Memorize and Rank: Elevating Large Language Models for Clinical Diagnosis Prediction

Mingyu Derek Ma[1], **Xiaoxuan Wang**[1], **Yijia Xiao**[1], **Anthony Cuturrufo**[1], **Vijay S Nori**[2], **Eran Halperin**[1,2], **Wei Wang**[1]

[1]University of California, Los Angeles
[2]Optum AI
{ma, xw27, yijia.xiao, acc, weiwang}@cs.ucla.edu, vijay.nori@optum.com, eran.halperin@uhg.com

Abstract

Clinical diagnosis prediction models, when provided with a patient's medical history, aim to detect potential diseases early, facilitating timely intervention and improving prognostic outcomes. However, the inherent scarcity of patient data and large disease candidate space often pose challenges in developing satisfactory models for this intricate task. The exploration of leveraging Large Language Models (LLMs) for encapsulating clinical decision processes has been limited. We introduce MERA, a clinical diagnosis prediction model that bridges pertaining natural language knowledge with medical practice. We apply hierarchical contrastive learning on a disease candidate ranking list to alleviate the large decision space issue. With concept memorization through fine-tuning, we bridge the natural language clinical knowledge with medical codes. Experimental results on MIMIC-III and IV datasets show that MERA achieves the state-of-the-art diagnosis prediction performance and dramatically elevates the diagnosis prediction capabilities of generative LMs.

1 Introduction

Electronic Health Records (EHR), containing patient status and diagnoses, embody valuable domain expertise and clinical operation patterns (Caufield et al. 2019). Clinicians make diagnosis judgments based on their extensive medical knowledge, acquired through years of education from textbooks and literature, as well as their accumulated experience derived from clinical practice. Clinical diagnosis prediction aims to predict patients' diseases that are highly likely to be diagnosed in the upcoming hospital admission by analyzing the patients' past diagnoses. The input and output are both presented in sequences of medical codes, which do not directly convey semantic information nor disease property. The resulting AI-enhanced diagnosis system (Morid, Sheng, and Dunbar 2023) may enable early warning of diseases (Rochefort, Buckeridge, and Forster 2015), optimized clinical resource allocation (Yadav et al. 2013), and better risk estimation for sustainable insurance (Hsu et al. 2016).

Two primary challenges in diagnosis prediction have driven various research efforts (Wornow et al. 2023b) but remain unsolved. First, what would be the best practice to incorporate clinical knowledge into the model? Existing works initialize concept embeddings from natural language descriptions (Wu et al. 2023b; Bornet et al. 2023), or enrich patient representation with external disease ontologies (An et al. 2023; Cheong et al. 2023). However, a significant gap persists between the primary knowledge modality, *i.e.* natural language, and the model's hidden representation. Second, how can we handle the large candidate space when making predictions and exploit the supervisory signals induced from this candidate space? The commonly used International Classification of Diseases (ICD) coding system encodes 13k+ diseases (Cartwright 2013). Existing works typically treat the task as k-way classification where k is the number of possible diseases, and then apply cross entropy loss for each disease individually. These approaches often overlook the dependencies among diseases and the structural nuances within the diagnosis coding system.

Generative Language Models (LM), especially the Large Language Models (LLM), are trained to predict the next token, adhere to task instructions (Brown and et al. 2020; Ma et al. 2024a), and align with human preferences (Ouyang and et al 2022). These models exhibit superior capabilities in language understanding and reasoning, as shown by their performance on science-based benchmarks (Ma et al. 2024b; Wu et al. 2023a; Zhang et al. 2024). During the pretraining stage, LLMs assimilate a large amount of knowledge extracted from literature and online corpora. However, there remains an underexplored domain in using LLM for clinical diagnosis prediction, due to the aforementioned gap between natural language and medical code, as well as the disparity between the token-level optimization process and the large candidate outcome space. These challenges impede the effective application of generative LMs to diagnosis prediction tasks, even as the state-of-the-art models predominantly rely on graph neural networks without fully harnessing natural language knowledge (Yang et al. 2023b; Wu et al. 2023b; An et al. 2023). Fine-tuning generative LM LLaMA2 (Touvron and et al. 2023) directly on diagnosis prediction yields almost 20-point lower recall@20 than GNN-based existing best model (Yang et al. 2023b) as shown in Table 1. There are some studies that use transformer-based LM for clinical outcome prediction, but they either do not support structured data as input (Niu et al. 2024; Wang et al. 2023a), not compatible with mainstream LLMs (Rupp, Peter, and Pattipaka 2023; Guo et al. 2023), or only work for narrow output space

with few classes (Wang et al. 2023a; Shoham and Rappoport 2023).

To tackle these challenges, we propose MERA, an LLM designed for clinical diagnosis prediction that incorporates a comprehensive understanding of clinical knowledge by leveraging relationships among medical codes and offers extensive supervision over the output space. The patient's historical diagnosis results are formulated as linear sequences and the LLM is tasked with generating a probability distribution for the diagnosis results in the subsequent visit. Compared with the ordinary paradigm that optimizes the probability of generating the correct token, we optimize the outcome directly. To enhance the inter-visit causal reasoning, we employ contrastive learning to compel the model to distinguish true diagnoses from false ones. The contrastive learning process is extended to multiple levels in the hierarchical organization of the medical codes within the ICD coding ontology. The model is learned to distinguish the true diagnoses from a pool of potential diagnoses while the pool is increasingly relevant to the true ones. To regularize the diagnosis predictions to follow intra-visit diagnosis patterns, we develop a teaching-forcing strategy to optimize the medical code ranking, assuming partial diagnoses of the visit are known. To allow the model to grasp the comprehensive clinical semantics and diagnosis property of each medical code, we fine-tune the LM to "memorize" the mapping between medical codes and their natural language definitions. Consequently, this process bridges the gap between raw codes and their contextual medical meanings and equips the LM to capture the intricate code dependencies that are crucial for precise diagnosis assessments.

We validate the effectiveness of MERA in general diagnosis and heart failure prediction tasks on the patient records in MIMIC-III (Johnson et al. 2016) and MIMIC-IV (Johnson et al. 2023) datasets. MERA yields significant improvements over the existing state-of-the-art models across tasks on all datasets while having almost perfect memorization of bidirectional medical code-definition mapping. An extensive analysis of leading LLM's medical code understanding and diagnosis prediction capabilities is conducted, and we observe that GPT-4 is still far behind fine-tuned models on both tasks. We further conduct ablation studies to validate the effectiveness of the proposed novel design choices.

2 Preliminaries

2.1 Task Formulations

MERA can be applied for any task whose output is a collection of candidates belonging to a pre-defined decision space. We introduce widely used diagnosis prediction settings as typical testbeds for MERA (Yang et al. 2023b).

Tasks. The first task is a general **diagnosis prediction** task, in which we aim to predict the diagnoses for the patient's potential next visit V_{T+1} given patient's history diagnoses by selecting a set of medical codes from the medical code ontology O, which can be formally described as $f_{DP} : \{V_1, V_2, \ldots, V_T\} \rightarrow V_{T+1}$. The second task is a disease-specific **heart failure prediction** task, which can be described as a binary classification function $f_{HF} : \{V_1, V_2, \ldots, V_T\} \rightarrow 0, 1$. We are more focused and aim to predict whether a patient would encounter heart failure (ICD-9 codes with head `428`) in any of the future visits.

Input patient record. Given an EHR collection of n patients $\{P_1, P_2, \ldots, P_n\}$, patient historical diagnosis can be represented as a sequence of admissions in chronological order $P = \{V_1^P, V_2^P, \ldots, V_T^P\}$ where T is the number of existing visits. For a particular visit V, the medical judgment made by clinicians as a result of the visit is an unordered set of diagnoses $V = \{d_1^V, d_2^V, \ldots, d_{|V|}^V\}$ in the format of $|V|$ unique medical code ($d \in O$). The task input has two variants, including 1) history diagnosis *codes* only, and 2) additionally providing patient profile (gender, race, medication and family history) as a *natural language sentence*.

Medical code ontology as the decision space. The International Classification of Diseases (ICD) (Cuadrado 2019) provides a comprehensive ontology O diseases, symptoms and diagnoses. Each leaf node represents a unique disease/diagnosis and is assigned a unique medical code $c \in \{c_1, c_2, \ldots, c_{|O|}\}$ where $|O|$ is the total number of codes. Diseases are organized into disease groups at multiple levels, represented by non-leaf nodes forming a tree hierarchy $G = \{G_{\text{level=0}}, G_{\text{level=1}}, \ldots, G_{\text{level=}depth(O)}\}$. Assuming the root of O is at level 0, at level $j > 0$, there are $|G_{\text{level=}j}|$ disjoint disease groups, *i.e.* $G_{\text{level=}j} = \{g_1, \ldots, g_{|G_{\text{level=}j}|}\}$. There is also a one-to-one mapping between a code c and its natural language definition def_c. For example, in version 9 of ICD, the medical code `250.23` stands for "Diabetes with hyperosmolarity, type I [juvenile type], uncontrolled". It belongs to the first-level group for all "Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders", and further belongs to the fine-grained disease group "type I uncontrolled diabetes". We use both ICD-9 and ICD-10 coding systems with 13k+ and 68k+ unique codes in this work.

2.2 Existing Paradigm of Generative LMs

The ordinary formulation of generative LMs takes the input sequence $seq_{in} = t_1^{in}, \ldots, t_{|seq_{in}|}^{in}$ and is expected to generate the ground-truth output $seq_{out} = t_1^{out}, \ldots, t_{|seq_{out}|}^{out}$. It produces a probability distribution $P\left(c \mid t_{1:|seq_{in}|}^{in}, \hat{t}_{1:k}^{out}\right)$ over the possible next token ($c \in V$) conditioned on both the input sequence and k generated tokens. Discrete tokens at each autoregressive decoding step are produced by Equation 1. The LM is optimized to minimize the cross-entropy loss shown in Equation 2 applied on the probability of the *gold* next token conditioned on the *gold* output tokens in the previous segment in a teacher-forcing manner, assuming the $|seq_{out}|$-th token marks the end of the decoding.

$$\hat{t}_{k+1}^{out} = \text{argmax}_{c \in V} P\left(c \mid t_{1:|seq_{in}|}^{in}, \hat{t}_{1:k}^{out}\right) \quad (1)$$

$$\mathcal{L}_{CE} = \sum_{k=0}^{|seq_{out}|} -\log P\left(t_{k+1}^{out} \mid t_{1:|seq_{in}|}^{in}, t_{1:k}^{out}\right) \quad (2)$$

3 MERA: Learning to Memorize and Rank

MERA builds upon a large language model LM after pre-training on a natural language corpus, instruction tuning, and potential alignment process. MERA is designed to be compatible with numerous generative LM architectures and inherit knowledge obtained through pre-training, including encoder-decoder LM and decoder-only LM. There are three steps involved as a pipeline: 1) Fine-tuning the model to memorize medical codes used to represent the diagnoses; 2) Further optimizing the model to learn inter-visit causal and temporal relations between patient visits as well as intra-visit patterns from patient history records; 3) During inference, performing autoregressive generation to produce diagnosis predictions given an unseen patient history input.

3.1 Medical Code Memorization

State-of-the-art LLMs struggle to associate medical codes with their correct definitions accurately. GPT-4 can only recall 45% of ICD-9 codes given corresponding definitions (row 3 of Table 2), which may be attributable to the absence of medical codes in the pre-training dataset. MERA explicitly teaches LM the semantic information associated with the medical codes and the relationships within the coding system. We consider all codes in O as special tokens, each unique medical code has a dedicated token embedding and can be represented by a single token. This design reduces the noise of the learning objectives as the diagnosis probability is equivalent to the token probability. The memorization process parameterizes embeddings of the special tokens and further equips the LM with the necessary external knowledge to facilitate downstream diagnosis prediction. To integrate information about medical codes in O and the natural language knowledge contained in LM, we fine-tune LM on synthetic question-answering pairs.

Bidirectional code and definition memorization. For each code c and the natural language definition def_c, we create two input-output pairs. The first pair includes "What is the definition of ICD-9 code c" as seq_{in} and the target answer "def_c" as seq_{out} to train the model to recall its definition given a code. The second pair helps the model memorize the inverse mapping. The question-answer pairs are created according to the O ontology being for the downstream task.

Decision space structure memorization. We further embed code dependencies collectively in LM by training with separate code-category instances. The curated pairs connect a code to its disease groups at various levels $1, \ldots, depth(O)$ in the code ontology O. For example, seq_{in} is "What is the chapter level disease group of the ICD-9 code 998.51?", and seq_{out} is "Injury and Poisoning".

3.2 Seq2seq Data Construction

The second phase aims to equip LM with a temporal and causal understanding of the diagnoses across multiple visits. We train the LM with a collection of sequence-to-sequence training instances $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_{\text{patient}}}\}$ based on n_{patient} patient records, where $\mathbf{X}_i$ is a set of (diagnosis history, future diagnosis) pairs created based on patient record P_i. Given the history of a patient containing T visits $P_i = \{V_1^{P_i}, \ldots, V_T^{P_i}\}$, we extract $T-1$ pairs of patient history and the expected diagnoses in the next visit to have maximum utilization of the patient records. For each pair, an input sequence is verbalized from 1-to-k visits following $seq_{in} = instruction, vb(V_1^{P_i}), \ldots, vb(V_k^{P_i})$ ($k \in [1, T-1]$). Additional patient profile sentences can be inserted following the instructions. A ground-truth output sequence is converted from expected diagnoses in the $(k+1)$-th visit following $seq_{out} = vb(V_{k+1}^{P_i})$. The verbalizer function vb concatenates the diagnosis codes within each visit to form a token segment for a specific visit and further prepend the starting prompt phrase ("The diagnosis codes for this visit are: ") and append a special token `EOV` representing "the end of the visit".

Diagnoses order perturbation. The order of patient visits is crucial to convey the dependent relations as a diagnosis in a later visit is conditioned on the previous diagnoses. However, the order of diagnosis codes *within* a particular visit does not carry cognitive rationale as indicated by EHR dataset documentation and papers (Johnson et al. 2023). An ideal model should preserve the inter-visit orders while ignoring the intra-visit orders. To achieve this goal with a sequential LM, we propose to create n_{perturb} variants of the input patient history sequences and output diagnosis sequences respectively, leading to n_{perturb}^2 diverse combinations. Each variant keeps the same visit order but randomly shuffles the diagnosis codes within each visit. By observing the data instances with shuffled orders and the same target distribution, we teach the LM to ignore the order of diagnosis codes with a model-agnostic design. To summarize, the training sequence-to-sequence data $\mathbb{X}$ contains data instances $\mathbf{X}$ generated according to n_{patient} patient history records. $\mathbf{X}$ contains $T-1$ groups of data instances with different patient history lengths, each group contains combinations among n_{perturb} perturbed input sequences and n_{perturb} perturbed output sequences.

3.3 Learning Inter-visit Reasoning

Up to this point, the created seq2seq data instances can be used to conduct supervised fine-tuning of LM following token-level optimization used in conventional generative LM reiterated in §2.2. However, as we analyze theoretically (in §1) and demonstrate empirically (line 14/15 of Table 1), vanilla generative LM does not handle the diagnosis prediction task well. We propose multiple specialized learning objectives to learn the *inter-visit reasoning* to infer upcoming diagnoses and capture *intra-visit diagnosis patterns*. We bridge the sequential modeling capabilities and LM's internal knowledge with the task property and decision space structure (*e.g.* , ICD hierarchy) for diagnosis prediction.

After encoding seq_{in} containing information on existing hospital visits, the LM starts to generate its prediction of the upcoming visit $\hat{seq}_{out}$. As an immediate step, it produces a probability distribution over the possible next token t_1^{out} conditioned on seq_{in}, reflecting the possibility of different tokens in the vocabulary as one of the diagnoses for visit V_{T+1}. Legit candidate tokens for t_1^{out} are the special code

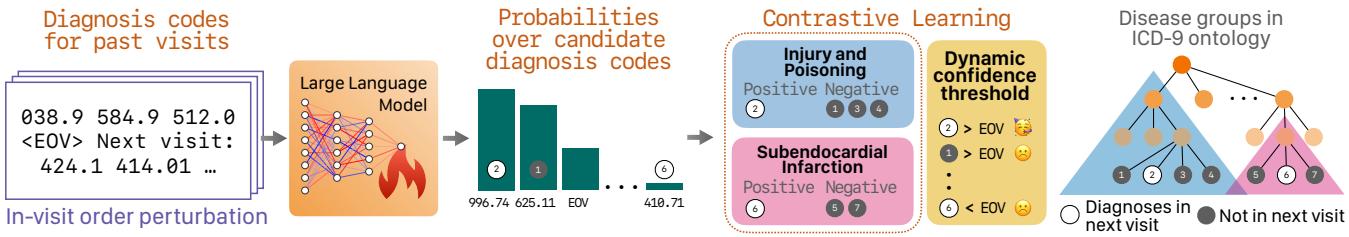

Figure 1: The model design of MERA. The diagnosis probability distribution is induced from token probabilities. It is optimized with hierarchical contrastive learning and dynamic cross-entropy losses.

tokens, including $\{c_1, c_2, ..., c_{|O|}\}$. We select the probabilities of all code tokens and then apply softmax, resulting in the probability distribution over the candidate codes

$$P\left(c \mid t^{in}_{1:|seq_{in}|}\right) = \{p_{c_1}, p_{c_2}, ..., p_{c_{|O|}}\}, c \in O. \quad (3)$$

Hierarchical contrastive learning. We design training objectives to identify the real diagnoses among a group of similar candidate diagnoses. With such a design, the model is forced to understand the subtle differences among neighbor diseases in O and learn to infer upcoming diagnoses from a candidate pool under the same disease group.

For a training instance $\mathbf{X}_i$, we first identify all disease groups that the diagnoses of the next visit belong to $G_{\mathbf{X}_i} = \{G_{\text{level}=0}, G_{\text{level}=1}, \ldots, G_{\text{level}=depth(O)}\}$. Then, for each group g_k at level j ($g_k \in G_{\text{level}=j}$), we identify positive diagnosis codes $g^{pos}_k = \{c^{pos}_1, \ldots, c^{pos}_{|g^{pos}_k|}\}$, which are the diseases in g_k that are diagnosed in the next visit. We then use all remaining diseases in g_k as negative codes $g^{neg}_k = g_k - g^{pos}_k = \{c^{neg}_1, \ldots, c^{neg}_{|g^{neg}_k|}\}$. Then, we calculate an InfoNCE loss (Oord, Li, and Vinyals 2018; Ma et al. 2021; Meng et al. 2021) term for each group in $G_{\mathbf{X}_i}$ and aggregate all the terms to be the aggregated objective $\mathcal{L}_{CL}$.

$$\mathcal{L}^{g_k}_{CL} = -\log \frac{\sum_{c^{pos}_m \in g^{pos}_k} P\left(c^{pos}_m \mid t^{in}_{1:|seq_{in}|}\right)}{\sum_{c_m \in g_k} P\left(c_m \mid t^{in}_{1:|seq_{in}|}\right)} \quad (4)$$

$$\mathcal{L}_{CL} = \frac{1}{|\mathbb{X}|} \sum_{\mathbf{X}_i \in \mathbb{X}} \sum_{G_{\text{level}=j} \in G_{\mathbf{X}_i}} \sum_{g_k \in G_{\text{level}=j}} \mathcal{L}^{g_k}_{CL} \quad (5)$$

The loss term for higher-level groups (where j is smaller) is used to enable the model to recognize disease scopes across a broad spectrum. Optimizing the high-level loss mimics the clinician's training process of making differential diagnoses, the "rough guesses" of possible diseases. Loss terms for lower-level groups focus on nuanced comparisons among diseases within the same family, increasing the model's ability to distinguish rare diseases. The proposed contrastive learning approach is efficient and capable in comparison to in-batch contrastive learning for two reasons: 1) The loss is calculated on the token probability distribution, essential for the typical decoding of generative LM, with no need for additional architecture or forward/backward passes. This ensures efficiency and maximum compatibility with the pre-trained LM. 2) The contradiction for loss calculation pertains to token probabilities, allowing the integration of prediction confidence for each disease into the optimization. This design differs significantly from in-batch contrastive learning, where forward and backward passes must be run for multiple data instances, and batch size significantly limits the size of positive and/or negative samples.

Dynamic confidence threshold. To produce a short list of *confident* diagnoses among the full ranking of *all* diagnosis codes, we learn a dynamic confidence threshold to select the most likely predictions. Existing works apply a fixed threshold to the probability distribution, which is often determined as a hyperparameter observed through the performance of the validation set (Morid, Sheng, and Dunbar 2023; Rasmy et al. 2021). This widely used strategy makes shortlisting less flexible, and the model tends to play it safe and produces more diagnoses than it should. To model the confidence threshold dynamically, we use a special token EOV to mark the confidence threshold within the token probability ranking list. EOV was appended at the end of the diagnosis sequence of each visit as introduced in §3.2.

The model LM learns the placement of the EOV in two ways. Implicitly, the visit segments in the input sequence demonstrate that the special token EOV represents the end of a visit segment, implying the model should stop generating more diagnosis codes. Training with EOV-ended visit sequence segment, LM naturally learns to assign EOV a higher probability than other code tokens when the model is not confident to make more diagnoses and chooses to generate EOV to end the diagnosis sequence of a particular visit. Explicitly, we design a learning objective to train the LM to place the EOV token at the proper rank of the token probability distribution $P\left(c \mid t^{in}_{1:|seq_{in}|}\right)$. We identify the positive medical codes that do appear in the target visit as O^{pos} and the ones not included as O^{neg} ($O^{pos} + O^{neg} = O$). The $\mathcal{L}_{DCE}$ is essentially a dynamic cross-entropy loss that regularizes the probability of each positive code to be not smaller than the probability of EOV and further make sure the probability of each negative code is not larger than $P\left(\text{EOV} \mid t^{in}_{1:|seq_{in}|}\right)$. The optimization of the dynamic confidence threshold applies fine-grained supervision to the probability distribution, enabling effective and efficient diagnosis capability learning with sparse patient data.

$$\mathcal{L}_{DCE} = \sum_{c \in O^{pos}} \log\left(ReLU(P\left(\text{EOV} \mid t^{in}_{1:|seq_{in}|}\right) - P\left(c \mid t^{in}_{1:|seq_{in}|}\right))\right) + \sum_{c \in O^{neg}} \log\left(ReLU(P\left(c \mid t^{in}_{1:|seq_{in}|}\right) - P\left(\text{EOV} \mid t^{in}_{1:|seq_{in}|}\right))\right) \quad (6)$$

3.4 Learning Intra-visit Diagnosis Patterns

Besides training the model to reason between visits, there are many implicit rules and latent dependencies buried in the large pool of diagnoses within each visit. For example, within a group of similar diseases, the clinicians normally only choose the most representative code for the patient's status; some diseases might suppress or correlate with other diagnoses. Modeling the intra-visit dependencies enables us to incorporate real-life clinic operation patterns into realistic diagnosis predictions. The prediction made for a specific visit should consider other diagnoses of the same visit.

To model the intra-visit dependencies, we apply the objectives over the token probability distribution introduced in §3.3 to multiple training instance variants with partial output sequences as conditions. This enables teacher-forcing training. For each (seq_{in}, seq_{out}) pair in $\mathbf{X}_i$ for patient record P_i where the seq_{out} expresses all diagnoses in the visit $V^{P_i}_{k+1}, k \in [1, T-1]$, we create $|V^{P_i}_{k+1}|$ variants to move partial diagnosis results in seq_{out} to be part of the input of LM together with seq_{in}. Given the new input including the patient history and m known diagnoses in the upcoming visit, LM produces probability over the candidate medical code $P\left(c \mid t^{in}_{1:|seq_{in}|}, t^{out}_{1:m}\right)$. Since the m known diagnoses have been part of the input sequence, we remove the corresponding medical codes from the positive code set for the calculation of $\mathcal{L}_{DCE}$ and $\mathcal{L}_{CL}$ to prevent the model from generating duplicated codes. Formally, the conditions for probability P in Equation 3, 4, and 6 are $t^{in}_{1:|seq_{in}|}, t^{out}_{1:m}$ instead of $t_{1:|seq_{in}|}$. The m known diagnoses in $V^{P_i}_{k+1}$ are removed from g^{pos}_k, O^{pos} and added to g^{neg}_k and O^{neg}.

3.5 Training and Inference Pipeline

Training objectives. For code memorization, LM is trained with the ordinary cross-entropy loss in Equation 2. The hierarchical contrastive learning loss (Equation 5) is additionally applied to the instances whose output is a medical code. For the diagnosis prediction task, the LM fine-tuned from the memorization task is further optimized with the hierarchical contrastive learning loss (Equation 5) and the dynamic cross-entropy loss (Equation 6) on $|V^{P_i}_{k+1}|$ teaching force variants. Unlike language modeling, no loss has been applied to the reconstruction of the input segment for both fine-tuning stages. We perform full-parameter fine-tuning.

Autoregressive decoding. The produced LM can be used for inference on unseen patient history. Given seq_{in}, LM performs autoregressive decoding to output discrete diagnosis code with the highest probability in the ranking list for each output step until the `EOV` token is generated.

4 Experiments

4.1 Experimental Setup

Datasets. We use MIMIC-III (Johnson et al. 2016) and MIMIC-IV (Johnson et al. 2023) EHR datasets containing patient records to train and evaluate. The MIMIC-III dataset focuses on patients eventually admitted to the ICU, while the MIMIC-IV dataset includes both ICU patients and other patients. We conduct data preprocessing following previous works (Lu, Han, and Ning 2022) and split the train/dev/test sets by patients to avoid information leak.

Metrics. We report the weighted F1 and recall@k, where k is the number of top-ranked predictions, and AUC and F1 for diagnosis prediction and heart failure, respectively.

Baselines. *RNN/CNN and attention-based models:* **RETAIN** (Choi et al. 2016), **Dipole** (Ma et al. 2017), **Timeline** (Bai et al. 2018), **HiTANet** (Luo et al. 2020), and **Deepr** (Nguyen et al. 2017). *Graph-based models:* **GRAM** (Choi et al. 2017), **G-BERT** (Shang et al. 2019), **CGL** (Lu et al. 2021), **Chet** (Lu, Han, and Ning 2022), and **MCDP** (Li and Gao 2022). **KGxDP** (Yang et al. 2023b) formulates each patient as a personalized medical KG, combining medical KGs with patient admission history. Note that additional medical notes are used by CGL, and additional Unified Medical Language System resource (Bodenreider 2004) is used as external knowledge by KGxDP. *Transformer-based models:* We adapt two encoder-only LM. **RoBERTa** (Liu et al. 2019) with 125M and **MedBERT** (Rasmy et al. 2021) with 109M parameters and append a $|O|$-way classification head. We choose MedBERT among other similar encoder-only architectures for medical sequence (Pang et al. 2021; Li et al. 2023; Rupp, Peter, and Pattipaka 2023) because other models require additional input information such as lab test results which is not available under our setting. **Seq2seq** uses ordinary generative LM's formulation introduced in §2.2 to fine-tune a LM to generate diagnosis codes as output. We include definition sentences in the prompt following each code, so these baselines are exposed to the same external knowledge used by MERA.

Base LMs. We use BioMistral (Labrak et al. 2024) trained on PubMed Central, LLaMA2 (Touvron and et al. 2023), GPT-2 (Radford et al. 2019), T5 (Raffel et al. 2023) and Flan-T5 (Chung and et al. 2022) as the base LMs.

4.2 Performance of Diagnosis Prediction

We show the performance comparison on the diagnosis prediction and heart failure prediction tasks (described in §2.1) using ICD-9 as decision space with history diagnosis code as input in Table 1 and the influence of base pre-trained LM selection in Table 2. We further show that MERA can be generalized to richer input with natural language patient profile, and the larger ICD-10 decision space in Table 3.

Encoder-only & vanilla generative LM perform poorly. The encoder-only LMs exhibit limited performance (rows 12-13 of Table 1), possibly because they do not account for the specialized modeling of intra-visit order and the extensive output space. When employing a vanilla generative LM (rows 14-15), the performance is further diminished. This is attributed to sparse supervision distributed in token-level loss. For each pass, only the probability of the single ground-truth token is optimized following Equation 2, while MERA optimizes the probabilities of all candidate diagnoses.

Gap between zero-shot and fine-tuned LMs. There remains a 20-point deficit in recall@20 comparing the best

#	Model	Diagnosis Prediction						Heart Failure			
		MIMIC-III			MIMIC-IV			MIMIC-III		MIMIC-IV	
		w-F1	R@10	R@20	w-F1	R@10	R@20	AUC	F1	AUC	F1
	RNN/CNN and attention-based models										
1	Deepr	18.87	24.74	33.47	24.08	26.29	33.93	81.36	69.54	88.43	61.36
2	Dipole	19.35	24.98	34.02	23.69	27.38	35.48	82.08	70.35	88.69	66.22
3	Timeline	20.46	25.75	34.83	25.26	29.00	37.13	82.34	71.03	87.53	66.07
4	RETAIN	20.69	26.13	35.08	24.71	28.02	34.46	83.21	71.32	89.02	67.38
5	HiTANet	21.15	26.02	35.97	24.92	27.45	36.37	82.77	71.93	88.10	68.21
	Graph-based models										
6	G-BERT	19.88	25.86	35.31	24.49	27.16	35.86	81.50	71.18	87.26	68.04
7	GRAM	21.52	26.51	35.80	23.50	27.29	36.36	83.55	71.78	89.61	68.94
8	CGL	21.92	26.64	36.72	25.41	28.52	37.15	84.19	71.77	89.05	69.36
9	MCDP	-	28.30	39.60	-	25.80	36.10	-	-	-	-
10	Chet	22.63	28.64	37.87	26.35	30.28	38.69	86.14	73.08	90.83	71.14
11	KGxDP	27.35	30.98	41.29	30.38	34.19	43.47	86.57	74.74	95.66	79.87
	Transformer-based models										
12	RoBERTa	17.39	22.84	32.07	22.54	24.89	32.38	79.74	68.28	87.03	60.21
13	MedBERT	19.01	23.68	34.39	24.13	25.88	33.81	81.06	69.96	88.73	61.81
14	Seq2seq (LLaMA2-7B)	18.05	18.38	23.56	20.47	20.77	24.19	77.62	66.06	85.98	59.14
15	Seq2seq (BioMistral-7B)	19.14	19.83	24.97	22.11	22.03	26.24	78.57	67.87	87.04	61.07
16	MERA (LLaMA2-7B)	32.77	35.94	47.48	34.64	38.16	46.94	89.49	77.21	97.26	82.31
17	MERA (BioMistral-7B)	**33.24**	**36.73**	**49.01**	**36.16**	**39.57**	**49.09**	**90.78**	**79.13**	**98.74**	**84.03**

Table 1: Diagnosis prediction comparison with baselines using ICD-9 as the decision space with code-only input (%).

zero-shot LLM (row 3 of Table 2) to the fine-tuned model. This underscores the importance of leveraging patient data.

MERA is the state-of-the-art diagnosis prediction model. Finally, MERA achieves significantly better performance in both diagnosis and heart failure prediction tasks on both MIMIC datasets. MERA exhibits a 5.89 point higher weighted F1 score and almost 8 points higher recall@20 for MIMIC-III compared to the existing best model (row 17 vs 11 of Table 1). In Table 2, we showcase the diagnosis prediction performance using different pre-trained LMs, noting that even MERA with GPT-2 large (row 10) achieves comparable performance with the existing best KGxDP.

4.3 Performance on Medical Code Memorization

Table 2 shows the evaluation of the memorization results for the ICD-9 medical code system while using various base LMs. We report code and definition accuracy, indicating the proportion of correct output full ICD codes/definitions given their definitions/ICD codes as input by exact match. We observed that **1) Almost perfect medical code recall using large-enough 7B LM. 2) Pre-trained LLMs alone do not know medical codes well.** GPT models exhibit better memorization of medical codes compared to LLaMA2 (rows 1-3 of Table 2), but they still lag far behind the fine-tuned models (line 3 vs 12). **3) Model scaling-up boosts memorization.** Increasing models' parameters significantly enhances their memorization capabilities, as evidenced by an 80-point improvement in code accuracy from GPT-2 medium to large. However, this does not fully translate into improvement of the same magnitude in diagnosis prediction (row 9 vs 10 in Table 2). **4) Encoder-decoder vs decoder-only.** Comparing rows 4-7 with rows 8-12 in Table 2, we observe that encoder-decoder LMs tend to perform well on definition-to-code mapping while performing significantly worse on producing the accurate definition given the code. However, the observation is different for decoder-only LMs who can handle code-to-definition mapping at the early stage. Derived from these observations, it is optimal to use a large-size decoder-only LM as the backbone for diagnosis prediction.

#	Model	Med. Code Mem.		Diagnosis Pred.	
		Code Acc	Def Acc	w-F1	R@20
	Zero-shot LM				
1	LLaMA2	4.69	0.61	5.62	15.64
2	GPT-3.5	33.50	9.31	6.11	17.07
3	GPT-4	45.16	48.48	6.46	21.56
	Fine-tuned encoder-decoder LM				
4	T5 base	81.71	1.26	20.53	30.13
5	T5 large	85.28	2.32	23.19	33.85
6	Flan-T5 base	88.58	0.19	21.01	32.24
7	Flan-T5 large	89.97	0.29	25.32	35.25
	Fine-tuned decoder-only LM				
8	GPT-2 base	0.00	95.68	23.29	32.06
9	GPT-2 medium	0.00	98.30	25.50	34.59
10	GPT-2 large	80.05	98.56	29.59	40.96
11	LLaMA2 7B	**99.87**	99.12	32.77	47.48
12	BioMistral 7B	99.61	**99.58**	**33.24**	**49.01**

Table 2: Memorization and diagnosis prediction (after fine-tuning on the memorization task) results on MIMIC-III data using different pre-trained LMs.

Model	w NL info	w/o NL info
Chet	17.51	17.51
Seq2seq (BioMistral 7B)	16.31	13.47
MERA (BioMistral 7B)	43.66	40.39

Table 3: Diagnosis prediction results (recall@20, %) on the MIMIC-IV dataset using ICD-10 as the decision space with or without additional natural language patient profile.

#	Method Variant	w-F1	R@20
	Knowledge injection approach		
1	No external knowledge	-2.33	-3.54
2	Code definition in the prompt	-1.69	-2.46
	Training objectives		
3	w/o hierarchical contrastive learning	-10.34	-10.27
4	- w/o 0-th level CL loss only	-9.24	-8.4
5	- w/o chapter level CL loss only	-5.86	-4.08
6	- w/o finest level CL loss only	-7.74	-6.81
7	w/o dynamic confidence threshold	-4.10	-2.57
	Outputting strategies MERA = decode (our losses)		
8	Decode (cross-entropy loss)	-10.31	-17.33
9	Rank (cross-entropy loss)	-6.72	-13.32
10	Rank (our losses)	-2.63	-3.16

Table 4: Ablation study on model design choices compared with full MERA (row 16 of Table 1) on MIMIC-III dataset.

4.4 Ablation Studies on Method Design

Knowledge injection approach. In rows 1-2 of Table 4, we observed that simply training the medical code sequence without providing meanings of the codes (row 1) leads to a 3.5-point lower recall@20. Providing the natural language definition of medical code in the input prompt along with the history diagnosis code (row 2 vs 1) is also helpful. However, the NL prompt method suffers from incomplete patient history due to the LM's input length limit, resulting in a 2.5-point lower recall@20 compared to memorization. Fine-tuning for concept memorization is the most effective knowledge injection approach.

Training objectives. Results in row 3-7 of Table 4 show that removing hierarchical contrastive learning leads to more than a 10-point drop in F1. Among the contrastive terms for disease groups categorized by different granularities, the 0-th level loss (row 4) is the most beneficial, which provides comparisons among the most involved diseases. The finest level loss (row 6) is the second most important, as the chapter-level disease is relatively easier to mine from data, while the fine-grained diagnosis decision involves distinguishing diseases that are similar in manifestation or etiology. Dynamic confidence threshold (row 7) also contributes more than 4-point F1 score improvement.

Outputting strategies. In rows 8-10 of Table 4, we explore optimal approaches to produce the diagnosis prediction set. LM can conduct autoregressive *decoding* to generate diagnosis codes as an output sequence. Alternatively, we can obtain the *ranking* list based on the token probability over the vocabulary of the first output token. Using decoding trained with sparse correct token cross-entropy loss (§2.2, row 8) compromises performance by 17 points in recall@20. The confusing in-visit diagnosis code order makes producing the result from the first token ranking list (row 9) a better choice than decoding along. When applying rich supervision with contrastive learning and dynamic confidence threshold, we observe a 10-point higher recall@20 with ranking output (row 10 vs 9). The comparison between row 10 and full MERA validates the effectiveness of intra-visit modeling, yielding a 3-point higher recall@20, where we decode token-by-token conditioned on other diagnoses but with specialized trained token probability for *each* decoding step.

5 Related Works

Diagnosis prediction. Existing works leverage structured diagnosis data (Morid, Sheng, and Dunbar 2023). They use sequential models like RNN and LSTM (Choi et al. 2016; Bai et al. 2018) to model the longitudinal patient history and GNNs to encapsulate spatial features (Proios et al. 2023; Lu, Han, and Ning 2022). To inject external knowledge, they conduct multi-task or transfer learning to borrow supervision from other tasks or domains (Yang et al. 2023a; Zhou et al. 2023), use pre-trained embedding to incorporate natural language into initial features (Wu et al. 2023b; Bornet et al. 2023), or utilizing external knowledge graphs or ontologies (An et al. 2023; Cheong et al. 2023; Li et al. 2020). We propose to use the capable LLM architecture to learn patterns from patient history sequences and inject external knowledge with a unified and shared architecture across the pipeline. Existing works apply contrastive learning on intermediate latent for KG relations (An et al. 2023) or patient embedding (Jeong et al. 2023), while we apply contrastive learning on diagnosis output space directly.

Transformer models for medical event prediction. Existing works either handle NL medical notes and other modalities (Niu et al. 2024; Zhou et al. 2023; Wang et al. 2023b; Liu et al. 2023), or they use a non-unified architecture that cannot inherit the pretrained knowledge (Rupp, Peter, and Pattipaka 2023; Li et al. 2023; Pang et al. 2021; Guo et al. 2023) or needs adaptation for downstream tasks (Steinberg et al. 2023; Lai, Zhai, and Ji 2023; Ma et al. 2023; Xu, Ma, and Chen 2023). (Wang et al. 2023a; Shoham and Rappoport 2023; Wornow et al. 2023a) fine-tune the generative LM for classification tasks. We develop a model that is compatible with mainstream LLMs to use the pretrained knowledge and specializes in producing predictions from large diagnosis decision space.

6 Conclusion

MERA stands out by seamlessly integrating clinical knowledge and addressing the challenges associated with a large candidate space. Contrasting learning, tailored to the coding system's hierarchical structure, enables effective distinguishing between accurate and inaccurate diagnosis codes. Through validation on MIMIC datasets, MERA emerges as a leading approach to diagnosis prediction.

Acknowledgments

The work is partially supported by Optum AI, NSF 2200274, 2106859, 2312501, and NIH U54HG012517, U24DK097771.

References

An, Y.; Tang, H.; Jin, B.; Xu, Y.; and Wei, X. 2023. KAMP-Net: Multi-Source Medical Knowledge Augmented Medication Prediction Network with Multi-Level Graph Contrastive Learning. *BMC Medical Informatics and Decision Making*.

Bai, T.; Zhang, S.; Egleston, B. L.; and Vucetic, S. 2018. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. In *SIGKDD*.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1): D267–D270.

Bornet, A.; Proios, D.; Yazdani, A.; Jaume-Santero, F.; Haller, G.; Choi, E.; and Teodoro, D. 2023. Comparing Neural Language Models for Medical Concept Representation and Patient Trajectory Prediction.

Brown, T. B.; and et al. 2020. Language Models Are Few-Shot Learners. arXiv:2005.14165.

Cartwright, D. J. 2013. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Advances in Wound Care*.

Caufield, J. H.; Zhou, Y.; Bai, Y.; Liem, D. A.; Garlid, A. O.; Chang, K.-W.; Sun, Y.; Ping, P.; and Wang, W. 2019. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*, 19009118.

Cheong, C. W.; Yin, K.; Cheung, W. K.; Fung, B. C. M.; and Poon, J. 2023. Adaptive Integration of Categorical and Multi-relational Ontologies with EHR Data for Medical Concept Embedding. *ACM Transactions on Intelligent Systems and Technology*.

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*.

Chung, H. W.; and et al. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

Cuadrado, M. T. 2019. Icd-9-cm: International classification of diseases, ninth revision, clinical modification.

Guo, L. L.; Steinberg, E.; Fleming, S. L.; Posada, J.; Lemmon, J.; Pfohl, S. R.; Shah, N.; Fries, J.; and Sung, L. 2023. EHR Foundation Models Improve Robustness in the Presence of Temporal Distribution Shift. *Scientific Reports*.

Hsu, W.; Han, S. X.; Arnold, C. W.; Bui, A. A.; and Enzmann, D. R. 2016. A data-driven approach for quality assessment of radiologic interpretations. *Journal of the American Medical Informatics Association*, 23(e1): e152–e156.

Jeong, H.; Oufattole, N.; Balagopalan, A.; Mcdermott, M.; Chandak, P.; Ghassemi, M.; and Stultz, C. 2023. Event-Based Contrastive Learning for Medical Time Series. arXiv:2312.10308.

Johnson, A. E. W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; Lehman, L.-w. H.; Celi, L. A.; and Mark, R. G. 2023. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Scientific Data*.

Johnson, A. E. W.; 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. *Scientific Data*.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373.

Lai, T. M.; Zhai, C.; and Ji, H. 2023. KEBLM: Knowledge-Enhanced Biomedical Language Models. *Journal of Biomedical Informatics*.

Li, R.; and Gao, J. 2022. Multi-Modal Contrastive Learning for Healthcare Data Analytics. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*.

Li, Y.; Mamouei, M.; Salimi-Khorshidi, G.; Rao, S.; Hassaine, A.; Canoy, D.; Lukasiewicz, T.; and Rahimi, K. 2023. 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*.

Li, Y.; Qian, B.; Zhang, X.; and Liu, H. 2020. Knowledge Guided Diagnosis Prediction via Graph Spatial-Temporal Network. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*, Proceedings.

Liu, S.; Wang, X.; Hou, Y.; Li, G.; Wang, H.; Xu, H.; Xiang, Y.; and Tang, B. 2023. Multimodal Data Matters: Language Model Pre-Training Over Structured and Unstructured Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lu, C.; Han, T.; and Ning, Y. 2022. Context-Aware Health Event Prediction via Transition Functions on Dynamic Disease Graphs. *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, M. D.; Chen, M.; Wu, T.-L.; and Peng, N. 2021. HyperExpan: Taxonomy Expansion with Hyperbolic Representation Learning. In *EMNLP Findings 2021*.

Ma, M. D.; Taylor, A.; Wang, W.; and Peng, N. 2023. DICE: Data-Efficient Clinical Event Extraction with Generative

Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *ACL*.

Ma, M. D.; Wang, X.; Kung, P.-N.; Brantingham, P. J.; Peng, N.; and Wang, W. 2024a. STAR: Boosting Low-Resource Information Extraction by Structure-to-Text Data Generation with Large Language Models. *AAAI*, 38(17).

Ma, M. D.; Ye, C.; Yan, Y.; Wang, X.; Ping, P.; Chang, T.; and Wang, W. 2024b. CliBench: A Multifaceted and Multigranular Evaluation of Large Language Models for Clinical Decision Making. arXiv:2406.09923.

Meng, Y.; Xiong, C.; Bajaj, P.; Bennett, P.; Han, J.; Song, X.; et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34: 23102–23114.

Morid, M. A.; Sheng, O. R. L.; and Dunbar, J. 2023. Time Series Prediction Using Deep Learning Methods in Healthcare. *ACM Transactions on Management Information Systems*.

Nguyen, P.; Tran, T.; Wickramasinghe, N.; and Venkatesh, S. 2017. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*.

Niu, S.; Ma, J.; Bai, L.; Wang, Z.; Guo, L.; and Yang, X. 2024. EHR-KnowGen: Knowledge-enhanced Multimodal Learning for Disease Diagnosis Generation. *Information Fusion*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint*.

Ouyang, L.; and et al. 2022. Training Language Models to Follow Instructions with Human Feedback.

Pang, C.; Jiang, X.; Kalluri, K. S.; Spotnitz, M.; Chen, R.; Perotte, A.; and Natarajan, K. 2021. CEHR-BERT: Incorporating Temporal Information from Structured EHR Data to Improve Prediction Tasks. In *Proceedings of Machine Learning for Health*.

Proios, D.; Yazdani, A.; Bornet, A.; Ehrsam, J.; Rekik, I.; and Teodoro, D. 2023. Leveraging Patient Similarities via Graph Neural Networks to Predict Phenotypes from Temporal Data. In *IEEEDSAA*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction. *npj Digital Medicine*.

Rochefort, C. M.; Buckeridge, D. L.; and Forster, A. J. 2015. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. *Implementation Science*, 10(1): 1–9.

Rupp, M.; Peter, O.; and Pattipaka, T. 2023. ExBEHRT: Extended Transformer for Electronic Health Records. In Chen, H.; and Luo, L., eds., *Trustworthy Machine Learning for Healthcare*, Lecture Notes in Computer Science.

Shang, J.; Ma, T.; Xiao, C.; and Sun, J. 2019. Pre-Training of Graph Augmented Transformers for Medication Recommendation. In *IJCAI*.

Shoham, O. B.; and Rappoport, N. 2023. CPLLM: Clinical Prediction with Large Language Models.

Steinberg, E.; Xu, Y.; Fries, J.; and Shah, N. 2023. MOTOR: A Time-To-Event Foundation Model For Structured Medical Records. arXiv:2301.03150.

Touvron, H.; and et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, H.; Gao, C.; Dantona, C.; Hull, B.; and Sun, J. 2023a. DRG-LLaMA : Tuning LLaMA Model to Predict Diagnosis-related Group for Hospitalized Patients. arXiv:2309.12625.

Wang, X.; Luo, J.; Wang, J.; Yin, Z.; Cui, S.; Zhong, Y.; Wang, Y.; and Ma, F. 2023b. Hierarchical Pretraining on Multimodal Electronic Health Records. arXiv:2310.07871.

Wornow, M.; Thapa, R.; Steinberg, E.; Fries, J. A.; and Shah, N. H. 2023a. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. arXiv:2307.02028.

Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.; Steinberg, E.; Fleming, S.; Pfeffer, M. A.; Fries, J.; and Shah, N. H. 2023b. The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records. *npj Digital Medicine*.

Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023a. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. arXiv:2304.14454.

Wu, J.; He, K.; Mao, R.; Li, C.; and Cambria, E. 2023b. MEGACare: Knowledge-guided Multi-View Hypergraph Predictive Framework for Healthcare. *Information Fusion*.

Xu, J.; Ma, M. D.; and Chen, M. 2023. Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *ACL*.

Yadav, K.; Sarioglu, E.; Smith, M.; and Choi, H.-A. 2013. Automated outcome classification of emergency department computed tomography imaging reports. *Academic Emergency Medicine*, 20(8): 848–854.

Yang, K.; Xu, Y.; Zou, P.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023a. KerPrint: Local-Global Knowledge Graph Enhanced Diagnosis Prediction for Retrospective and Prospective Interpretations. *AAAI*.

Yang, Z.; Lin, Y.; Xu, Y.; Hu, J.; and Dong, S. 2023b. Interpretable Disease Prediction via Path Reasoning over Medical Knowledge Graphs and Admission History. *Knowledge-Based Systems*.

Zhang, Y.; Hou, S.; Ma, M. D.; Wang, W.; Chen, M.; and Zhao, J. 2024. CLIMB: A Benchmark of Clinical Bias in Large Language Models. arXiv:2407.05250.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A Transformer-Based Representation-Learning Model with Unified Processing of Multimodal Input for Clinical Diagnostics. *Nature Biomedical Engineering*.