DearLLM: Enhancing Personalized Healthcare via Large Language Models-Deduced Feature Correlations

Yongxin Xu[1,2 *], **Xinke Jiang**[1,2 *], **Xu Chu**[1,2,4,5†], **Rihong Qiu**[1,2], **Yujie Feng**[7], **Hongxin Ding**[1,2], **Junfeng Zhao**[1,2,6], **Yasha Wang**[2,3,5], **Bing Xie**[1,2],

[1] School of Computer Science and School of Software & Microelectronics, Peking University, Beijing, China
[2] Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[3] National Engineering Research Center For Software Engineering, Peking University, Beijing, China
[4] Center on Frontiers of Computing Studies, Peking University, Beijing, China
[5] Peking University Information Technology Institute (Tianjin Binhai)
[6] Nanhu Laboratory, Jiaxing, China
[7] Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.
{xuyx, xinkejiang}@stu.pku.edu.cn; chu_xu@pku.edu.cn

Abstract

Exploring the correlations between medical features is essential for extracting patient health patterns from electronic health records (EHR) data, and strengthening medical predictions and decision-making. To constrain the hypothesis space of pure data-driven deep learning in the context of limited annotated data, a common trend is to incorporate external knowledge, especially knowledge priors related to personalized health contexts, to optimize model training. However, most existing methods lack flexibility and are constrained by the uncertainties brought about by fixed feature correlation priors. In addition, in utilizing knowledge, these methods overlook the knowledge informative for personalized healthcare. To this end, we propose DearLLM, a novel and effective framework that leverages feature correlations deduced by large language models (LLMs) to enhance personalized healthcare. Concretely, DearLLM captures and learns quantitative correlations between medical features by calculating the conditional perplexity of LLMs' deduction based on personalized patient backgrounds. Then, DearLLM enhances healthcare predictions by emphasizing knowledge that carries unique patient information through a feature-frequency-aware graph pooling method. Extensive experiments on two real-world benchmark datasets show significant performance gains brought by DearLLM. Furthermore, the discovered findings align well with medical literature, offering meaningful clinical interpretations.

1 Introduction

With the widespread adoption and deployment of electronic medical systems, the availability of electronic health records (EHR) for medical services and clinical research is continuously increasing. Therefore, as a leading approach that has achieved record-breaking achievements in various fields, deep learning technology has been applied to extract patients' health patterns from EHR data, and to optimize patient treatment and resource allocation through various healthcare tasks (Choi et al. 2016; Ma et al. 2017, 2020; Xu et al. 2024b). In these processes, extracting the correlations between input features is highly beneficial for learning compact patient health representations and enhancing the performance of downstream tasks (Kosambi 2016; Ma et al. 2022).

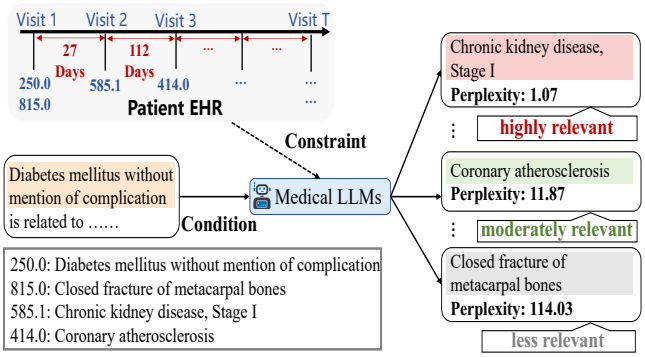

Figure 1: Using conditional perplexity to detect correlation degree between medical features. Lower perplexity indicates a stronger correlation, as perplexity measures the likelihood of predicting one feature based on another. In the personalized healthcare context of this patient, *Diabetes mellitus without mention of complication (250.0)* is more related to *Chronic kidney disease, Stage I (585.1)*.

However, the sparse nature of EHR data poses challenges for these data-hungry deep learning models in learning feature correlations with limited training samples (Ren, Wang, and Zhao 2022). To tackle this essential obstacle, there are some research works delving into leveraging external medical knowledge as priors to narrow down the hypothesis space for model learning (Ma et al. 2018; Lu et al. 2021; Tan et al. 2022). Particularly, extracting information relevant to the patient's personalized health context from the entire knowledge graph has already shown potential (Ye et al. 2021a; Yang et al. 2023; Xu et al. 2023a; Jiang et al. 2023, 2024a).

*These authors contributed equally.
†Corresponding Author.

However, they still face the following challenges:

Challenge I: When extracting knowledge, they are limited by generality and extra uncertainty. On the one hand, existing methods either suffer from the inability of the hierarchical ontologies they use to capture the various relational information between medical features of different categories, or they may suffer from information loss due to the reliance on heterogeneous encoding transformations of medical features and external knowledge, making them not flexible enough for ubiquitous medical tasks. On the other hand, existing methods typically only extract fixed triples as feature correlation priors. However, in real-world medical applications, the correlation strength between medical features can often be compared and evaluated. For example, although *Type 2 Diabetes* has comorbidity relationships with both *Hypertension* and *Cardiovascular Disease*, the correlation between *Type 2 Diabetes* and *Hypertension* is closer due to similar risk factors (Petrie, Guzik, and Touyz 2018). The uncertainty in the learning process can be further reduced if quantitatively comparable correlation priors are introduced.

Challenge II: When utilizing knowledge, they are limited by the dilution of the knowledge informative for personalized health. Most methods treat each type of knowledge equally by mean-pooling the feature representations. However, features (i.e., diagnosis codes included in each visit) that are prevalent in large patient populations may not provide meaningful perspective into a patient's personalized health status. Features containing unique patient information and critical for personalized health insights might occur less frequently within the overall data distribution, yet they hold the potential to deliver more profound insights for robust personalized health representations.

To solve the aforementioned challenges, in light of the impressive natural language understanding capabilities (Zhao et al. 2023), the powerful reasoning abilities (Wei et al. 2022; Huang and Chang 2022; Feng et al. 2023, 2024b,a; Ma et al. 2024), and the potential to serve as knowledge bases of large language models (LLMs) (OpenAI 2023; Touvron et al. 2023; Singhal et al. 2022), we propose a novel framework called DearLLM, which comprehensively mines the precise feature correlations deduced by LLMs as external knowledge priors in a more universal manner. By reducing the uncertainty in knowledge extraction and increasing the information density during knowledge utilization, DearLLM further enhances the efficacy of personalized healthcare. Specifically, for solving the **Challenge I**, our key idea is based on the knowledge injected into LLMs in domain adaptation, using **natural language as the key** to detect the degree of correlations between medical features. As shown in Figure 1, inspired by the success of LLMs-based generative capabilities in ranking tasks (Mao et al. 2023; Sachan et al. 2022), we argue that the **conditional perplexity**, calculated when guiding LLMs to deduce feature correlations, serves as a **direct and significant signal** for measuring feature correlations. This is because it naturally evaluates the degree of correlation between the "hypothetical text" (e.g., *Diabetes mellitus without mention of complication* is related to *Chronic kidney disease, Stage I*) and the medical knowledge encoded in LLMs within the personalized healthcare contexts (Jelinek et al. 1977; Mao et al. 2023). Therefore, we propose the adoption of this direct and crucial signal to precisely quantify the effectiveness of prior knowledge, aiming to further reduce the uncertainty inherent in fixed qualitative priors. Meanwhile, this method of stimulating the semantic comprehension capabilities of LLMs greatly enhances the universality of knowledge mining and utilization. Addressing **Challenge II**, after utilizing graph neural networks (GNNs) to model and learn these LLMs-deduced quantitative correlation priors based on global topological structure information, we propose a feature-frequency-aware graph pooling approach to optimize the weight allocation during knowledge utilization by comprehensively considering both the frequency of features in the personalized health context and their distribution across the entire training set. This allows the model to focus more on features that carry unique patient information and are informative and valuable for personalized health in the final aggregation phase. Our main contributions are as follows:

- DearLLM is the first work utilizing LLMs to deduce quantitative correlation priors between medical features, further reducing the learning difficulty and uncertainty in the process of integrating external knowledge.
- We propose a feature-frequency-aware graph pooling method, which highlights knowledge valuable for personalized healthcare for the process of utilizing knowledge.
- We conduct experiments on two real-world EHR datasets to verify the performance. The results demonstrate the efficacy of DearLLM. Ablation studies and further analysis substantiate the reasonableness and interpretability of the proposed framework.

2 Task Definition

Definition 1 (Diagnosis Codes). Let $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ signify all distinct diagnosis codes, and $|\mathcal{C}|$ is the total count.

Definition 2 (EHR Dataset). Each patient's visit sequence is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, with the t-th visit indicated by a multi-hot vector $\mathbf{x}_t \in \{0,1\}^{|C|}$. In any visit vector, the i-th element is assigned 1 if the visit includes diagnosis code c_i.

Definition 3 (LLMs-Deduced Feature Prior Graph). We aim to utilize the medical LLMs (denoted as $\mathcal{M}$) to deduce correlations between medical features in the personalized healthcare contexts and to form a patient-specific feature prior graph. It can be denoted in the form of triples as $\mathcal{G} = \{(c_i, \alpha_{ij}, c_j) \mid c_i, c_j \in \mathcal{C}\}$, where c_i and c_j are diagnosis codes, and α_{ij} represents the strength of the correlation from c_i to c_j deduced by $\mathcal{M}$.

Definition 4 (Mortality prediction). We outline our predictive goal as a mortality prediction task. Given a patient's visit sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, the task is to predict the survival status $y \in \{0, 1\}$.

3 Methodology

3.1 Overview

As in Figure 2, DearLLM includes these modules:

- **Feature Extractor** takes the EHR data $\mathbf{X}$ as the input and encodes $\mathbf{X}$ into a hidden representation vector.

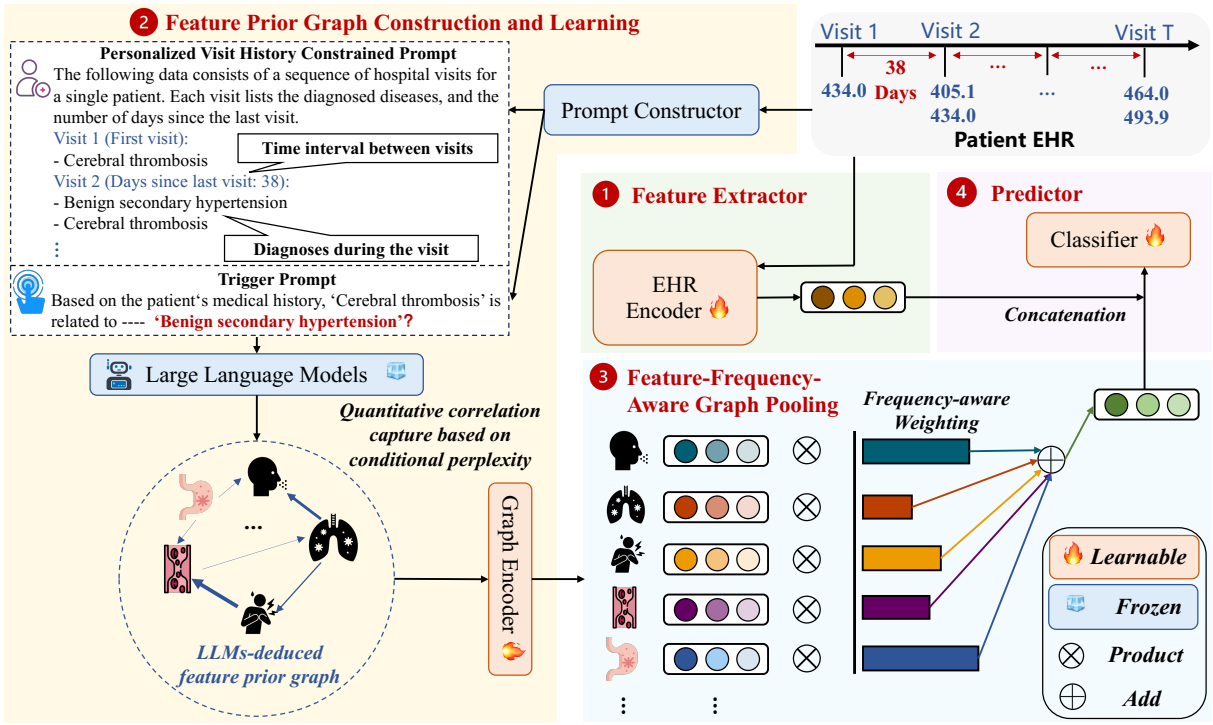

Figure 2: Illustration of DearLLM.

- **Feature Prior Graph Construction and Learning** utilizes the conditional perplexity during LLMs' inference to quantify the degree of feature correlations in personalized healthcare contexts, and employs a graph encoder to model and learn these correlations.
- **Feature-Frequency-Aware Graph Pooling** aggregates knowledge in the constructed feature prior graph via frequency-aware weighting, focusing on features carrying unique patient information for personalized health.
- **Predictor** combines hidden representations with weighted feature knowledge for specific prediction tasks.

3.2 Feature Extractor

DearLLM is a general framework that can be built on various existing feature extractors. These feature extractors are based on efficient deep learning architectures, such as recurrent neural networks (RNNs) (Cho et al. 2014) and Transformer (Vaswani et al. 2017), and are capable of extracting temporal patterns from EHR data. Concretely, considering a sequence of medical records $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, we use the hidden representations before the final layer of these feature extractor backbones (i.e., EHR Encoder) to represent the patient's health status:

$$\mathbf{h}_T = \text{Backbone}\left([\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]\right), \quad (1)$$

where $\mathbf{h}_T \in \mathbb{R}^e$ is the hidden vector of dimension e, encompassing historical visit information.

3.3 Feature Prior Graph Construction&Learning

Feature Prior Graph Construction To enhance the generality of the knowledge infusion process and further reduce uncertainty, we propose using natural language as the key to guide LLMs in deducing quantitative correlation priors between medical features, thereby facilitating the learning of downstream tasks. In our study, we specifically focus on the application of LLMs tailored for medical domains (Chen et al. 2023; Bao et al. 2023; Zhang et al. 2023) as they are trained under huge and various domain-specific knowledge, such as high-quality medical literature corpus and doctor-patient dialogues, etc. Additionally, these models can be easily accessed and deployed locally, which benefits the versatility of our framework by ensuring better data privacy and security compared to using closed-source LLMs with commercialized APIs. Specifically, DearLLM employs HuatuoGPT-II (Chen et al. 2023) (denoted as $\mathcal{M}$) to deduce correlations among medical features, which employs a one-stage domain adaptation protocol and achieves outstanding performance across various benchmarks.

However, a key challenge in achieving precise personalized healthcare predictions is how to transform the broad insights of $\mathcal{M}$ into personalized inferences that are directly meaningful to individual patients. Specifically, for different patients, even for the same disease, the causes may vary significantly (Tan et al. 2023; Xu et al. 2023b). Therefore, it is necessary to perform personalized inferences on the disease progression for different patients. Secondly, as diseases evolve over time, the time-effectiveness of medical feature correlations in different contexts should also be considered (Schechtman and Shelef 2018). To address the aforementioned challenges and achieve more accurate personalized inference, in Figure 2, we propose to construct *Person-*

alized Visit History Constrained Prompt $\mathcal{P}_p$ to obtain natural language inputs representing the personalized healthcare context. First, we outline the format of the input data, such as "*The data format includes ...*". Then, since diagnosis codes (e.g., ICD-9) also represent diseases or symptoms related to natural language names, it is natural to consider converting these longitudinal records into natural language descriptions acceptable to $\mathcal{M}$, based on $\mathbf{X}$. In addition, we include information about the time interval between visits in the prompt, e.g., "*Days since last visit: ...*". Through *Personalized Visit History Constrained Prompt*, DearLLM can effectively mine $\mathcal{M}$'s medical knowledge within the constraints of the personalized healthcare context.

Next, inspired by the success of LLMs-based generative capabilities in ranking tasks (Mao et al. 2023; Sachan et al. 2022), we propose a new method for measuring the degree of correlations between different features under the constraints of the personalized healthcare context. Specifically, this method is implemented by calculating the conditional perplexity of $\mathcal{M}$ when predicting one feature (e.g., c_j) given another feature (e.g., c_i). This is based on the premise that perplexity is a metric for assessing the probability of output sequences, defined as the exponentiated average negative log-likelihood when LLMs generate a particular sequence (Jelinek et al. 1977). Furthermore, as shown in Figure 2, we design the *Trigger Prompt* $\mathcal{P}_t$, which seamlessly follows $\mathcal{P}_p$ by adding the text "*Based on the patient's medical history*". It adds the probe (i.e., "*is related to ...*"), guiding $\mathcal{M}$ to deduce the correlation from c_i to c_j. Then, we calculate the conditional perplexity of $\mathcal{M}$'s predictions, denoted as *cppl*, as quantitative correlation priors:

$$cppl_{ij} = \exp\Big(-\frac{1}{n}\sum_{k}^{n}\log p_\theta(q_k|\mathcal{P}_p, \mathcal{P}_t, q_{<k})\Big), \tag{2}$$

where θ represents the parameters of $\mathcal{M}$ (i.e., the implicit medical knowledge), $p_\theta(\cdot)$ denotes the probability output from $\mathcal{M}$, q_k denotes the k-th token of the natural language name of c_j, and n is the token length. It's important to note that the lower the value of $cppl_{ij}$, the stronger the correlation from c_i to c_j in the personalized healthcare context.

After obtaining the quantitative correlations between each pair of features (i.e., diagnosis codes), to further capture and learn complex and high-order relationships, we convert these quantitative correlations into a patient-specific feature prior graph $\mathcal{G}$ (Scarselli et al. 2008), where the graph nodes denote medical features, and their connected edges denote the correlation weights between features. To enhance the effectiveness of correlation modeling and learning, inspired by the directional nature among diseases in real medical scenarios (Angold, Costello, and Erkanli 1999; Bayliss et al. 2003; Jiang et al. 2024b) (namely, even in the context of comorbidities, some diseases act as primary risk factors and play a dominant role), we model $\mathcal{G}$ as a directed graph to more accurately reflect these dynamic and dominant relationships. Specifically, we can represent the graph using a $N \times N$ adjacency matrix $\mathbf{A}$, whose element $\alpha_{ij} \in [0, 1]$ denotes the correlation weight from feature c_i to c_j, and N is the number of types of diagnosis codes in patient's medical history. To avoid the impact of outliers in *cppl* with large-scale numerical scales on learning effectiveness and further reduce the learning complexity of the fully connected dense structure (Patro and Sahu 2015), we propose a pruning and normalization strategy oriented to conditional perplexity, which calculates the boundaries for the minimum and maximum values by obtaining the upper and lower bounds of the δ (e.g., 90%) confidence interval of the *cppl* collection for all patients in the training set. These boundaries are denoted as $cppl_{max}$ and $cppl_{min}$, respectively, and *cppl* values outside the confidence interval are clipped to these boundaries. Meanwhile, we make lower *cppl* values correspond to higher correlation weights:

$$\alpha_{ij} = 1 - \frac{cppl_{ij} - cppl_{min}}{cppl_{max} - cppl_{min}}. \tag{3}$$

Next, we cut off most of the useless connections between features with weak values. Consequently, the receptive field of each feature node is confined to features with higher correlation, further reducing the complexity of learning (Jiang et al. 2024c; Zhang et al. 2024; Ma et al. 2023).

Graph Encoder Next, DearLLM uses Graph Convolutional Network (GCN) (Kipf and Welling 2016) to enhance feature correlation learning by leveraging global topological structure information. Specifically, for the l-th layer of the GCN, it employs a parameter matrix $\mathbf{W}^{(l)}$ for feature transformation, and facilitates the interaction of features with all adjacent nodes based on the adjacency matrix $\mathbf{A}$:

$$\mathbf{H}^{(0)} = \mathbf{Z}, \quad \mathbf{H}^{(l+1)} = \mathrm{ReLU}(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times o}$ is the initial representation of diagnosis codes, $\mathbf{W}^{(l)} \in \mathbb{R}^{o\times o}$, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$, and $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$. After message propagation through L layers of GCN, DearLLM extracts global correlations and achieves the final knowledge representation $\mathbf{H}^{(L)} \in \mathbb{R}^{N\times o}$.

3.4 Feature-Frequency-Aware Graph Pooling

To highlight knowledge valuable for personalized healthcare in the process of utilizing knowledge to enhance embedding learning, inspired by the successful application of Term Frequency-Inverse Document Frequency (TF-IDF) in the field of natural language processing for measuring the relevance of words to specific documents (Ramos et al. 2003; Qaiser and Ali 2018), we propose a feature-frequency-aware graph pooling method. Specifically, we treat each patient's visit history as a singular document, transforming the collection of these records into a *Code Corpus*. To calculate the significance of each feature in the feature prior graph for personalized healthcare, we first define Code Frequency $\mathrm{CF}(i; j)$ to measure the frequency of a specific feature (i.e., diagnosis code) c_i in the entire visit history $\mathbf{X}^{(j)}$ of the current patient j. Then, we define Inverse Patient Frequency $\mathrm{IPF}(i)$ to measure the distribution of c_i across the entire training set, and then deemphasize features that are prevalent throughout the training set:

$$\begin{aligned}\mathrm{CF}(i; j) &= \frac{\#\text{ of times } c_i \text{ appears in } \mathbf{X}^{(j)}}{\#\text{ of codes in } \mathbf{X}^{(j)}}\\ \mathrm{IPF}(i) &= \log\Big(\frac{\#\text{ of patients}}{1 + \#\text{ of patients with code } c_i}\Big).\end{aligned} \tag{5}$$

We prioritize features (i.e., diagnosis codes) that frequently occur within personalized medical history but are less common across training set by multiplying these two indicators, thereby assigning higher weights to discriminative codes:

$$\text{CF-IPF}(i) = \text{CF}(i; j) \times \text{IPF}(i). \tag{6}$$

Next, we perform frequency-aware weighting based on CF-IPF scores of each feature by using a softmax layer on all N feature nodes in patient-specific feature prior graph $\mathcal{G}$, and aggregate knowledge representations in $\mathcal{G}$ based on weights:

$$\begin{aligned} \pi_1, \cdots, \pi_N &= \text{Softmax}\left(\text{CF-IPF}(1), \cdots, \text{CF-IPF}(N)\right), \\ \mathbf{h}_s &= \sum_{i=1}^{N} \pi_i \mathbf{H}_i^{(L)}, \end{aligned} \tag{7}$$

where $\mathbf{h}_s \in \mathbb{R}^o$ is hidden vector, $\mathbf{H}_i^{(L)} \in \mathbb{R}^o$ is the final layer embedding from Graph Encoder for c_i.

3.5 Predictor

Following previous sections, we now have the patient health status representation $\mathbf{h}_T$ and the aggregated feature knowledge representation $\mathbf{h}_s$. To get the feature for the final prediction, we integrate $\mathbf{h}_T$ and $\mathbf{h}_s$ through concatenation. Then, the construction of the predictor can be achieved through the implementation of a fully connected layer. The predicted probability can be calculated:

$$\hat{y} = \text{Sigmoid}\left(\mathbf{W}_y \left(\mathbf{h}_T \| \mathbf{h}_s\right) + \mathbf{b}_y\right), \tag{8}$$

where $\mathbf{W}_y \in \mathbb{R}^{1\times(e+o)}$ and $\mathbf{b}_y \in \mathbb{R}^1$ are learnable parameters of linear transformation, and $\|$ denotes concatenation.

3.6 Training

Notice that DearLLM employs the predicted probability $\hat{y}$ to calculate the Binary Cross-Entropy (BCE) Loss:

$$\min_{\Theta} \mathcal{L} = -\frac{1}{B} \sum_{u=1}^{B} \left(y_u^\top \log\left(\hat{y}_u\right) + \left(1 - y_u\right)^\top \log\left(1 - \hat{y}_u\right) \right) + \eta ||\Theta||^2, \tag{9}$$

where B is the batch size, $\hat{y}_u \in [0, 1]$ is the predicted probability, and $y_u \in \{0, 1\}$ is the ground truth. Θ represents all trainable parameters (i.e., the parameters of Graph Encoder, EHR Encoder, and Predictor) of DearLLM. L_2 regularization with η on Θ is conducted to prevent over-fitting.

4 Experiment

In this section, we provide detailed information on the experimental setup, further analysis to validate the performance, rationality, and interpretability of DearLLM. The code is provided in [1].

4.1 Experimental Setup

Datasets In this paper, experiments are conducted on two publicly available real-world EHR datasets: MIMIC-III (Johnson et al. 2016) and MIMIC-IV (Johnson et al. 2023). Following (Ye et al. 2021a) and (Jiang et al. 2024a), we choose patients with at least two visits, using the outcome of the last visit as the mortality label and the remaining visits as the input data. It's worth mentioning that MIMIC-IV covers admissions over a decade from 2008 to 2019, which overlaps with the time range of the MIMIC-III data (between 2001 and 2012). Therefore, following (Lu, Han, and Ning 2022), we filtered out duplicates. Both the MIMIC-III and MIMIC-IV datasets are imbalanced, with their mortality (positive) rates being 19.36% and 4.85%, respectively.

To achieve a better balance between performance and efficiency, we construct a feature prior graph based on the patient's most recent five visits and employ vLLM (Kwon et al. 2023) to expedite the inference process. During both prediction and graph construction, we utilize ICD diagnosis codes.

Baselines We compare DearLLM with several state-of-the-art methods from two perspectives. Firstly, DearLLM is a general framework that can be combined with various EHR feature extractors. To validate the predictive performance of the proposed DearLLM, without losing generality, we choose four baseline models: Gated Recurrent Units (GRU) (Cho et al. 2014), StageNet (Gao et al. 2020), HiTANet (Luo et al. 2020) and Transformer (Vaswani et al. 2017). They are widely applied in various EHR analysis tasks, and can serve both as baseline models and as feature extractors for DearLLM:

- **GRU** (Cho et al. 2014) is a typical deep learning model for capturing temporal dependencies.
- **StageNet** (Gao et al. 2020) improves long-short term memory with personalized disease progression stages.
- **HiTANet** (Luo et al. 2020) introduces time-awareness into the self-attention module.
- **Transformer** (Vaswani et al. 2017) is a deep learning model architecture based on self-attention mechanisms.

Next, to validate the effectiveness in extracting and utilizing knowledge, we also compare several SOTA methods for incorporating external knowledge, and evaluate performance based on the aforementioned four types of backbones:

- **MedRetriever** (Ye et al. 2021b) enhances prediction performance by retrieving unstructured medical texts.
- **RAM-EHR** (Xu et al. 2024a) enhances the local model by retrieving knowledge from multiple sources and applying consistency regularization.
- **GraphCare** (Jiang et al. 2024a) acquires personalized knowledge by directing LLMs to produce triples.

Evaluation Metrics and Strategy We employ three widely used evaluation metrics to measure the performance, namely, the area under the precision-recall curve (AUPRC), the area under the receiver operating characteristic curve (AUROC), and the F1-score. Higher scores in these metrics indicate better predictive performance of the model. Both MIMIC-III and MIMIC-IV datasets are randomly divided into training, validation and testing sets in a 8:1:1 ratio. We select the best model based on its performance on validation set, and run the algorithm 5 times, reporting the average and standard deviation in Table 1.

4.2 Experimental Results

Performance Comparison Table 1 shows the performance of DearLLM and baselines accross two datasets.

[1] https://github.com/Artessay/DearLLM

Dataset	Methods	GRU			StageNet			HiTANet			Transformer		
	Metric	AUPRC	AUROC	F1-Score	AUPRC	AUROC	F1-Score	AUPRC	AUROC	F1-Score	AUPRC	AUROC	F1-Score
MIMIC-III	Vanilla	.2087(.017)	.4998(.033)	.1330(.026)	.2492(.028)	.5704(.034)	.2210(.044)	.2472(.037)	.5573(.031)	.2028(.050)	.2372(.011)	.5679(.012)	.1669(.048)
	MedRetriever	.2571(.014)	.5624(.033)	.2045(.052)	.2622(.034)	.5857(.035)	.2268(.008)	.2436(.017)	.5669(.018)	.2207(.029)	.2588(.018)	.5957(.019)	.2349(.016)
	RAM-EHR	.2215(.034)	.5892(.043)	.2111(.054)	.2437(.026)	.5537(.024)	.2103(.021)	.2538(.019)	.5712(.020)	.2243(.029)	.2262(.034)	.6250(.018)	.2358(.014)
	GraphCare	.2435(.046)	.5448(.032)	.2050(.076)	.2643(.029)	.5918(.015)	.2438(.023)	.2509(.010)	.5733(.018)	.2250(.032)	.2541(.016)	.5943(.009)	.2481(.025)
	DearLLM	**.3079(.015)**	**.6096(.011)**	**.2251(.033)**	**.3147(.018)**	**.6493(.017)**	**.2660(.034)**	**.2999(.017)**	**.6393(.023)**	**.2631(.037)**	**.2958(.019)**	**.6574(.008)**	**.2685(.021)**
MIMIC-IV	Vanilla	.0685(.042)	.6043(.058)	.1192(.041)	.1173(.038)	.6792(.032)	.1346(.016)	.1193(.018)	.6624(.037)	.1354(.029)	.1461(.023)	.7262(.020)	.1182(.043)
	MedRetriever	.1478(.026)	.7544(.031)	.1181(.011)	.1219(.018)	.6682(.024)	.1319(.023)	.1380(.012)	.6907(.039)	.1483(.011)	.1532(.007)	.7223(.024)	.1260(.017)
	RAM-EHR	.1362(.026)	.7457(.035)	.1514(.028)	.1670(.013)	.7039(.022)	.1433(.014)	.1624(.015)	.7389(.012)	.1280(.022)	.1667(.011)	.7368(.019)	.1420(.025)
	GraphCare	.1401(.032)	.7114(.027)	.1553(.031)	.1664(.006)	.7133(.014)	.1544(.023)	.1461(.008)	.7284(.012)	.1559(.016)	.1698(.013)	.7380(.010)	.1312(.020)
	DearLLM	**.1735(.013)**	**.7659(.013)**	**.1684(.015)**	**.1806(.014)**	**.7788(.010)**	**.1736(.015)**	**.1862(.009)**	**.7557(.003)**	**.1694(.022)**	**.1996(.015)**	**.7603(.022)**	**.1573(.022)**

Table 1: Performance comparisons of four feature extractor backbones incorporating external knowledge on MIMIC-III and MIMIC-IV datasets. The best performance is in **boldface** and the second runners are underlined.

Dataset	MIMIC-III		MIMIC-IV	
Methods	AUPRC	AUROC	AUPRC	AUROC
DearLLM	**0.3079**	**0.6096**	**0.1735**	**0.7659**
DearLLM$_{p-}$	0.2893	0.5859	0.1633	0.7542
DearLLM$_{f-}$	0.2822	0.5788	0.1619	0.7491
DearLLM$_{a-}$	0.1909	0.4902	0.0593	0.5952

Table 2: Ablation study results of our proposed DearLLM.

Overall, across all evaluation metrics on the two datasets, especially AUPRC, which is the most informative primary evaluation metric when dealing with highly imbalanced datasets, DearLLM **significantly surpasses the current state-of-the-art methods**. Firstly, for the four feature extractor baselines, we initially evaluate the vanilla model, and then integrate the LLMs-Deduced Feature Prior Graph on this basis, forming the corresponding DearLLM model. Compared to vanilla model, we observe that DearLLM consistently improves the predictive performance. This reveals the necessity of incorporating external knowledge to reduce model's learning hypothesis space in the context of limited training samples. Secondly, compared to other methods that also incorporate external medical knowledge, DearLLM outperforms the aforementioned methods. This indicates the effectiveness of utilizing LLMs to deduce quantitative feature correlation priors and emphasizing knowledge carrying unique patient information in reducing learning uncertainty.

Ablation Study We conduct following ablation studies to examine DearLLM by implementing several variants of DearLLM: DearLLM$_{p-}$, DearLLM$_{f-}$, and DearLLM$_{a-}$. In DearLLM$_{p-}$, we remove *Personalized Visit History Constrained Prompt* when guiding LLMs to deduce feature correlations. In DearLLM$_{f-}$, we replace *Feature-Frequency-Aware Graph Pooling* module with mean pooling. Furthermore, to validate the effectiveness of the quantitative correlation priors deduced by LLMs, we add DearLLM$_{a-}$, which randomly permutes α_{ij} in each patient-specific feature prior graph. Without loss of generality, we use GRU for DearLLM in our analysis. As shown in Table 2, the performance degradation of DearLLM$_{p-}$ relative to DearLLM indicates the necessity of introducing personalized healthcare context as a constraint when guiding LLMs in reasoning. Secondly, the superior performance of DearLLM over DearLLM$_{f-}$ proves the effectiveness of highlighting knowledge valuable for personalized healthcare in the process of utilizing knowledge. In particular, for DearLLM$_{a-}$, its performance drops significantly on both datasets, indicating that our proposed LLMs-deduced quantitative correlation priors are crucial in reducing uncertainty and improving predictions.

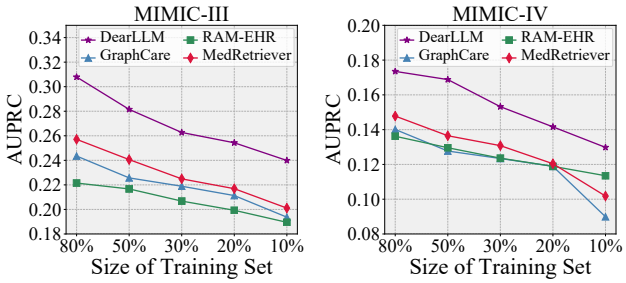

Figure 3: Performance comparisons of DearLLM and baselines (GRU as the backbone) on MIMIC-III (left) and MIMIC-IV (right) datasets under different training data size.

4.3 Analysis

Robustness Against Data Insufficiency To investigate the impact and effectiveness of using quantitative correlation prior knowledge derived from LLMs in enhancing robustness under conditions of scarce training samples, we simulate scenarios with extremely scarce training data and conduct comprehensive experiments. Specifically, we adjust the sample size of the training set (i.e., the number of patients) for MIMIC-III and MIMIC-IV datasets from originally comprising 80% of the total dataset to 50%, 30%, 20%, and 10%, while keeping the validation and test data unchanged. Without loss of generality, we use GRU for DearLLM and other methods incorporating external knowledge, and evaluate the performance under these various settings. As shown in Figure 3, DearLLM outperforms all selected baselines in all settings, indicating that our proposed LLMs-deduced quantitative correlation priors are capable of enhancing robustness under data scarcity by reducing external prior uncertainty.

Case Study for Model Reasonableness and Interpretability To illustrate the reasonableness, we provide a case study to explain the frequency-aware weights and the quantitative feature correlations deduced by LLMs in the personalized disease progression process. Figure 4 shows the feature prior graph constructed by DearLLM for a positive (expired)

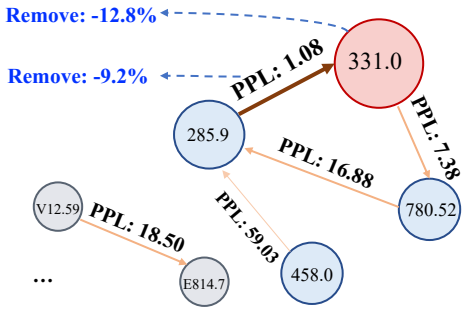

Figure 4: A positive sample from MIMIC-III test set. Nodes with higher frequency-aware weights are warmer in color and larger in size, while edges with lower perplexity (PPL) (i.e., higher correlation weight) are darker and thicker.

Code	Meanings and frequency-aware weights
331.0	Alzheimer's disease (0.148)
285.9	Anemia, unspecified (0.076)
458.0	Orthostatic hypotension (0.047)
780.52	Insomnia, unspecified (0.041)
E814.7	Motor vehicle traffic accident involving collision with pedestrian injuring pedestrian (0.014)
V12.59	Personal history of other diseases of circulatory system (0.011)

Table 3: ICD-9 codes appearing in Figure 4, and their meanings and frequency-aware weights. The numbers in the brackets represent weights, and the most important feature knowledge is in red.

patient, as well as the quantitative feature correlations calculated by DearLLM based on LLMs' inference. To facilitate understanding, we provide the codes appearing in the figure along with their meanings and frequency-aware weights in Table 3. Utilizing DearLLM, the predicted probability of the patient's unfortunate death is 0.912. We observe that important node associated with mortality prediction, *Alzheimer's disease (331.0)*, has been allocated the highest importance score, aligning with medical research (Ganguli et al. 2005; James et al. 2014). However, if we remove this node, the probability of being positive drops to 0.784. Furthermore, DearLLM is highly focused on the associative impact of *Anemia, unspecified (285.9)* on *Alzheimer's disease (331.0)*. According to medical research (Faux et al. 2014; Beard et al. 1997), the reduction of hemoglobin is strongly associated with complications of Alzheimer's disease, which increases the mortality risk for patients. If we eliminate this crucial correlation, the probability of being positive reduces from 0.912 to 0.820. These observation indicate that DearLLM can capture important medical feature correlations like doctors, demonstrating its reasonableness and interpretability.

4.4 Related Work

Healthcare predictive models focusing on capturing sequence correlations. Due to the longitudinal property of EHR data, one category of methods focuses on capturing contextual dependencies between visits through deep temporal models. Firstly, some of these works adopt architectures based on recurrent neural networks (RNNs). For example, RETAIN (Choi et al. 2016) captures important visits and key diagnoses by adding a two-level attention mechanism to the RNNs. Dipole (Ma et al. 2017) introduces three types of attention mechanisms based on bidirectional RNNs. Secondly, SAnD (Song et al. 2018), LSAN (Ye et al. 2020), and HiTANet (Luo et al. 2020) capture the correlations between visits and features based on the self-attention module of the Transformer structure. Furthermore, T-LSTM (Baytas et al. 2017) and StageNet (Gao et al. 2020) further consider the impact of time intervals in EHR data.

Incorporating medical knowledge in healthcare prediction. To reduce the hypothesis space for model learning, some other works try to integrate external medical knowledge. For example, GRAM (Choi et al. 2017) and KAME (Ma et al. 2018) enhance learning by combining the hierarchical information of medical knowledge ontologies. MetaCare++ (Tan et al. 2022) and CGL (Lu et al. 2021) combine patient-disease interactions and medical domain knowledge. MedRetriever (Ye et al. 2021b) enhances predictions based on a dynamically updated text memory bank. To achieve better personalized diagnoses, MedPath (Ye et al. 2021a), KerPrint (Yang et al. 2023), and SeqCare (Xu et al. 2023a) construct a personalized knowledge graph for each patient by extracting information related to the medical features of each patient from a large-scale medical knowledge graph. RAM-EHR (Xu et al. 2024a) retrieves multi-source knowledge and enhances the relevance of external knowledge through LLMs-based summarization. GraphCare (Jiang et al. 2024a) extracts personalized knowledge by guiding LLMs to directly generate triples and sampling relevant subgraphs from knowledge graphs. Despite their effectiveness, most existing methods lack flexibility and are limited by uncertainties from fixed feature correlation priors. Additionally, they overlook valuable information for personalized healthcare when utilizing knowledge.

5 Conclusions and Future Works

In this paper, we propose a novel and general framework, named DearLLM, which leverages feature correlations deduced by LLMs to compensate for the mismatch between low sample complexity and high learning complexity in EHR data. Specifically, DearLLM first guides LLMs to deduce the correlations between medical features based on personalized patient context, and accurately quantifies the effectiveness of prior knowledge by calculating the conditional perplexity of predictions. Then, DearLLM models these quantitative feature correlations as graph structure, capturing the global correlations via GNNs. In the process of aggregating and utilizing knowledge, DearLLM emphasizes important knowledge that contains unique patient information through a feature-frequency-aware graph pooling method. Extensive experimental results on two public benchmark datasets demonstrate the clear advantages of DearLLM over the SOTA baselines, validate its reasonableness and interpretability. Although vertical domain LLMs generally provide benefits, hallucinations may occur. In the future, we will dedicate efforts to further mitigate potential hallucination during the knowledge enhancement process.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20468).

References

Angold, A.; Costello, E. J.; and Erkanli, A. 1999. Comorbidity. *J Child Psychol Psychiatry*.

Bao, Z.; Chen, W.; Xiao, S.; Ren, K.; Wu, J.; Zhong, C.; Peng, J.; Huang, X.; and Wei, Z. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. arXiv:2308.14346.

Bayliss, E. A.; Steiner, J. F.; Fernald, D. H.; Crane, L. A.; and Main, D. S. 2003. Descriptions of barriers to self-care by persons with comorbid chronic diseases. *Ann Fam Med*.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware LSTM networks. In *SIGKDD*.

Beard, C. M.; Kokmen, E.; O'Brien, P. C.; Anía, B. J.; and Melton III, L. J. 1997. Risk of Alzheimer's disease among elderly patients with anemia: population-based investigations in Olmsted County, Minnesota. *Ann Epidemiol*.

Chen, J.; Wang, X.; Gao, A.; Jiang, F.; Chen, S.; Zhang, H.; Song, D.; Xie, W.; Kong, C.; Li, J.; et al. 2023. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *SIGKDD*.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*.

Faux, N. G.; Rembach, A.; Wiley, J.; Ellis, K. A.; Ames, D.; Fowler, C. J.; Martins, R. N.; Pertile, K. K.; Rumble, R. L.; Trounson, B.; et al. 2014. An anemia of Alzheimer's disease. *Mol. Psychiatry*.

Feng, Y.; Chu, X.; Xu, Y.; Lu, Z.; Liu, B.; Yu, P. S.; and Wu, X.-M. 2024a. Tasl: Task skill localization and consolidation for language model continual learning. *arXiv preprint arXiv:2408.05200*.

Feng, Y.; Chu, X.; Xu, Y.; Shi, G.; Liu, B.; and Wu, X.-M. 2024b. TaSL: Continual Dialog State Tracking via Task Skill Localization and Consolidation. In *ACL*.

Feng, Y.; Lu, Z.; Liu, B.; Zhan, L.; and Wu, X.-M. 2023. Towards LLM-driven Dialogue State Tracking. In *EMNLP*.

Ganguli, M.; Dodge, H. H.; Shen, C.; Pandav, R. S.; and DeKosky, S. T. 2005. Alzheimer disease and mortality: a 15-year epidemiological study. *Arch. neurol. neurosci.*

Gao, J.; Xiao, C.; Wang, Y.; Tang, W.; Glass, L. M.; and Sun, J. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *WWW*.

Huang, J.; and Chang, K. C.-C. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

James, B. D.; Leurgans, S. E.; Hebert, L. E.; Scherr, P. A.; Yaffe, K.; and Bennett, D. A. 2014. Contribution of Alzheimer disease to mortality in the United States. *Neurology*.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *J. Acoust*.

Jiang, P.; Xiao, C.; Cross, A.; and Sun, J. 2024a. GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. In *ICLR*.

Jiang, X.; Fang, Y.; Qiu, R.; Zhang, H.; Xu, Y.; Chen, H.; Zhang, W.; Zhang, R.; Fang, Y.; Chu, X.; et al. 2024b. TC-RAG: Turing-Complete RAG's Case study on Medical LLM Systems. *arXiv preprint arXiv:2408.09199*.

Jiang, X.; Qiu, R.; Xu, Y.; Zhang, W.; Zhu, Y.; Zhang, R.; Fang, Y.; Chu, X.; Zhao, J.; and Wang, Y. 2024c. RAGraph: A General Retrieval-Augmented Graph Learning Framework. *NeurIPS*.

Jiang, X.; Zhang, R.; Xu, Y.; Qiu, R.; Fang, Y.; Wang, Z.; Tang, J.; Ding, H.; Chu, X.; Zhao, J.; et al. 2023. HyKGE: A Hypothesis Knowledge Graph Enhanced Framework for Accurate and Reliable Medical LLMs Responses. *arXiv preprint arXiv:2312.15883*.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data*.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kosambi, D. 2016. Statistics in function space. *Mathematics and Statistics*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.

Lu, C.; Han, T.; and Ning, Y. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *AAAI*.

Lu, C.; Reddy, C. K.; Chakraborty, P.; Kleinberg, S.; and Ning, Y. 2021. Collaborative Graph Learning with Auxiliary Text for Temporal Event Prediction in Healthcare. In *IJCAI*.

Luo, J.; Ye, M.; Xiao, C.; and Ma, F. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *SIGKDD*.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *SIGKDD*.

Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*.

Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*.

Ma, X.; Chu, X.; Wang, Y.; Lin, Y.; Zhao, J.; Ma, L.; and Zhu, W. 2023. Fused Gromov-Wasserstein graph mixup for graph-level classifications. In *NeurIPS*.

Ma, X.; Chu, X.; Yang, Z.; Lin, Y.; Gao, X.; and Zhao, J. 2024. Parameter Efficient Quasi-Orthogonal Fine-Tuning via Givens Rotation. In *ICML*.

Ma, X.; Wang, Y.; Chu, X.; Ma, L.; Tang, W.; Zhao, J.; Yuan, Y.; and Wang, G. 2022. Patient Health Representation Learning via Correlational Sparse Prior of Medical Features. *TKDE*.

Mao, Z.; Wang, H.; Du, Y.; and Wong, K.-F. 2023. UniTRec: A Unified Text-to-Text Transformer and Joint Contrastive Learning Framework for Text-based Recommendation. In *ACL*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Patro, S.; and Sahu, K. K. 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Petrie, J. R.; Guzik, T. J.; and Touyz, R. M. 2018. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Can J Cardiol*.

Qaiser, S.; and Ali, R. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput*.

Ramos, J.; et al. 2003. Using tf-idf to determine word relevance in document queries. In *FICML*.

Ren, H.; Wang, J.; and Zhao, W. X. 2022. Generative Adversarial Networks Enhanced Pre-training for Insufficient Electronic Health Records Modeling. In *SIGKDD*.

Sachan, D.; Lewis, M.; Joshi, M.; Aghajanyan, A.; Yih, W.-t.; Pineau, J.; and Zettlemoyer, L. 2022. Improving Passage Retrieval with Zero-Shot Question Generation. In *EMNLP*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Trans. Neural Netw*.

Schechtman, E.; and Shelef, A. 2018. Correlation and the time interval over which the variables are measured–A non-parametric approach. *PLOS ONE*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2022. Large language models encode clinical knowledge.

Song, H.; Rajan, D.; Thiagarajan, J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*.

Tan, Y.; Yang, C.; Wei, X.; Chen, C.; Liu, W.; Li, L.; Zhou, J.; and Zheng, X. 2022. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *SIGIR*.

Tan, Y.; Zhou, Z.; Yu, L.; Liu, W.; Chen, C.; Ma, G.; Hu, X.; Hertzberg, V. S.; and Yang, C. 2023. Enhancing Personalized Healthcare via Capturing Disease Severity, Interaction, and Progression. In *ICDM*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Xu, R.; Shi, W.; Yu, Y.; Zhuang, Y.; Jin, B.; Wang, M. D.; Ho, J. C.; and Yang, C. 2024a. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *ACL*.

Xu, Y.; Chu, X.; Yang, K.; Wang, Z.; Zou, P.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023a. SeqCare: Sequential Training with External Medical Knowledge Graph for Diagnosis Prediction in Healthcare Data. In *WWW*.

Xu, Y.; Jiang, X.; Chu, X.; Xiao, Y.; Zhang, C.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2024b. Protomix: Augmenting health status representation learning via prototype-based mixup. In *SIGKDD*, 3633–3644.

Xu, Y.; Yang, K.; Zhang, C.; Zou, P.; Wang, Z.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023b. VecoCare: Visit Sequences-Clinical Notes Joint Learning for Diagnosis Prediction in Healthcare Data. In *IJCAI*.

Yang, K.; Xu, Y.; Zou, P.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023. KerPrint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *AAAI*.

Ye, M.; Cui, S.; Wang, Y.; Luo, J.; Xiao, C.; and Ma, F. 2021a. MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths. In *WWW*.

Ye, M.; Cui, S.; Wang, Y.; Luo, J.; Xiao, C.; and Ma, F. 2021b. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *CIKM*.

Ye, M.; Luo, J.; Xiao, C.; and Ma, F. 2020. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *CIKM*.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Li, J.; Chen, G.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. 2023. HuatuoGPT, towards Taming Language Model to Be a Doctor. *arXiv preprint arXiv:2305.15075*.

Zhang, R.; Jiang, X.; Fang, Y.; Luo, J.; Xu, Y.; Zhu, Y.; Chu, X.; Zhao, J.; and Wang, Y. 2024. Infinite-horizon graph filters: Leveraging power series to enhance sparse information aggregation. *arXiv preprint arXiv:2401.09943*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.