CPLLM: CLINICAL PREDICTION WITH LARGE LANGUAGE MODELS

Ofir Ben Shoham
Department of Software and
Information Systems Engineering
Ben-Gurion University of the Negev
benshoho@post.bgu.ac.il

Nadav Rappoport
Department of Software and
Information Systems Engineering
Ben-Gurion University of the Negev
nadavrap@bgu.ac.il

ABSTRACT

We present Clinical Prediction with Large Language Models (CPLLM), a method that involves fine-tuning a pre-trained Large Language Model (LLM) for clinical disease and readmission prediction. We utilized quantization and fine-tuned the LLM using prompts. For diagnosis prediction, we predict whether patients will be diagnosed with a target disease during their next visit or in the subsequent diagnosis, leveraging their historical diagnosis records. We compared our results to various baselines, including RETAIN, and Med-BERT, the current state-of-the-art model for disease prediction using temporal structured EHR data. In addition, We also evaluated CPLLM for patient hospital readmission prediction and compared our method's performance with benchmark baselines. Our experiments have shown that our proposed method, CPLLM, surpasses all the tested models in terms of PR-AUC and ROC-AUC metrics, showing state-of-the-art results for diagnosis prediction and patient hospital readmission prediction. Such a method can be easily implemented and integrated into the clinical process to help care providers estimate the next steps of patients.

1 INTRODUCTION

Large Language Models (LLMs) are a type of artificial intelligence (AI) that have been shown to be effective at a variety of Natural Language Processing tasks (Zhao et al., 2023). LLMs are trained on large amounts of textual data, which allows them to learn the statistical relationships between words and phrases. LLMs are used for different types of tasks, including natural language understanding, natural language generation, knowledge-intensive tasks, reasoning, and more (Yang et al., 2023b). This makes them well-suited for tasks that require understanding the meaning of a text, such as text classification (Gasparetto et al., 2022; Sun et al., 2023) and even clinical predictions in the medical domain (Thirunavukarasu et al., 2023; Steinberg et al., 2021).

Clinical predictions are used to estimate a patient's susceptibility to disease, gauge the likelihood of treatment response, or prognosticate the course of a patient's medical condition. (Laupacis et al., 1997; Wasson et al., 1985). These predictions have been executed via classical models such as Logistic Regression (Hosmer Jr et al., 2013) and Random Forest. However, these traditional methods do not model the order of the medical concept events (diagnoses, procedures, medications, etc.). Instead, they rely solely on the absence or presence of these events (features).

Modern event order prediction models, which are more advanced than the mentioned traditional prediction models, are based on RNNs or transformers, where the latter were shown to be superior (Vaswani et al., 2017). Specifically, BERT-Style Models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Deberta (He et al., 2020). Another transformer-based architecture is GPT-style language model. GPT models are trained to generate the next word in a sequence. GPT models are used in a wide range of downstream tasks such as summarization, translation, question answering, and more (Floridi & Chiriatti, 2020). To name a few GPT models: LLaMA (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), Bloom (Scao et al., 2022), and GPT4 (OpenAI, 2023). The flexibility and versatility of decoder-only models seem to be advantageous (Yang et al., 2023b).

The significance of the mentioned language models for handling sequential data is emphasized, particularly within the context of clinical prediction models relying on Electronic Health Record (EHR) data. Structured EHR data encompasses a patient's clinical history, notable for its irregular temporal sequence of events and observations (Steinberg et al., 2021). Previous works deal with modeling EHR diagnosis data as a sequence, such as BEHRT (Li et al., 2020; 2022a; Shoham & Rappoport, 2023; Meng et al., 2021), Med-BERT (Rasmy et al., 2021) and Medic-BERT (Hansen et al., 2023) (for length of stay prediction), using BERT models. However, these models represent each diagnosis code as an index and do not address the textual description of the ICD code. In addition, These models are pre-trained using clinical data, and have a limited sequence length input.

There is limited research on developing clinical prediction models using pre-trained LLM as a starting point and fine-tune it. One of the main focus of applications of LLM in the clinic is on chat capability of these models (Singhal et al., 2023; Thirunavukarasu et al., 2023) or using an LLM for medical texts-based tasks like text generation (Lu et al., 2022; Agrawal et al., 2022) and text comprehension (Yang et al., 2022; Sivarajkumar & Wang, 2022; Li et al., 2022b; Jiang et al., 2023). In addition, Chen et al. (2023) proposed a method called ClinTaT for cancer prediction. Their focus was on cancer prognostic prediction using few-shot learning, and their data modeling was not designed for structured EHR data that consists of a sequence of diagnoses. However, we want to harness the power of LLMs in understanding sequences of tokens derived from structured EHR data, specifically to train prediction models. We represent the structured data as a text by representing each medical concept corresponding to a word, admissions are treated as visits, and patient history is considered a document. The objectives of this study are to develop a novel method for using LLMs to train clinical predictors and to evaluate the performance of this method on real-world datasets.

Our proposed method uses an LLM to predict future diagnoses and readmission of patients by fine-tuning LLMs. The medical concepts are represented by text descriptions. Fine-tuning is performed using a prompt that feeds the model with training samples. We used two different LLMs, Llama2, which is a general LLM (Touvron et al., 2023) and BioMedLM, which was trained on biological and clinical text (Venigalla et al., 2022). We used four prediction tasks and two datasets and compared the performance to baseline models.

The proposed method outperforms the state-of-the-art methods. Our generic method can be used for a variety of tasks and is not specific to any particular LLM. Moreover, our method is also suitable for different clinical domains such as demographics, diagnoses, laboratory test results, measurements, procedures, and more.

Contributions: **(1)** We propose CPLLM, a novel method for clinical prediction with LLM that outperforms state-of-the-art models for disease prediction and patient readmission prediction for structured EHR data. In addition, CPLLM does not require pre-training on clinical data and achieves better performance than alternative approaches. Moreover, Our method has a longer sequence length limit compared to the baseline methods. **(2)** We show that adding additional tokens to the pre-trained tokenizer of the LLM before fine-tuning improves the performance of the clinical prediction model. **(3)** Our code is flexible for any LLM, available to use, and easily adaptable to various clinical prediction tasks.

2 METHODS

2.1 DISEASE PREDICTION - PROBLEM DEFINITION

Formally, for a given patient p, let n denote the total number of diagnoses in their medical history. Thus, the patient's sequence of diagnoses is represented as $\{D_{p,1}, D_{p,2}, D_{p,3}, \ldots, D_{p,n}\}$, where each $D_{p,i}$ $(1 \leq i \leq n)$ corresponds to a medical diagnosis in the patient's history. We considered two types of binary diagnosis prediction: next diagnosis and next visit diagnosis.

Next diagnosis prediction: Given a patient's medical history, we predict whether the patient's next diagnosis will be the target disease of interest. More formally: we predict whether patient p will be diagnosed with a specific disease D_x (a text that describes the disease) as the $D_{p,i+1}$ diagnosis given previous diagnoses. Our model relys on the patient's medical records up to the i-th diagnosis, denoted as $\{D_{p,1}, D_{p,2}, \ldots, D_{p,i}\}$. Where $D_{p,i}$ $(1 \leq i < n)$ indicates the most

recent diagnosis observed for patient p. The predictive model utilizes this patient-specific historical medical information to determine whether patient p's next diagnosis is a specific disease or not.

Next visit diagnosis prediction: Sometimes we can not predict the next diagnosis for a patient. Predicting the next diagnosis requires knowledge of the precise timing of each diagnosis. However, these data may occasionally be unavailable, such as when diagnoses are documented at the end of an admission. Therefore, we define the next visit diagnosis prediction task. Next visit diagnosis prediction is defined as predicting, based on a patient's medical history, whether the patient will be diagnosed with the disease of interest during their next admission visit. Consequently, in the context of the MIMIC-IV dataset, we undertake the task of forecasting whether a patient will receive a specific diagnosis in his subsequent admission.

2.2 Patient Hospital readmission prediction

Based on a patient's medical history, including procedures, diagnoses, and medications, our objective is to forecast whether the patient will experience hospital readmission within the next X days. We follow the definition of X as specified by the PyHealth benchmark (Yang et al., 2023a). In our experiments with the MIMIC-IV dataset, we predict hospital readmission within a 15-day window, and for the eICU-CRD dataset, the prediction time-frame is 5 days (see section 2.3).

2.3 Data

In this study, we used data from the eICU-CRD database (Pollard et al., 2018) and data from the MIMIC-IV database (Johnson et al., 2020). Our datasets include ICD-9-CM (eICU-CRD) and ICD-10-CM (MIMIC-IV) diagnoses and their descriptions. In the eICU-CRD database, each diagnosis is associated with a timestamp. Consequently, we arranged the diagnoses in chronological order based on their respective diagnosis times. Our disease prediction task aims to anticipate whether the forthcoming diagnosis will correspond to a specific disease. Unlike the eICU-CRD dataset, the MIMIC-IV data lacks information on the exact time of each diagnosis assignment. However, it provides the start time for admission and the discharge times for each patient. As a result, our prediction task for this dataset revolves around determining whether a patient will be diagnosed with a specific disease during his subsequent visit.

Med-BERT adopts a pre-training strategy and trains BERT using Masked Language Modeling (MLM) and Length of stay (LOS) prediction tasks (Rasmy et al., 2021). Therefore, we extracted the necessary data from the databases, including the diagnosis codes for each patient. Additionally, we also include information on the LOS of each admission and the number of visits of each patient. On the other hand, in our approach, we did not conduct an additional pre-training step, as we focused on fine-tuning an LLM. In our proposed method, it is not required to note at which visit each diagnosis was given. Furthermore, the duration of hospital stay is not required. Notably, our method attains superior results even in the absence of these particulars. This aspect holds significance, since in certain situations, this data may not be accessible. For example, when a patient has not been admitted to the hospital but is under the care of a family doctor.

Data Preprocessing: For readmission prediction, we follow PyHealth's data preprocessing methodology. We include drugs, procedures, and diagnosis codes alongside their respective descriptions. Additionally, we incorporate both ICD-9 and ICD-10 codes and convert them to Clinical Classification Software (CCS) codes (Elixhauser, 2009). For drugs, we convert the codes to ATC codes (Nahler & Nahler, 2009). For procedures, we include ICD-9 and ICD-10 procedure codes and convert them to CCS codes using PyHealth. For diagnosis prediction, for the MIMIC-IV dataset, we excluded patients with only one visit, as there is no medical history in such a case. Similarly, for the eICU-CRD dataset, patients with just one diagnosis were removed. We also excluded patients who have the disease we are trying to predict at the first visit (or the first diagnosis for eICU-CRD data). We converted our ICD-10 codes to their corresponding CCS categories for MIMIC-IV, while for eICU-CRD, we retained the ICD-9 codes as they were. This decision was motivated by the higher number of ICD-10 codes compared to ICD-9 codes (Manchikanti et al., 2013). Based on the sequence of diagnoses for each patient, we determined whether the patient exhibited a specific diagnosis based on ICD diagnosis codes related to the specific disease according to the relevant CCS category (Elixhauser et al., 2014). Table 1 provides an overview of the number of patients, the

count of final patients after preprocessing, average diagnoses, and average visits for each disease prediction task.

2.3.1 CLINICAL OUTCOMES

We evaluated our model for four prediction tasks: patient hospital readmission prediction and three diagnosis predictions covering Chronic kidney disease, Acute and unspecified renal failure, and Adult respiratory failure. The first two diagnoses were derived from the MIMIC-IV dataset, and the last was derived from the eICU-CRD dataset. The corresponding CCS codes for these diseases are 157 for Acute and unspecified renal failure, 158 for Chronic kidney disease, and 131 for Adult respiratory failure. For each prediction task, patients with specific disease ICD codes were assigned a positive label, and their diagnosis history encompassed all diagnostic codes recorded until the specific code indicated the outcome of interest.

Table 1: Task statistics of the prediction tasks. Visit and diagnosis counts are calculated from the patient's medical history after preprocessing. IQR - Interquartile range.

Dataset	Task	# of patients	Final # of patients	Median # of visits (IQR)	Median # of diagnoses (IQR)
MIMIC-IV	Chronic kidney disease	84,453	26,161	1 (1-2)	11 (7-19)
MIMIC-IV	Acute and unspecified renal failure	84,453	26,736	1 (1-2)	11 (7-19)
eICU-CRD	Adult respiratory failure	132,677	56,419	1 (1-1)	1 (1-2)

2.4 BASELINE METHODS

We conducted a rigorous performance assessment of the CPLLM against three baseline methods. For diagnosis prediction task, we used the next baseline models. First, Med-BERT with a classification layer (Rasmy et al., 2021). Second, with Logistic Regression (Hosmer Jr et al., 2013). Furthermore, we compared our method to RETAIN - a disease prediction model featuring double GRUs and attention modules (Choi et al., 2016). We compared CPLLM with these baseline methods to gain valuable insights into its performance in clinical prediction downstream tasks. The comparison was conducted using two metrics: the area under the precision-recall curve (PR-AUC) and the area under the receiver operating characteristic curve (ROC-AUC). Disease prediction tasks are typically imbalanced; therefore ROC-AUC is less suitable for binary classifiers with imbalanced data (Davis & Goadrich, 2006). Therefore, our main evaluation metric is PR-AUC, but we also report ROC-AUC for consistency with the baseline methods. For readmission prediction, as mentioned earlier, we compared CPLLM with PyHealth baselines. The models we compared with include ConCare (Ma et al., 2020), RETAIN (Choi et al., 2016), deeper (Nguyen et al., 2016) and GRASP (Zhang et al., 2021).

2.5 OUR PROPOSED METHOD

We propose a method called Clinical Prediction with Large Language Models (CPLLM). This method involves fine-tuning a LLM using prompts tailored to medical concept sequences. Through fine-tuning using prompts (inputs for LLM guidance), we direct the LLM to grasp intricate relationships among medical concepts.

We utilized two LLMs: Llama2 (13B parameters) (Touvron et al., 2023) and BioMedLM (also called PubMedGPT, 2.7B parameters) (Venigalla et al., 2022). To enhance the time and memory efficiency of fine-tuning these LLMs, we used QLoRA (Dettmers et al., 2023) and PEFT (Houlsby et al., 2019). QLoRA is a PEFT approach that decreases the number of parameters requiring fine-tuning and also performs quantization (Dettmers et al., 2023). This combined approach effectively optimized the models' efficiency, enabling single-GPU fine-tuning for both BioMedLM and Llama2 models.

We performed separate fine-tuning of each LLM, leveraging specific prompts tailored to our patients' medical codes and their corresponding labels. In Figure 1, we present an example of the prompts utilized during the fine-tuning process for both the Llama2 and BioMedLM. We also indicated in the prompt the target disease, and the prompts were designed to incorporate the patients' individual medical code histories, with the goal of improving the models' performance. For readmission prediction, the prompt was very similar, but it included in addition drugs and procedures. For diagnosis prediction tasks, we added tokens of diagnosis descriptions missing from the original

tokenizer vocabulary of the LLM. We performed an ablation study that compared the performance with and without changing the vocabulary of the pre-trained tokenizer.

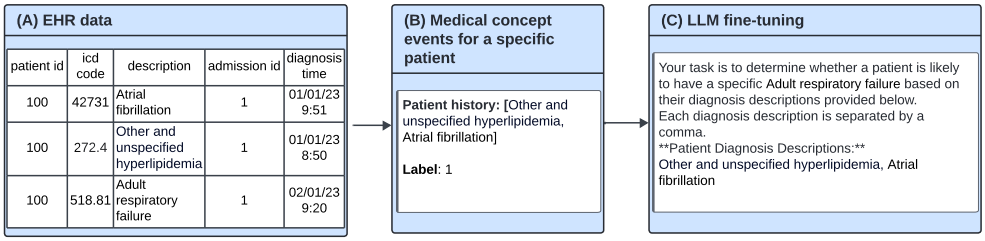

Figure 1: Illustration of the fine-tuning process for diagnosis prediction. (A) An example of EHR structured data. The patient has three diagnoses. (B) Patient's historical data is extracted from the EHR, and decoded to a textual list of descriptions. (C) The decoded textual data is then injected into a designed prompt for fine-tuning the LLM. Fine-tuning prompts consist of a general description, the patient's diagnosis history, and a label. The label is set to 1 when the patient is diagnosed with the outcome of interest (e.g., Adult Respiratory Failure in the subsequent diagnosis or during the next admission, depending on the task.

For the clinical prediction downstream task, we performed fine-tuning as depicted in Figure 1. Each sample in our training data consists of a prompt (text) and a label. We used prompts to ask the LLMs to generate a single binary token (0 or 1) in response, by adding a fully connected classification layer as the final layer of the LLM, corresponding to the number of labels. We used QLora (Dettmers et al., 2023) for our fine-tuning process and froze all layers except the linear layers of the LLM. By training the models with all patients' data using Binary Cross Entropy loss for the specified number of epochs, we obtained the fine-tuned LLM tailored to our specific clinical prediction task.

3 Experiments

3.1 Experimental Setup

For readmission prediction, we compared our method to the PyHealth benchmark. For the diagnosis prediction tasks, We compare our method to three baseline models. The first is a simple Logistic Regression that does not model the data as a sequence but as simple independent, unordered variables (Manogaran & Lopez, 2018). For the input to Logistic Regression, we used one-hot encoding because it can not handle text input directly. The second is RETAIN which is a two-level neural attention model (Choi et al., 2016). The third baseline is Med-BERT, which is the state-of-the-art for structured EHR data for disease prediction. RETAIN was the baseline of Med-BERT. We split our data using an 70-10-20 ratio for train, validation, and test sets accordingly. For Med-BERT, we trained the pre-training model with the MLM and LOS tasks on the train split of the MIMIC-IV dataset, with the TensorFlow package (Abadi et al., 2015). The training of the Med-BERT's MLM phase was performed according to the fixed number of steps in the original implementation. The training took about 1.5 days on an RTX1080 GPU. Subsequently, we performed fine-tuning on the pre-trained model for the specific clinical prediction downstream tasks. The RETAIN and Med-BERT baselines trained for 500 epochs with early stopping based on the PR-AUC value derived from the validation set, using a maximum number of epochs without improvement of 5 epochs (Prechelt, 2002). During the training of the baselines, we experimented with various batch sizes $\{32, 100\}$ and different learning rates $\{1e^{-5}, 2e^{-5}\}$. For each prediction task, we selected the hyper-parameters that achieved the best results on the validation set. For Logistic Regression training, we utilized the scikit-learn package (Pedregosa et al., 2011) and trained the model on a CPU. To determine the optimal hyper-parameters for Logistic Regression, we conducted a grid search encompassing *penalty* (L1 and L2 regularization), *C*, *solver*, and the maximum number of iterations. We explored values of $\{0.1, 1, 10\}$ for *C*, {'liblinear', 'saga'} for *solver*, and $\{100, 200, 500\}$ for the number of iterations. We took the best hyper-parameters based on the validation PR-AUC for each prediction task.

For CPLLM experiments, we fine-tuned two LLMs Llama2 (13B) and BioMedLM (2.7B) using HuggingFace (Wolf et al., 2019). (Dettmers et al., 2023). Specifically, we used a learning rate of $2e^{-5}$, Lora alpha of 32, Lora dropout of 0.1, and bias of none. Given the resource constraints, we meticulously determined and employed the maximum batch size that our GPU memory could accommodate. We fine-tuned each model over six epochs (and four epochs for readmission due to the larger dataset), selecting the best checkpoint based on validation PR-AUC. Fine-tuning Llama2 for six epochs required about a day of training on an RTX 6000 GPU, while BioMedLM took about two hours on the same hardware. Our fine-tuning process used PEFT, and we did not perform additional pre-training in the clinical domain, yet our CPLLM method outperformed the baseline models.

3.2 Results

3.2.1 Diagnosis Prediction Results

We consider various models for the clinical prediction task: Logistic Regression, Med-BERT with a classification layer, RETAIN, and our proposed method called CPLLM. To examine the statistical significance of the results, we ran each model three times. Table 2 shows the mean and 95% confidence interval of PR-AUC and ROC-AUC of these models.

Our findings demonstrate that our method, CPLLM, outperforms all tested models, including RETAIN, Med-BERT, and Logistic Regression, across both PR-AUC and ROC-AUC metrics. Specifically, in the context of the Adult respiratory failure task, CPLLM-Llama2 achieved a noteworthy PR-AUC value of 35.962%, signifying an absolute improvement of 0.912% over the best-performing baseline model, Logistic Regression, which obtained a PR-AUC score of 35.05%. This improvement corresponds to a relative enhancement of 2.6% in PR-AUC. Additionally, our method exhibits a relative increase of 5.1% in PR-AUC when compared to RETAIN and a 3.31% increase when compared to Med-BERT. Regarding ROC-AUC performance, CPLLM outperforms the baseline models. Furthermore, CPLLM-Llama2 demonstrates superior performance in this specific task compared to CPLLM-BioMedLM. Logistic Regression outperforms RETAIN in both PR-AUC (35.05%) and ROC-AUC (74.664%), but it also outperforms Med-BERT in PR-AUC, albeit not in ROC-AUC (74.664% compared to 75.407% for Med-BERT).

For Chronic kidney disease using the MIMIC-IV dataset, RETAIN had the worst performance in both metrics. Med-BERT outperformed Logistic Regression and RETAIN. CPLLM-Llama2 had the highest PR-AUC score of 33.992%, followed by CPLLM-BioMedLM with 33.984% and Med-BERT with 33.37%. However, in ROC-AUC, CPLLM-BioMedLM outperformed all models with a score of 83.404%, followed by CPLLM-Llama2 with 83.034% and Med-BERT with 83.12%.

For Acute and unspecified renal failure, CPLLM-Llama2 achieved the highest measurements, boasting a PR-AUC score of 45.442% and an ROC-AUC score of 78.504%. This signifies a notable improvement of 4.22% in PR-AUC compared to the leading baseline model, RETAIN, in this task. Additionally, it demonstrates a 1.31% improvement in ROC-AUC compared to the best-performing baseline, which is Logistic Regression with an ROC-AUC score of 77.486%. Furthermore, it is worth highlighting that in this specific task, RETAIN outperforms Med-BERT in terms of PR-AUC but not ROC-AUC. Additionally, CPLLM-Llama2 demonstrates superior performance compared to CPLLM-BioMedLM. We found that CPLLM-Llama2 outperformed CPLLM-BioMedLM and therefore the rest of the analysis will be based on CPLLM-Llama2.

3.2.2 Hospital Readmission Prediction Results

To demonstrate the robustness of CPLLM, we expanded our analysis beyond diagnosis to include procedures and drugs. We compared CPLLM against several baseline methods from the PyHealth benchmark. Table 3 presents the results for patient hospital readmission prediction. In the case of MIMIC-IV, CPLLM with LLama2-13B achieved a PR-AUC of 68.986%, outperforming ConCare, the second-best performing model, by 1.46% (absolute). For eICU-CRD, CPLLM exhibited the highest PR-AUC among the baselines, achieving a PR-AUC of 94.115%. Additionally, CPLLM achieved the highest ROC-AUC in both datasets.

Table 2: Performances of various models assessed across multiple tasks and datasets. The highest score per task is highlighted in bold.

Task	Model	PR-AUC	ROC-AUC
Adult respiratory failure	Logistic Regression	35.050	74.664
	RETAIN	34.22 ± 0.299	74.454 ± 0.173
	Med-BERT	34.81 ± 0.208	75.407 ± 0.073
	CPLLM-Llama2	**35.962 ± 0.380**	**76.407 ± 0.262**
	CPLLM-BioMedLM	35.494 ± 0.352	75.975 ± 0.214
Chronic kidney disease	Logistic Regression	32.230	83.016
	RETAIN	31.407 ± 1.379	81.692 ± 0.899
	Med-BERT	33.37 ± 0.891	83.12 ± 0.173
	CPLLM-Llama2	**33.992 ± 1.262**	83.034 ± 0.511
	CPLLM-BioMedLM	33.984 ± 1.077	**83.404 ± 0.429**
Acute and unspecified renal failure	Logistic Regression	42.075	77.486
	RETAIN	43.603 ± 0.409	77.364 ± 0.394
	Med-BERT	42.237 ± 0.408	77.427 ± 0.185
	CPLLM-Llama2	**45.442 ± 0.839**	**78.504 ± 0.684**
	CPLLM-BioMedLM	45.161 ± 1.622	78.484 ± 0.403

Table 3: PR-AUC and ROC-AUC performances of hospital readmission prediction task for MIMIC-IV and eICU-CRD datasets.

Dataset	Model	PR-AUC	ROC-AUC
MIMIC-IV	CPLLM-Llama2	**68.986 ± 0.499**	**68.155 ± 0.38**
	ConCare	67.523 ± 0.697	67.242 ± 0.269
	RETAIN	67.343 ± 0.558	66.893 ± 0.421
	deeper	66.891 ± 0.604	66.575 ± 0.371
	GRASP	65.656 ± 2.929	65.302 ± 3.369
eICU-CRD	CPLLM-Llama2	**94.115 ± 0.704**	**77.916 ± 1.026**
	ConCare	93.429 ± 0.733	77.024 ± 1.156
	RETAIN	93.615 ± 0.340	77.149 ± 1.048
	deeper	93.814 ± 0.422	77.814 ± 0.385
	GRASP	93.677 ± 1.824	77.515 ± 3.899

3.3 Ablation Study

We conducted an ablation study to investigate the impact of the added tokens to the pre-trained tokenizer of the LLMs before fine-tuning. Table 4 provides a comprehensive overview of the PR-AUC and ROC-AUC, comparing scenarios with and without adding extra tokens. For the task of predicting Acute and unspecified renal failure, adding the tokens yields enhancements in both PR-AUC and ROC-AUC for CPLLM-Llama2 (0.499% absolute increase in PR-AUC and a 0.554% absolute increase in ROC-AUC). Similarly, CPLLM-BioMedLM shows substantial improvements with a 1.631% absolute increase in PR-AUC, representing a relative enhancement of 3.746%, and a 0.414% absolute increase in ROC-AUC. In contrast, for the prediction of Chronic kidney disease, the inclusion of extra tokens does not significantly impact PR-AUC and ROC-AUC in the case of CPLLM-Llama2. However, CPLLM-BioMedLM demonstrates improvements, specifically an absolute enhancement of 0.686% in ROC-AUC and an increase in PR-AUC from 32.638% to 33.984%. It is worth noting that the PR-AUC of BioMedLM exhibits less stability, as evidenced by a larger confidence interval when no additional tokens are employed (4.358%). Nevertheless, we conducted two additional runs to get a better estimate of the PR-AUC. Subsequently, we observed that the PR-AUC for these five experiments amounted to 33.078%, and the confidence intervals were reduced to 1.773%. For Adult respiratory failure prediction, the presence of additional tokens results in improved PR-AUC and ROC-AUC for CPLLM-Llama2, whereas it enhances PR-AUC but does not influence ROC-AUC for CPLLM-BioMedLM. In summary, the findings of this ablation study suggest that, in the majority of cases (9 out of 12 measurements across three prediction tasks), incorporating the added tokens leads to enhanced performance in clinical prediction tasks.

Task	Model	Added Tokens	PR-AUC	ROC-AUC
Acute and unspecified renal failure	CPLLM-Llama2	+	**45.442 ± 0.839**	**78.504 ± 0.684**
		-	44.943 ± 1.268	77.95 ± 0.814
	CPLLM-BioMedLM	+	**45.161 ± 1.622**	**78.484 ± 0.403**
		-	43.53 ± 1.101	78.07 ± 0.625
Chronic kidney disease	CPLLM-Llama2	+	33.992 ± 1.262	83.034 ± 0.511
		-	**34.563 ± 1.578**	**83.178 ± 1.02**
	CPLLM-BioMedLM	+	**33.984 ± 1.077**	**83.404 ± 0.429**
		-	32.638 ± 4.358	82.718 ± 1.191
Adult respiratory failure	CPLLM-Llama2	+	**35.962 ± 0.38**	**76.407 ± 0.262**
		-	35.683 ± 0.164	75.776 ± 0.085
	CPLLM-BioMedLM	+	35.494 ± 0.352	**75.975 ± 0.214**
		-	**35.714 ± 0.516**	75.794 ± 0.194

Table 4: PR-AUC and ROC-AUC for CPLLM-Llama2 and CPLLM-BioMedLM, across three distinct medical tasks. Added Tokens column indicates whether additional tokens were incorporated into the pre-trained tokenizer. "+" and "-" - additional tokens were or were not added accordingly.

4 DISCUSSION

Our proposed CPLLM method outperformed the baselines on all four tasks (3 diagnosis prediction and readmission prediction) across two different datasets. We used MIMIC-IV and eICU-CRD datasets to assess the model's ability to handle two diagnoses coding systems (ICD9 and ICD10) and two data types (homogeneous data from the same hospital in MIMIC-IV and multi-center data in eICU-CRD). CPLLM was superior to all baselines. CPLLM-Llama2 was the best model overall, and only for Chronic kidney disease did CPLLM-BioMedLM outperform CPLLM-Llama2, but only in terms of ROC-AUC. Using CPLLM-Llama2, we achieved PR-AUC relative improvements of 3.309%, 1.864%, and 7.588% over Med-BERT on the three tasks, and ROC-AUC relative improvements of 1.326% and 1.391% on the Adult respiratory failure and Acute and unspecified renal failure prediction tasks. For hospital readmission prediction, CPLLM achieved relative improvements of 2.17% compared to ConCare in PR-AUC for MIMIC-IV. For eICU-CRD readmission prediction, CPLLM showed a relative improvement of 0.31% compared to the second-best result, deeper.

We hypothesize that CPLLM's superior performance compared to the baselines is due to its larger number of parameters and the substantial amount of training tokens used during pre-training. For instance, CPLLM-Llama2 was pre-trained on 2 trillion tokens and has 13 billion parameters (Touvron et al., 2023). This reasoning may also explain why CPLLM-Llama2 outperformed CPLLM-BioMedLM in nearly all tasks. The greater parameter count and more extensive training data of CPLLM-Llama2, in comparison to BioMedLM's 2.7 billion parameters and 34.6 billion tokens, provide a substantial advantage, despite BioMedLM being pre-trained on PubMed abstracts and full articles (Venigalla et al., 2022).

In addition, We found that including additional tokens in the LLM's tokenizer before fine-tuning improves the measurement of the prediction model in most cases. For instance, as Llama2 was not initially pre-trained on clinical data, supplementing it with missing description codes can enhance its understanding of the medical domain.

Regarding the comparison between Med-BERT and RETAIN, in the original Med-BERT paper, improvements over RETAIN were demonstrated in terms of ROC-AUC for three disease prediction tasks (Rasmy et al., 2021). We also found that Med-BERT consistently outperformed RETAIN in all prediction tasks based on ROC-AUC. However, it is worth noting that, as previously mentioned, ROC-AUC may not be an optimal metric for imbalanced datasets (Davis & Goadrich, 2006). In contrast, when considering PR-AUC, Med-BERT exhibited superior performance compared to RETAIN in two out of three tasks, although it did not outperform RETAIN in the prediction of Acute and unspecified renal failure (with PR-AUC values of 43.603% for RETAIN and 42.237% for Med-BERT), despite achieving a higher ROC-AUC than RETAIN.

4.1 Strengths and Limitations

CPLLM has several advantages compared to existing approaches.

First, Unlike existing approaches that necessitate pre-training with medical concept sequences, our method eliminates the need for additional pre-training tasks. For instance, Med-BERT entails both MLM and LOS prediction tasks using patient sequences of medical concepts. Based on our findings and results, it is evident that LLMs possess the capability to adeptly represent sequential clinical data without the need for specific pre-training based on clinical sequences. Beyond that, our method can be used even without the LOS data of each patient's hospitalizations, which is required for Med-BERT pre-training. Sometimes, these data are not available, for example, when there is no hospitalization, but rather data collected among patients who visited a physician in outpatient settings, or when LOS data is not available like in claims data.

Second, our proposed method lies in its remarkable capacity to handle longer sequences compared to the current state-of-the-art models for structured EHR data. With maximum sequence lengths of 1024 tokens for CPLLM-BioMedLM and 4096 tokens for CPLLM-Llama2, our approach far surpasses the limitations imposed by Med-BERT and BEHRT (Li et al., 2020). Both Med-BERT and BEHRT are constrained by BERT's maximum of 512 tokens, which significantly restricts their ability to handle longer inputs (Devlin et al., 2018). Without the need for additional training, our method also handles longer sequences compared to Hi-BEHRT, which is specially trained and designed to handle sequences with a maximum of 1220 tokens (Li et al., 2022a).

Third, during the fine-tuning training of CPLLM, it is not necessary to know which diagnoses were given in which visit but only the diagnoses as a sequence. This differs from Med-BERT, which relies on this information for fine-tuning. Notably, we achieved superior performance even without these specific details.

Fourth, CPLLM demonstrated flexibility for various input types and clinical prediction outcomes beyond disease prediction. This was evident in the readmission prediction experiment, where our approach seamlessly incorporated diagnoses, drugs, and procedures into the sequence with minimal adjustments to the prompt text.

While our method demonstrates promising results in utilizing LLMs for clinical prediction tasks, it is important to acknowledge several limitations. We pre-trained Med-BERT on the MIMIC-IV dataset rather than a large corpus as described in the original paper, due to our lack of access to larger datasets and the unavailability of pre-trained Med-BERT weights, which are not publicly accessible because of patient privacy concerns. In addition, while our method accommodates sequences of up to 4096 tokens for CPLLM-Llama2 and 1024 tokens for CPLLM-BioMedLM, our tests did not include exceptionally long sequences that could fully explore the implications of this extended token limit. That is because the datasets we used do not contain very long observations or many diagnoses of a single patient. Moreover, due to the greater number of parameters in LLMs, our method demands more computational resources, inference time, and training time. Specifically, CPLLM-Llama2 had a longer training time than Med-BERT. However, CPLLM-BioMedLM requires less training time compared to Med-BERT (Section 3.1). That's because CPLLM-BioMedLM does not require additional pre-training, unlike necessity for MLM and LOS pre-training in Med-BERT. In addition, in our method, there is a necessity to use a specific prompt, a requirement that does not apply to the baseline models. As a result, sometimes the prompt needs to be adapted according to a base model.

4.2 Future Work

We hypothesize that combining a retrieval augmentation (Mialon et al., 2023; Zakka et al., 2024), with LLM can improve performance. This is because it allows to include general updated knowledge about the diseases that the patient has been diagnosed with in their medical history. Additionally, this approach can incorporate general knowledge and known risk factors into research on the disease we are trying to predict.

5 CONCLUSION

In this work, we presented CPLLM, a novel method for clinical disease prediction and patient hospital readmission prediction based on the clinical history of patients. CPLLM has practical application potential. By surpassing the state-of-the-art in clinical task prediction, our method enables more accurate and robust disease forecasting, as well as patient hospital readmission forecasting. CPLLM demonstrated superior performance across all three four on two datasets (MIMIC-IV and eICU-CRD). It processes ICD9 and ICD10 diagnoses, procedures, and drugs. We showcased its robustness in dealing with homogeneous and multi-center data. Our method's advantage lies in eliminating the need for additional pre-training tasks, unlike Med-BERT. Furthermore, our method remains adaptable the length of stay data is unavailable, making it suitable for a broader range of healthcare scenarios, including those involving non-hospitalized patients. In addition, CPLLM's fine-tuning process requires patients' diagnoses as a sequence, without the need for which diagnoses were given in which visit. Notably, our method can handle much longer sequences than existing state-of-the-art models.

We believe that CPLLM has significant practical applications. For instance, healthcare stakeholders are increasingly seeking methods to enhance patient care without compromising data privacy. The two LLMs we tested can be deployed and utilized on-premise or in secure environments, eliminating the need to share personal data over the web.

6 REPRODUCIBILITY

Our code is available at the following link: https://github.com/nadavlab/CPLLM. Implementation details can be found in the Experimental Setup section 3.1. To execute the baseline code, we used the source code published as part of the Med-BERT paper (Rasmy et al., 2021).

For our experiments, we used the MIMIC-IV v2.0 dataset (Johnson et al., 2020), accessible at https://physionet.org/content/mimiciv/2.0/, as well as the eICU-CRD multi-center dataset (Pollard et al., 2018), which can be found at https://physionet.org/content/eicu-crd/2.0/.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, 2022.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. Falcon-40b: an open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773, 2023.

Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Boosting transformers and language models for clinical prediction in immunotherapy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 332–340, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.32. URL `https://aclanthology.org/2023.acl-industry.32`.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Anne Elixhauser. Clinical classifications software (ccs) 2009. *http://www. hcug-us. ahrq. gov/toolssoft-ware/ccs/ccs. jsp*, 2009.

Anne Elixhauser, Claudia Steiner, and L Palmer. Clinical classifications software (ccs). *US agency for healthcare research and quality*, 2014, 2014.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83, 2022.

Emil Riis Hansen, Thomas Dyhre Nielsen, Thomas Mulvad, Mads Nibe Strausholm, Tomer Sagi, and Katja Hose. Patient event sequences for predicting hospitalization length of stay. In *International Conference on Artificial Intelligence in Medicine*, pp. 51–56. Springer, 2023.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, pp. 1–6, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Andreas Laupacis, Nandita Sekar, et al. Clinical prediction rules: a review and suggested modifications of methodological standards. *Jama*, 277(6):488–494, 1997.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022a.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5436–5443, 2022.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 833–840, 2020.

Laxmaiah Manchikanti, Frank JE Falco, and Joshua A Hirsch. Ready or not! here comes icd-10. *Journal of neurointerventional surgery*, 5(1):86–91, 2013.

Gunasekaran Manogaran and Daphne Lopez. Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms*, 10 (1-2):118–132, 2018.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129, 2021.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. *ArXiv*, 2023.

Gerhard Nahler and Gerhard Nahler. Anatomical therapeutic chemical classification system (atc). *Dictionary of Pharmaceutical Medicine*, pp. 8–8, 2009.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30, 2016.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pp. 55–69. Springer, 2002.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Ofir Ben Shoham and Nadav Rappoport. Federated learning of medical concepts embedding using behrt. *arXiv preprint arXiv:2305.13052*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pp. 1–9, 2023.

Sonish Sivarajkumar and Yanshan Wang. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, pp. 972. American Medical Informatics Association, 2022.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, pp. 1–11, 2023.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Venigalla, J Frankle, and M Carbin. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23(3):2, 2022.

John H Wasson, Harold C Sox, Raymond K Neff, and Lee Goldman. Clinical prediction rules: applications and methodological standards. *New England Journal of Medicine*, 313(13):793–799, 1985.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin Danek, and Jimeng Sun. PyHealth: A deep learning toolkit for healthcare predictive modeling. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2023*, 2023a. URL `https://github.com/sunlabuiuc/PyHealth`.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*, 2023b.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 715–723, 2021.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.