Text-Attributed Knowledge Graph Enrichment with Large Language Models for Medical Concept Representation

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Chen Chen, Dongjie Wang, Zijun Yao*
University of Kansas, USA
{mohsen.nayebi, a.hadizadehm, chenchen, wangdongjie, zyao}@ku.edu

Abstract

In electronic health record (EHR) mining, learning high-quality representations of medical concepts (e.g., standardized diagnosis, medication, and procedure codes) is fundamental for downstream clinical prediction. However, robust concept representation learning is hindered by two key challenges: (i) clinically important cross-type dependencies (e.g., diagnosis-medication and medication-procedure relations) are often missing or incomplete in existing ontology resources, limiting the ability to model complex EHR patterns; and (ii) rich clinical semantics are often missing from structured resources, and even when available as text, are difficult to integrate with KG structure for representation learning. To address these challenges, we present MEDCO, an LLM-empowered graph learning framework for medical concept representation. MEDCO first builds a global knowledge graph (KG) over medical codes by combining statistically reliable associations mined from EHRs with type-constrained LLM prompting to infer semantic relations. It then utilizes LLMs to enrich the KG into a text-attributed graph by generating node descriptions and edge rationales, providing semantic signals for both concepts and their relationships. Finally, MEDCO jointly trains a LoRA-tuned LLaMA text encoder with a heterogeneous GNN, fusing text semantics and graph structure into unified concept embeddings. Extensive experiments on MIMIC-III and MIMIC-IV show that MEDCO consistently improves prediction performance and serves as an effective plug-in concept encoder for standard EHR pipelines.

1 Introduction

Electronic health records (EHRs) encode a patient's medical history as a high-dimensional and sparse sequence of diagnosis (dx), medication (rx), and procedure (px) concepts, which enable a broad range of clinical prediction tasks (Choi et al., 2016b; Nayebi Kerdabadi et al., 2023; Moghaddam et al., 2024; Nayebi Kerdabadi et al., 2025). A cornerstone in EHR mining is to learn medical concept representations that can capture the meaningful but complex dependency patterns (e.g., dx-rx treatment/contraindication, dx-px indication/care pathways, or dx-dx comorbidity) to benefit downstream outcome prediction. Therefore, utilizing knowledge graphs to model heterogeneous concepts and their relations becomes an appealing direction to improve concept representation learning.

However, the KGs most commonly used in practice are largely derived from existing medical ontologies, which either encode within-type parent-child hierarchies (e.g., ICD) or provide limited cross-type semantics (e.g., UMLS)[1]. Although large language models (LLMs) have recently emerged as a promising tool for enriching KGs by proposing missing relations and generating structured semantic descriptions, unconstrained prompting can produce plausible but unsupported edges and inconsistent outputs for the same concepts. Furthermore, because many clinically meaningful relations are context-dependent and vary across cohorts, care settings, and time, LLM-based KG enrichment must validate proposed relations against what is actually observed in the target EHR dataset, ensuring that edges are not only clinically interpretable but also empirically supported. These requirements create a key tension: although LLMs encode broad biomedical knowledge, KG induction for clinical modeling must remain evidence-grounded, type-aware, and globally consistent.

Beyond *which* relations a KG should contain, a second challenge is *how* to incorporate node and edge information into concept embedding. Since each medical concept can be associated with rich

*Corresponding author.

[1] A broader comparison of existing medical ontologies is provided in Appendix A.1, Table 6.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16544–16560
July 2-7, 2026

LLM-generated semantics, such as indications, mechanisms of action, roles in care, and contextual constraints, it is natural to model the KG as a text-attributed graph in which nodes and edges are paired with descriptive text. This representation has the potential to combine the relational expressiveness of KGs with the semantic richness of LLM-generated knowledge. However, effectively learning from an LLM-enriched text-attributed graph is nontrivial. Graph neural networks (GNNs) excel at aggregating graph structure but are not designed to interpret long-form medical language. In contrast, LLMs can encode nuanced textual semantics but do not explicitly enforce or exploit global relational constraints. This challenge motivates a co-learning mechanism in which (i) an LLM encodes node- and edge-level textual context, while (ii) a relation-aware GNN propagates and refines these representations over the KG structure. In this way, textual semantics and graph structure can be jointly captured and aligned under task supervision.

In this work, we propose MEDCO, which addresses the above challenges by constructing a clinically interpretable and empirically supported KG and learning text-attributed concept embeddings through a KG-LLM co-learning framework. Specifically, we first build a global heterogeneous KG over diagnosis, medication, and procedure codes by combining (i) statistically reliable associations from EHR data, such as intra-visit co-occurrences and cross-visit temporal transitions, with (ii) structured, type-constrained LLM prompting that assigns directed relation types and calibrated confidence scores. Using the resulting text-attributed KG, enriched with node descriptions and edge rationales, we then jointly train a LoRA-tuned LLaMA text encoder and a heterogeneous GNN to fuse text-derived clinical semantics with graph-derived relational structure into unified concept embeddings for downstream EHR prediction. Our contributions are as follows:

- We introduce an LLM-assisted pipeline to construct a global heterogeneous KG by combining EHR-derived co-occurrences and temporal statistics with type-constrained relation inference, yielding clinically meaningful relations among diagnosis, medication, and procedure codes.
- We enrich the KG into a text-attributed graph with LLM-generated node and edge semantics, including textual rationales, relation labels, confidence scores, and EHR-derived statistics.
- We propose MEDCO, a KG-LLM co-learning framework that jointly fine-tunes a LLaMA text encoder and a heterogeneous GNN to learn unified medical concept embeddings.
- We show that MEDCO serves as an effective plug-in concept encoder for standard EHR backbones through extensive experiments on MIMIC-III and MIMIC-IV for sequential diagnosis prediction.

2 Methodology

2.1 Notation and Problem Definition

EHRs. We represent the EHR record for patient $j \in \mathcal{J}$ as a sequence of T_j clinical visits, denoted by $\mathcal{X}_j = \{V_{j,t}\}_{t=1}^{T_j}$. Each visit $V_{j,t}$ is a set of $N_{j,t}$ medical codes, $V_{j,t} = \{c_i\}_{i=1}^{N_{j,t}}$, where each code c_i corresponds to a diagnosis (*dx*), a medication (*rx*), or a procedure (*px*). For brevity, we omit the patient index j and visit index t.

Healthcare Predictive Tasks. Given a patient's visit sequence $\mathcal{X}_j = \{V_1, V_2, \ldots, V_{T_j}\}$, the objective is to predict clinical outcomes. These may involve binary tasks (e.g., mortality) or multi-label classification (e.g., diagnosis prediction). In this paper, we focus on predicting diagnosis codes for the next visit V_{T_j+1}, a comprehensive task aimed at identifying potential diseases for future encounters.

2.2 Method Summary

We present MEDCO, which constructs an evidence-grounded, LLM-guided heterogeneous KG and learns plug-in medical concept embeddings through an LLM-GNN co-learning framework. As illustrated in Figure 1, the framework consists of four steps: **(1) Evidence extraction.** Compute co-occurrence and next-visit transition statistics, and retain statistically significant code pairs. **(2) KG induction.** Use type-constrained, evidence-conditioned LLM prompting to assign directed relation types, confidence scores, and rationales, thereby constructing a heterogeneous KG. **(3) KG enrichment.** Enrich the KG with LLM-generated node descriptions and edge metadata, including relations, supporting evidence, confidence scores, and rationale embeddings. **(4) Co-learning.** Jointly train a LoRA-tuned LLaMA encoder and a relation-aware GNN to fuse textual semantics and graph structure into concept embeddings for downstream EHR prediction.

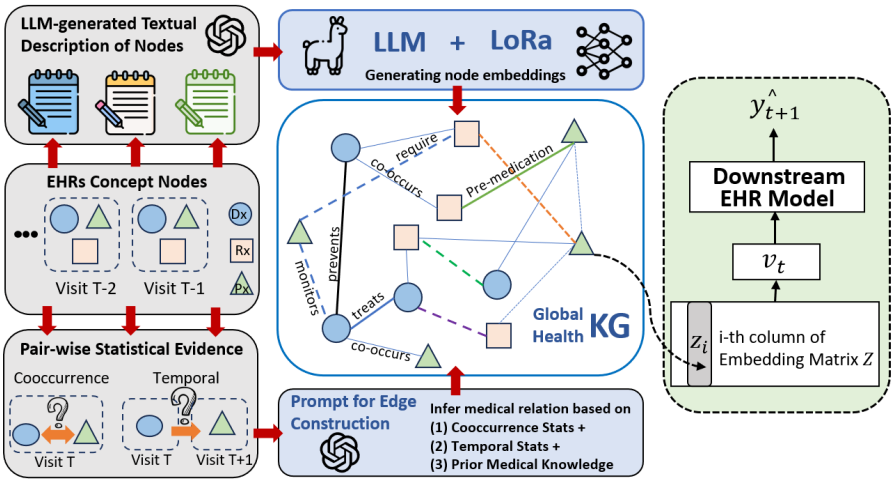

Figure 1: Overview of MEDCO. (1) Extract intra-visit co-occurrence and next-visit transition evidence from EHRs, and retain well-supported code pairs. (2) Use type-constrained, evidence-conditioned LLM prompting to assign directed relation types (with confidence and rationales), thereby constructing a heterogeneous KG. (3) Enrich nodes with LLM-generated concept descriptions and edges with relation metadata. (4) Jointly train a LoRA-tuned LLaMA encoder and a relation-aware GNN to learn medical concept embeddings for downstream prediction.

2.3 LLM-Guided Construction of Global Health Knowledge Graph

We construct a clinically grounded heterogeneous KG by inducing relations among diagnosis, procedure, and medication codes using longitudinal EHRs. The pipeline combines EHR-derived co-occurrence and temporal evidence with evidence-conditioned LLM prompting to infer directed relation types. It proceeds in two stages: (1) extract pair-level statistical evidence and (2) perform evidence-supported LLM relation inference.

2.3.1 Pairwise Statistical Evidence Extraction from EHRs

We extract empirical evidence for candidate code pairs from EHRs in two steps: (i) computing intra-visit co-occurrence and next-visit transition statistics, and (ii) filtering to retain well-supported, clinically meaningful pairs. Each patient is represented as a sequence of visits, where each visit is a de-duplicated set of diagnosis (dx), procedure (px), and medication (rx) codes. These visit-level code sets are used to compute the corresponding marginal frequencies and transition statistics.

Pair-level evidence. For a directed pair $(c_i \to c_j)$, we compute three signals: a smoothed conditional probability, pointwise mutual information (PMI), and a chi-square test of dependence. We use a common formulation for both intra-visit co-occurrence and next-visit transitions. Let $x(\cdot)$ denote the appropriate count function: for co-occurrence, $x(c_i)$ is the number of visits containing c_i, and $x(c_i, c_j)$ is the number of visits in which c_i and c_j co-occur; for transitions, $x(c_i)$ is the number of source-visit occurrences of c_i, and $x(c_i, c_j)$ is the number of observed transitions from c_i in V_t to c_j in V_{t+1}. Let $\mathcal{C}$ denote the set of distinct codes and $|\mathcal{C}|$ its cardinality.

(1) Smoothed conditional probability.

$$P(c_j \mid c_i) \;=\; \frac{x(c_i, c_j) + \alpha}{x(c_i) + \alpha|\mathcal{C}|},$$

where α is a Laplace smoothing constant used to mitigate sparsity and avoid zero probabilities.

(2) PMI-style association.

$$\mathrm{PMI}(c_i, c_j) \;=\; \log_2\!\left(\frac{p(c_i, c_j)}{p_{\mathrm{src}}(c_i)\, p_{\mathrm{tgt}}(c_j)}\right),$$

where $p(\cdot)$ denotes empirical probabilities estimated from counts, and p_{src} and p_{tgt} are the corresponding marginal probabilities. For intra-visit co-occurrence, $p_{\mathrm{src}} = p_{\mathrm{tgt}}$.

(3) Statistical significance. We compute a chi-square p-value from a 2×2 contingency table to test dependence between c_i and c_j for both intra-visit co-occurrence and next-visit transition settings.

Evidence Filtering. For each unordered pair (c_i, c_j), we compute intra-visit and temporal evidence (support counts, conditional probabilities, PMI, and χ^2 p-values) and consolidate them into a unified table covering both directions $(c_i \to c_j)$ and $(c_j \to c_i)$. We discard pairs with low support, weak

association (low conditional probability/PMI), or non-significant dependence (e.g., $p > 0.05$), yielding a set of statistically reliable candidates. These pairs are then passed to the LLM for relation-type assignment to construct a clinically meaningful, evidence-grounded KG. A detailed breakdown is provided in Appendix A.2.

2.3.2 Evidence-Supported LLM Prompting for Medical Relationship Induction

Given the final set of code pair candidates, the last stage of our pipeline assigns each pair a clinically meaningful semantic relationship using an LLM. We design a structured, evidence-supported prompting framework that guides the LLM to choose the most appropriate relation between two medical concepts from a predefined, type-specific relation pool, explicitly grounding its decisions in clinical knowledge and EHR-derived evidence.

Type-Constrained Relationship Pool. Each code-type combination (dx–dx, rx–dx, px–dx, rx–rx, px–px, px–rx) is associated with a curated set of candidate relations grounded in clinical practice (Appendix A.3). The resulting relation inventory spans key semantic axes, including causality (e.g., *causes*), temporal progression (e.g., *leads_to*), therapeutic intent (e.g., *treats*, *combination_therapy_with*), diagnostic usage (e.g., *diagnostic_of*), workflow structure (e.g., *prerequisite_for*), and safety considerations (e.g., *contraindicated_for*). When neither clinical priors nor statistical evidence supports a confident assignment, the relation inventory provides conservative abstention labels (*no_significant_relation*, *cannot_decide*). Collectively, these relations define a concise, interpretable semantic space for modeling clinically meaningful dependencies in EHRs.

Structured Prompting. For each code pair, we construct a prompt (Figure 2) that includes: (i) the code identifiers, textual names, parent categories, and marginal frequencies for both codes; (ii) directional statistical evidence from intra-visit co-occurrence and next-visit transitions (8 metrics) together with a brief metrics glossary; (iii) candidate relations permitted by the corresponding code-type combination (e.g., dx-rx); and (iv) explicit decision rubrics and confidence-calibration guidelines. The LLM is instructed to prioritize clinical plausibility, using statistical signals as supporting evidence, and to return a strict JSON object containing: (1) a single relation label, (2) an oriented

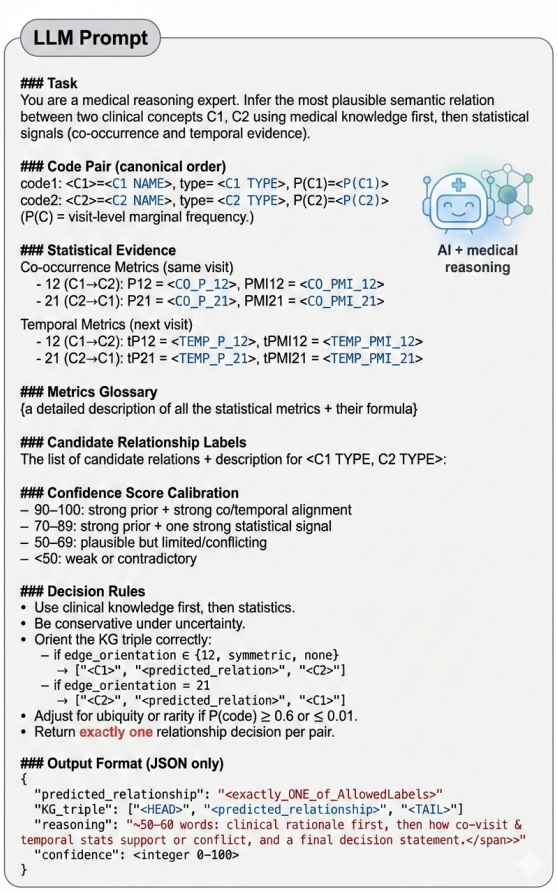

Figure 2: Prompt for Medical Relationship Induction.

KG triple, (3) a calibrated confidence score, and (4) a 50–60 word reasoning explaining the clinical rationale and how the co-occurrence and temporal statistics support or challenge the decision. This inference stage transforms medical knowledge encoded in the LLM, guided by statistically validated evidence, into a structured, relation-typed, and clinically interpretable knowledge graph that unifies EHR-derived evidence with clinical reasoning.

Clinical expert audit of induced edges. To assess potential hallucination in LLM-inferred relations, we conducted a targeted clinical expert audit of the constructed KG. We sampled 50 edges by selecting 5 edges from each of the 10 most frequent predicted relation types. Two independent frontline clinicians from the University of Kansas Health System reviewed each edge under general clinical knowledge and rated its correctness on a 1–5 scale. The mean score, averaged across the two reviewers for each edge, was 4.84 ± 0.29, indicating high agreement with the clinical validity of the sampled relations. Full protocol details and the relation-wise summary are provided in Appendix A.5.

2.4 LLM-Based Contextual KG Enrichment

Starting from the constructed medical KG, we further enrich nodes and edges with contextual information derived from a large language model (LLM). The LLM is treated as a high-coverage medical knowledge base: it provides clinically detailed descriptions for individual medical concepts and augments edge semantics using relation reasoning and EHR-derived evidence. The resulting features supply a semantic and knowledge-grounded context for downstream graph learning.

Node-Level Enrichment. For each medical code (diagnosis, procedure, or medication), we use a type-specific prompt to generate a concise, clinically focused description using an LLM (e.g., typical presentation, indications, role in care, and key nuances; see Appendix A.6). We attach this description to the corresponding KG node as textual attributes, which are subsequently encoded by MEDCO during joint KG-LLM co-learning.

Edge-Level Enrichment. For each candidate code pair, the previous subsection yields (i) an oriented relation label, (ii) an LLM-generated confidence score and free-text rationale, and (iii) eight EHR-derived statistical metrics (co-occurrence and temporal transition statistics). We combine these signals to construct informative edge features.

2.5 MEDCO: Co-Learning of LLM and GNN for Medical Concept Representations

Given the constructed KG, we learn unified node representations with MEDCO, a co-learning medical concept representation framework that jointly fine-tunes a large language model (LLM) and a heterogeneous graph neural network (GNN). The LLM encodes each node's textual description into an initial embedding, which is then passed to the GNN for neighborhood message passing over the KG. We train both components end-to-end; for the LLM, we use LLaMA-1B and fine-tune it with low-rank adaptation (LoRA). This co-learning setup combines the LLM's semantic knowledge with the GNN's ability to incorporate relational structure and EHR-derived evidence.

Architecture. Let $G = (V, E)$ denote the global heterogeneous medical knowledge graph, where V is the set of medical concepts and E is the set of directed edges. Each node $i \in V$ represents a medical concept with node type $\tau(i) \in \mathcal{T}$, where $\mathcal{T}$ denotes the set of node types (diagnosis, medication, and procedure), and is associated with a textual description s_i (Appendix A.6). Each directed edge $(j \to i) \in E$ is associated with a relation type $r_{ji} \in \mathcal{R}$, where $\mathcal{R}$ is the set of relation types, and an edge feature vector $\mathbf{e}_{ji} \in \mathbb{R}^{d_e}$.

A LLaMA-based text encoder f_θ, fine-tuned with LoRA adapters, maps the description s_i to a text embedding, which is then projected into the GNN hidden space using a type-specific linear map $\mathbf{W}_{\tau(i)} \in \mathbb{R}^{d \times d_L}$:

$$\begin{aligned} \mathbf{z}_i^{\text{text}} &= f_\theta(s_i) \in \mathbb{R}^{d_L}, \\ \mathbf{h}_i^{(0)} &= \mathbf{W}_{\tau(i)}\, \mathbf{z}_i^{\text{text}} \in \mathbb{R}^{d}. \end{aligned} \tag{1}$$

These embeddings initialize a relation-aware heterogeneous GNN g_ϕ that incorporates both relation types and edge features during message passing. Let $\mathcal{N}(i) = \{j \in V : (j \to i) \in E\}$ denote the set of in-neighbors of node i. For layer $\ell \in \{0, \ldots, L-1\}$, node states are updated as

$$\begin{aligned} \mathbf{u}_{ji}^{(\ell)} &= \mathrm{MSG}^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \mathbf{e}_{ji}, r_{ji}\right), \\ \mathbf{m}_i^{(\ell)} &= \mathrm{AGG}^{(\ell)}\left(\left\{\mathbf{u}_{ji}^{(\ell)} : j \in \mathcal{N}(i)\right\}\right), \\ \mathbf{h}_i^{(\ell+1)} &= \mathrm{UPDATE}^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{m}_i^{(\ell)}\right). \end{aligned} \tag{2}$$

Here, $\mathrm{MSG}^{(\ell)}(\cdot)$ denotes the message function that computes the edge-specific message $\mathbf{u}_{ji}^{(\ell)}$ from node j to node i; $\mathrm{AGG}^{(\ell)}(\cdot)$ denotes a permutation-invariant aggregation operator over incoming messages; and $\mathrm{UPDATE}^{(\ell)}(\cdot)$ denotes the node-state update function at layer ℓ. Intuitively, the LLM provides semantic information from textual concept descriptions, while the heterogeneous GNN incorporates relational structure and EHR-derived evidence. Joint end-to-end training fuses these two sources into clinically meaningful concept representations.

LLM-GNN Training Strategy. We jointly train the LLM-based node encoder and the GNN, but a practical challenge arises during mini-batch training. Each batch contains multiple patients, and the union of diagnosis, medication, and procedure codes appearing across their visit sequences can be large. If we update the LLM representations for all such codes at every iteration, the cost becomes prohibitively expensive, especially given the size of the medical code vocabulary. A naive workaround is to update only a random subset of batch-activated

codes per iteration, but this has two drawbacks: (i) rare yet clinically important codes may receive very few or no updates, and (ii) highly frequent codes may be updated disproportionately often, leading to unbalanced adaptation of the LLM.

To address this issue, we maintain lightweight per-code statistics that track how often each code appears in the training data and how many times its LLM representation has been updated. We then adopt a two-phase sampling schedule for LoRA updates. In the early phase, we prioritize codes with the fewest updates ("least-updated-first"), encouraging broad coverage and ensuring that all observed codes receive at least a few LLM updates. In the later phase, each iteration mixes a subset of the least-updated codes with a subset of the most frequently occurring codes from the current batch, allowing the model to further refine high-impact nodes without neglecting the long tail. This strategy amortizes LLM computation while enabling controlled, coverage-aware fine-tuning of the LLM representations used by the GNN.

2.6 Integrating MEDCO into EHR Models

As a medical concept encoder, MEDCO can be integrated into standard EHR models and trained end-to-end; it improves concept representation learning and consequently downstream predictive performance. Given the learned concept (code) embeddings $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, we form the embedding matrix $\mathbf{Z} \in \mathbb{R}^{d \times N}$, where $\mathbf{z}_i$ is the i-th column of $\mathbf{Z}$. This matrix is then used by a downstream task model. Considering sequential diagnosis prediction, which maps a sequence of visits to the diagnoses of the next visit: $f : \{V_1, V_2, \ldots, V_t\} \rightarrow \hat{\mathbf{y}}_{t+1}$, where $\hat{\mathbf{y}}_{t+1} \in \mathbb{R}^{N_{dx}}$ is a multi-hot vector and N_{dx} denotes the total number of diagnosis codes:

$$\begin{aligned} \mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] &\leftarrow \text{MEDCO}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \\ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t &= \mathbf{Z}[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_t] \\ \mathbf{h}_p &= \text{Model}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t) \\ \hat{\mathbf{y}}_{t+1} &= \text{Sigmoid}(\mathbf{W}\mathbf{h}_p + \mathbf{b}) \end{aligned} \tag{3}$$

For each visit V_t, we compute $\mathbf{v}_t = \mathbf{Z}\mathbf{u}_t$, where $\mathbf{u}_t \in \{0, 1\}^N$ is a multi-hot code indicator and $\mathbf{Z}$ is the learned concept embedding matrix. The sequence $\{\mathbf{v}_1, \ldots, \mathbf{v}_t\}$ is encoded by a backbone model to produce a patient representation $\mathbf{h}_p$. The final prediction is obtained by a linear classification layer followed by a sigmoid activation. We train with multi-label cross-entropy between $\hat{\mathbf{y}}_{t+1}$ and $\mathbf{y}_{t+1}$ at each timestep.

3 Experimental Setting

Datasets. We evaluate on two public EHR benchmarks, MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023). We map ICD diagnosis/procedure codes to CCS categories and medications from National Drug Codes (NDC) to the Anatomical Therapeutic Chemical (ATC) classification. Table 2 reports cohort statistics. The task is next-visit diagnosis prediction over imbalanced label spaces (515 codes in MIMIC-III; 562 in MIMIC-IV). We use `Llama-3.2-1B` (Grattafiori and et al., 2024; Meta, 2024) as the text encoder and fine-tune it with LoRA using rank r=8 and α=32. The heterogeneous GNN encoder has 2 layers with 1 attention head per layer and dropout 0.4. The source code is publicly available.[2]

Evaluation Metrics. We report mean results over 5 folds for **AUPRC** (area under the precision-recall curve; also stratified by label frequency to assess performance across code rarity levels), **Acc@k** (top-k accuracy normalized by $\min(k, |\mathbf{y}_{t+1}|)$), and **F1** (harmonic mean of precision and recall).

4 Evaluation Results

To assess MEDCO, we address the following research questions: **RQ1:** Does MEDCO improve downstream EHR prediction when used as a plug-in medical concept encoder? **RQ2:** How does MEDCO compare with existing medical code encoders? **RQ3:** How do individual components of MEDCO and different KG edge categories contribute to overall performance? **RQ4:** How well does MEDCO mitigate data scarcity and improve performance on rare conditions?

4.1 RQ1: Plug-in Enhancement Evaluation

We hypothesize that integrating MEDCO, our medical concept encoder, into standard EHR backbones improves downstream performance by enhancing concept representations. To test this hypothesis, we incorporate MEDCO into four representative models: (1) **AdaCare** (Ma et al., 2020), an explainable architecture that models multi-scale biomarker variations; (2) **Transformer** (Vaswani, 2017), based on self-attention; (3) **RETAIN** (Choi et al., 2016b), an RNN with reverse-time, two-level attention; and (4) **TCN** (Bai et al., 2018), which uses causal convolutions for temporal modeling. We compare each backbone with and without MEDCO. As shown

[2] https://github.com/mohsen-nyb/MedCo.git

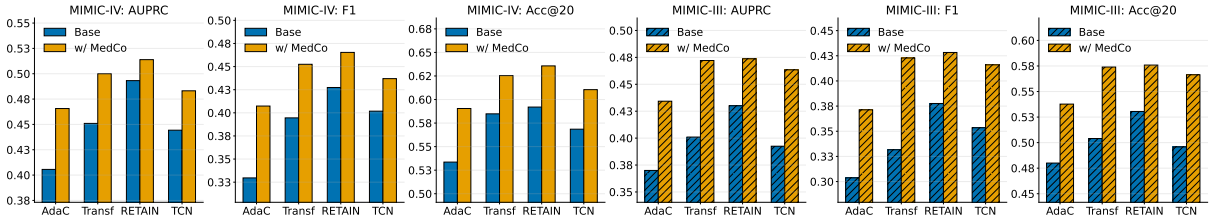

Figure 3: Performance gains from integrating MEDCO into four diagnosis prediction backbones (plug-in analysis) on MIMIC-III and MIMIC-IV.

Table 1: Performance comparison on MIMIC-III and MIMIC-IV. We report overall AUPRC, F1, and Acc@k, and stratified AUPRC by diagnosis label frequency quartiles (0–25%, 25–50%, 50–75%, 75–100%). The first row corresponds to the base Transformer, and subsequent rows denote plug-in variants (e.g., GRAM = Base + GRAM).

	Model	General Performance					Label Category Performance (AUPRC)			
		AUPRC	F1	Acc@15	Acc@20	Acc@30	0–25%	25–50%	50–75%	75–100%
MIMIC-III	Base	41.00	33.16	47.20	50.40	58.80	40.60	47.30	72.80	78.40
	GRAM	41.70	34.60	48.60	51.90	59.80	42.20	50.10	74.10	79.30
	MMORE	42.60	35.30	49.20	52.60	60.40	42.80	51.20	74.80	80.00
	KAME	42.18	35.07	48.97	52.06	60.06	41.84	49.76	74.06	79.14
	G-BERT	42.47	35.38	49.26	52.33	60.27	42.06	50.04	74.24	79.28
	HAP	42.36	35.17	49.11	52.18	60.18	41.95	49.92	74.13	79.24
	ADORE	42.58	35.43	49.33	52.47	60.32	42.18	50.09	74.29	79.31
	KAMPNet	43.06	35.96	49.88	53.07	60.86	42.63	50.83	74.77	79.89
	GraphCare	43.35	35.46	52.76	56.00	62.75	44.80	58.16	70.97	65.72
	LINKO	44.91	38.20	52.30	55.62	62.80	44.80	55.20	**76.43**	83.41
	MEDCO	**47.21**	**42.28**	**54.20**	**57.40**	**64.39**	**47.67**	**66.29**	76.02	**86.70**
MIMIC-IV	Base	45.10	39.10	54.80	58.10	64.70	46.80	53.60	53.90	77.20
	GRAM	46.30	40.00	55.60	58.90	65.70	47.70	54.60	54.80	77.90
	MMORE	46.90	40.60	56.10	59.40	66.10	48.20	55.10	55.70	78.40
	KAME	45.97	39.66	55.27	58.57	65.26	47.16	54.24	54.28	77.63
	G-BERT	46.37	40.07	55.68	58.96	65.66	47.58	54.73	54.87	78.04
	HAP	46.33	39.88	55.57	58.92	65.58	47.52	54.57	54.83	77.92
	ADORE	46.52	40.23	55.84	59.09	65.79	47.74	54.84	55.06	78.13
	KAMPNet	47.08	40.96	56.46	59.77	66.47	48.26	55.57	56.16	78.68
	GraphCare	46.90	40.12	55.89	59.14	65.71	48.31	**60.31**	73.70	66.33
	LINKO	48.14	42.05	57.78	61.21	67.97	49.61	56.87	57.58	80.27
	MEDCO	**50.00**	**45.25**	**59.22**	**62.54**	**68.71**	**52.01**	57.70	**59.35**	**81.19**

Table 2: Data statistics for MIMIC-III and MIMIC-IV.

Metric	MIMIC-III	MIMIC-IV
# Patients	7,515	18,829
# Visits (samples)	12,430	25,028
# Labels/sample	12.31	10.56
# Unique conditions (ICD)	515	562
# Conditions/sample	26.85	59.50
# Drugs/sample	70.10	118.16
# Unique drugs	471	510
# Procedures/sample	6.49	5.41
# Unique procedures	280	322

Table 3: Component-wise ablation of MEDCO on MIMIC-IV and MIMIC-III.

	Model	PRAUC	F1	Acc@20
MIMIC-IV	Base (Transformer)	45.10	39.10	58.10
	Base + KG	48.55	42.60	60.90
	Base + KG + Edge feat.	48.72	42.85	61.10
	Base + KG + Edge feat. + LLM (freeze)	49.38	43.71	61.47
	Base + KG + Edge feat. + LLM (LoRA)	**50.00**	**45.25**	**62.54**
MIMIC-III	Base (Transformer)	41.00	33.16	50.40
	Base + KG	45.79	38.90	55.60
	Base + KG + Edge feat.	45.91	39.20	55.85
	Base + KG + Edge feat. + LLM (freeze)	46.10	40.45	56.44
	Base + KG + Edge feat. + LLM (LoRA)	**47.21**	**42.28**	**57.40**

in Figure 3, MEDCO consistently improves performance across all models, demonstrating its effectiveness as a plug-in concept encoder.

4.2 RQ2: Baseline Comparison

We compare MEDCO against widely used medical concept encoders, including the **base Transformer** (Vaswani, 2017) (i.e., no concept encoder), hierarchy-driven ontology methods such as **GRAM** (Choi et al., 2017), **MMORE** (Song et al., 2019), **KAME** (Ma et al., 2018), and **HAP** (Zhang et al., 2020), as well as KG-based approaches that leverage external knowledge resources. Among these, **ADORE** (Cheong et al., 2023) relies on SNOMED, while **GraphCare** (Jiang et al., 2023) leverages UMLS-derived edges together with LLM-derived edges. **KAMPNet** (An et al., 2023) and **LINKO** (Nayebi Kerdabadi et al., 2025) construct cross-type connections but do not explicitly model rich semantic relation types. All encoders are integrated into the same downstream backbone for a controlled comparison. As shown in Table 1 (General Performance), MEDCO achieves the best overall performance, indicating that our evidence-grounded, heterogeneous, text-attributed KG provides richer and more task-relevant semantics than existing medical concept baselines.

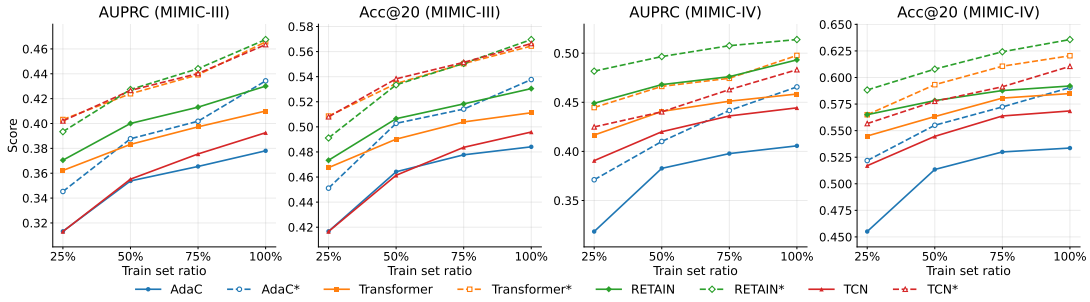

Figure 4: Performance evaluation across different training set sizes using the MIMIC-III and MIMIC-IV datasets. An asterisk (*) next to each encoder indicates the integration of MEDCO to that model, e.g., Transformer* = Transformer + MEDCO.

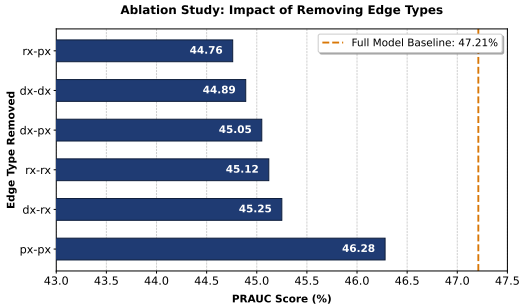

Figure 5: Ablation study showing the impact of removing each edge type on PRAUC for MIMIC-III. The dashed vertical line indicates the full-model baseline.

Furthermore, to evaluate performance under data scarcity, we stratify diagnosis codes into four frequency quartiles (0–25%, 25–50%, 50–75%, 75–100%), where 0–25% contains the rarest codes. This analysis probes whether MEDCO can better propagate informative signals to low-frequency concepts. Table 1 (Label Category Performance) shows that MEDCO improves AUPRC across all bands, with pronounced gains on rare codes.

4.3 RQ3: Ablation Study

Table 3 shows progressive gains as components are added to MEDCO. Starting from the base Transformer, adding the induced KG with a relation-aware GNN provides the largest boost, highlighting the value of explicit cross-domain structure. Edge features yield additional improvements by injecting LLM clinical reasoning and EHR-derived association evidence into message passing. Adding LLM-based node descriptions (frozen encoder) consistently boosts performance, indicating that clinical semantics from natural language complements graph signals. Finally, enabling LoRA fine-tuning delivers the best results, showing that adapting the LLM to the cohort and task produces more informative node representations and maximizes downstream accuracy. Figure 5 presents an edge-category ablation of the induced KG. For each bar, we remove all edges belonging to a given relation family (e.g., rx-px removes every medication-procedure edge) while keeping the rest of the graph unchanged, and report the resulting PRAUC. Larger drops indicate edge categories that contribute more to downstream performance.

LLM backbone	AUPRC	Acc@20
Llama-3.2-1B	50.00	62.54
Gemma-2-2B	50.31	62.89
Qwen2.5-1.5B	50.16	62.67
Qwen2.5-0.5B	49.31	61.44
SmolLM2-1.7B	50.11	62.61

Table 4: Sensitivity of MEDCO to the choice of LLM backbone on MIMIC-IV.

4.4 Sensitivity to the LLM Backbone

We further evaluate whether MEDCO is sensitive to the choice of LLM backbone for node-text encoding. While our main experiments use `Llama-3.2-1B`, the framework is model-agnostic and can be instantiated with other open-source LLMs. Under the same LoRA fine-tuning setup, we additionally test `Gemma-2-2B` (Gemma Team, 2024; Google, 2024), `Qwen2.5-1.5B` (Hui et al., 2024; Qwen Team, 2024b), `Qwen2.5-0.5B` (Hui et al., 2024; Qwen Team, 2024a), and `SmolLM2-1.7B` (Allal et al., 2025; Hugging Face TB, 2025) on MIMIC-IV. As shown in Table 4, performance remains consistently strong across backbones, indicating that MEDCO is not tied to a specific LLM family. Larger backbones provide only modest gains, suggesting that downstream performance is driven primarily by the evidence-grounded KG construction and graph propagation components.

4.5 RQ4: Data Insufficiency Analysis

To evaluate MEDCO under data scarcity, we subsample the training set to simulate limited-data regimes. In the plug-in setting, Figure 4 shows that MEDCO delivers substantial gains for downstream

Variant	AUPRC	Acc@20	Train(s)	Infer(s)	Peak train mem (MiB)	Peak infer mem (MiB)
Base (no KG, no LLM/GNN)	45.10	58.10	22.0	1.30	169	106
GNN only (KG, no LLM)	48.72	61.10	42.7	3.19	1,273	518
LLM frozen (K=0) + GNN	49.38	61.47	40.0	3.19	1,273	518
LLM LoRA + GNN (K=5)	49.70	62.35	183.4	3.19	14,477	518
LLM LoRA + GNN (K=10)	50.00	62.54	275.4	3.19	24,260	518

Table 5: Cost-accuracy trade-off on MIMIC-IV with batch size 128. Experiments were run on a machine with an Intel Xeon Silver 4214R CPU (24 cores), 256 GiB RAM, and one RTX A6000 GPU (48 GB). For compactness, *Train(s)* denotes training time per epoch, and *Infer(s)* denotes inference time per epoch for the test set.

EHR models even when training data are reduced, highlighting its robustness to data insufficiency.

5 Cost-Accuracy Trade-off

Although jointly training the LLM and GNN yields the best performance, it can be computationally expensive. MEDCO mitigates this cost in two ways: (i) LoRA fine-tunes only a small set of adapter parameters while freezing the backbone LLM, and (ii) node representations are cached and only a selected subset of codes is updated per epoch. At inference time, the LLM is removed from the loop, and prediction uses the cached concept representations together with the GNN and downstream EHR encoder, making deployment much lighter than training. Table 5 summarizes the efficiency-accuracy trade-off on MIMIC-IV. The results show a clear frontier: KG-based variants substantially improve performance over the base model with moderate additional cost, while LoRA-based joint training achieves the highest accuracy at greater training cost. Inference remains lightweight and identical across KG-based variants because the LLM is not used at test time. This suggests that practitioners can prefer frozen-LLM or smaller-K settings when efficiency matters most, and larger-K LoRA training when accuracy is the priority.

6 Related Work

The growing availability of EHR data has driven rapid progress in clinical prediction, evolving from early sequential architectures (Choi et al., 2016a) to attention-based methods (Choi et al., 2016b; Hu et al., 2025), transformer models (Li et al., 2020; Nayebi Kerdabadi et al., 2023), and graph neural networks (Lu et al., 2021b; Xu et al., 2022; Yang et al., 2023; Poulain and Beheshti, 2024).

A line of work strengthens code embeddings by injecting hierarchical ontology structure into models. GRAM (Choi et al., 2017) represents each concept by mixing it with its ancestors, while MMORE (Song et al., 2019) extends this idea with multiple parent representations to better handle mismatch between ontologies and observed EHR patterns. KAME (Ma et al., 2018) incorporates ontology knowledge beyond embedding learning into the prediction pipeline. To better exploit hierarchy, HAP (Zhang et al., 2020) propagates information through top-down and bottom-up attention. Beyond pure hierarchies, G-BERT (Shang et al., 2019) combines ontology graphs with BERT-style encoders, ADORE (Cheong et al., 2023) leverages relational ontologies such as SNOMED to integrate heterogeneous code systems, and KAMPNet (An et al., 2023) adopts graph contrastive learning for EHR representations.

More recent work augments structured EHRs with external signals such as clinical text and web knowledge. GCL (Lu et al., 2021a) and RAM-EHR (Xu et al., 2024) incorporate unstructured text and retrieved medical knowledge, while retrieval-augmented frameworks such as KARE (Jiang et al., 2024) and GraphCare (Jiang et al., 2023) integrate multi-source knowledge for representation learning. LINKO (Nayebi Kerdabadi et al., 2025) leverages multiple ontologies with intra- and inter-ontology propagation via multi-level graph attention.

7 Conclusion

We presented MEDCO, a framework that integrates evidence-grounded KG construction, LLM-based semantic enrichment, and joint LLM-GNN co-learning for medical concept representation. MEDCO builds a heterogeneous diagnosis–medication–procedure KG from EHR-derived evidence, enriches it with LLM-inferred relations and textual semantics, and learns unified concept embeddings through a LoRA-tuned LLM encoder and a heterogeneous GNN. Experiments on MIMIC-III and MIMIC-IV show that MEDCO consistently improves sequential diagnosis prediction across multiple backbones and outperforms strong baselines, especially on rare labels and in limited-data settings. Additional analyses show MEDCO's strong clinical validity, robustness across diverse LLM encoders, and high inference efficiency.

8 Limitations

While MEDCO amortizes LLM computation via a coverage-aware update schedule, joint KG–LLM training can still be resource-intensive, and scaling to larger vocabularies, longer contexts, or larger LLM backbones may increase computational cost.

9 Ethical Considerations

We comply with the ACL Ethics Policy throughout this study. All datasets used are publicly available and contain de-identified patient records, providing strong privacy protections. We do not send any patient-level data to external or public LLM services; all LLM interactions are restricted to general, concept-level information (e.g., medical code descriptions and relation prompts).

10 Acknowledgments

We thank Zhan Ye, MD, and Guangfan Zhang, MD, for their expert clinical validation of our methodology and results. This research was supported by National Science Foundation (Award No. 2531881).

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Yang An, Haocheng Tang, Bo Jin, Yi Xu, and Xiaopeng Wei. 2023. Kampnet: multi-source medical knowledge augmented medication prediction network with multi-level graph contrastive learning. *BMC Medical Informatics and Decision Making*, 23(1):243.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arxiv. *arXiv preprint arXiv:1803.01271*, 10.

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

Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. 2023. Adaptive integration of categorical and multi-relational ontologies with ehr data for medical concept embedding. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–20.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

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

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Google. 2024. Gemma-2-2b model card. https://huggingface.co/google/gemma-2-2b. Hugging Face model card, accessed 2026-04-13.

Aaron Grattafiori and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Healthcare Cost and Utilization Project (HCUP). 2025. Hcup clinical classifications software (ccs) for icd-9-cm. https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed 2025-12-18.

Jinxiang Hu, Mohsen Nayebi Kerdabadi, Xiaohang Mei, Joseph Cappelleri, Richard Barohn, and Zijun Yao. 2025. Recurrent neural networks and attention scores for personalized prediction and interpretation of patient-reported outcomes. *Journal of Biopharmaceutical Statistics*, pages 1–11.

Hugging Face TB. 2025. Smollm2-1.7b model card. https://huggingface.co/HuggingFaceTB/SmolLM2-1.7B. Hugging Face model card, accessed 2026-04-13.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

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.

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.

Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. 2021a. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. *arXiv preprint arXiv:2105.07542*.

Chang Lu, Chandan K Reddy, and Yue Ning. 2021b. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4):2124–2136.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 825–832.

Meta. 2024. Llama-3.2-1b model card. https://huggingface.co/meta-llama/Llama-3.2-1B. Hugging Face model card, accessed 2026-04-13.

Arya Hadizadeh Moghaddam, Mohsen Nayebi Kerdabadi, Mei Liu, and Zijun Yao. 2024. Contrastive learning on medical intents for sequential prescription recommendation. *arXiv preprint arXiv:2408.10259*.

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Bin Liu, Mei Liu, and Zijun Yao. 2023. Contrastive learning of temporal distinctiveness for survival analysis in electronic health records. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1897–1906.

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Dongjie Wang, and Zijun Yao. 2025. Multi-ontology integration with dual-axis propagation for medical concept representation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2201–2211.

Raphael Poulain and Rahmatollah Beheshti. 2024. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations*.

Qwen Team. 2024a. Qwen2.5-0.5b model card. https://huggingface.co/Qwen/Qwen2.5-0.5B. Hugging Face model card, accessed 2026-04-13.

Qwen Team. 2024b. Qwen2.5-1.5b model card. https://huggingface.co/Qwen/Qwen2.5-1.5B. Hugging Face model card, accessed 2026-04-13.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Lihong Song, Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. 2019. Medical concept embedding with multiple ontological representations. In *IJCAI*, volume 19, pages 4613–4619.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

World Health Organization. 2025. International classification of diseases (icd). https://www.who.int/standards/classifications/classification-of-diseases. WHO Family of International Classifications.

World Health Organization Collaborating Centre for Drug Statistics Methodology. 2025. Anatomical therapeutic chemical (atc) classification system. https://www.whocc.no/. WHOCC; first published 1976.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2022. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, pages 259–278. PMLR.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pages 4075–4085.

Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 249–256.

A Appendix

A.1 Existing Medical Ontologies and Coding Resources

This appendix summarizes widely used medical ontologies and coding resources and highlights why they are insufficient, on their own, for learning heterogeneous medical concept representations over diagnosis (*dx*), medication (*rx*), and procedure (*px*) codes. While these resources provide valuable standardization, they often lack high-coverage, actionable cross-domain relations (e.g., *dx-rx*, *dx-px*, *px-rx*) and/or do not reflect cohort- and workflow-specific dependencies observed in real-world EHR data.

ICD and CCS. ICD (World Health Organization, 2025) is one of the most widely adopted clinical coding systems, and CCS (Healthcare Cost and Utilization Project (HCUP), 2025) provides clinically meaningful groupings of ICD codes for downstream analysis. Both are primarily hierarchical: they organize diagnoses and procedures through parent-child structure (or category groupings in CCS), which is effective for standardization and coarse abstraction. However, they do not provide explicit semantics linking diagnoses, medications, and procedures, and therefore offer limited support for modeling cross-domain clinical dependencies needed for predictive tasks.

ATC. The Anatomical Therapeutic Chemical (ATC) classification system (World Health Organization Collaborating Centre for Drug Statistics Methodology, 2025) organizes drug concepts into a hierarchical taxonomy based on therapeutic and chemical characteristics. This structure is useful for grouping medications and supporting within-drug semantic similarity. However, ATC does not explicitly encode care relationships that connect medications to diagnoses or procedures (e.g., treatment vs. contraindication links, peri-procedural medication dependencies), limiting its ability to capture heterogeneous clinical mechanisms across *dx/rx/px* domains.

SNOMED CT. SNOMED CT (Stearns et al., 2001) is a richer clinical ontology that supports description-logic–based concept definition and classification through a large hierarchy and a constrained set of definitional attributes governed by its concept model (MRCM). These attributes can include, for example, finding site, causative agent, method, and active ingredient, enabling structured definitional semantics and fine-grained clinical concept modeling. Nevertheless, SNOMED CT is not designed to systematically encode guideline-oriented, cross-domain care relationships—such as drug-disease treatment/contraindication or procedure-disease indication and workflow dependencies—at the breadth and granularity typically required for clinical prediction from EHRs.

UMLS. The UMLS Metathesaurus (Bodenreider, 2004) provides an integration layer across major biomedical vocabularies (e.g., SNOMED CT, RxNorm, ICD-10-CM, CPT/HCPCS) by aligning synonymous meanings via shared Concept Unique Identifiers (CUIs). This normalization capability is highly valuable for cross-system alignment and harmonization. However, UMLS largely functions as a mapping and aggregation resource rather than a comprehensive repository of actionable clinical relations. Although it contains some curated cross-type links (e.g., drug-disease associations via medication-focused sources such as MED-RT), it does not consistently provide broad, high-coverage procedure-disease indications or procedure-drug semantics at the granularity needed for predictive modeling.

Summary. In summary, commonly used resources provide strong standardization, hierarchical organization, and normalization, but they do not fully capture the heterogeneous, cross-domain dependencies that drive real-world care trajectories in EHRs. Many clinically important *dx-rx*, *px-dx*, and *px-rx* relations remain missing or fragmented, motivating approaches that can induce a more complete relational structure grounded in empirical EHR evidence. Table 6 provides a compact comparison of each resource's structure/goal, scope, cross-domain coverage, and key limitations.

A.2 Statistics of Final Evidence-Supported Code Pairs

Figures 6 and 7 summarize the categorical distribution of the final set of statistically supported

Resource	Structure / Goal	Scope	Cross-Type	Key Limitations
ICD / CCS	Coding taxonomy	dx, px	No	Primarily within-dx or within-px hierarchical categories with only parent-child relations; lacks explicit cross-domain relations.
ATC	Drug classification taxonomy	rx	No	Drug taxonomy; primarily within-rx hierarchical categories with only parent-child relations; lacks explicit cross-domain relations.
SNOMED CT	Definitional concept model	dx, px, rx	Limited	Rich definitional semantics (e.g., finding site, causative agent, method, active ingredient); limited coverage of within-domain relations (dx-due-to-dx); no coverage of cross-domain relations at predictive granularity.
UMLS	Metathesaurus integration / mapping	dx, rx, px	Limited	Strong normalization/mapping; cross-type relations are incomplete/uneven (e.g., limited coverage for px-dx indications and px-rx semantics).

Table 6: Common medical resources (ICD (World Health Organization, 2025), CCS (Healthcare Cost and Utilization Project (HCUP), 2025), ATC (World Health Organization Collaborating Centre for Drug Statistics Methodology, 2025), SNOMED CT (Stearns et al., 2001), UMLS (Bodenreider, 2004)) and their limitations.

code-code pairs extracted from the MIMIC-III and MIMIC-IV cohorts. These pairs represent the output of the evidence extraction and filtering pipeline described in Section 2.3.1, and serve as input to the LLM-based relation inference stage.

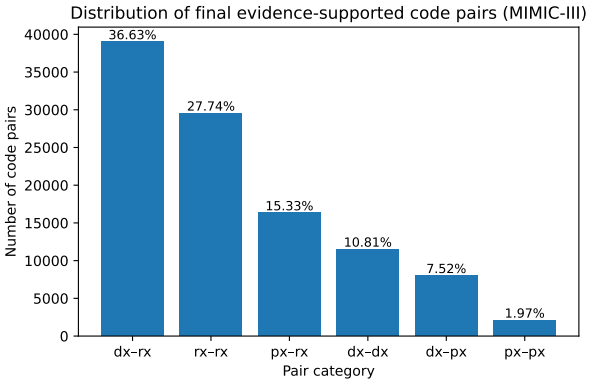

Figure 6: Distribution of final evidence-supported code pairs by category (MIMIC-III). Percentage values are shown on top of each bar.

A.3 Clinical Relation Inventory for Code–Code Edges

We develop a type-constrained relationship inventory comprising 28 clinically meaningful relationships that span all pairs of code categories (diagnosis, medication, procedure). Tables 8-13 enumerate the full relation sets and their definitions. These relations were selected to capture the major semantic patterns observed in clinical care, including etiologic links, disease progression, therapeutic and diagnostic intent, procedural workflow, pharmacologic interactions, safety considerations, and contextual co-occurrence patterns.

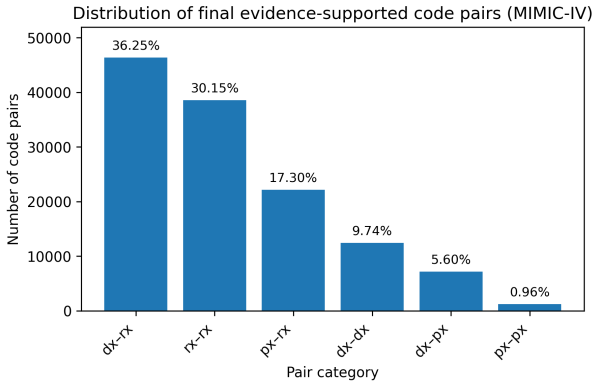

Figure 7: Distribution of final evidence-supported code pairs by category (MIMIC-IV). Percentage values are shown on top of each bar.

This relationship inventory is designed to be (i) clinically grounded, (ii) interpretable for downstream KG reasoning tasks, and (iii) sufficiently expressive to cover the majority of meaningful inter-code relationships in structured EHR data, while maintaining conservative abstention categories (*no_significant_relation*, *cannot_decide*) for ambiguous or low-evidence cases.

A.4 LLM-inferred relation distribution.

Figures 8 and 9 summarize the distribution of relation labels predicted by the LLM over the evidence-filtered candidate code pairs, using the type-constrained relation inventory in our prompt. The resulting histogram is strongly long-tailed (shown on a log-scaled y-axis): a small number of generic relations account for most assignments, while many clinically specific relations occur much less frequently. This pattern is expected in real-

world EHR data, where broad statistical association is common, but high-specificity semantics (e.g., contraindications, substitutions, or rare adverse interactions) require stronger or more distinctive evidence and therefore appear sparsely. Importantly, MIMIC-III and MIMIC-IV exhibit highly similar distributional shapes and rank orderings of the dominant relation types, suggesting that the prompting strategy yields stable and reproducible relation labeling across dataset versions. Any moderate shifts in mid-frequency labels are plausibly attributable to differences in cohort size, coding practices, and temporal coverage rather than prompt instability.

A.5 Clinical Expert Audit of LLM-Inferred Edges

To assess the clinical validity of LLM-inferred relations and quantify hallucination risk, we conducted a targeted expert audit of edges in the constructed knowledge graph. We sampled 50 edges from the final KG by selecting 5 edges from each of the 10 most frequent predicted relation types. The audit was performed under general clinical knowledge rather than patient-specific context.

Each sampled edge was independently reviewed by two frontline clinicians from our university health system. Reviewers rated the correctness of each edge on a 5-point scale: *1 = wrong*, *2 = somewhat wrong*, *3 = not sure / depends on context*, *4 = somewhat correct*, and *5 = correct*. For each edge, we computed the mean of the two clinician ratings, and we then report mean $\pm$ standard deviation across all audited edges and within each relation type.

Overall, the sampled edges received a mean rating of 4.84 ± 0.29, suggesting that the induced relations are largely clinically valid. As shown in Table 7, all audited relation types achieved a mean score of at least 4.40, and several relation types (*post_procedure_medication_for*, *risk_factor_for*, and *treats*) achieved perfect mean scores in the audited sample. The lowest-scoring category was *no_significant_relation* (4.40 ± 0.42), which is unsurprising because abstention-style labels can be more context-sensitive than strongly expressed clinical relations.

These findings are consistent with the safeguards built into our KG induction pipeline. First, candidate code pairs are proposed only when supported by strong EHR association signals after statistical filtering. Second, the LLM is constrained to a predefined, type-specific relation inventory and is allowed to abstain using *cannot_decide* or *no_significant_relation*. Third, we apply post-processing checks to remove invalid labels, directionality inconsistencies, and low-confidence outputs. While this audit is limited in scale, it provides an initial expert validation that the induced KG relations are clinically meaningful and that the evidence-grounded prompting strategy substantially mitigates unsupported edge generation.

Predicted relationship	**Mean**	**Std**
causes_adverse_event	4.70	0.27
co_occurs_with	4.80	0.45
co_prescribed_with	4.80	0.27
complicates	4.90	0.22
interacts_with	4.90	0.22
leads_to	4.90	0.22
no_significant_relation	4.40	0.42
post_procedure_medication_for	5.00	0.00
risk_factor_for	5.00	0.00
treats	5.00	0.00
Overall	**4.84**	**0.29**

Table 7: Clinical expert audit of sampled LLM-inferred edges. We sampled 50 edges by selecting 5 edges from each of the 10 most frequent predicted relation types. Each edge was independently rated by two frontline clinicians on a 1-5 correctness scale, and the reported values are computed from the per-edge mean of the two ratings.

A.6 LLM Prompts for Node-Level Clinical Descriptions

Figure 10 illustrates the high-level structure of the prompt used to generate node-level clinical descriptions. Below, we provide the exact template used for all diagnosis (dx), procedure (px), and medication (rx) codes. Placeholders in angle brackets (e.g., `<CODE_ID>`) are programmatically filled for each node.

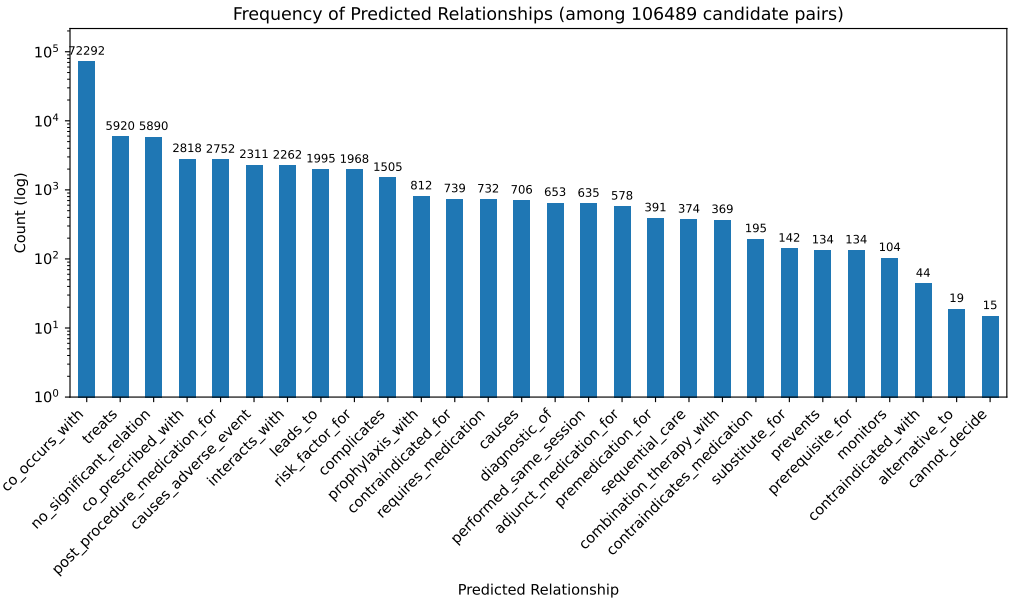

Figure 8: Frequency of LLM-predicted relation labels over evidence-filtered candidate code pairs in MIMIC-III. The y-axis is log-scaled to highlight the long-tailed distribution of relation types.

Table 8: Diagnosis-Diagnosis (DX-DX) relation types. Here, *codeA* and *codeB* are diagnosis codes.

Relation	**Description**
`causes`	codeA is an etiologic cause of codeB (direct causal or pathophysiologic link).
`risk_factor_for`	codeA increases the likelihood of codeB (epidemiologic association; not necessarily causal).
`leads_to`	codeA typically precedes codeB as a downstream condition or stage (temporal progression without requiring strict causality).
`complicates`	codeA occurs as a complication during the course of codeB (arises secondary to codeB or its treatment).
`co_occurs_with`	codeA and codeB are frequently observed together without implied direction or clear causality (contextual association).
`no_significant_relation`	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
`cannot_decide`	Insufficient or conflicting evidence; the model should abstain from assigning a relation.

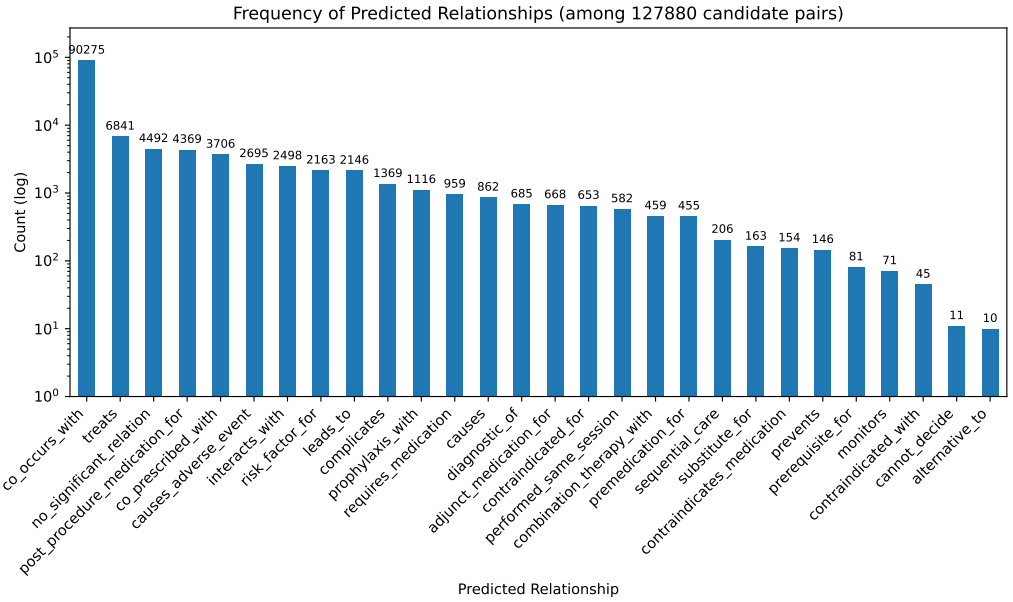

Figure 9: Frequency of LLM-predicted relation labels over evidence-filtered candidate code pairs in MIMIC-IV. The y-axis is log-scaled to highlight the long-tailed distribution of relation types.

Table 9: Medication-Diagnosis (RX-DX) relation types. *codeA* is a medication and *codeB* is a diagnosis.

Relation	Description
treats	codeA treats or manages codeB (therapeutic use).
prevents	codeA reduces risk or recurrence of codeB (prophylaxis).
causes_adverse_event	codeA may induce codeB as an adverse reaction.
contraindicated_for	codeA should be avoided when codeB is present.
co_occurs_with	codeA and codeB often co-occur without a clear causal link.
no_significant_relation	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
cannot_decide	Insufficient or conflicting evidence; the model should abstain from assigning a relation.

Table 10: Procedure-Diagnosis (PX-DX) relation types. *codeA* is a procedure and *codeB* is a diagnosis.

Relation	Description
diagnostic_of	codeA is performed to diagnose, confirm, or rule in/out codeB (diagnostic evaluation).
treats	codeA is a procedure used to treat, correct, or palliate codeB (therapeutic intervention).
monitors	codeA is performed to monitor, follow, or assess disease activity or treatment response for codeB.
contraindicated_for	codeA should be avoided when codeB is present due to safety or unfavorable risk-benefit considerations.
co_occurs_with	codeA and codeB frequently appear together without a clear diagnostic, therapeutic, or causal link (contextual association).
no_significant_relation	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
cannot_decide	Insufficient or conflicting evidence; the model should abstain from assigning a relation.

Table 11: Medication-Medication (RX-RX) relation types. *codeA* and *codeB* are medications.

Relation	Description
co_prescribed_with	codeA and codeB are intentionally prescribed together in practice.
contraindicated_with	Concomitant use of codeA and codeB is contraindicated due to serious safety risk.
interacts_with	codeA and codeB have a clinically meaningful drug-drug interaction (pharmacokinetic and/or pharmacodynamic) that may impact safety or efficacy.
substitute_for	codeA is commonly used as a therapeutic alternative to codeB for similar indications (typically not co-prescribed).
combination_therapy_with	codeA and codeB are used together as an established combination therapy.
co_occurs_with	codeA and codeB frequently appear together in data without a known therapeutic relationship or interaction.
no_significant_relation	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
cannot_decide	Insufficient or conflicting evidence; the model should abstain from assigning a relation.

Table 12: Procedure-Procedure (PX-PX) relation types. *codeA* and *codeB* are procedures.

Relation	Description
sequential_care	codeA is typically followed by codeB as the next procedural step (ordered workflow; not necessarily causal).
prerequisite_for	codeA is commonly required or preparatory for performing codeB (e.g., access, imaging guidance, setup).
alternative_to	codeA and codeB are procedural alternatives for a similar clinical purpose (mutually substitutable options).
performed_same_session	codeA and codeB are commonly performed during the same procedural session or time block within an episode (intentional bundling).
co_occurs_with	codeA and codeB occur within the same episode without implied order or intentional bundling (contextual association).
no_significant_relation	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
cannot_decide	Insufficient evidence: no strong clinical prior and statistical signals are low-support, unstable, or conflicting; the model should abstain from assigning a relation.

Table 13: Procedure-Medication (PX-RX) relation types. *codeA* is a procedure and *codeB* is a medication.

Relation	Description
`requires_medication`	codeA routinely requires administration of codeB as part of the procedure (e.g., sedation, anesthesia, anticoagulation).
`premedication_for`	codeB is typically given before codeA to enable or optimize the procedure (e.g., anxiolysis, antibiotic prophylaxis).
`post_procedure_medication_for`	codeB is commonly given after codeA for recovery, prophylaxis, or symptom control (e.g., analgesia, anticoagulation).
`prophylaxis_with`	codeB is used to prevent complications associated with codeA (e.g., peri-procedural antibiotics, DVT prophylaxis).
`adjunct_medication_for`	codeB is an adjunct given with codeA to improve efficacy, safety, or tolerability (non-essential but commonly used).
`contraindicates_medication`	When codeA is planned or present, codeB should be avoided due to safety or risk-benefit concerns.
`co_occurs_with`	codeA and codeB are observed together in data without a clear procedural rationale or causal link (contextual association).
`no_significant_relation`	Clinical prior and available data indicate no clinically meaningful association between codeA and codeB.
`cannot_decide`	Insufficient or conflicting evidence; the model should abstain from assigning a relation.

LLM Prompt for Node Description

You are a medical reasoning assistant. Return **JSON only. Write **one dense paragraph (~250 words)** giving most important/accurate clinical context about a medical concept (code) for embedding. Use established medical knowledge.**

Code
- id: <CODE_ID> - label: <CODE_LABEL> - type: <CODE_TYPE> - marginal P(code):<P(CODE)>

Focus (ideal information to provide, if exists)
{guidance} for type: <CODE_TYPE>

Constraints
- Write one cohesive paragraph (~250 words); prefer well-established clinical knowledge.
- Widely accepted, class-level clinical facts appropriate to the label/type. If label is broad/ambiguous, provide a generic overview and state the ambiguity; do not guess
- Avoid numeric dosing, exact thresholds, or hospital-specific policies.
- NO lists, headings, bullets, numbered steps, citations. NO markdown—return pure JSON.

Forbidden content
- Specific product names, URLs, dosing, numeric lab thresholds, hospital policies, billing rules.
- ATC/CCS subdivisions not supplied in the prompt.
- Speculative claims, local statistics, or unverified niche practices.

Output (JSON only)
{
"code_id": "<code_id>",
"description": "<one dense paragraph ~250 words; no lists/URLs/doses/thresholds>"
}

Type-specific Guidance

guidance_dx = {
Define the condition succinctly; typical clinical presentation and key red flags; common etiologies and high-level pathophysiology; major risk factors and prevalent comorbidities; core diagnostic approach (history/exam cues and hallmark labs/imaging—no numeric cutoffs); severity/staging concepts; brief management overview (first-line modalities, supportive care, when to escalate/consult); complications to watch for and longitudinal outcomes; closely related or easily confused diagnoses and one cue to distinguish them. Keep to class-level statements.
}
guidance_px = {
State the procedure's purpose and primary indications; prerequisites/contraindications and perioperative prep; a high-level sense of how it is performed (steps/components, anesthesia—no device brands); immediate and delayed risks/complications and how they are mitigated; typical monitoring/aftercare; common alternatives or non-procedural options and when preferred; how this fits in the care pathway. Keep to class-level statements.
}
guidance_rx = {
Identify the pharmacologic class and plain-language mechanism; core indications and common clinically meaningful off-label uses (no marketing claims); major contraindications and black-box concerns; high-signal interaction patterns (drug–drug, drug–condition); adverse effects emphasizing serious/frequent events and monitoring; practical clinical pearls (onset, adherence, organ-adjustment considerations) without doses or lab thresholds; briefly relate to nearby classes. Keep to class-level statements; do not list specific products.
}

Figure 10: Prompt template used to generate LLM-based node-level clinical descriptions.