Learning What to Ignore: Mitigating Negative Transfer in Medical Knowledge Fusion via Clinical Task-Adaptive Selection

Xinyan Deng[1], **Shoubin Dong**[1,*], **Xiaorou Zheng**[1]
[1]School of Computer Science and Engineering, South China University of Technology
csstu_dengxy@mail.scut.edu.cn, {sbdong, zhengxrsc}@scut.edu.cn

Abstract

Integrating external medical knowledge into longitudinal electronic health record modeling is a prevailing paradigm to mitigate clinical data sparsity. However, existing approaches face a reliability-timeliness dilemma, struggling to balance the structural authority of static ontologies with the reasoning flexibility of large language models. Furthermore, most frameworks overlook the risk of relative negative transfer, where indiscriminately fusing task-irrelevant knowledge can introduce noise or even cause conflicts that weakens patient-specific signals. In this paper, we propose TrustKE, a Trustworthy Knowledge Enhancement framework. First, we construct a dual-layer knowledge graph that anchors dynamic, evidence-based chain-of-thought reasoning from medical literature within the stable structure of medical knowledge graph. Second, we introduce a task-adaptive knowledge selection mechanism that dynamically optimizes the graph, retaining only task-specific signals. Extensive experiments on MIMIC-III and MIMIC-IV across four clinical tasks show that TrustKE outperforms state-of-the-art baselines. Our analysis confirms that TrustKE effectively mitigates negative transfer while offering transparent reasoning for clinical decision-making.

1 Introduction

Deep learning has revolutionized the modeling of longitudinal Electronic Health Records (EHRs) for critical clinical tasks, ranging from mortality prediction to drug recommendation (Rasmy et al., 2021; Chen et al., 2023; Li and Zhou, 2025). However, real-world clinical data often exhibits significant sparsity and long-tail distributions (Zhao et al., 2024; Li et al., 2024a), yielding suboptimal generalization in few-shot scenarios (Zhao et al., 2025a,b). Consequently, integrating external medical knowledge to augment patient representations has become a prevailing paradigm.

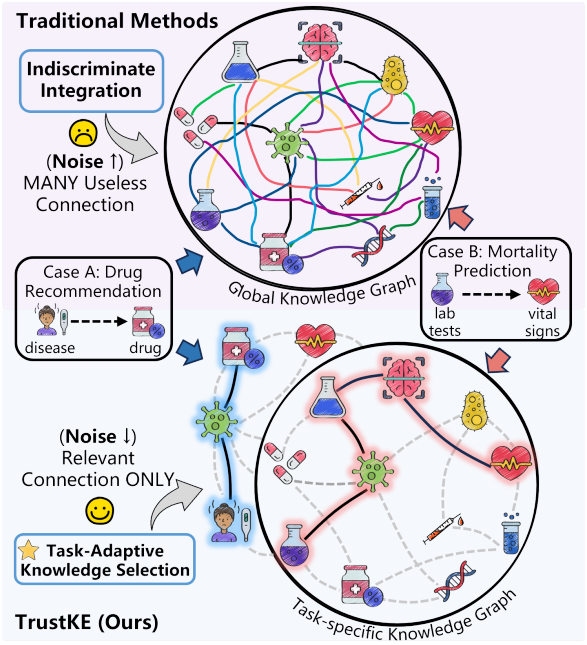

Figure 1: Unlike multi-source knowledge fusion may accumulate noise and cause conflicts, TrustKE employs Task-Adaptive Knowledge Selection. It dynamically optimizes the graph structure to retain only task-specific signals (e.g., Cases A and B), effectively filtering noise.

Despite progress, current knowledge fusion approaches face a **reliability-timeliness dilemma**. Traditional methods rely on structured Knowledge Graphs (KGs) like UMLS (Bodenreider, 2004), which provide authoritative facts but remain static, failing to capture emerging medical nuances or complex patient pathways (Lu et al., 2022; Kang et al., 2024). Conversely, Large Language Models (LLMs) offer comprehensive reasoning but often suffer from hallucinations or retrieve unstructured noise that lacks grounding in the patient's specific physiological state (Thirunavukarasu et al., 2023; Liu et al., 2025; Lu et al., 2025). Relying exclusively on either source limits system robustness.

Critically, a fundamental challenge remains overlooked: **knowledge enhancement does not inherently guarantee information gain**. Most existing

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23295–23309

frameworks indiscriminately fuse a fixed, global KG or retrieved context into patient representations. However, as illustrated in Figure 1, distinct clinical tasks possess unique knowledge boundaries. For instance, mortality prediction relies heavily on physiological signals (e.g., lab tests), whereas drug recommendation depends on qualitative drug co-occurrence patterns. Indiscriminately incorporating irrelevant knowledge introduces spurious correlations, leading to relative **negative transfer** (Yang et al., 2023; Li et al., 2024b), where task-irrelevant knowledge causes signal dilution, weakening the patient-specific history compared to an optimally pruned graph.

To address these challenges, we propose **TrustKE**, a framework balancing credibility and adaptivity. First, we construct a dual-layer knowledge graph by embedding dynamic, evidence-based chain-of-thought (CoT) reasoning from LLMs into a stable UMLS structure to resolve the reliability-timeliness dilemma. Second, we design a task-adaptive knowledge selection mechanism that dynamically refines the graph structure based on the task, retaining only relevant connections to mitigate negative transfer. Finally, a patient-guided fusion module aligns the refined knowledge with patient history.

Our contributions are threefold: (1) We identify the negative transfer phenomenon caused by task-irrelevant knowledge fusion in clinical modeling. (2) We propose a dual-layer knowledge construction method that merges the structural authority of ontologies with the flexible reasoning of LLMs. (3) We develop a task-adaptive knowledge selection mechanism, achieving state-of-the-art (SOTA) performance on MIMIC-III/IV across four distinct clinical tasks while offering interpretability.

2 Related Works

Longitudinal EHR Modeling. Deep learning has become the bedrock for modeling time-series data in healthcare (Luo et al., 2020; Wu et al., 2022; Ren et al., 2022). Standard approaches employ specialized attention mechanisms or recurrent units to capture temporal dynamics from patient history. For instance, some methods (Ma et al., 2020a,b) introduce multi-scale calibration and cross-head attention to capture long-term biomarker variations for risk prediction, while Kim et al. (2024) utilizes a selective attention mechanism to filter out unimportant visits for clinical prediction. However, these methods operate under a closed-world assumption, relying exclusively on statistical correlations within the training corpus. They lack the ability to consult external medical knowledge, leading to poor generalization for patients with rare conditions or sparse records. **To address this, TrustKE augments patient representations with a dual-layer KG to handle data sparsity.**

Knowledge-Enhanced Clinical Representation. To mitigate the limitations of data-driven models, integrating KGs has become a standard paradigm to enrich patient representations. Early works (Shang et al., 2019; Lu et al., 2021) pre-train on medical ontologies to refine code embeddings, while recent structure-aware models (Lu et al., 2022; Yang et al., 2023; Zou et al., 2024; Zheng et al., 2025) construct dynamic clinical graphs to capture evolving patient health status. Despite their contributions, these approaches often suffer from negative transfer. Most methods employ global fusion strategies that indiscriminately input generic knowledge. **In contrast, TrustKE employs a task-adaptive knowledge selection module, dynamically optimizing the graph to retain task-relevant connections (e.g., prioritizing lab tests for mortality prediction).**

LLM-Driven Clinical Reasoning. The frontier of clinical AI is shifting towards interpretability and advanced reasoning capabilities. Recent studies (Zhao et al., 2025a; Li et al., 2024b; Kang et al., 2024) have explored diverse strategies to enhance robustness, particularly by introducing discrete clues and causal interventions to handle long-tail distributions. While LLM agents like Zhu et al. (2024) leverage Retrieval-Augmented Generation (RAG) to enhance their knowledge base, others like Wang et al. (2025) use multi-agent collaboration to simulate clinical decision-making processes. While promising, these methods face the reliability-timeliness dilemma. Purely agent-based frameworks are computationally expensive and prone to hallucinations, whereas causal models often rely on opaque latent embeddings. **TrustKE addresses these challenges by anchoring dynamic evidence-based CoT reasoning within a authoritative base graph, offering a framework that is both clinically robust and transparent.**

3 Methodology

We propose TrustKE, a framework integrating interpretable medical reasoning into clinical sequence

modeling (Figure 2). TrustKE comprises three modules: (1) **Dual-Layer Knowledge Construction** builds a global graph by augmenting EHR with ontologies and evidence-based CoT reasoning; (2) **Task-Adaptive Knowledge Selection** mitigates negative transfer by dynamically refining the graph structure guided by downstream tasks; and (3) **Patient-Guided Heterogeneous Fusion** queries subgraph based on patient context to generate the final representation for prediction.

3.1 Problem Formulation

We formalize the task of knowledge-enhanced clinical prediction over longitudinal EHRs.

Patient Representation. Let $\mathcal{C} = \mathcal{C}_d \cup \mathcal{C}_p \cup \mathcal{C}_m \cup \mathcal{C}_s \cup \mathcal{C}_l$ denote the heterogeneous medical vocabulary, representing diagnoses, procedures, medications, symptoms, and lab tests. A patient is represented as a temporal sequence of visits $\mathcal{P} = \{v_1, \ldots, v_T\}$. Each visit $v_t \subseteq \mathcal{C}$ is a multi-hot set of medical codes recorded at time step t.

Knowledge Graph Definition. An external KG $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ models the latent relationships between medical concepts. Here, $\mathcal{V} \supseteq \mathcal{C}$ denotes the entity set, and $\mathcal{E}$ represents the relation edges.

Prediction Tasks. Given patient $\mathcal{P}$ and graph $\mathcal{G}$, the goal is to learn a mapping function $f(\mathcal{P}, \mathcal{G}) \rightarrow \mathcal{Y}$. We generalize across two task categories: **1) Binary Risk Prediction** (e.g., Mortality, Readmission): The target is a scalar $y \in \{0, 1\}$ indicating an adverse event. **2) Multi-label Prediction** (e.g., Disease Prediction, Drug Recommendation): Let $\mathcal{Y}_{task} \subset \mathcal{C}$ denote the task-specific label space. The target is a binary vector $\mathbf{y} \in \{0, 1\}^{|\mathcal{Y}_{task}|}$, where each dimension indicates the presence of a label.

3.2 Dual-Layer Knowledge Construction

To resolve the reliability-timeliness dilemma, where static ontologies lack coverage and generative models hallucinate, we construct a global KG $\mathcal{G}$ via a dual-layer strategy. This anchors dynamic, evidence-based reasoning within a stable base.

Layer 1: Base Graph via Ontology Alignment. We first map heterogeneous EHR codes (ICD, NDC) to the UMLS to identify the seed entity set $\mathcal{V}_{seed}$. To mitigate knowledge explosion in general ontologies, we construct a concise active subgraph. Let $\mathcal{T}_{umls}$ denote the set of all UMLS triples. We instantiate the base edge set $\mathcal{E}_{base}$ as:

$$\mathcal{E}_{base} = \{(u, r, v) \in \mathcal{T}_{umls} \mid u, v \in \mathcal{V}_{seed}\} \quad (1)$$

where $\mathcal{V}_{seed}$ denotes high-frequency medical concepts essential for connectivity (filtering thresholds detailed in **Appendix B.1**). This constraint ensures a dense, clinically relevant structural base.

Layer 2: Enrichment via RAG-Based Reasoning. Static base graphs often lack explicit causal chains found in recent literature. We enrich the graph using a rigorous RAG, expanding it dynamically: $\mathcal{G} = (\mathcal{V}_{top} \cup \mathcal{V}_{rag}, \mathcal{E}_{base} \cup \mathcal{E}_{rag})$, where $\mathcal{V}_{top}$ are the top K most frequent $\mathcal{V}_{seed}$.

1) Adaptive Evidence Retrieval. Raw clinical terms are first normalized via a hybrid rule LLM pipeline (details in **Appendix B.2**). To ensure evidence quality, we employ a cascading retrieval policy. The system initially queries for systematic reviews; if insufficient, it adaptively degrades to standard articles. This ensures the graph is built primarily on authoritative consensus.

2) Evidence-Based CoT Extraction. To guarantee transparency and reduce hallucinations, we propose an evidence-based CoT mechanism. Instead of directly predicting triples, we prompt the LLM to follow a strict reasoning path: (1) Locate sentences mentioning the target entity; (2) Infer the specific clinical relation (e.g., Treats vs. Causes); and (3) Quote the exact text span as supporting evidence. To ensure structural validity, we apply a self-loop filter ($u \neq v$) to remove trivial self-loops. Formally, the extracted relation set is defined as:

$$\begin{aligned}\mathcal{E}_{rag} = \{(u, r, v, e_{uv}) \mid \text{LLM}(u, D_u) \\ \rightarrow (u, r, v, e_{uv}) \wedge u \neq v\}\end{aligned} \quad (2)$$

where e_{uv} represents the textual evidence grounded in the document D_u. This evidence-aware tuple (u, r, v, e_{uv}) enables the model to justify its reasoning. Implementation details and comparisons with standard extraction are in **Appendix B.3**.

During the RAG process, if the LLM identifies valid medical concepts not in our initial UMLS mapping, we generate a unique identifier TXT_{Hash} and initialize their embeddings with Bio-ClinicalBERT(Alsentzer et al., 2019). This enables dynamic graph expansion beyond the structured ontology's closed vocabulary.

3.3 Task-Adaptive Knowledge Selection

The base graph from Section 3.2 anchors reasoning but remains task-agnostic. Direct utilization of such a static structure often introduces irrelevant noise, leading to negative transfer. To resolve this,

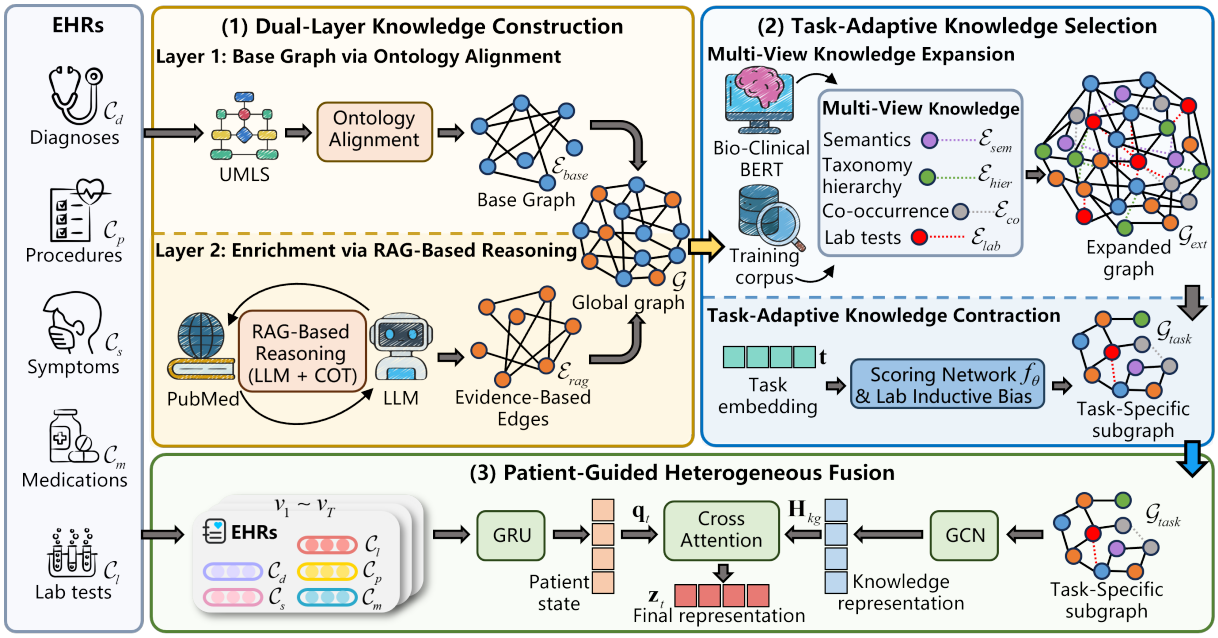

Figure 2: Overall architecture of our proposed TrustKE.

we propose a select-from-expanded graph strategy that operates in two phases: (1) expanding the base graph into a multi-view expanded graph to maximize coverage, and (2) applying task-adaptive knowledge contraction to extract the optimal task-specific subgraph.

3.3.1 Multi-View Knowledge Expansion

We extend the base graph into a comprehensive heterogeneous network $\mathcal{G}_{ext} = (\mathcal{V}, \mathcal{E}_{ext})$. We integrate three distinct views to capture implicit, empirical, and physiological associations:

1) Implicit Semantic View. Ontologies often miss latent associations between semantically similar concepts (e.g., renal insufficiency and kidney failure). We recover these implicit links by encoding entity descriptions using Bio-ClinicalBERT to obtain context-aware embeddings $\mathbf{H}_{bert}$. We establish semantic edges $\mathcal{E}_{sem}$ between pairs exceeding a high-confidence similarity threshold $\tau = 0.9$, creating a dense layer for message passing between structurally distant nodes.

$$\mathcal{E}_{sem} = \{(u, v) \mid \cos(\mathbf{h}_u, \mathbf{h}_v) > \tau\} \quad (3)$$

2) Empirical Structural View. To incorporate statistical priors, we construct two graph types from the training corpus: a taxonomy hierarchy ($\mathcal{E}_{hier}$) built by truncating medical codes to parent categories to capture vertical generalization, and a co-occurrence graph ($\mathcal{E}_{co}$) retaining edges where the conditional probability $P(v|u) > 0.05$ to capture horizontal comorbidities.

3) Lab-Aware Heterogeneity. A critical design in our framework is the explicit segregation of laboratory tests. Unlike qualitative diagnostic codes, lab results (e.g., abnormal lactate) act as critical precursors to adverse events. We assign lab-related connections to a distinct relation type $\mathcal{E}_{lab}$, separate from standard diagnostic correlations $\mathcal{E}_{co}$. This heterogeneity allows the model to learn specialized attention weights for vital signals.

3.3.2 Task-Adaptive Knowledge Contraction

The expanded graph (1.3 million edges in MIMIC-III) introduces significant noise and redundancy. To extract task-relevant knowledge, we propose task-adaptive knowledge contraction, which dynamically compresses the graph based on task-specific embedding $\mathbf{t}$ through node-aware semantics.

Task-Conditioned Edge Scoring. Unlike global relation attention, our knowledge contraction evaluates the relevance of each specific edge instance. We design a scoring network f_θ that evaluates the relevance of each specific edge instance conditioned on both node-aware semantics and the task embedding $\mathbf{t}$. The relevance score $s_{uv} \in [0, 1]$ is computed as:

$$\mathbf{x}_{uv} = [\mathbf{W}_p\mathbf{h}_u \parallel \mathbf{W}_p\mathbf{h}_v \parallel \mathbf{r}_{rel} \parallel \mathbf{t}] \quad (4)$$

$$s_{uv} = \sigma(\text{MLP}(\mathbf{x}_{uv})) \quad (5)$$

where $\parallel$ denotes concatenation, $\mathbf{r}_{rel}$ is the learnable relation embedding, and $\mathbf{W}_p$ projects BERT

features into the task space. This mechanism allows the model to capture fine-grained dependencies (e.g., retaining a specific interaction only when relevant to the target outcome).

Hybrid Fusion with Lab Inductive Bias. To balance data-driven learning with empirical priors, we fuse the learned score s_{uv} with the static structural weight w_{base} (derived from similarity or probability). Crucially, to prevent sparse but vital laboratory signals from being overwhelmed by frequent edges during early training, we introduce a lab-prior mechanism. We enforce a minimum activation floor $\epsilon = 0.5$ based on experience for the lab relation set $\mathcal{R}_{lab}$:

$$\alpha_{uv} = \begin{cases} \max(s_{uv}, w_{base}, \epsilon) & \text{if } r \in \mathcal{R}_{lab} \\ \max(s_{uv}, w_{base}) & \text{otherwise} \end{cases} \quad (6)$$

This strategy introduces a domain-specific inductive bias, architecturally separating quantitative physiological signals from qualitative semantic noise to ensure physiological indicators are preserved. During inference, we apply $\alpha > 0.1$ to prune low-weight edges, yielding a compact, interpretable, and task-specific subgraph.

3.4 Patient-Guided Heterogeneous Fusion

This module aligns the task-adaptive knowledge $\mathcal{G}_{task}$ with the patient's evolving status using a dual-stream architecture.

3.4.1 Dual-Stream Representation Learning

1) Knowledge Stream (Graph Encoder). We employ a relation-weighted GCN over the pruned adjacency matrix $\mathbf{A}_{task}$ (from Eq. 6) to update BioBERT-initialized node features. Crucially, because the learnable relation embedding r_{rel} directly dictates the final scalar weights in A_{task} (as defined in Eq. 4 and 5), the subsequent message passing is strictly modulated by the relation type, effectively preserving graph heterogeneity. This yields a knowledge-enriched representation matrix $\mathbf{H}_{kg} \in \mathbb{R}^{N \times d}$, capturing task-specific structures.

2) Patient Stream (Sequence Encoder). We use separate GRUs to model heterogeneous modalities (diagnoses, procedures, medications, labs) in parallel. Let $\mathbf{h}_t^m$ be the hidden state for modality m. To synthesize a unified patient state $\mathbf{q}_t$, we employ a View-Interaction Attention to weigh clinical views dynamically:

$$\alpha_t^m = \text{softmax}(\mathbf{w}_v^T \tanh(\mathbf{W}_a[\mathbf{h}_t^{diag} \parallel \cdots \parallel \mathbf{h}_t^{lab}])) \quad (7)$$

$$\mathbf{q}_t = \sum_m \alpha_t^m \mathbf{h}_t^m \quad (8)$$

This ensures the model focuses on the most informative modality (e.g., prioritizing symptoms for early diagnosis) at each step.

3.4.2 Semantic Alignment and Fusion

We first align both streams via a lightweight residual projector that aligns both streams into a shared semantic manifold: $\tilde{\mathbf{q}}_t = \text{MLP}(\mathbf{q}_t) + \mathbf{q}_t$ and $\tilde{\mathbf{H}}_{kg} = \text{MLP}(\mathbf{H}_{kg}) + \mathbf{H}_{kg}$.

We then implement a patient-query mechanism, treating the patient state $\tilde{\mathbf{q}}_t$ as the query and knowledge nodes $\tilde{\mathbf{H}}_{kg}$ as keys/values. This allows dynamically retrieve relevant knowledge from the global graph:

$$\mathbf{z}_t = \text{LayerNorm}\left(\tilde{\mathbf{q}}_t + \text{Attn}(\tilde{\mathbf{q}}_t, \tilde{\mathbf{H}}_{kg})\right) \quad (9)$$

The retrieved context $\mathbf{z}_t$ serves as the patient's final representation.

3.5 Prediction and Optimization

The final representation $\mathbf{z}_t$ is passed to task-specific decoders optimized end-to-end.

Task-Specific Decoding. For binary risk prediction, we employ a MLP to capture high-order interactions, outputting a scalar probability $\hat{y}_t = \sigma(\text{MLP}(\mathbf{z}_t))$. For multi-label prediction, we utilize a linear projection head to map representations to the label space $\mathbb{R}^{|\mathcal{Y}_{task}|}$.

Optimization Objective. Binary tasks are optimized via Binary Cross-Entropy (BCE). For multi-label tasks, the model is optimized using a combination of BCE and margin ranking loss to ensure valid labels ($i \in \mathcal{Y}^+$) consistently score higher than invalid ones ($j \in \mathcal{Y}^-$):

$$\mathcal{L}_{rank} = \sum_{i \in \mathcal{Y}^+, j \in \mathcal{Y}^-} \frac{\max(0, 1 - (\hat{y}_i - \hat{y}_j))}{|\mathcal{Y}_{task}|} \quad (10)$$

$$\mathcal{L} = \gamma \mathcal{L}_{bce} + (1 - \gamma)\mathcal{L}_{rank} \quad (11)$$

where $\gamma = 0.97$ is a hyperparameter balancing $\mathcal{L}_{bce}$ and $\mathcal{L}_{rank}$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TrustKE on two real-world datasets, MIMIC-III and MIMIC-IV (Johnson et al., 2016, 2018). Following standard preprocessing pipelines (Yang et al., 2023; Sun et al., 2022), we

Model	MIMIC-III Mortality			MIMIC-III Readmission			MIMIC-IV Mortality			MIMIC-IV Readmission		
	AUPRC	AUROC	F1	AUPRC	AUROC	F1	AUPRC	AUROC	F1	AUPRC	AUROC	F1
AdaCare	0.4493	0.8796	0.3576	0.2845	0.5896	0.3052	0.5336	0.9471	0.4810	0.4627	0.6159	0.3502
ConCare	0.4554	0.8635	0.4471	0.2910	0.6014	0.3327	0.5319	0.9458	0.4177	0.4328	0.6073	0.3325
GRASP	0.4155	0.8772	0.3378	0.2849	0.5865	0.3028	0.5150	0.9484	0.5204	0.4417	0.6105	0.3455
Chet	0.2524	0.5232	0.1604	0.1802	0.4946	0.2069	0.4833	0.6127	0.3917	0.1940	0.5362	0.2967
VITA	0.3566	0.8005	0.3958	0.2151	0.4921	0.2194	0.4629	0.7345	0.4759	0.3516	0.6047	0.2627
EMERGE	0.5610	0.8356	0.4713	0.3474	0.6262	0.3926	0.6182	0.9542	0.6297	0.4967	0.6438	0.5137
ColaCare	0.4674	0.8893	0.4323	0.3436	0.6045	0.3586	0.5247	0.8439	0.4583	0.4743	0.6247	0.3708
UDC	0.4992	0.8906	0.5804	0.2479	0.5973	0.2585	0.5985	0.9612	0.5842	0.4715	0.6256	0.3619
TrustKE	**0.6502**	**0.9160**	**0.6271**	**0.3640**	**0.6760**	**0.4244**	**0.6230**	**0.9681**	**0.6306**	**0.5192**	**0.6716**	**0.5423**

Table 1: Performance comparison on mortality and readmission. We report the average performance for each model after 10 runs. The best results are highlighted in **bold**, and the second best are underlined.

Model	MIMIC-III Drug Rec			MIMIC-IV Drug Rec		
	AUPRC	F1	Jaccard	AUPRC	F1	Jaccard
MoleRec	0.7748	0.6841	0.5301	0.7002	0.6180	0.4625
StratMed	0.7779	0.6861	0.5321	0.7023	0.6122	0.4560
DAI-Net	0.7717	0.6798	0.5253	0.7035	0.6114	0.4552
VITA	0.7635	0.6785	0.5261	0.6989	0.6200	0.4816
RAREMed	0.7811	0.6853	0.5360	0.7295	0.6182	0.4751
CausalMed	0.7831	0.6898	0.5367	0.7236	0.6258	0.4757
TrustKE	**0.7888**	**0.6932**	**0.5404**	**0.7454**	**0.6506**	**0.4982**
Model	**MIMIC-III Disease Pred**			**MIMIC-IV Disease Pred**		
	w-F1	re@10	re@20	w-F1	re@10	re@20
Timeline	20.46	25.75	34.83	25.26	29.00	37.13
G-BERT	19.88	25.86	35.31	24.49	27.16	35.86
HiTANet	21.15	26.02	35.97	24.92	27.45	36.37
CGL	21.92	26.64	36.72	25.41	28.52	37.15
Chet	22.63	**28.64**	**37.67**	25.10	**30.28**	**38.69**
UDC	19.12	24.65	33.69	23.08	26.72	34.52
TrustKE	**22.79**	26.94	37.34	**27.93**	28.75	37.73

Table 2: Performance comparison on drug recommendation and disease prediction. We report the average performance for each model after 10 runs. The best results are highlighted in **bold**, and the second best are underlined.

extract longitudinal visits and identify symptoms. For binary tasks, we predict mortality (in-hospital death) and readmission (within 30 days). For multi-label tasks, we perform drug recommendation and disease prediction. Dataset statistics are detailed in **Appendix A.1**.

Baselines. Our baselines include AdaCare (Ma et al., 2020a), ConCare (Ma et al., 2020b), GRASP (Zheng et al., 2025), and Chet (Lu et al., 2022). For multi-label prediction tasks (medication and disease), we additionally consider specialized frameworks including Timeline (Bai et al., 2018), G-BERT (Shang et al., 2019), HiTANet (Luo et al., 2020), CGL (Lu et al., 2021), VITA (Kim et al., 2024), MoleRec (Yang et al., 2023), RAREMed (Zhao et al., 2024), StratMed (Li et al., 2024a), DAI-Net (Zou et al., 2024), and CausalMed (Li et al., 2024b). Furthermore, we benchmark against recent language model-based approaches, including LLM and multi-agent approaches, such as, EMERGE (Zhu et al., 2024), UDC (Zhao et al., 2025a), and ColaCare (Wang et al., 2025). Implementation is detailed in **Appendix A.2**.

Evaluation Metrics. We adopt standard metrics aligned with specific task requirements. For binary risk prediction (mortality/readmission), we report Area Under the ROC Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and F1-score. For drug recommendation, we utilize Jaccard Similarity (Jaccard), F1-score, and AUPRC. For disease prediction, following previous works (Lu et al., 2022; Zhao et al., 2025a), we employ Weighted-F1 (w-F1) and Recall@k ($k = 10, 20$) to evaluate top-k performance. Details in **Appendix A.3**.

Implementation Details. TrustKE implement in Python 3.10.13 and PyTorch 2.4.0. The text encoder is initialized with Bio-ClinicalBERT, with Qwen2.5-7B-Instruct (Yang et al., 2024) parameters frozen for CoT reasoning and relation extraction. The graph module uses a 2-layer GNN with a hidden dimension of 128. We set the retrieval threshold $K = 1000$ based on sensitivity analysis. Training uses the AdamW optimizer with learning rates of 2e-5 for the text encoder and 2e-4 for other modules. The batch size is 64, and we train for 30 epochs. All runs are conducted on a single NVIDIA V100 GPU (32GB VRAM) on a server with 256GB system RAM, running CUDA 12.2.

4.2 Experimental Results

Table 1 details mortality and readmission prediction performance. TrustKE consistently consistently leads on both datasets. Notably, on the MIMIC-III mortality task, it surpasses ColaCare (AUPRC 0.4674) by about 19%. This indicates that while multi-agent frameworks offer flexible inference capabilities, they tend to be unstable in sparse clinical settings. In contrast, TrustKE's dual-layer knowledge construction provides a robust struc-

Variant	Mortality		Readmission		Drug Rec		Disease Pred	
	AUPRC	Δ	AUPRC	Δ	Jaccard	Δ	w-F1	Δ
TrustKE (Full)	**0.6502**	-	**0.3640**	-	**0.5404**	-	**22.79**	-
1) Knowledge Source								
EHR Only	0.6004	-4.98%	0.3245	-3.95%	0.5170	-2.34%	20.59	-2.20
UMLS Only	0.6062	-4.40%	0.3349	-2.91%	0.5191	-2.13%	20.70	-2.09
2) Dual-Layer Structure								
w/o Evidence-Based CoT	0.6313	-1.89%	0.3631	-0.09%	0.5138	-2.66%	20.54	-2.25
3) Task-Adaptive Selection								
w/o Knowledge Expansion	0.6124	-3.78%	0.3325	-3.15%	0.5173	-2.31%	20.47	-2.32
w/o Knowledge Contraction	0.6295	-2.07%	0.3460	-1.80%	0.5197	-2.07%	22.43	-0.36
w/o Task Adaptive	0.6392	-1.10%	0.3457	-1.83%	0.5221	-1.83%	21.83	-0.96
4) Fusion Strategy								
w/o Cross-Attention	0.6447	-0.55%	0.3491	-1.49%	0.5162	-2.42%	21.24	-1.55

Table 3: Ablation study on MIMIC-III. We report AUPRC for binary tasks, Jaccard for drug recommendation, and w-F1 for disease prediction. Δ denotes the performance drop compared to the full TrustKE model.

tural foundation, ensuring reliable generalization. TrustKE also significantly outperforms the RAG-based EMERGE (e.g., +9% AUPRC on MIMIC-III mortality). While EMERGE retrieves clinical evidence, it suffers from task-irrelevant knowledge. In contrast, TrustKE's task-adaptive selection effectively prunes irrelevant signals, demonstrating that learning what to ignore is as critical as what to expand in complex clinical predictions.

Table 2 presents the performance of the multi-label task. TrustKE particularly outperforming the baseline CausalMed on the MIMIC-IV dataset, with F1 improvement of approximately 2.5%. While CausalMed relies on a static causal graph, our task-adaptive mechanism effectively mitigates the negative migration effect by pruning irrelevant knowledge. In disease prediction, TrustKE performs robustly. We observe that Chet achieves a higher Recall@k, likely due to its disease-level temporal learning with transition functions specifically designed for co-occurrence ranking. However, TrustKE, as a unified framework, still significantly outperforms recent language model-based methods such as UDC (a 3.2% improvement on MIMIC-IV re@20), offering a better balance between structural interpretability and generalizability.

4.3 Ablation Studies

We validated the contribution of each component on the MIMIC-III dataset (Table 3). Similar trends were observed on MIMIC-IV (see Appendix A.4).

1) Knowledge Sources. The model performance was lowest when using only the EHR dataset, confirming the data sparsity problem. Adding the UMLS dataset resulted in a slight improvement in performance (AUPRC increased by 0.58%), indicating that static ontologies lack causal depth.

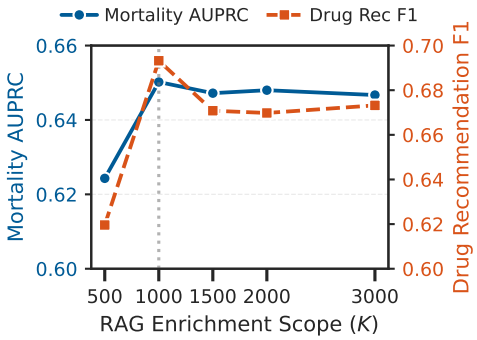

Figure 3: Performance by different RAG sizes.

2) Dual-Layer Structure. Removing the evidence-based CoT decreases performance, confirming that our reasoning with the evidence-based CoT can capture latent causal chains not present in static databases. This effectively addresses the trade-off between reliability and timeliness.

3) Task-Adaptive Selection. This module effectively mitigates the negative transfer.

w/o Knowledge Expansion: Disable multi-view expansion leads to a sharp decline in mortality prediction (-3.78%), demonstrating that a sparse base graph alone is insufficient.

w/o Knowledge Contraction: Using the complete expanded graph also reduces optimal performance. This validates that indiscriminate fusion introduces noise, and learning what to ignore is crucial.

w/o Task Adaptive: Removing the task embedding vector $\mathbf{t}$ leads to a decrease in performance for multi-label tasks, demonstrating that different clinical tasks require different views.

4) Fusion Strategy. Replacing the cross-attention with concatenation reduces model performance, confirming that knowledge must be used

dynamically based on the patient's changing state.

4.4 Parameter Sensitivity Analysis

We investigate how the scope K of RAG enrichment affects model performance. As described in the methodology, we select the top-K high-frequency entities ($\mathcal{V}_{top}$) from the base graph to undergo LLM-based relation extraction from PubMed. We varied K within {500, 1000, 1500, 2000, 3000} and report the results in Figure 3.

Performance significantly improves as K reaches 1000, confirming that the static UMLS graph lacks recent causal chains, which RAG bridges effectively. However, performance plateaus or degrades slightly when K exceeds 1500, indicating that critical reasoning is concentrated within this range. Extending RAG to lower-frequency entities yields diminishing returns and may introduce literature noise or irrelevant associations. Thus, we set $K = 1000$ as a cost-effective threshold for knowledge enrichment.

4.5 Task-Adaptive Knowledge Analysis

Figure 4 visualizes the learned attention weights, revealing distinct selection patterns that align with medical intuition. Mortality prediction prioritizes Lab-Medical Entity edges (0.46). This reflects the reliance on acute physiological derangements (e.g., abnormal lactate) captured by lab tests rather than static hierarchies. Disease prediction is dominated by Medical Entity-Medical Entity edges (0.72), utilizing clinical patterns (e.g., drug-disease) to infer potential diagnoses. Drug recommendation shifts to Semantic edges (0.61) to exploit implicit relations for medication matching. Readmission prediction relies heavily on Hierarchy edges (0.48). Unlike mortality, readmission relates to chronic complexity; the hierarchy helps generalize specific codes to broader conditions for robust long-term prediction. These results confirm that TrustKE actively decouples the KG into task-specific views.

Crucially, beyond merely amplifying relevant signals, this differential weighting explicitly validates our paradigm of learning what to ignore. For instance, semantic edges are aggressively suppressed to a near-zero weight (0.01) in mortality prediction—in stark contrast to their prominence (0.61) in drug recommendation. This targeted suppression prevents noisy associations (e.g., diabetes is similar to obesity) from interfering with acute mortality signals.

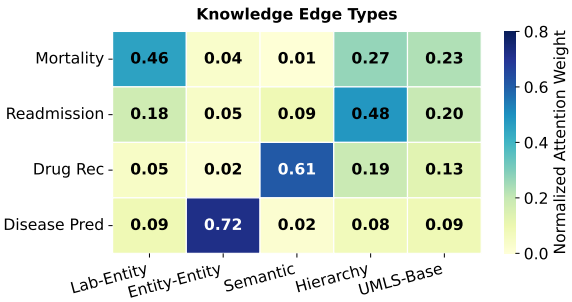

Figure 4: Visualization of Task-Adaptive Knowledge Selection.

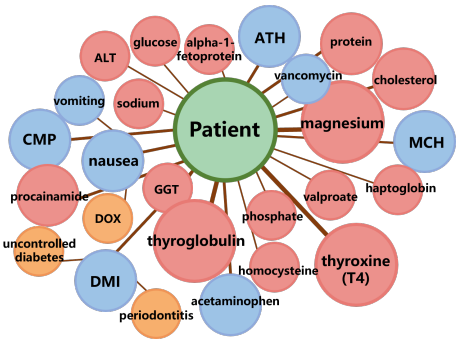

Figure 5: Visualization of the reasoning subgraph for a high-risk patient (Patient #14, Mortality=1). Red nodes denote lab tests, blue nodes denote medical concepts (diseases/drugs), and orange blue nodes represent PubMed knowledge.

4.6 Case Study

To demonstrate TrustKE's interpretability, we visualize the inference subgraph for a representative high-risk patient in Figure 5. This patient was admitted with multiple organ dysfunctions and eventually died. TrustKE correctly predicted a high mortality risk. Consistent with our statistical analysis in Section 4.6, the model assigns the highest attention weights (visualized by node size and edge width) to lab tests (red nodes), such as Magnesium, Glucose, and Thyroxine. These nodes as direct signals for acute physiological instability, which are critical predictors in ICU settings. Crucially, while the subgraph for this patient contains dense connections related to chronic history, TrustKE explicitly assigns them negligible attention weights. By suppressing these irrelevant edges, the model prevents the signal dilution effect.

Beyond direct associations, TrustKE exhibits reasoning capabilities. For instance, while attending to the explicit abnormal record of Glucose, the model activates the 2-hop neighbor Uncontrolled Diabetes (orange node). This indicates that the model does

Task	MIMIC-III (ms)	MIMIC-IV (ms)	Avg. Time (ms)
Mortality	3.11	3.13	3.12
Readmission	2.38	2.83	2.61
Drug Rec	18.14	18.73	18.44
Disease Pred	5.88	5.58	5.73

Table 4: Empirical online inference latency per patient.

not merely memorize numerical values but understands the underlying pathology driving the abnormality. Similarly, the connection between Nausea and related symptom concepts verifies the model's ability to perform semantic expansion. Specifically, the node-aware semantics effectively ensures the reasoning remains dynamically grounded in task-specific evidence. Such a transparent reasoning process builds clinical trust.

4.7 Complexity and Efficiency Analysis

To ensure the practical feasibility of TrustKE in real-world clinical deployments, we explicitly decouple the system into offline knowledge construction and online inference, analyzing their computational complexity and empirical latency separately.

Offline Knowledge Construction (One-Time Cost). The computationally intensive retrieval and LLM reasoning are one-off pre-computation processes. The complexity of this phase is dominated by the Layer 2 Enrichment via RAG-Based Reasoning (Section 3.2), bounded by $\mathcal{O}(K \cdot C_{LLM})$, where C_{LLM} is the inference cost of the LLM per query. As we set $K = 1000$, this produces a static global graph with a highly controllable cost. Empirically, on a single NVIDIA V100 GPU ((32GB VRAM)), the end-to-end construction takes approximately 2 hours for MIMIC-III and 3 hours for MIMIC-IV (including network latency).

Online Inference (Real-Time Efficiency). During deployment, the generative LLM is not involved. The online complexity is determined by three lightweight components:

1) Task-Adaptive Knowledge Selection: Computing edge scores takes $\mathcal{O}(|\mathcal{E}_{ext}| \cdot d)$, where $|\mathcal{E}_{ext}|$ is the number of edges in the expanded graph and d is the hidden dimension.

2) Graph Encoder: The message passing complexity is $\mathcal{O}(L(|\mathcal{E}_{task}|d + |\mathcal{V}|d^2))$, where L is the number of layers. Crucially, our task-adaptive contraction explicitly filters out low-weight edges, ensuring $|\mathcal{E}_{task}| \ll |\mathcal{E}_{ext}|$. This significantly reduces the computational burden compared to full-graph.

3) Sequence Encoder: The GRU-based modeling takes $\mathcal{O}(T \cdot d^2)$, where T is the sequence length.

Empirical Latency. To validate the real-time capability, we measured the actual inference latency per patient. As shown in Table 4, the online inference is strictly dominated by the GCN on the task-specific graph, consistently requiring less than 20 ms per patient across all distinct clinical tasks.

5 Conclusion

In this work, we identify and address two critical impediments in knowledge-enhanced clinical sequence modeling, the trade-off between knowledge reliability and timeliness, and the phenomenon of relative negative transfer caused by indiscriminate fusion. We propose TrustKE, a unified framework that synergizes the structural rigor of curated KGs with the generative reasoning of LLMs through a novel dual-layer knowledge construction. Crucially, by incorporating a task-adaptive knowledge selection mechanism, TrustKE prove that learning what to ignore is as vital as knowledge expansion in complex clinical settings. Empirical results on multiple benchmarks validate that our approach not only achieves superior predictive performance but also provides grounded explanations for its decisions.

Limitations

While TrustKE achieves robust performance, we identify three aspects for future exploration:

1) Efficiency-Accuracy Trade-off. To achieve deep semantic reasoning, TrustKE incorporates a comprehensive graph construction and retrieval process. While this design ensures high predictive precision essential for critical care, it naturally incurs a higher computational cost compared to shallow baselines. In resource-constrained deployment scenarios, future iterations could employ knowledge distillation to compress the reasoning capabilities into lighter student models.

2) Specialization for Intensive Care Settings. Our framework is explicitly optimized for ICUs data(e.g., MIMIC-III/IV), leveraging the dense temporal granularity of laboratory tests to enhance representation learning. While highly effective in this domain, extending the framework to outpatient scenarios with irregular records represents a distinct challenge. We plan to investigate adaptation mechanisms for low-frequency data in subsequent work.

3) Dependency on Backbone Capabilities. The

quality of the constructed knowledge graph correlates with the capabilities of the underlying LLM and the retrieval corpus. While our evidence-based constraint mechanism effectively filters noise, the system's reasoning ceiling is partially determined by the generative backbone. As LLMs continue to evolve, replacing the backbone with more advanced models is expected to further elevate the quality of knowledge extraction without architectural changes.

Ethics Statement

This research focuses on clinical prediction using the MIMIC-III and MIMIC-IV datasets. Both datasets are publicly available and strictly deidentified. No private patient information was exposed during this study.

Furthermore, while TrustKE integrates medical knowledge to enhance interpretability, it is intended as a research framework to assist clinical decision-making, not to replace it. Due to the potential for errors in deep learning models and LLM-generated content, any deployment in real-world clinical settings would require rigorous prospective validation and human-in-the-loop oversight to ensure patient safety.

Acknowledgments

This work was supported by Innovation Foundation of High-end Scientific Research Institutions in Zhongshan of China and Hong Kong Scholars Program.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 43–51.

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

Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. 2023. Boosting transformers and language models for clinical prediction in immunotherapy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 332–340.

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.

Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. 2018. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Yan Kang, Jingyu Zheng, Mingjian Yang, and Ning An. 2024. Inter-structure and intra-semantics graph contrastive learning for disease prediction. *Knowledge-Based Systems*, 300:112059.

Taeri Kim, Jiho Heo, Hongil Kim, Kijung Shin, and Sang-Wook Kim. 2024. Vita:'carefully chosen and weighted less' is better in medication recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8600–8607.

Xiang Li, Shunpan Liang, Yulei Hou, and Tengfei Ma. 2024a. Stratmed: Relevance stratification between biomedical entities for sparsity on medication recommendation. *Knowledge-Based Systems*, 284:111239.

Xiang Li, Shunpan Liang, Yu Lei, Chen Li, Yulei Hou, Dashun Zheng, and Tengfei Ma. 2024b. Causalmed: Causality-based personalized medication recommendation centered on patient health state. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 1276–1285.

Xiang Li and Xiao-Hua Zhou. 2025. Temporal visiting-monitoring feature interaction learning for modelling structured electronic health records. *Knowledge-Based Systems*, 327:114155.

Zihang Liu, Jiawei Guo, Hao Zhang, Hongyang Chen, Jiajun Bu, and Haishuai Wang. 2025. Long-form hallucination detection with self-elicitation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4082–4100.

Chang Lu, Tian Han, and Yue Ning. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574.

Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. 2021. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3529–3535. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. 2025. Doctorrag: Medical rag fusing knowledge with patient analogy through textual gradients. *arXiv preprint arXiv:2505.19538*.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020a. 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.

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

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

Yongjian Ren, Yuliang Shi, Kun Zhang, Xinjun Wang, Zhiyong Chen, and Hui Li. 2022. A drug recommendation model based on message propagation and ddi gating mechanism. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3478–3485.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5953–5959. International Joint Conferences on Artificial Intelligence.

Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2022. Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. In *Advances in Neural Information Processing Systems*, volume 35, pages 27837–27849. Curran Associates, Inc.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, and 1 others. 2025. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261.

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM web conference 2022*, pages 935–945.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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.

Chuang Zhao, Hui Tang, Jiheng Zhang, and Xiaomeng Li. 2025a. Unveiling discrete clues: Superior healthcare predictions for rare diseases. In *Proceedings of the ACM on Web Conference 2025*, pages 1747–1758.

Chuang Zhao, Hui Tang, Hongke Zhao, and Xiaomeng Li. 2025b. Diffmv: A unified diffusion framework for healthcare predictions with random missing views and view laziness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3933–3944.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. Leave no patient behind: Enhancing medication recommendation for rare disease patients. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–542.

Haoran Zheng, Jieming Shi, and Renchi Yang. 2025. Grasp: Simple yet effective graph similarity predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22884–22892.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3549–3559.

Xin Zou, Xiao He, Xiao Zheng, Wei Zhang, Jiajia Chen, and Chang Tang. 2024. Dai-net: Dual adaptive interaction network for coordinated medication recommendation. *IEEE Journal of Biomedical and Health Informatics*.

A Experimental Setup

A.1 Dataset Statistics

Our experiments are conducted on two widely recognized real-world electronic health record

(EHR) datasets: MIMIC-III and MIMIC-IV. These datasets are standard benchmarks for analyzing longitudinal clinical data, encompassing diagnostic codes, procedures, medications, and laboratory tests. For laboratory tests, we utilize the explicit flag column (e.g., abnormal) provided in the MIMIC labevents table to determine the input abnormality state without introducing subjective thresholds.

We follow a standard preprocessing pipeline. Patients with fewer than two visits are filtered out to ensure longitudinal validity. The datasets are split into training, validation, and test sets with a ratio of 0.75:0.1:0.15, ensuring no information leakage across splits. Table 5 presents the detailed statistics of the processed datasets.

Item	MIMIC-III
Patients	6,313
Visits	14,599
Dis. / Proc. / Sym. / Med. / Lab	1,957 / 1,416 / 434 / 131 / 278
Lab Avg. / Max Visits	2.31 / 29
Avg. / Max Dis. per Visit	10.67 / 128
Avg. / Max Proc. per Visit	3.89 / 50
Avg. / Max Sym. per Visit	7.88 / 78
Avg. / Max Med. per Visit	11.66 / 65
Avg. / Max Lab per Visit	18.23 / 109
Item	**MIMIC-IV**
Patients	54,845
Visits	138,824
Dis. / Proc. / Sym. / Med. / Lab	2,000 / 1,500 / 564 / 131 / 325
Lab Avg. / Max Visits	2.53 / 69
Avg. / Max Dis. per Visit	8.89 / 268
Avg. / Max Proc. per Visit	2.19 / 63
Avg. / Max Sym. per Visit	7.22 / 118
Avg. / Max Med. per Visit	7.04 / 72
Avg. / Max Lab per Visit	18.35 / 110

Table 5: Dataset statistics comparison between MIMIC-III and MIMIC-IV.

A.2 Baselines Details

We categorize the baseline models based on their target tasks. All baseline methods followed their optimal experimental settings.

1) Mortality/Readmission Prediction

- **AdaCare**: Uses multiple time scales to capture long-term temporal dependencies in biomarker changes, interpreting them as health status within clinical characteristics.
- **ConCare**: Introduces a cross-head attention mechanism to explicitly capture the interdependencies between different clinical characteristics, forming personalized health contexts.
- **GRASP**: A graph similarity computation method that enhances node features through positional encoding to learn general patient representations.
- **Chet**: Utilizes a context-aware learning framework with transformation functions on a dynamic disease graph to design dynamic subgraphs for each patient's visit, modeling health events in a single step.
- **VITA**: Proposes a selective attention mechanism to filter out noisy or uninformative patient visit records, thereby enhancing the robustness of long-sequence modeling of patient visits.
- **EMERGE**: A Retrieval-Enhanced Generative (RAG) framework for retrieving clinical evidence from guidelines to support predictions.
- **ColaCare**: A multi-agent framework where different LLM agents simulate roles to collaboratively improve clinical decision-making.
- **UDC**: Utilizes language models to generate discrete clinical cues to enhance sparse patient representations.

2) Drug Recommendation

- **MoleRec**: Explicitly models the molecular substructure of drugs for finer-grained drug recommendations.
- **RAREMed**: Specifically designed to handle long-tailed distribution problems in drug recommendations, employing relative relation modules.
- **StratMed**: Utilizes a correlation-based pyramidal hierarchical approach to enhance the expressive power of sparse data, thereby obtaining patient representations.
- **DAI-Net**: Utilizes a dual-view attention mechanism to capture dynamic patient states and static drug interactions.
- **VITA**: Proposes a selective attention mechanism to filter out noisy or uninformative patient visit records, thereby enhancing the robustness of long-sequence modeling of patient visits.

Variant	Mortality		Readmission		Drug Rec		Disease Pred	
	AUPRC	Δ	AUPRC	Δ	Jaccard	Δ	w-F1	Δ
TrustKE (Full)	**0.6230**	-	**0.5192**	-	**0.4982**	-	**27.93**	-
1) Knowledge Source								
EHR Only	0.5943	-2.87%	0.5036	-1.56%	0.4777	-2.05%	27.11	-0.82
UMLS Only	0.6129	-1.01%	0.5027	-1.65%	0.4767	-2.15%	26.45	-1.48
2) Dual-Layer Structure								
w/o Evidence-Based CoT	0.6090	-1.4%	0.5032	-1.6%	0.4730	-2.52%	25.18	-2.75
3) Task-Adaptive Selection								
w/o Knowledge Expansion	0.5997	-2.33%	0.5021	-1.71%	0.4744	-2.38%	26.75	-1.18
w/o Knowledge Contraction	0.5968	-2.62%	0.5067	-1.25%	0.4764	-2.18%	27.18	-0.75
w/o Task Adaptive	0.5974	-2.56%	0.5035	-1.57%	0.4747	-2.35%	27.24	-0.69
4) Fusion Strategy								
w/o Cross-Attention	0.5887	-3.43%	0.5047	-1.45%	0.4782	-2.00%	26.14	-1.79

Table 6: Ablation study on MIMIC-IV. We report AUPRC for binary tasks, Jaccard for drug recommendation, and w-F1 for disease prediction. Δ denotes the performance drop compared to the full TrustKE model.

- **RAREMed**: Proposes a pre-training and fine-tuning framework tailored for rare diseases to enhance patient representations.

- **CausalMed**: A causal graph learning framework that eliminates spurious relevance in drug recommendations by identifying invariant substructures.

3) Disease Prediction

- **Timeline**: Learns fine-grained time decay factors for each medical code by explicitly designing an attention mechanism with attention weights for visit intervals and each medical code.

- **G-BERT**: Integrates a pre-trained model that fully considers the hierarchical information of the disease as part of the pre-training process, achieving performance improvements.

- **HiTANet**: Employs a hierarchical temporal attention mechanism to calculate the impact of each visit and its corresponding time point on the final disease prediction result.

- **CGL**: Designs a collaborative graph learning model that combines disease knowledge and free text data to explore patient-disease interactions, providing more reference information for disease prediction.

- **Chet**: Constructs a global disease co-occurrence matrix to calculate the neighbors and global neighbors of each disease category in each visit, thereby obtaining the internal network of diseases between different patients and between different visits of the same patient, which helps predict the risk of future disease development.

- **UDC**: Utilizes a language model to generate discrete clinical cues to enhance sparse patient representations.

A.3 Evaluation Metrics Details

We adopt standard metrics aligned with specific task requirements.

1) Binary Risk Prediction (e.g., Mortality, Readmission): We report the AUROC, AUPRC, and F1.

2) Multi-label Prediction (e.g., Disease Prediction, Drug Recommendation): For drug recommendation, we evaluate using Jaccard, F1, and AUPRC. For disease prediction, following, we additionally employ w-F1 and Recall@k to assess top-k retrieval performance. Formally, let Y_i be the true label set and $\hat{Y}_i$ be the predicted label set for patient i. The key metrics are defined as:

$$\text{Jaccard} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \tag{12}$$

$$\text{Recall@k} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i^{(k)}|}{|y_i|} \tag{13}$$

where $\hat{y}_i^{(k)}$ denotes the top-k predicted labels ranked by probability.

$$\text{w-F1} = \sum_{c\in\mathcal{C}}\frac{N_c}{N_{total}}\cdot \text{F1}_c \tag{14}$$

where N_c is the number of samples for class c, N_{total} is the total samples, and F1_c is the F1-score for class c.

Raw EHR Term	Cleaning Method	Final Query
PNEUMONIA, ORGANISM NOS	Rule-Based	Pneumonia Organism
ACUTE MI UNSPEC SITE	Rule-Based	Acute MI
GIB	LLM-Rewriting	Gastrointestinal Bleeding
CABG	LLM-Rewriting	Coronary Artery Bypass Grafting

Table 7: Examples of the hybrid query optimization results.

A.4 Ablation Studies in MIMIC-IV

Due to space limitations, we present the ablation study on the larger MIMIC-IV dataset in Table 6. The trend is consistent with MIMIC-III, where removing the Task-Adaptive Selection module leads to a significant performance drop. This further validates the necessity of learning what to ignore to mitigate negative transfer.

B Knowledge Construction Details

B.1 Base Graph Construction & Filtering

Constructing a graph with the entire UMLS introduces significant noise and sparsity. Therefore, we define an active entity sets $\mathcal{V}_{seed}$ consisting of medical concepts that appear in the MIMIC-III/IV dataset with a frequency greater than a threshold $\delta = 2$. For every entity $u \in \mathcal{V}_{seed}$, we query the UMLS semantic network for neighbors. A relation edge (u, r, v) is added to the base graph if and only if both nodes belong to the active set:

$$u, v \in \mathcal{V}_{seed} \tag{15}$$

This constraint ensures the base graph remains dense and clinically relevant, effectively filtering out obscure associations and rare, low-confidence entities.

B.2 Hybrid Query Optimization & Retrieval

Hybrid Term Cleaning. Raw EHR descriptions are often administrative or abbreviated. We employ a two-stage cleaning pipeline:

1) Rule-Based Filtering: We systematically remove administrative suffixes (e.g., "NEC", "NOS") and strip non-semantic punctuation.

2) LLM-Based Rewriting: For cryptic terms remaining after filtering, we employ a specific prompt designed to standardize medical terminology.

The specific prompt template used for query optimization is presented in Figure 6. The prompt enforces strict rules such as abbreviation expansion and administrative suffix removal to ensure the generated query is compatible with PubMed's indexing system.

Prompt for Relation-Agnostic Extraction (Ablation Study)

You are a medical terminology expert. Your task is to convert the following raw EHR/ICD code description into a standard, searchable medical term for PubMed. Here is a raw term description. Please follow the rules to normalize it.
Raw Term: *{raw_term}*
Rules:

1. Expand abbreviations (e.g., "CHF" → "Congestive Heart Failure").
2. Remove administrative suffixes like "NEC", "NOS", "w/o".
3. Fix capitalization (Title Case).
4. Keep it concise.
5. Output ONLY the standardized term. No explanation.

Example:

- Input: "ACUTE MI UNSPEC SITE"
- Output: Acute Myocardial Infarction

Figure 6: The prompt template for normalizing raw EHR terms into standard medical search queries.

Table 7 shows concrete examples of this optimization process.

Cascading Retrieval Strategy. To handle the variance in medical literature availability, we implement an adaptive cascading search via the PubMed API. One is a strict strategy used to filter *Guideline*, *Systematic Review*, or *Review*. This prioritizes high-evidence summaries. The other is an alternative strategy. If strict strategy yields 0 results, the system relaxes the filter to standard *Journal Articles*, sorted by relevance.

For both strategies, we strictly require the presence of a structured abstract and limit the context to the top-3 results.

B.3 Evidence-Based CoT Prompt Design

We leverage the instruction-following capability of Qwen-2.5-7B to perform explicit medical reason-

Prompt for Relation-Agnostic Extraction (Ablation Study)

Role: System
Content: You are a strict Medical Knowledge Extraction Expert.

Role: User
Content: Entity: *{entity_name}*
Text: {context_text}
Task: Extract clinical relations involving "{entity_name}".
Instructions:

1. Identify sentences mentioning "{entity_name}".
2. Extract specific relationships supported by the text. Relations must be one of: [`Treats`, `Causes`, `Has_Symptom`, `Contraindicates`, `Associated_With`].
3. **Ignore self-loops (Target != Tail)**.
4. **QUOTE** the sentence as 'evidence'.
5. Output JSON list. No markdown formatting, no explanations, no <think> tags.

Output Format:

```
[
    {
        "relation": "Treats",
        "tail_entity": "Aspirin",
     "evidence": "Exact quote from text..."
    }
]
```

Figure 7: The **Evidence-Based CoT** prompt template. The instructions explicitly require the model to ground its extraction in text quotes and distinguish specific relation types, contrasting with simple co-occurrence extraction.

ing. Unlike standard relation extraction which acts as a black box, our Evidence-Based CoT prompt enforces a three-step cognitive process: context identification, relation verification, and evidence citation. Figure 7 presents the specific prompt template used in our proposed framework.

To validate the necessity of this fine-grained reasoning, we also designed a general prompt (used in ablation studies), as shown in Figure 8, which simply extracts entities appearing in the same context as "Associated With" without analyzing the specific causal mechanism.

Prompt for Relation-Agnostic Extraction (Ablation Study)

Role: System
Content: You are a medical term extractor.

Role: User
Content: Target Entity: "*{entity_name}*"
Text: {context_text} ...

Task: Identify other medical entities in the text that are mentioned in the same context as "{entity_name}". **Do NOT analyze the specific relationship type** (e.g., whether it treats or causes). Simply identify that they are semantically related or co-occur.

Constraints:

1. Output valid JSON ONLY. No markdown, no explanations.
2. **ALWAYS use "Associated_With" as the relation type for all entries.**
3. Ignore self-loops (Target != Tail).

Output Format:

```
[
    {
        "relation": "Associated_With",
        "tail_entity": "Aspirin",
      "evidence": "Found in same sentence"
    }
]
```

Figure 8: The general prompt template.