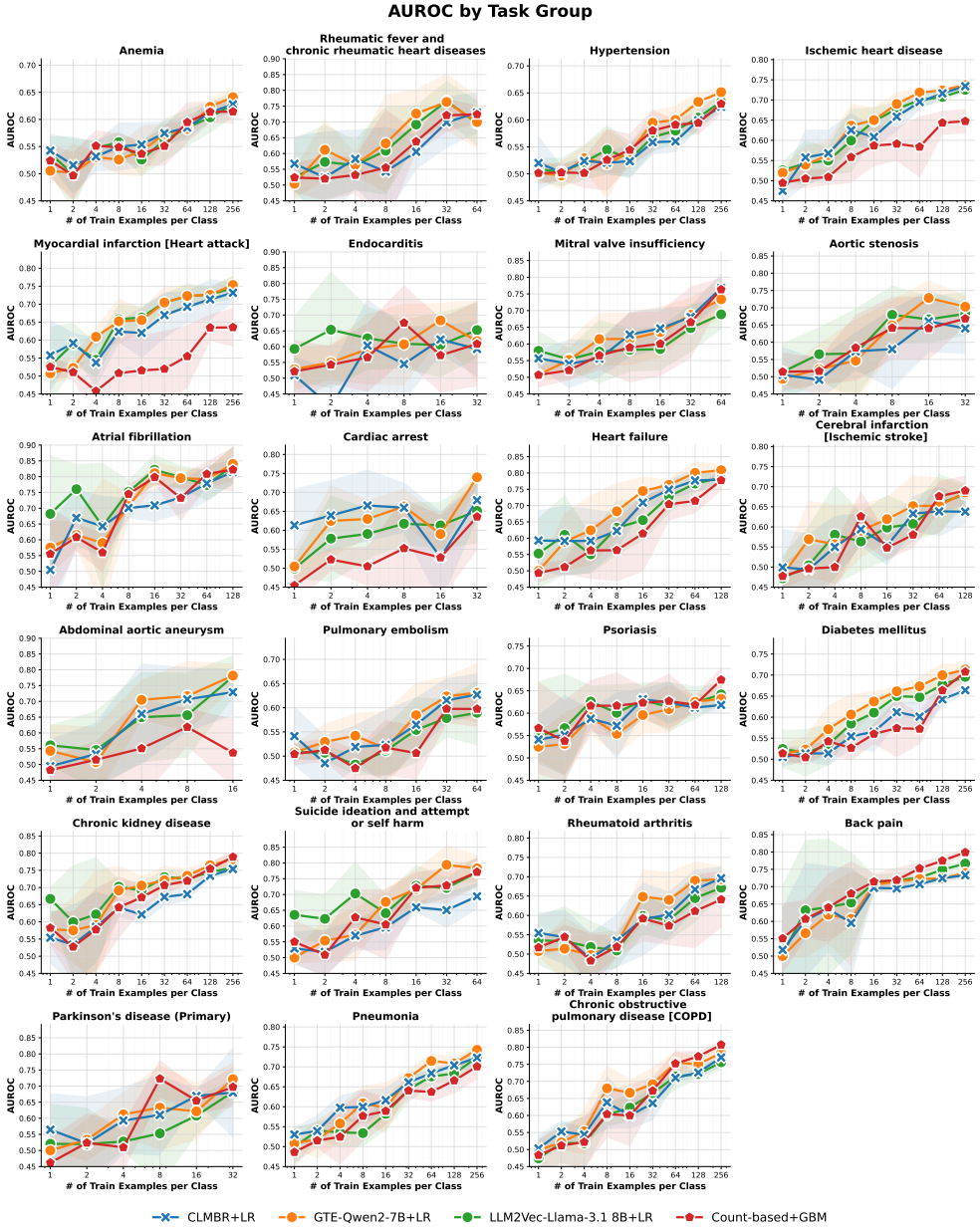

Fig. 12 Disease onset AUROC performance for UKB. Macro-averaged area under receiver operating characteristic curve (AUROC) performance across five CV rounds for all diseases from the assignment of new diagnoses tasks.

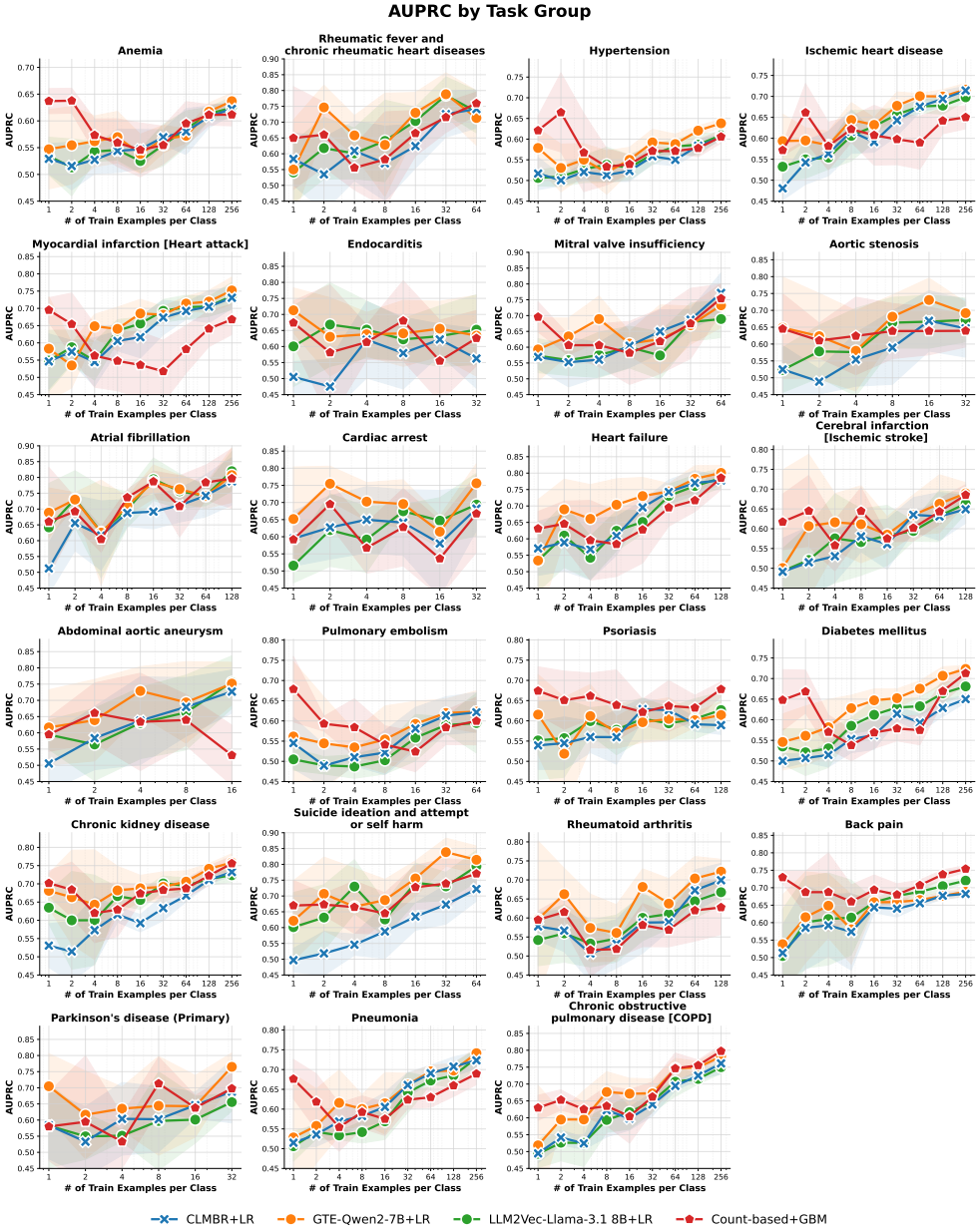

Fig. 13 Disease onset AUPRC performance for UKB. Macro-averaged area under the precision-recall curve (AUPRC) performance across five CV rounds for all diseases from assignment of new diagnoses tasks.