RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records

Ran Xu[♡*], **Wenqi Shi**[♠*], **Yue Yu**[♠], **Yuchen Zhuang**[♠], **Bowen Jin**[♣]
May D. Wang[♠], **Joyce C. Ho**[♡], **Carl Yang**[♡]
[♡] Emory University [♠] Georgia Institute of Technology
[♣] University of Illinois at Urbana Champaign
{ran.xu,joyce.c.ho,j.carlyang}@emory.edu,
{wshi83,yczhuang,yueyu,maywang}@gatech.edu, bowenj4@illinois.edu

Abstract

We present RAM-EHR, a Retrieval AugMentation pipeline to improve clinical predictions on Electronic Health Records (EHRs). RAM-EHR first collects multiple knowledge sources, converts them into text format, and uses dense retrieval to obtain information related to medical concepts. This strategy addresses the difficulties associated with complex names for the concepts. RAM-EHR then augments the local EHR predictive model co-trained with consistency regularization to capture complementary information from patient visits and summarized knowledge. Experiments on two EHR datasets show the efficacy of RAM-EHR over previous knowledge-enhanced baselines (3.4% gain in AUROC and 7.2% gain in AUPR), emphasizing the effectiveness of the summarized knowledge from RAM-EHR for clinical prediction tasks. The code will be published at https://github.com/ritaranx/RAM-EHR.

1 Introduction

Electronic Health Records (EHRs), encompassing detailed information about patients such as symptoms, diagnosis, and medication, are widely used by physicians to deliver patient care. Recently, a vast amount of deep learning models have been developed on EHR data (Choi et al., 2020; Gao et al., 2020; Wang et al., 2023a) for various downstream prediction tasks (e.g., disease diagnosis, risk prediction) to facilitate precision healthcare.

To further improve the downstream predictive performance, several works attempt to augment the EHR visits with external knowledge. For example, van Aken et al. (2021) and Naik et al. (2022) incorporate additional clinical notes, although these clinical notes can be noisy and contain irrelevant contents for clinical predictions; another solution is to leverage external clinical knowledge graphs (KGs), such as UMLS (Chandak et al., 2023), which contain rich medical concepts (e.g., disease, medications) and their corresponding relationships. Integrating KGs with EHRs has been shown to boost model performance (Xu et al., 2023b; Gao et al., 2023). However, these works mostly rely on knowledge from a single source and medical KGs mainly focus on specific types of relations (e.g., hierarchical relations), which do not comprehensively capture the semantic information for medical codes (e.g., phenotype). Besides, it is non-trivial to align medical codes in EHRs with KGs due to the non-uniformity of surface names (e.g., abbreviations or colloquial terms) (Hao et al., 2021; Zhang et al., 2022). There also exist methods that use knowledge generated from large language models (LLMs) to assist EHR prediction (Jiang et al., 2024), but LLMs may not always provide the most relevant knowledge for target tasks and face the risk of hallucination. Effectively leveraging external knowledge to facilitate EHR predictive tasks remains a significant challenge.

In this work, we propose RAM-EHR, a retrieval-augmented framework tailored for clinical predictive tasks on EHRs. Instead of leveraging a single knowledge source, RAM-EHR collects multiple knowledge sources (e.g., KGs, scientific literature) and converts them to text corpus, which enjoys the merits of a more comprehensive coverage of knowledge in a unified format. Then, to obtain unified representations for different knowledge sources, we leverage dense retrieval (DR) (Karpukhin et al., 2020; Lin et al., 2023) to encode corpus and medical codes as dense vectors, intuitively capturing the semantics of medical codes and addressing the alignment issue between EHR and external knowledge. Finally, to reduce irrelevant information, we utilize an LLM to summarize the top-retrieved passages into concise and informative knowledge summaries relevant to downstream tasks for each medical code. This process enhances the relevance and utility of the retrieved knowledge for clinical tasks.

[*] Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 754–765

To leverage external knowledge to assist clinical prediction, we introduce a retrieval-augmented model alongside the local EHR predictive model, which relies solely on patient visit information. The augmented model concatenates summarized passages and medical codes, feeding them into a moderate-size, pre-trained language model. We then co-train the local model and the augmented model with a consistency regularization, which captures the *complementary information* from patient visits and summarized knowledge and helps the model with better generalization (Wan, 2009).

We verify the effectiveness of RAM-EHR by conducting experiments on two EHR datasets and show that RAM-EHR outperforms strong knowledge-enhanced predictive baselines by 3.4% in AUROC and 7.2% in AUPR on average. Our analysis further confirms the advantage of leveraging multi-source external knowledge as well as retrieval augmentation as plugins to assist vanilla EHR predictive models based on visits only. Additional studies justify the usefulness of summarized knowledge for assisting clinical prediction tasks.

2 Methodology

2.1 Problem Setup

The EHR data consists of a group of patients $\mathcal{P}$ with corresponding hospital visits $V = \{v_1, v_2, ..., v_{|V|}\}$. Each visit v_i includes a set of medical codes $C_i \subset \mathcal{C}$, where $\mathcal{C}$ is the total set of medical codes for $\mathcal{P}$. In this study, $\mathcal{C}$ contains multiple types of medical codes including *diseases*, *medications*, and *procedures*. Each medical code $c_i \in C_i$ is a *clinical concept*, and it is associated with a name s_i in the form of *short text snippets*. Given the clinical record v_i with the involved medical codes $\mathcal{C}_i$, we aim to predict the patient's clinical outcome y_i (a binary label).

Figure 1 presents a comprehensive workflow of RAM-EHR, with a specific focus on dense retrieval from multiple knowledge sources and consistency regularization with co-training.

2.2 Retrieval Augmentation w/ Medical Codes

Existing approaches often treat each visit as context-free vectors, which fail to capture the concrete semantics of medical codes. Being aware of this, we aim to create the summarized knowledge for each medical code c_i using its surface name s_i via retrieval augmentation with additional contexts.

Multi-source Corpus Creation. Retrieval augmentation requires additional corpora as external knowledge. To ensure the coverage of clinical knowledge, we collect a diverse external resources $\mathcal{M} = \{d_1, d_2, \ldots, d_{|\mathcal{M}|}\}$. We represent each knowledge unit as a raw text to facilitate retrieval. The detailed information of $\mathcal{M}$ is in Appendix C.

Passage Retrieval. Given a collection of $|\mathcal{M}|$ passages, the objective of the retriever is to transform passages in a *dense* vector, so that it can efficiently retrieve the most relevant information to the input query. In our work, we adopt Dragon (Lin et al., 2023), a dual-encoder model with strong performance across domains as the retriever. Specifically, we first use the passage encoder $R_D(\cdot)$ to build an index for corpus $\mathcal{M}$ to support retrieval. Then, at runtime, we use the query encoder $R_Q(\cdot)$ to map the input to an embedding (same dimension as the passage embedding) and calculate the similarity as $f(q, d) = R_Q(q)^\top R_D(d)$. For the medical code c_i with the surface name s_i, we retrieve top-k ($k = 5$ in this work) passages $\mathcal{T}_i$ from the corpus $\mathcal{M}$ as

$$\mathcal{T}_i = \underset{d \in \mathcal{M}}{\text{Top-}k}\ f(s_i, d). \tag{1}$$

The top retrieved passages are considered as the external knowledge for the medical code c_i.

Summarized Knowledge Generation. Although $\mathcal{T}_i$ contains the most relevant information for c_i from $\mathcal{M}$, directly using them to assist predictions can be suboptimal, as simply concatenating these passages often leads to long contexts, and some of the retrieved passages can also be irrelevant (Yu et al., 2023). Motivated by the fact that LLMs have strong capabilities in text summarization (Zhang et al., 2024), we propose to use the off-the-shelf LLM (`gpt-3.5-turbo-0613`) to generate the summarized knowledge e_i for medical code c_i as

$$e_i = \text{LLM}([\text{Prompt}, t_{i,1}, \cdots, t_{i,k}]), \tag{2}$$

where $t_i \in \mathcal{T}_i$ stands for the retrieved passages in Eq.(1). We incorporate information related to the downstream task within our prompt to ensure the generated summaries are task-specific. Detailed prompt designs can be found in Appendix F.

Remark. The retrieval step is efficient as the corpus indexing only needs to be done *once* before applying to prediction tasks. It only needs one extra ANN retrieval operation per query, which is efficiently supported by FAISS (Johnson et al., 2021). Besides, we cache the summarized knowledge for each medical code to avoid redundant operations.

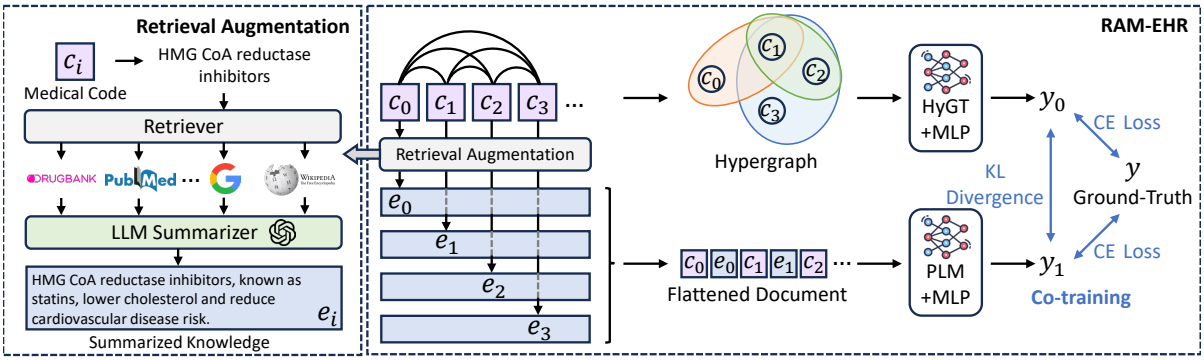


Figure 1: An overview of retrieval augmentation framework (left) and a detailed workflow of RAM-EHR (right). RAM-EHR initially gathers multiple knowledge sources and converts them into textual format. We then use dense retrieval to obtain information related to medical concepts. Next, we design an additional module to augment the local EHR predictive model co-trained with consistency regularization, capturing complementary information from both patient visits and summarized knowledge.

2.3 Augmenting Patient Visits with Summarized Knowledge via Co-training

Recall that patient visits and summarized knowledge encode complementary information for clinical prediction tasks — visits capture *cooccurence relationships*, while summarized knowledge encodes *semantic information*. To effectively aggregate these two types of information, we design a co-training approach, detailed as follows.

Augmented Model g_ϕ with Summarized Knowledge. For patient p_i having the hospital visit v_i with involved medical codes $\mathcal{C}_i$, we decompose $\mathcal{C}_i$ into three subsets: $\mathcal{C}_i^{\mathrm{d}}$ for diseases, $\mathcal{C}_i^{\mathrm{m}}$ for medications, and $\mathcal{C}_i^{\mathrm{p}}$ for procedures. For each type of medical code, we flatten the visit into a document by concatenating all the codes and their summarized knowledge in a reversed sequential order. For example, for disease code $\mathcal{C}_i^{\mathrm{d}}$, the flattened document can be $X_i^{\mathrm{d}} = \{[\mathrm{CLS}], D_t, D_{t-1}, \ldots, D_1\}$, where $D_i = ||_{c \in D_i}(c, e)$ is the concatenation of disease code and its summarized knowledge (Eq. 2) within the i-th visit. We then use a pre-trained language model (PLM) with a multi-layer perceptron (MLP) classification head as g_ϕ for prediction with flattened documents as inputs:

$$\boldsymbol{h}_i^k = \mathrm{PLM}(X_i^k), \quad \widehat{y}_{i,1} = \mathrm{MLP}\left(||_{k \in \mathcal{S}} \boldsymbol{h}_i^k\right). \quad (3)$$

Here $\mathcal{S} = \{\mathrm{p}, \mathrm{m}, \mathrm{d}\}$, $\boldsymbol{h}_i$ is the representation of `[CLS]` token of X_i, $\widehat{y}_{i,1}$ is the prediction for the target task. We share PLM weights for three types of medical codes to improve efficiency.

Local Model f_θ with Visit Information. To harness the visit-level information, various deep learning architectures have been proposed. In principle, g_ϕ can be combined with any f_θ to improve performance. In main experiments, we use a hypergraph transformer (HyGT, Cai et al. (2022); Xu et al. (2023a)) due to its strong ability to capture high-order relationships between visits and medical codes. It first builds hypergraphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by treating medical codes as nodes and patients as hyperedges, then leverages self-attention for aggregating neighborhood information. The details for HyGT are in Appendix E. We obtain the prediction $\widehat{y}_{i,2}$ with f_θ as

$$\boldsymbol{e}_i = \mathrm{HyGT}(\mathcal{G}, V_i), \widehat{y}_{i,2} = \mathrm{MLP}(\boldsymbol{e}_i), \quad (4)$$

where $\boldsymbol{e}_i$ is the representation of patient i after hypergraph transformer.

Co-training. We integrate the two predictors into a co-training framework, with the learning objective:

$$\begin{aligned} \mathcal{L}_{\mathrm{aug}} &= \mathbb{E}_{(V_i, y_i) \sim \mathcal{P}}\, \ell(\widehat{y}_{i,1}, y_i) + \lambda \mathcal{D}_{\mathrm{KL}}(\widehat{y}_{i,1}, \widetilde{y}), \\ \mathcal{L}_{\mathrm{loc}} &= \mathbb{E}_{(V_i, y_i) \sim \mathcal{P}}\, \ell(\widehat{y}_{i,2}, y_i) + \lambda \mathcal{D}_{\mathrm{KL}}(\widehat{y}_{i,2}, \widetilde{y}), \end{aligned} \quad (5)$$

where $\ell(\cdot)$ is the binary cross-entropy loss, $\widetilde{y} = \beta \widehat{y}_{i,1} + (1-\beta)\widehat{y}_{i,2}$, λ, β are two hyperparameters. Two losses in Eq. 5 are designed to encourage f_θ and g_ϕ regularize each other, which can stabilize the learning for two models. During the **inference stage**, we directly use the $\widetilde{y}_j$ as the final prediction for the j-th test example p_j.

3 Experiments

3.1 Experiment Setups

◇ **Datasets.** We conduct experiments on the public MIMIC-III dataset (Johnson et al., 2016) and a private CRADLE dataset collected from a large healthcare system in the United States. We perform a 25-label phenotypes prediction task on MIMIC-III, and a cardiovascular disease (CVD) endpoints

Table 1: The statistics of MIMIC-III and CRADLE.

Stats	MIMIC-III	CRADLE
# of diagnosis	846	7915
# of medication	4525	489
# of procedure	2032	4321
# of health records	12353	36611

prediction task for diabetes patients on CRADLE. We randomly split them into train/validation/test sets by 7:1:2. We present the detailed statistics of MIMIC-III and CRADLE in Table 1. Please refer to Appendix B for details.

◇ **Evaluation Metrics.** Following Choi et al. (2020), we employ Accuracy, AUROC, AUPR, and Macro-F1 as evaluation metrics, where AUROC is the main metric. For accuracy and F1 score, we use a threshold of 0.5 after obtaining predicted results.

◇ **Baselines.** We consider three groups of baselines: (a) Predictive models with *visit information only*: (1) **Transformers** (Li et al., 2020); (2) **GCT** (Choi et al., 2020); (3) **HyGT** (Cai et al., 2022); (b) Predictive models with *external knowledge*: (4) **MedRetriever** (Ye et al., 2021); (5) **GraphCare** (Jiang et al., 2024); (c) Predictive models with *clinical notes*: (6) **CORE** (van Aken et al., 2021); (7) **BEEP** (Naik et al., 2022). See Appendix D for more details.

◇ **Implementation Details.** In this work, we use Dragon (Lin et al., 2023) as the dense retriever, with the passage encoder $R_D(\cdot)$[1] and the query encoder $R_Q(\cdot)$[2]. We use $k = 5$ during the retrieval stage without tuning. We choose UMLS-BERT (Michalopoulos et al., 2021) for RAM-EHR and relevant baselines as g_ϕ, with a maximum length of 512, and HyGT (Cai et al., 2022) as f_θ in main experiments, but RAM-EHR can be adapted to multiple g_ϕ and f_θ (Sec 3.3). We set the learning rate to 5e-5 for g_ϕ and 1e-4 for f_θ, batch size to 32, and the number of epochs to 5. We select β, λ based on the performance of the validation set, and present the parameter study in Appendix G. For the model training, all the experiments are conducted on a Linux server with one NVIDIA A100 GPU.

3.2 Main Experimental Results

Table 2 exhibits the experiment results of RAM-EHR and baselines. **First**, we observe RAM-EHR surpasses baselines lacking external knowledge, highlighting the benefits of retrieval augmentation. **Second**, RAM-EHR outperforms knowledge-enhanced baselines due to the diverse collection of external knowledge as well as the co-training scheme that leverages information from both visit and semantic perspectives. **Third**, directly using medical notes leads to inferior outcomes due to potential irrelevance, whereas combining medical codes with summarized knowledge as RAM-EHR proves more effective for prediction tasks.

3.3 Additional Studies

Ablation Study. On the bottom of Table 2, we inspect different components in RAM-EHR and observe that removing any of them hurts the performance, which justifies the necessity of our designs. Besides, we observe that using the summarized knowledge with g_ϕ already achieves strong performance, highlighting the benefit of capturing the semantics of medical codes.

Effect of f_θ and g_ϕ. With various f_θ and g_ϕ, we demonstrate the flexibility of RAM-EHR in Figure 2(a) and 2(b) by the consistent performance gain across different models. Notably, even with a lightweight f_θ (Clin-MobileBERT) having only 25M parameters, RAM-EHR reaches close performance to UMLS-BERT, providing an efficient option for EHR predictive modeling.

Effect of Information Source $\mathcal{M}$. We then evaluate the effectiveness of each knowledge source within $\mathcal{M}$. Figure 2(c) indicates that incorporating all corpus yields the highest performance, highlighting the value of diverse corpora. Besides, using Drugbank alone contributes minimally, likely due to its limited scope of medication information. Moreover, we observe that leveraging knowledge bases (e.g., MeSH) is more beneficial than literature sources, as they offer broader and more generic information conducive to clinical prediction tasks.

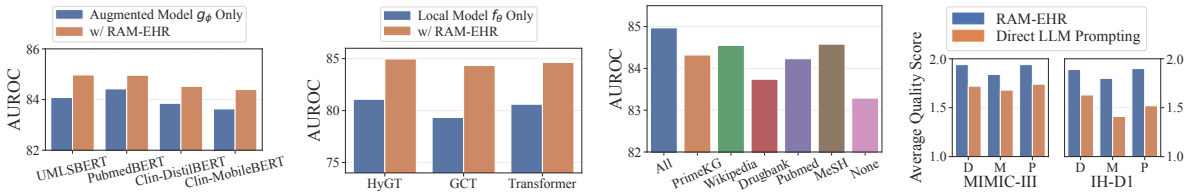
Parameter Study. In Figure 3, we conduct parameter studies on both datasets for β and λ in Eq. 5. Figure 3(a) demonstrates that the model achieves the best performance when β is set to 0.2 and 0.4 on MIMIC-III and CRADLE, respectively, while the gain diminishes at the extremes. This highlights the contribution of combining the predictions from both the augmented model and the local model on the performance gain. In addition, λ is set to 1 and 5 on MIMIC-III and CRADLE, respectively, according to Figure 3(b). The positive values of λ indicate that the consistency loss enhances model performance.

[1] https://huggingface.co/facebook/dragon-plus-query-encoder

[2] https://huggingface.co/facebook/dragon-plus-context-encoder

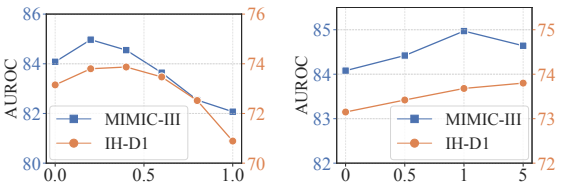
Table 2: Performance on two EHR datasets compared with baselines. The result is averaged over five runs. We use * to indicate statistically significant results ($p < 0.05$). For 'w/o Retrieval', we directly use LLM to generate summarized knowledge. For 'w/o LLM Summarization', we concatenate top-k retrieved documents as summarized knowledge. 'w/ g_ϕ only' means we set $\lambda = 0, \beta = 1$ (i.e., only use the prediction from g_ϕ as the final prediction).

Model	MIMIC-III				CRADLE			
	ACC	AUROC	AUPR	F1	ACC	AUROC	AUPR	F1
Transformer (Li et al., 2020)	76.18	80.61	67.12	42.75	78.10	69.49	40.14	58.23
GCT (Choi et al., 2020)	77.20	78.62	64.87	37.57	76.51	68.31	37.55	44.10
HyGT (Cai et al., 2022)	78.07	81.09	68.08	44.93	79.45	70.59	41.04	60.00
MedRetriever (Ye et al., 2021)	77.15	80.14	68.45	39.29	78.95	70.07	42.19	57.96
GraphCare (Jiang et al., 2024)	80.11	82.26	71.19	44.33	79.09	71.12	43.98	59.00
CORE (van Aken et al., 2021)	79.63	82.05	70.79	43.76	77.11	67.84	40.74	61.12
BEEP (Naik et al., 2022)	79.90	82.67	71.58	44.15	79.29	68.59	41.93	60.95
RAM-EHR	**81.59*** (1.8%)	**84.97*** (2.8%)	**74.64*** (4.3%)	**48.19*** (7.2%)	**80.41*** (1.2%)	**73.80*** (3.8%)	**48.40*** (10.1%)	**63.98*** (4.7%)
w/o Retrieval	80.68	83.29	72.95	44.65	79.83	73.06	47.05	63.25
w/o LLM Summarization	80.08	82.14	71.35	41.49	77.30	69.71	42.58	61.70
w/ Augmented Model g_ϕ Only	81.04	83.80	73.41	46.83	79.70	73.15	47.62	63.33



(a) Effect of g_ϕ (b) Effect of f_θ (c) Effect of $\mathcal{M}$ (d) Human Study

Figure 2: Results for Additional Studies. (a), (b), (c) is for MIMIC dataset, the results on CRADLE is in Appendix G.



(a) β in Eq. 5 (b) λ in Eq. 5

Figure 3: Parameter studies of β and λ on both datasets.

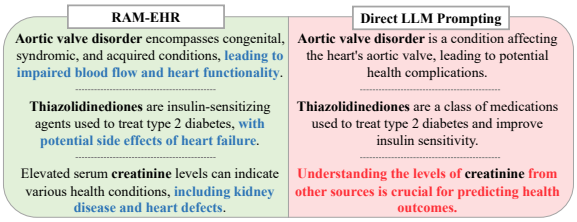


Figure 4: Comparing knowledge summarized by RAM-EHR and directly generated by LLM prompting. **Bold** denotes disease, medication and procedure concepts. **Blue** and **Red** indicate useful and irrelevant knowledge.

3.4 Case Study

Figure 4 presents a case study on CRADLE to compare knowledge summarized by RAM-EHR and directly generated by LLM prompting. We observe that RAM-EHR provides more relevant information for the downstream task, particularly regarding the CVD outcome in this case, compared to direct LLM prompting. This also aligns with the *human study* evaluating the quality of 40 randomly sampled knowledge per type of code on a scale of [0,1,2] in Figure 2(d). The study on hyperparameters and retrieval components is in Appendix G.

4 Conclusion

We propose RAM-EHR, which uses dense retrieval with multiple knowledge sources and consistency regularization to enhance EHR prediction tasks. Experiments on two EHR datasets show the efficacy of RAM-EHR over baselines with a gain of 3.4% in AUROC and 7.2% in AUPR. In addition, we conduct human studies to confirm the utility of generated knowledge.

Acknowledgement

We thank the anonymous reviewers and area chairs for valuable feedbacks. This research was partially supported by the Emory Global Diabetes Center of the Woodruff Sciences Center, Emory University. Research reported in this publication was supported by the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number K25DK135913. The research also receives partial support by the National Science Foundation under Award Number IIS-2145411. The content is solely the responsibility of the authors and does not necessarily rep-

resent the official views of the National Institutes of Health. We also thank Microsoft for providing research credits under the Accelerating Foundation Models Research Program.

Limitations

In this work, we propose RAM-EHR to unify external knowledge in text format and adapt it for EHR predictive tasks. Despite its strong performance, we have listed some limitations of RAM-EHR:

Building Multi-source Corpus $\mathcal{M}$. In this study, we construct a multi-source corpus $\mathcal{M}$ by manually selecting five relevant sources within the clinical domain. In real-world scenarios, the grounding corpora usually require customization according to query domains and user needs. Therefore, effectively selecting grounding corpora and efficiently evaluating their relative contributions remains an unresolved issue. Furthermore, retrieved evidence may contain noise that could potentially degrade model performance, highlighting the importance of developing fine-grained filtering or re-ranking modules as a crucial area for future research.

Efficiency. The integration of the augmented model g_ϕ can result in additional time complexity. In our main experiment setups (using UMLS-BERT), co-training usually takes $1.5\times$ to $2\times$ more times than using the local model alone. One potential solution is to use a lightweight model (e.g., Clin-MobileBERT) to improve efficiency.

Ethical Considerations

One potential ethical consideration concerns the use of credential data (MIMIC-III and CRADLE) with GPT-based online services. We have signed and strictly adhered to the PhysioNet Credentialed Data Use Agreement[3] for the legal usage of the MIMIC-III dataset. To prevent sensitive information from being shared with third parties through APIs, we carefully follow the guidelines[4] for the responsible use of MIMIC data in online services. Specifically, we have requested to opt out of human review of the data by filling out the Azure OpenAI Additional Use Case Form[5] in order to utilize the Azure Open AI service while ensuring that Microsoft does not have access to the patient data. The utilization of LLMs in our framework is strictly for the purpose of building medical concept-specific KGs. In addition, the building of medical concept-specific KGs *does not involve direct interaction* with any individual patient information. We iterate through all concepts in the medical coding system (e.g., CCS and ICD) to generate their respective KGs using LLMs, and these KGs are stored locally.

[3] https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150/

[4] https://physionet.org/news/post/gpt-responsible-use

[5] https://aka.ms/oai/additionalusecase

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2206–2240. PMLR.

Derun Cai, Chenxi Sun, Moxian Song, Baofeng Zhang, Shenda Hong, and Hongyan Li. 2022. Hypergraph contrastive learning for electronic health records. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 127–135. SIAM.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 606–613.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction.

Junheng Hao, Chuan Lei, Vasilis Efthymiou, Abdul Quamar, Fatma Özcan, Yizhou Sun, and Wei Wang. 2021. Medto: Medical data to ontology matching using hybrid graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2946–2954.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2024. Graphcare: Enhancing healthcare predictions with open-world personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 629–639.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online. Association for Computational Linguistics.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023a. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2852, Singapore. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023b. Augmenting black-box llms with medical textbooks for clinical question answering.

David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906.

Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2023a. Hypergraph transformers for ehr-based clinical predictions. *AMIA Summits on Translational Science Proceedings*, 2023:582.

Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023b. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2414–2423.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models.

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

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Related Works

Retrieval Augmented Learning and its Application in Clinical Domain. Retrieval augmented learning, which collects additional contextual information from external corpus, has shown effectiveness on diverse tasks including language modeling (Borgeaud et al., 2022), knowledge-intensive NLP (Lewis et al., 2020; Shi et al., 2023), commonsense reasoning (Wang et al., 2021), code genereation (Parvez et al., 2021), and few/zero-shot learning (Izacard et al., 2023). Compared to the general domain, the application of retrieval augmented learning to clinical tasks is still underexplored. Some efforts have been paid to retrieval augmented clinical language models (Zakka et al., 2024) as well as clinical question answering (Wang et al., 2023b; Jeong et al., 2024). The most relevant works are Ye et al. (2021); Naik et al. (2022), which leverage clinical literature to augment clinical predictive models. Compared to these works, our contribution lies in two folds: (1) we design a retrieval augmentation for *structured* EHRs with a diverse collection of external knowledge, which provides more relevant information for target clinical prediction tasks; (2) we incorporate a co-training scheme to leverage both the visit-level information and external knowledge for predictions.

Knowledge-enhanced EHR Predictive Models. Many studies attempt to harness external knowledge for clinical prediction tasks. The majority of them leverage structured knowledge, such as medical ontology (Choi et al., 2017; Panigutti et al., 2020), to capture hierarchical relationships among medical codes, or employ personalized knowledge graphs (Xu et al., 2023b; Jiang et al., 2024) to integrate patient-specific information. However, these methods often suffer from limited coverage of all medical codes due to the complexity of surface names. Alternatively, some approaches utilize unstructured medical text for health prediction tasks (Ye et al., 2021). However, Ye et al. (2021) rely on a restricted corpus of approximately 30,000 passages as their external corpus, resulting in limited coverage.

B Task Information

MIMIC-III. The MIMIC-III dataset (Johnson et al., 2016) is a large, freely available database that contains de-identified health-related data from over 4,000 patients who stayed in critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. We conduct the phenotyping prediction task proposed by (Harutyunyan et al., 2019). It aims to predict whether the 25 pre-defined acute care conditions (see Table 3) are present in a patient's next visit, based on the information from their current visit. The problem is formulated as a 25-label binary classification, considering that multiple phenotypes may exist in a single visit. For data preprocessing, we focus on patients with multiple hospital visits, identified based on their admission information. We extract pairs of consecutive visits for each patient. For each pair, we extract diseases, medications, and procedures from the health records in the former visit as input, and identify the phenotypes in the latter visit as labels, using Clinical Classifications Software (CCS) from the Healthcare Cost and Utilization Project (HCUP)[6].

Algorithm 1 Overview of RAM-EHR.

Input: $\mathcal{P}$: patients; V: corresponding hospital visits of patients.
Initializing multi-source external knowledge $\mathcal{M}$;
for $i = 1, \cdots, |V|$ **do**
  for $c_i \in v_i$ **do**
    Get the medical code c_i and the corresponding textual name s_i included in visit v_i;
    // Passage Retrieval
    Retrieve passages $\mathcal{T}_i$ via Eq. (1)
    // Knowledge Summarization (Accelerated with caching)
    Summarize knowledge e_i for c_i via Eq. (2);
  end for
  // Co-training
  Predict $\widehat{y}_{i,1}$ with knowledge-augmented model g_ϕ via Eq. (3);
  Predict $\widehat{y}_{i,2}$ with visit-based local model f_θ via Eq. (4);
  // Update Model Parameters
  Compute loss function $\mathcal{L}$ via Eq. (5);
  Update model parameters ϕ and θ;
end for
Output: Augmented model g_ϕ and local model f_θ; Final prediction $\widetilde{y}_j = \beta\widehat{y}_{j,1} + (1-\beta)\widehat{y}_{j,2}$ for the j-th test example p_j.

CRADLE. For the CRADLE dataset, we conduct a CVD outcome prediction task, which predicts whether patients with type 2 diabetes will experience CVD complications within 1 year after their initial diagnosis, including coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke. Diseases are identified by their ICD-9 or ICD-10 clinical codes. The usage of data has been approved by the Institutional Review Board (IRB).

[6] https://hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

Table 3: The 25 pre-defined phenotypes in MIMIC-III.

Phenotype	Type
Acute and unspecifed renal failure	acute
Acute cerebrovascular disease	acute
Acute myocardial infarction	acute
Cardiac dysrhythmias	mixed
Chronic kidney disease	chronic
Chronic obstructive pulmonary disease	chronic
Complications of surgical/medical care	acute
Conduction disorders	mixed
Congestive heart failure; nonhypertensive	mixed
Coronary atherosclerosis and related	chronic
Diabetes mellitus with complications	mixed
Diabetes mellitus without complication	chronic
Disorders of lipid metabolism	chronic
Essential hypertension	chronic
Fluid and electrolyte disorders	acute
Gastrointestinal hemorrhage	acute
Hypertension with complications	chronic
Other liver diseases	mixed
Other lower respiratory disease	acute
Other upper respiratory disease	acute
Pleurisy; pneumothorax; pulmonary collapse	acute
Pneumonia	acute
Respiratory failure; insufficiency; arrest	acute
Septicemia (except in labor)	acute
Shock	acute

C Knowledge Sources

C.1 Descriptions

- **PubMed**[7]: PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. It provides users with access to millions of scientific documents, including research papers, reviews, and other scholarly articles. We use the `Entrez` package to extract the PubMed articles[8], resulting in 230k documents.

- **DrugBank**[9] (Wishart et al., 2008): DrugBank is a comprehensive and freely accessible online database containing information on drugs and drug targets. It integrates detailed drug data (chemical, pharmacological, and pharmaceutical) with comprehensive information on drug targets (sequence, structure, and pathway). We use the data from the original database, which contains 355k documents.

- **Medical Subject Headings (MeSH)**[10]: Medical Subject Headings (MeSH) is a comprehensive controlled collection for indexing journal articles and books in the life sciences. It organizes information on biomedical and health-related topics into a hierarchical structure. The corpus contains 32.5k documents covering various medical concepts.

- **Wikipedia**[11] (Vrandečić and Krötzsch, 2014): Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project that is supported by the non-profit Wikimedia Foundation. We extract web pages that contain medical-related information by using the medical codes list (e.g., ICD10 and ATC), resulting in 150k documents.

- **KG**[12] (Chandak et al., 2023): We use PrimeKG in our experiments. It offers a comprehensive overview of diseases, medications, side effects, and proteins by merging 20 biomedical sources to detail 17,080 diseases across ten biological levels. For this study, we select knowledge triplets that contain medical codes within three types (disease, medication, procedure) used in this work, resulting in 707k triplets. We use the template in Appendix C.2 to transform these triplets into sentences.

C.2 Translating Format

We list the template to transform knowledge triplets into sentences in KG as follows:

```
candidate_relation =
["disease_phenotype_positive",
"disease_protein", "disease_disease",
"drug_effect", "drug_protein"]

relations = {
    "phenotype present": "[ent1] has the
    phenotype [ent2]",
    "carrier": "[ent1] interacts with the
    carrier [ent2]",
    "enzyme": "[ent1] interacts with the enzyme
    [ent2]",
    "target": "The target of [ent1] is [ent2]",
    "transporter": "[ent2] transports [ent1]",
    "associated with": "[ent2] is associated
    with [ent1]",
    "parent-child": "[ent2] is a subclass of
    [ent1]",
    "side effect": "[ent1] has the side effect
    of [ent2]"
}
```

[7] https://pubmed.ncbi.nlm.nih.gov/

[8] https://biopython.org/docs/1.75/api/Bio.Entrez.html

[9] https://go.drugbank.com/releases/latest

[10] https://www.ncbi.nlm.nih.gov/mesh/

[11] https://www.wikipedia.org/

[12] https://github.com/mims-harvard/PrimeKG

D Baseline Information

- **Transformer** (Li et al., 2020): It leverages the Transformer (Vaswani et al., 2017) architecture to model sequential EHR visits for clinical prediction tasks.
- **GCT** (Choi et al., 2020): It employs the Transformer model to learn the EHR's hidden structure via medical codes. Additionally, it introduces the Graph Convolutional Transformer, integrating graph neural networks to utilize the EHR structure for prediction.
- **HyGT** (Cai et al., 2022): It leverages hypergraph transformers that regard patients as hyperedges and medical codes as nodes for EHR predictive tasks.
- **MedRetriever** (Ye et al., 2021): It retrieves the most relevant text segments from a local medical corpus using string similarity. Then, it uses query features aggregated with EHR embeddings and disease-specific documents via self-attention.
- **GraphCare** (Jiang et al., 2024): It generates personalized knowledge graphs via prompting LLMs and leverages attention-based graph neural networks for healthcare predictions.
- **CORE** (van Aken et al., 2021): It integrates clinical knowledge with specialized outcome pre-training, and uses language models to predict clinical notes for prediction.
- **BEEP** (Naik et al., 2022): It augments the language models with the retrieved PubMed articles and fuses them with information from notes to predict clinical outcomes.

E Details for Hypergraph Transformer

First of all, we construct a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on EHR data, where each patient visit is represented as a hyperedge connecting to all medical codes associated with the visit as nodes. Then we utilize HyGT (Cai et al., 2022) to jointly learn the node and hyperedge embeddings. Specifically, The *hyperedge embeddings* aggregate information from nodes within each hyperedge, while the *node embeddings* aggregate information from hyperedges connecting the nodes. In the l-th neural network layer, the node and hyperedge embeddings are updated as

$$\boldsymbol{X}_v^{(l)} = f_{\mathcal{E}\to\mathcal{V}}\left(\mathcal{E}_{v,\boldsymbol{E}^{(l-1)}}\right), \tag{6}$$

$$\boldsymbol{E}_e^{(l)} = f_{\mathcal{V}\to\mathcal{E}}\left(\mathcal{V}_{e,\boldsymbol{X}^{(l-1)}}\right), \tag{7}$$

where $\boldsymbol{X}_v^{(l)}$ and $\boldsymbol{E}_e^{(l)}$ represent the embeddings of node v and hyperedge e in the l-th layer ($1 \le l \le L$), respectively. $\mathcal{E}_{v,\boldsymbol{E}}$ denotes the hidden representations of hyperedges that connect the node v, while $\mathcal{V}_{e,\boldsymbol{X}}$ is the hidden representations of nodes that are contained in the hyperedge e. The two message-passing functions $f_{\mathcal{V}\to\mathcal{E}}(\cdot)$ and $f_{\mathcal{E}\to\mathcal{V}}(\cdot)$ utilize multi-head self-attention (Vaswani et al., 2017) to identify significant neighbors during propagation as

$$f_{\mathcal{V}\to\mathcal{E}}(\boldsymbol{S}) = f_{\mathcal{E}\to\mathcal{V}}(\boldsymbol{S}) = \text{Self-Att}(\boldsymbol{S}),$$

where $\boldsymbol{S}$ is the input embedding for the attention layer, $\text{Self-Att}(\boldsymbol{S}) = \text{LayerNorm}(\boldsymbol{Y} + \text{FFN}(\boldsymbol{Y}))$. $\boldsymbol{Y}$ is the output from the multi-head self-attention block $\boldsymbol{Y} = \text{LayerNorm}(\boldsymbol{S} + \Big\|_{i=1}^{h}\text{SA}_i(\boldsymbol{S}))$, $\text{SA}_i(\boldsymbol{S})$ denotes the scaled dot-product attention:

$$\text{SA}_i(\boldsymbol{S}) = \text{softmax}\left(\frac{\boldsymbol{W}_i^Q(\boldsymbol{S}\boldsymbol{W}_i^K)^\top}{\sqrt{\lfloor d/h \rfloor}}\right)\boldsymbol{S}\boldsymbol{W}_i^V.$$

$\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$, and $\boldsymbol{W}_i^V$ are learnable parameters for the i-th head corresponding to queries, keys, and values, respectively. To interpret the above process, the input sequence $\boldsymbol{S}$ is projected into different h heads. The output of each head is then concatenated (denoted by $\Big\|$) to form the multi-head attention output. This output of multi-head attention layer $\boldsymbol{Y}$ is then fed into a feed-forward neural network (FFN), comprising a two-layer Multilayer Perceptron (MLP) with ReLU activation functions.

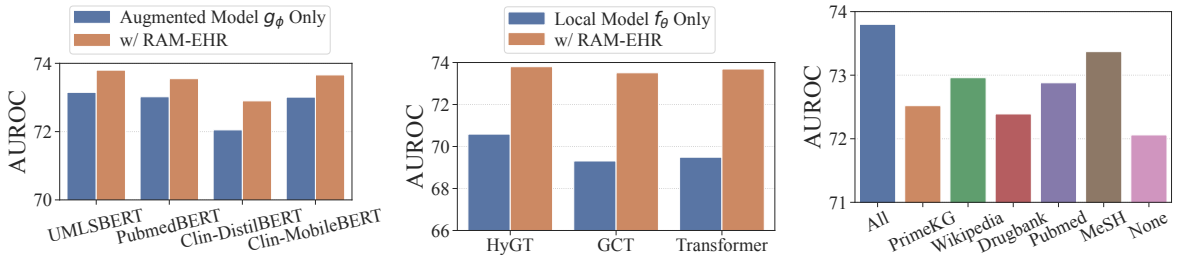
F Details for Prompt Design

We present the detailed design of the prompt template as follows:

Prompt for LLM Summarization

```
Suppose you are a physician working on a health
    -related outcome prediction task and need
    to get relevant information for the given
    <task>. Here is relevant information:
<medical code type> Name: <medical code name>
Retrieve Passage #1: <retrieved document 1>
Retrieve Passage #2: <retrieved document 2>

...
Based on the above information, Could you
    generate 1 sentence of around 10-20 words
    to summarize the knowledge for the
```



(a) Effect of g_ϕ (b) Effect of f_θ (c) Effect of $\mathcal{M}$

Figure 5: Results for additional studies on the effect of (a) f_θ, (b) g_ϕ, and (c) the information source $\mathcal{M}$ on the CRADLE dataset.

```
    <medical code type> that is useful for
    the <task>?
```

<task> is the brief description of the downstream task. <medical code type> is either "disease", "medication" or "procedure", depending on the input.

G Additional Experiment Results

We evaluate the effect of the augmented model g_ϕ, the local model f_θ, and the knowledge source $\mathcal{M}$ on CRADLE in Figure 5. The experimental results further demonstrate that both models and different knowledge sources contribute to the performance gain. Moreover, it is observed that RAM-EHR is flexible to be applied upon different models, with a comparable performance with RAM-EHR.

For the human studies in Section 3.3, we provide the following guidelines for the annotators to evaluate the quality of the generated knowledge.

```
The goal of this evaluation is to assess the
helpfulness of generated knowledge explaining
or relating to specific medical codes in the
context of target prediction tasks. Helpfulness
is defined by the relevance, accuracy, and
utility of the information in facilitating
understanding or decision-making related to
medical coding and its implications for
predictive tasks.

Please rate the following generated knowledge
with score 0, 1 or 2.

> 0: Irrelevant
Definition: The knowledge does not provide any
relevant information related to the medical
code in question. It might be factually
accurate but completely off-topic or not
applicable to the context of target prediction
tasks.

> 1: Partially Relevant and Useful
Definition: The knowledge provides some
relevant information but either lacks
completeness, specificity, or direct
applicability to target prediction tasks. It
might include general facts or insights that
are related to the medical code but does not
fully support decision-making or understanding
in a predictive context.

> 2: Very Useful
Definition: The knowledge directly addresses
the medical code with accurate, relevant, and
comprehensive information that is highly
applicable to target prediction tasks. It
should provide detailed understanding, or
specific examples that facilitate
decision-making, understanding, or application
in predictive modeling.
```

H Cost Information

Utilizing `GPT-3.5-turbo` as our base LLM model for generating summarized knowledge, we observe an average retrieval augmentation cost of $0.0025 per medical code in MIMIC-III and $0.0032 in CRADLE. Consequently, RAM-EHR does not result in excessive monetary expenses.