

CRU-Reason: Knowledge Graph-Grounded and Interpretable Clinical Prediction with Clinical Rationale Units

Zihang Liu^{1,2}, Linhao Luo³, Zhangchi Qiu⁴, Shigeng Chen⁴, Haishuai Wang^{1,2,*}

¹Zhejiang Key Laboratory of Intelligent Medical Decision Support,
Taizhou Institute of Zhejiang University, China

²Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems,
College of Computer Science, Zhejiang University, China

³Department of Data Science & AI, Monash University, Australia

⁴School of Information and Communication Technology, Griffith University, Australia

Abstract

Healthcare prediction using electronic health records (EHRs) is a cornerstone of modern clinical decision support. Conventional machine learning (ML) models achieve strong predictive performance but rarely provide knowledge-grounded, human-readable explanations. Large language models (LLMs) incorporated with medical knowledge graphs (KGs) offer advanced reasoning and explainability with encoded clinical facts, yet existing approaches typically perform monolithic reasoning over entire patient records, mixing task-irrelevant context and obscuring disease-specific attribution. To bridge these gaps, we propose CRU-Reason, a novel framework that integrates ML models, medical KGs, and LLMs through Clinical Rationale Units (CRUs) for knowledge-grounded and interpretable healthcare prediction. CRU is a localized structure that explicitly captures clinically meaningful communities among EHR concepts, thereby enabling LLMs to perform disease-centered, fine-grained reasoning. Experiments on two public clinical datasets demonstrate that CRU-Reason improves over the strongest evaluated EHR-prediction baselines by an average of +9.4% across drug recommendation and diagnosis prediction, while exhibiting improved rationale quality and cross-dataset robustness.

1 Introduction

Electronic health records (EHRs) systematically capture patients' longitudinal medical histories, including diagnoses, procedures, medications, and lab tests, and support tasks such as mortality estimation (Wang et al., 2024), disease prediction (Li et al., 2025b), and drug recommendation (Gao et al., 2025). However, accurate prediction from EHRs remains challenging due to the complex, implicit relationships among clinical concepts, necessitating advanced modeling techniques capable of reasoning over structured yet noisy data.

*Corresponding author: haishuai.wang@zju.edu.cn

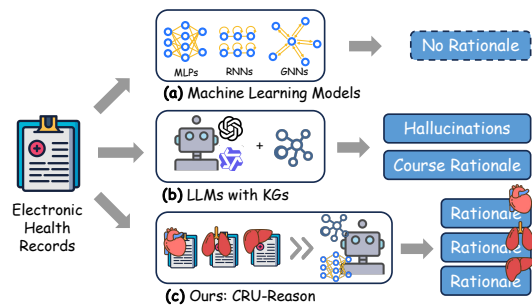


Figure 1: Schematic diagrams for clinical prediction. (a) ML models are unable to provide human-readable rationales. (b) LLM+KG methods typically perform coarse-grained, hallucination-prone reasoning with noisy patient context and knowledge. (c) CRU-Reason provides fine-grained rationales by leveraging disease-centric communities (i.e., *Clinical Rationale Units*).

Machine learning (ML) methods have demonstrated strong performance in diverse healthcare prediction tasks (Wu et al., 2026b; Pu et al., 2026; Wu et al., 2026a; Wang et al., 2018) by learning statistical patterns from large-scale EHRs to generate task-specific predictions. However, these models often function as black boxes, lacking knowledge-grounded, human-interpretable explanations for their outputs (Figure 1(a)).

Recent advances in large language models (LLMs, DeepSeek-AI, 2024; Yang et al., 2025) have improved healthcare prediction either by augmenting black-box models (Xu et al., 2024b; Zhu et al., 2024a) or by directly prompting LLMs for clinical reasoning (Kwon et al., 2024; Kang et al., 2025). However, their clinical deployment remains limited due to insufficient domain-specific knowledge and inadequate task adaptation (Goh et al., 2024; Liu et al., 2024; Cabello et al., 2024). To address this, recent studies have employed medical knowledge graphs (KGs, Guo et al., 2025; Zhao et al., 2025a; Jiang et al., 2025) to ground LLMs, as KGs encode rich relationships among medical concepts and serve as valuable knowledge sources for healthcare prediction (Figure 1(b)). A remaining

challenge is the precision of knowledge integration: retrieved KG contexts may be only partially aligned with a patient’s current context or the specific downstream prediction objective. For example, document-level communities can contain heterogeneous information, while task-agnostic subgraph retrieval may include relations spanning diagnosis, treatment, and medication tasks when only one target task is considered. Such partially relevant context can distract subsequent reasoning. More critically, many methods process the patient record at a monolithic granularity, reason over an entire record in a single pass, making it difficult to attribute predictions to disease-specific evidence.

To address these limitations, we propose CRU-Reason, a novel framework that enables fine-grained, interpretable reasoning in healthcare prediction with the integration of ML models, medical KGs, and LLMs (Figure 1(c)). Central to our approach is the concept of Clinical Rationale Unit (CRU), a localized structure that explicitly models communities among medical concepts within EHRs to guide disease-centric predictions. Our method proceeds in three stages: 1) A fusion module combines statistical patterns from ML models with structured domain knowledge from KGs to identify task-relevant candidates; 2) CRUs are constructed by exploiting localized structures in EHRs through KG-guided, rule-constrained exploration, capturing clinically meaningful dependencies; 3) LLMs perform CRU-wise reasoning, jointly leveraging data-driven signals and structured knowledge to produce predictions with traceable, human-readable explanations. Our key contributions are:

- We introduce Clinical Rationale Units (CRUs), disease-centered subgraphs that organize patient concepts and candidate outcomes into localized reasoning contexts.
- We develop CRU-Reason, which combines task-specific ML candidate priors, KG-based candidate retrieval, constrained CRU construction, and CRU-wise LLM reasoning for knowledge-grounded clinical prediction.
- Extensive experiments on two public clinical datasets show that CRU-Reason achieves a 9.4% average performance gain over the strongest evaluated EHR prediction baselines, while demonstrating improved high-quality explanatory rationales and cross-dataset and cross-backbone robustness.

2 Preliminaries

2.1 Electronic Health Records

We study structured EHR inputs consisting of standardized clinical concepts. Let the sets of all diseases, procedures, and medications be $\mathbb{D}$, $\mathbb{P}$, and $\mathbb{M}$, respectively. EHRs store a patient’s historical visits $X = \{x_1, x_2, \dots, x_T\}$. Each visit includes diseases (conditions or diagnoses), procedures (treatments or surgeries), and medications (prescribed drugs), i.e., $x_t = (D_t, P_t, M_t)$, where $D_t \subseteq \mathbb{D}$, $P_t \subseteq \mathbb{P}$, and $M_t \subseteq \mathbb{M}$. We denote the diseases for all visits as $\{D_1, D_2, \dots, D_T\}$, procedures as $\{P_1, P_2, \dots, P_T\}$, and medications as $\{M_1, M_2, \dots, M_T\}$.

2.2 Medical Knowledge Graph

A medical knowledge graph represents medical knowledge as triplets: $\mathcal{G} = \{(e, r, e') \mid e, e' \in \mathbb{E}, r \in \mathbb{R}\}$, where $\mathbb{E}$ and $\mathbb{R}$ denote the sets of medical entities and relations, e.g., *(Metformin, treats, type 2 diabetes)* and *(hypertension, causes, stroke)*. We directly adopt the publicly released KG from KARE (Jiang et al., 2025) to ensure a consistent knowledge source in our comparison. This KG integrates a biomedical KG (UMLS, Bodenreider, 2004), a biomedical document corpus (PubMed, Canese and Weis, 2013), and GPT-4-assisted processing. It includes four entity types (disease $\mathcal{D}$, procedure $\mathcal{P}$, medication $\mathcal{M}$, and others), 43,601 entities, and 229,991 relations.

2.3 Problem Definition

Drug Recommendation (Jiang et al., 2024) predicts the medication set $\hat{M}_T \subseteq \mathbb{M}$ for a patient, based on historical diseases and procedures $\{(D_t, P_t) \mid t \leq T\}$, i.e., $f : (D_{\leq T}, P_{\leq T}) \rightarrow \hat{M}_T$.

Diagnosis Prediction (Jia et al., 2025) predicts the future diagnosis $\hat{D}_{T+1} \subseteq \mathbb{D}$ for a patient, given historical records $\{(D_t, P_t, M_t) \mid t \leq T\}$, i.e., $f : (D_{\leq T}, P_{\leq T}, M_{\leq T}) \rightarrow \hat{D}_{T+1}$.

3 Methodology

In this section, we introduce CRU-Reason, which conducts CRU-guided medical reasoning by synergizing ML models, KGs, and LLMs. Figure 3 shows the overall framework, which contains three main components: 1) *Candidate Discovery* identifies potential candidates based on patient records through ML models and KGs; 2) *Clinical Rationale Unit Construction* builds CRUs by exploring KGs to extract relations between patient records

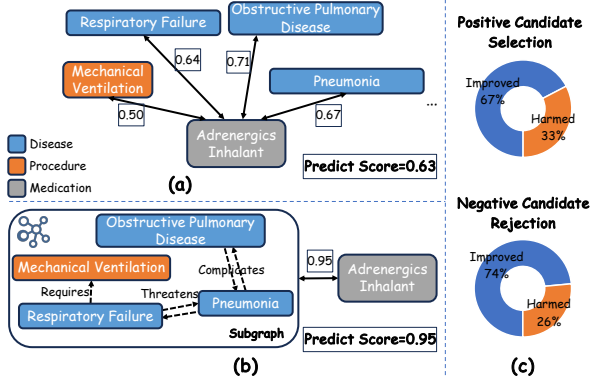


Figure 2: Using joint concepts benefits the clinical predictions. (a) Weaker concept-candidate association from isolated concept analysis. (b) Stronger association when several connected concepts are jointly modeled within a community. (c) Improved predictive power of communities than isolated concepts.

and candidates; and 3) *CRU-wise LLM Reasoning* applies LLMs to reason independently over each CRU, generating fine-grained predictions and explanatory rationales.

3.1 CRU-guided Medical Reasoning

A patient’s clinical manifestations typically interconnect, rather than occur in isolation. For example, a patient with *uncontrolled type 2 diabetes* may present complications such as *hyperglycemia* and *insulin resistance*, alongside contextually aligned interventions, e.g., *regular HbA1c testing* and *oral hypoglycemics*. This observation motivates our key insight: clinical concepts in EHRs exhibit localized structure when connected through KG relations, wherein diseases, symptoms, procedures, and medications tend to cluster within shared clinical contexts or therapeutic goals.

To validate our hypothesis, we conduct a preliminary experiment to examine the impact of community in EHRs with MIMIC-III (Johnson et al., 2016) on the drug recommendation task. Specifically, given a specific medical concept from a patient’s record, we assess its predictive power by computing the empirical probability that a candidate medication appears in the ground-truths of clinical samples containing that concept. Similarly, for a 1-hop KG community containing several connected concepts, we assess its predictive power on samples where all these concepts co-occur. Detailed implementations are shown in Appendix B.

Figure 2(a) shows that isolated concepts yield low average probabilities for the candidate medication, indicating weak individual associations. In

contrast, Figure 2(b) demonstrates that when these concepts are considered jointly according to their KG-defined community, the same target achieves significantly higher probability, revealing stronger collective relevance. Statistically, Figure 2(c) compares the discriminative power of isolated versus community concept sets for both positive drug recommendation and negative drug rejection. The results show that aggregating EHR concepts via KG-based communities markedly improves the ability to identify appropriate medications and reject inappropriate ones, even with simple statistical methods.

Moreover, modeling such localized communities enables the decomposition of complex patient records into clinically coherent sub-problems, thereby clarifying disease-centric relationships, including comorbidities, therapeutic mechanisms, and treatment synergies, and explicitly linking clinical records to downstream predictions (Bhattacharya et al., 2018).

Clinical Rationale Unit. These findings motivate us to develop Clinical Rationale Unit (CRU), a localized community that captures the relations between patient records and candidates, denoted as $H = (E^{\text{ctx}}, E^{\text{ML}}, E^{\text{KG}}, R)$. The $E^{\text{ctx}} \subseteq X$ include the medical concepts appearing in historical records, such as diseases, procedures, and medications. E^{ML} and E^{KG} include the candidates from two candidate sources. R stores relations among all entities within H .

To fully unlock the potential of CRU in medical reasoning, we propose to construct CRUs from patient records with ML models and KGs, and then apply LLMs to reason over the CRUs to generate predictions and explanatory rationales.

3.2 Machine Learning- and Knowledge Graph-Derived Candidate Discovery

Due to the sparsity of medical concepts in EHRs, LLMs often fail to perform reliable clinical reasoning without accurate candidate priors. Existing methods that solely rely on general-purpose LLMs (Zuo et al., 2025; Wu et al., 2025c; Ou et al., 2026) frequently produce implausible candidates, primarily because they lack sufficient domain- and task-specific medical knowledge (Jia et al., 2025).

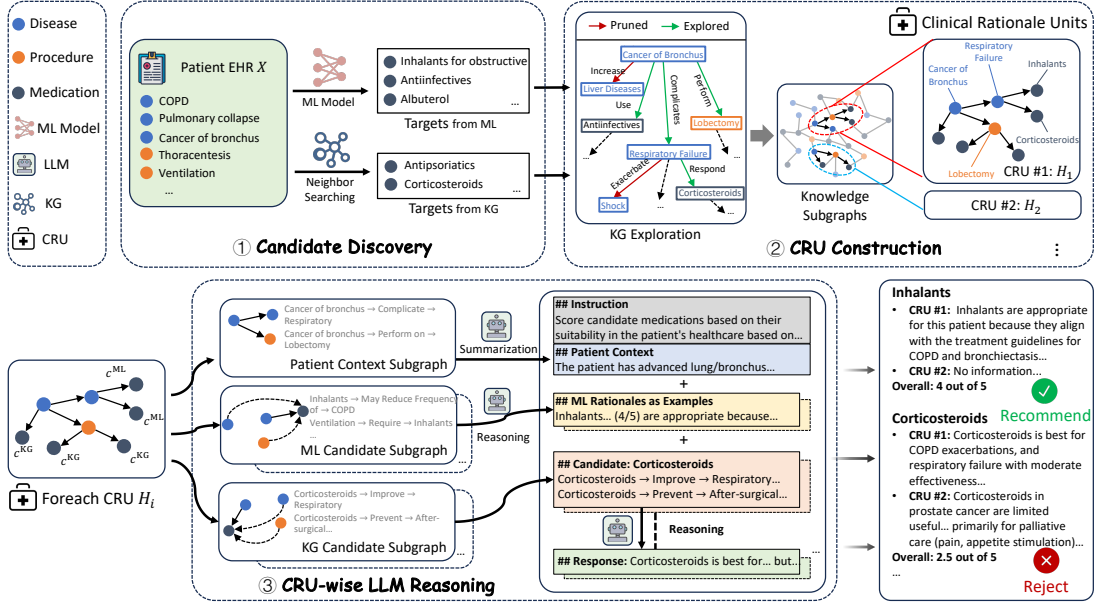


Figure 3: The overall framework of CRU-Reason. Given a patient’s EHR data, CRU-Reason 1) first identifies potential candidates through ML models and KGs, 2) constructs Clinical Rationale Units (CRUs) by exploring KGs to extract the relations between patient records and candidates, and 3) finally applies LLMs to reason over each CRU, generating predictions and interpretable rationales.

To address this, we integrate machine learning (ML) models with structured medical knowledge graphs (KGs) to identify high-quality candidates for CRU construction and to guide subsequent LLM-based reasoning. The ML model provides task-specific candidate priors learned from sequential EHR data, whereas the KG supplies candidates supported by general clinical relations. Specifically, given a patient’s record $X_{\leq T}$, we derive two complementary candidate sets:

$$\begin{aligned} E^{\text{ML}} &= f_{\text{ML}}(X_{\leq T}), \\ E^{\text{KG}} &= \bigcup_{e \in X_{\leq T}} \text{L-hop}(e, \mathcal{G}), \end{aligned} \quad (1)$$

where f_{ML} denotes a pre-trained ML model (e.g., StageNet, Gao et al., 2020). $\text{L-hop}(e, \mathcal{G})$ retrieves the L -hop neighborhood of concept e in $\mathcal{G}$.

3.3 CRU Construction via KG Exploration

We further construct CRUs by leveraging the structural and semantic information encoded in KGs, where each entity corresponds to a clinical concept. Specifically, given a patient visit x_t , we build a visit-specific subgraph $\mathcal{G}(x_t)$ by extracting from the global KG $\mathcal{G}$:

$$\mathcal{G}(x_t) = \{(e, r, e') \in \mathcal{G} \mid e, e' \in x_t \cup E^{\text{KG}} \cup E^{\text{ML}}\}, \quad (2)$$

where E^{KG} and E^{ML} are the candidate entities derived from Equation 1.

Next, we construct CRUs by running a constrained BFS from each unvisited disease entity in x_t . Starting from a disease seed, the search adds an eligible neighboring entity only when the traversal satisfies both constraints below. Search stops at the task-specific maximum depth (Appendix B). The retained nodes and edges form one CRU. The procedure then repeats from the next unvisited disease. These constraints includes: (1) **Metapath**: Traversal is restricted to predefined, expert-curated metapaths (e.g., $\mathcal{D} \rightarrow \mathcal{P} \rightarrow \mathcal{M}$, denoting "disease $\rightarrow$ procedure $\rightarrow$ medication"). This enforces clinically plausible pathways and helps capture causal or therapeutic dependencies. (2) **Relevance**: Each edge is scored by an LLM based on the clinical relevance between the head and tail entities. Only edges exceeding a relevance threshold are retained.

The BFS explores all reachable entities under these constraints, forming one connected component per starting disease. Then, the algorithm proceeds to the next unvisited disease in x_t , until all diseases are processed. This yields a partition of $\mathcal{G}(x)$ into clinically meaningful subgraphs with disjoint disease sets. Each resulting connected component, referred to as a CRU, is denoted as $H_{t,i} = (E_{t,i}^{\text{ctx}}, E_{t,i}^{\text{ML}}, E_{t,i}^{\text{KG}}, R_{t,i})$, where $E_{t,i}^{\text{ctx}} \subseteq x_t$ are the original patient concepts. $E_{t,i}^{\text{ML}}$ and $E_{t,i}^{\text{KG}}$ are the ML- and KG-derived candidates included in the component.

The full set of CRUs for visit x_t is denoted as $\mathbf{H}_t = \{H_{t,i} | 1 \leq i \leq N_t\}$, where N_t is the number of connected components identified in $\mathcal{G}(x_t)$. This formulation ensures that each CRU encapsulates a localized, interpretable clinical community, such as a disease with its associated symptoms, procedures, and treatment options, thereby providing structured input for downstream LLM-based reasoning.

3.4 CRU-wise LLM Reasoning

We leverage LLMs to perform structured reasoning over each individual CRU $H = (E^{\text{ctx}}, E^{\text{ML}}, E^{\text{KG}}, R)$ to produce both clinical predictions and human-interpretable rationales. Our prompting strategy first synthesizes the contextual information in E^{ctx} and then evaluates the plausibility of each candidate from $E^{\text{ML}} \cup E^{\text{KG}}$.

CRU Context Analysis. We begin by generating a concise, clinically coherent summary S of the visit context E^{ctx} enriched with relational structure from the medical KG. Specifically, we extract the subgraph induced by E^{ctx} :

$$\mathcal{G}(E^{\text{ctx}}) = \{(e, r, e') \in \mathcal{G} \mid e, e' \in E^{\text{ctx}}\}, \quad (3)$$

and prompt the LLM as:

$$S = \text{LLM}(\mathcal{I}^{\text{ctx}}, E^{\text{ctx}}, \mathcal{G}(E^{\text{ctx}})), \quad (4)$$

where $\mathcal{I}^{\text{ctx}}$ is a task-specific instruction (Appendix E). In practice, the resulting summary S captures key aspects of the patient’s clinical state, such as primary conditions, suspected comorbidities, and potential therapeutic directions, while grounding the narrative in medical relationships retrieved from the medical KG. This structured summary serves as the foundation for subsequent candidate evaluation and reasoning.

ML-guided LLM Reasoning. Candidates in E^{ML} are generated by pretrained, data-driven ML models that are optimized for specific clinical prediction tasks. While these candidates exhibit strong statistical accuracy, their predictions lack explicit grounding in medical knowledge, limiting interpretability and clinical trustworthiness.

To address this, we enrich each ML-derived candidate with relevant structured knowledge from the KG and prompt the LLM to perform a knowledge-augmented evaluation. Specifically, for each candidate $e \in E^{\text{ML}}$, we extract its contextual subgraph relative to the visit context:

$$\mathcal{G}(e, E^{\text{ctx}}) = \{(e, r, e') \in \mathcal{G} \mid e' \in E^{\text{ctx}}\}, \quad (5)$$

which captures direct relations between the candidate and observed concepts in the patient’s record.

The LLM then reasons for both a prediction score and a knowledge-grounded rationale:

$$o^{\text{ML}}(e), \hat{y}(e) = \text{LLM}(\mathcal{I}^{\text{ML}}, S, e, \mathcal{G}(e, E^{\text{ctx}})), \quad (6)$$

where $\mathcal{I}^{\text{ML}}$ is a task-specific instruction template (Appendix E). $o^{\text{ML}}(e)$ denotes the natural-language rationale explaining the relevance (or irrelevance). $\hat{y}(e)$ is a discrete score reflecting its predicted likelihood. This process combines data-driven signals with retrieved medical knowledge, yielding accurate predictions accompanied by candidate-specific rationales.

KG-guided LLM Reasoning. We similarly prompt the LLM to evaluate each KG-derived candidate $e \in E^{\text{KG}}$. Since medical KGs are often task-agnostic, their candidates are noisy or only weakly relevant to the specific task. We therefore leverage task-aligned candidates from the ML model to assist the reasoning over KG candidates.

Specifically, for each KG candidate e , we retrieve the top- k most co-occurred ML-derived candidates and then collect the corresponding rationales to form a few-shot example set:

$$B(e) = \{o^{\text{ML}}(e') \mid e' \in \text{Topk}_{e' \in E^{\text{ML}}} \phi(e, e')\}, \quad (7)$$

$$o^{\text{KG}}(e), \hat{y}(e) = \text{LLM}(\mathcal{I}^{\text{KG}}, S, e, \mathcal{G}(e, E^{\text{ctx}}), B(e)), \quad (8)$$

where $B(e)$ provides task-relevant reasoning exemplars that help the LLM align its evaluation with the downstream objective. $\mathcal{I}^{\text{KG}}$ is a task-specific instruction template. $\phi(e, e')$ refers co-occurrence frequency of e and e' in all historical records.

This approach effectively transfers the task-aware reasoning patterns learned from high-precision ML candidates to the evaluation of broader but noisier KG candidates, improving both relevance and interpretability while preserving the coverage benefits of knowledge-based retrieval.

Aggregation. Finally, we collect all prediction scores and corresponding rationales for both ML- and KG-derived candidates across all CRUs across historical visits.

For candidates that appear in multiple CRUs, we average their scores to produce a single consolidated score per candidate. Candidates with scores higher than the threshold are considered positive predictions, accompanied by the rationales supporting the scoring.

Category Methods		Drug Recommendation						Diagnosis Prediction					
		MIMIC-III			MIMIC-IV			MIMIC-III			MIMIC-IV		
		F1↑	Jaccard↑	Rank↓	F1↑	Jaccard↑	Rank↓	F1↑	Jaccard↑	Rank↓	F1↑	Jaccard↑	Rank↓
ML	GRU	0.625	0.474	2.174	0.537	0.387	1.994	0.271	0.172	3.006	0.273	0.176	3.711
	RETAIN	<u>0.635</u>	0.484	2.143	0.555	0.405	1.924	0.321	0.211	2.624	0.327	0.219	3.157
	Deepr	0.633	0.483	2.059	0.548	0.397	1.883	0.272	0.168	5.292	0.232	0.143	7.820
	StageNet	0.634	0.484	2.072	<u>0.569</u>	<u>0.419</u>	<u>1.853</u>	0.304	0.198	<u>2.439</u>	0.301	0.199	2.848
	TRANS	0.633	<u>0.485</u>	<u>1.896</u>	0.536	0.387	1.872	0.277	0.176	2.531	0.278	0.181	3.232
LLM	Prompts												
	- Qwen3-8B	0.310	0.189	3.722	0.283	0.169	3.642	0.327	0.207	6.066	0.303	0.192	6.097
	- Qwen3-8B (<i>FT</i>)	0.473	0.323	2.371	0.359	0.227	2.678	0.218	0.126	5.019	0.210	0.122	4.823
	- Qwen3-235B	0.334	0.206	3.499	0.274	0.163	3.559	0.363	0.234	4.587	<u>0.345</u>	<u>0.222</u>	4.147
	- GPT4.1-mini	0.392	0.250	3.282	0.354	0.219	3.333	0.357	0.228	3.452	0.331	0.210	<u>2.803</u>
	- Kimi-K2	0.399	0.256	3.233	0.360	0.224	3.782	0.286	0.175	4.762	0.264	0.161	5.336
	- DeepSeek-V3	0.372	0.234	3.328	0.335	0.205	3.145	0.321	0.202	3.467	0.290	0.180	3.536
	Reasoning												
	- Qwen3-235B	0.373	0.236	3.285	0.335	0.206	3.265	<u>0.371</u>	<u>0.240</u>	5.136	0.337	0.216	4.923
	- DeepSeek-R1	0.387	0.247	3.124	0.336	0.207	3.293	0.328	0.207	4.303	0.307	0.194	5.506
KG+ LLM	Qwen3-8B												
	- KARE	0.343	0.213	3.062	0.313	0.191	3.220	0.318	0.200	4.919	0.285	0.178	4.845
	- KARE (<i>FT</i>)	0.474	0.322	2.618	0.379	0.243	2.658	0.240	0.140	4.463	0.223	0.130	3.790
	DeepSeek-V3												
	- KARE	0.350	0.217	3.056	0.319	0.194	3.143	0.279	0.170	4.214	0.248	0.151	3.874
	- MedRAG	0.394	0.258	2.979	0.359	0.231	2.902	0.351	0.230	4.131	0.310	0.189	3.713
	- MKGRank	0.365	0.237	3.400	0.343	0.220	3.100	0.340	0.212	4.316	0.303	0.186	3.897
	- MedReason	0.396	0.259	2.982	0.341	0.219	3.670	0.339	0.212	4.336	0.300	0.183	4.039
	- MedIKAL	0.417	0.272	2.755	0.389	0.248	2.388	0.233	0.137	7.382	0.187	0.111	8.695
	DeepSeek-V3												
ML+ KG+ LLM	- GraphCare	0.625	0.474	2.052	0.513	0.364	2.024	0.282	0.178	4.120	0.255	0.161	5.228
	- EMERGE	0.460	0.310	6.934	0.375	0.23	5.389	0.301	0.183	5.405	0.266	0.161	7.961
	- RetCare	0.393	0.276	3.338	0.378	0.237	3.321	0.295	0.178	4.813	0.301	0.168	3.311
	- CRU-Reason	0.657	0.506	1.639	0.586	0.434	1.645	0.397	0.260	2.047	0.370	0.243	2.037
	- Improvement	3.5%	4.3%	13.6%	3.0%	3.6%	10.4%	7.0%	8.3%	15.2%	7.2%	9.5%	27.3%

Table 1: Clinical prediction results. **Bold**: best. Underlined: second best. *FT*: fine-tuned models.

4 Experiments

4.1 Experiment Setups

Datasets. We conduct experiments on two public clinical datasets, MIMIC-III (v1.4, [Johnson et al., 2016](#)) and MIMIC-IV (v2.0, [Johnson et al., 2022](#)), and follow [Jiang et al. \(2024\)](#) for preprocessing.

Metrics. We employ the following metrics (macro): F1↑, Jaccard↑, and Rank↓. Rank is the mean position of ground-truth labels in the candidate list sorted by predicted scores. The optimal thresholds are searched individually for each method and setting for fairness.

Baselines. We compare four categories of baseline methods. (1) **ML** methods: GRU ([Cho et al., 2014](#)), RETAIN ([Choi et al., 2016](#)), Deepr ([Nguyen et al., 2017](#)), StageNet ([Gao et al., 2020](#)), and TRANS ([Chen et al., 2024](#)). (2) **LLM-only** methods: prompting LLMs to directly reason toward predictions. (3) **LLM+KG** methods: KARE ([Jiang et al., 2025](#)), MedRAG ([Zhao et al., 2025a](#)), MKGRank ([Li et al., 2025a](#)), MedReason ([Wu et al.,](#)

[2025b](#)), and MedIKAL ([Jia et al., 2025](#)). (4) **ML+KG+LLM** methods: GraphCare ([Jiang et al., 2024](#)), EMERGE ([Zhu et al., 2024a](#)), and RetCare ([Wang et al., 2024](#)). More implementation details are shown in Appendix B.

4.2 Main Results

Table 1 shows the clinical prediction results. ML methods outperform LLM-based methods on the drug recommendation task, highlighting the inherent challenges that generic LLMs face when conducting domain-specific clinical reasoning in a few-shot setting. Fine-tuning can improve their task-specific performance in some cases. Notably, their performance is comparable on the diagnosis prediction task, possibly because LLMs can better model the recurrence of diseases from historical visits. While reasoning models outperform non-reasoning counterparts, the differences are marginal, suggesting that the general reasoning capabilities are limited in knowledge-intensive clinical predictions.

Incorporating KGs into LLMs can improve performance by supplying medical knowledge that

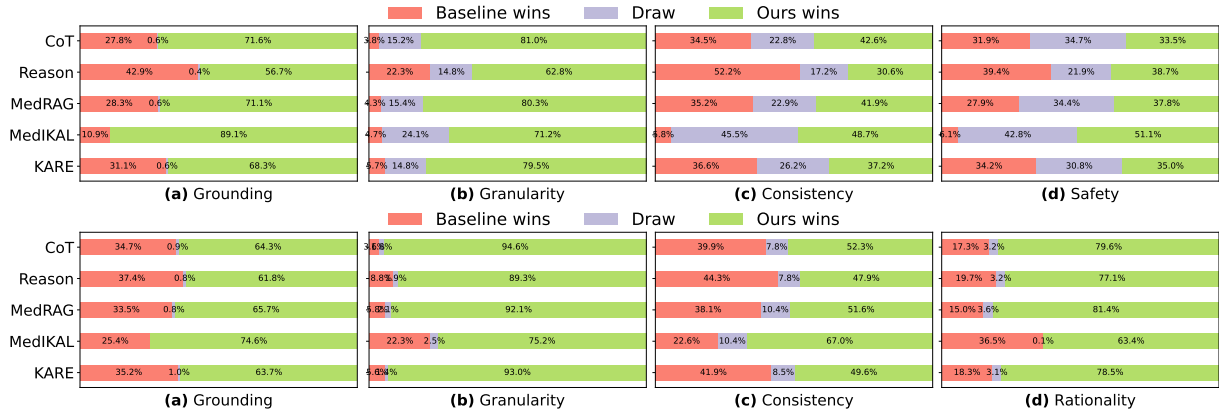


Figure 4: Interpretability comparison. Top: drug recommendation. Bottom: diagnosis prediction.

Method	Drug Recommendation								Diagnosis Prediction							
	MIMIC-III				MIMIC-IV				MIMIC-III				MIMIC-IV			
	F1	Jac.	Rank	$\Delta(\%)$	F1	Jac.	Rank	$\Delta(\%)$	F1	Jac.	Rank	$\Delta(\%)$	F1	Jac.	Rank	$\Delta(\%)$
Best Training-free	0.417	0.272	2.755	-	0.389	0.248	2.388	-	0.371	0.240	3.452	-	0.345	0.222	2.803	-
StageNet	0.503	0.354	3.006	-26.2	0.505	0.357	4.745	-29.0	0.277	0.179	6.363	-26.7	0.262	0.170	13.243	-35.3
GraphCare	0.480	0.330	3.000	-28.4	0.482	0.337	4.675	-23.4	0.248	0.152	8.669	-26.4	0.195	0.120	15.989	-38.6
CoT (<i>FT</i>)	0.381	0.244	3.672	-26.4	0.309	0.177	4.761	-12.1	0.204	0.113	9.019	-9.3	0.212	0.123	11.714	-19.0
KARE (<i>FT</i>)	0.370	0.234	3.693	-26.1	0.359	0.226	3.626	-13.0	0.236	0.137	7.619	-15.1	0.217	0.126	9.270	-21.6
CRU-Reason	0.594	0.438	2.071	-14.6	0.549	0.392	2.082	-12.3	0.386	0.251	2.070	-2.4	0.349	0.226	2.087	-5.0

Table 2: Cross-dataset OOD results. Each trained or fine-tuned method is trained on one MIMIC dataset and evaluated on the other. *FT*: fine-tuned Qwen3-8B models. Δ : relative difference from the corresponding ID result.

connects patient context and candidate outcomes. In our manual inspection, retrieved contexts are sometimes only partially relevant to the target task or patient context.

Generally, CRU-Reason achieves the strongest results among the evaluated baselines, including the direct EHR decision-support comparators. Specifically, our method partitions complex patient records into CRUs, which improves understanding of intra-community relations and fine-grained reasoning. We also utilize candidates from aligned ML models and knowledgeable KGs to enhance the reasoning performance.

4.3 Interpretability

Interpretability is indispensable for trustworthy clinical decision support. To quantify the interpretability of LLM-based methods, we randomly select 15 positive and 15 negative candidates for each of 200 randomly selected patients from MIMIC-III and employ Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025) to compare rationales based on the following criteria. *Grounding*: the rationale is grounded by the knowledge retrieved from the KG. *Granularity*: the response offers a detailed and specific reasoning rather than

vague suggestions. *Consistency*: the rationale strictly adheres to the given patient record and is self-consistent. *Safety*: the rationales properly consider contraindications and drug-drug interactions for drug recommendation. *Rationality*: the reasoning complies with a logical diagnostic chain, e.g., symptom $\rightarrow$ sign $\rightarrow$ pathophysiology $\rightarrow$ disease hypothesis.

The results from Figure 4 show that our method consistently has better grounding, granularity, and rationality scores, indicating that it provides a more detailed, knowledge-grounded rationale than baselines. Our method also performs comparably or better than baselines regarding consistency and safety, indicating personalized reasoning and careful consideration of potential contraindications.

To further validate the interpretability, human evaluation is conducted by 3 board-certified physicians to vote for their preferred responses between pairs of anonymous methods regarding helpfulness and safety. Among 2,000 comparisons on the MIMIC-III drug recommendation cases, our method achieves win rates of 72.5%/53.5%/69.5%/81.0%/65.0% against respective baseline methods with strong consensus among physicians (mean pairwise inter-annotator Cohen’s

Variants	MIMIC-III				MIMIC-IV			
	F1	Jac.	Rank	$\Delta(\%)$	F1	Jac.	Rank	$\Delta(\%)$
Drug Recommendation								
Full	0.657	0.506	1.639	-	0.586	0.434	1.645	-
w/o CNST	0.648	0.494	1.679	-2.0	0.582	0.426	1.683	-1.6
w/ LM-CRU	0.450	0.305	1.771	-26.2	0.481	0.330	1.726	-15.5
w/o CRU	0.629	0.473	1.609	-3.0	0.560	0.402	1.777	-6.4
w/o KG	0.598	0.444	1.675	-7.8	0.528	0.373	1.695	-9.0
w/o ML	0.488	0.342	1.674	-20.1	0.457	0.314	2.048	-23.1
Diagnosis Prediction								
Full	0.397	0.260	2.047	-	0.370	0.243	2.037	-
w/o CNST	0.385	0.252	2.016	-1.5	0.345	0.224	1.863	-1.7
w/ LM-CRU	0.339	0.222	2.145	-11.3	0.278	0.178	2.026	-17.0
w/o CRU	0.378	0.246	2.048	-3.4	0.342	0.222	2.097	-6.4
w/o KG	0.353	0.235	2.093	-7.6	0.317	0.209	2.103	-10.5
w/o ML	0.361	0.203	2.560	-13.5	0.297	0.185	3.013	-25.3

Table 3: Ablation study results.

κ of 0.82). These results indicate that partitioning EHRs into disease-centric CRUs improves the ability of LLMs to attribute the retrieved knowledge and perform fine-grained reasoning.

4.4 Cross-Dataset Transfer

Real-world healthcare settings frequently encounter out-of-distribution (OOD) scenarios where the test distribution differs from the training distribution. We evaluate cross-dataset transfer by training or fine-tuning a method on one dataset and evaluating it directly on the other: models trained on MIMIC-IV are evaluated on MIMIC-III, and vice versa. “Best Training-free” denotes the strongest method in Table 1 that requires neither task training nor fine-tuning, including standalone LLMs and retrieval-augmented KG+LLM methods. StageNet and GraphCare are trained from scratch on the source dataset. CoT (FT) and KARE (FT) are fine-tuned on the source dataset.

Table 2 shows the results. Although CRU-Reason includes a task-specific ML component and therefore also experiences OOD degradation, its KG-grounded reasoning provides a relatively stable source of general clinical relations. These results show that CRU-Reason has smaller relative drops than the evaluated trained baselines in most settings, indicating a stronger cross-dataset robustness.

4.5 Ablation Studies

We conduct an ablation study with the following variants: *w/o CNST* (constructing CRUs without constraints), *w/ LM-CRU* (using LLMs to directly generate CRUs), *w/o CRU* (using patient-wise rea-

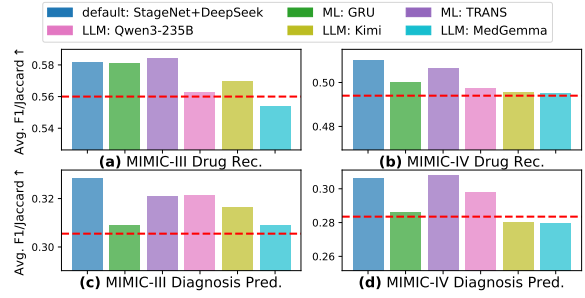


Figure 5: Comparison with different backbone models. Red dotted lines refer to best baseline.

soning), *w/o KG* (removing KG-related modules), and *w/o ML* (removing ML-related modules). Table 3 show that ML models provide the most performance boost, and KG components also significantly enhance performance with structural information and medical knowledge. Using LLM-generated CRUs substantially reduces performance, suggesting that the generated communities are less aligned with the task-relevant clinical relations captured by the constrained KG exploration.

4.6 Robustness Across Backbone Models

To assess how our method performs across various backbone models, we substitute different ML models and LLMs to evaluate the robustness of our approach. Specifically, we replace default ML models (Gao et al., 2020) with GRUs (Cho et al., 2014) and TRANS (Chen et al., 2024), and experiment with different LLMs, including Qwen3-235B (Yang et al., 2025) and Kimi (Moonshot-AI, 2025). We also include MedGemma-27b-text-it (Sellergren et al., 2026) as a representative model pre-trained on domain-specific corpora. Figure 5 shows that our method consistently outperforms baseline approaches across most backbone configurations, demonstrating the generalizability of our proposed framework.

4.7 Case Study

Figure 6 provides a case study. The patient’s record includes various diseases affecting different organs. Existing methods (e.g., MedReason) indiscriminately conduct coarse, patient-wise reasoning, resulting in vague rationales and insufficient utilization of knowledge. In contrast, our method partitions the patient’s record into CRUs, each centered on a major disease, allowing LLMs to provide fine-grained reasoning with clearly traceable rationales.

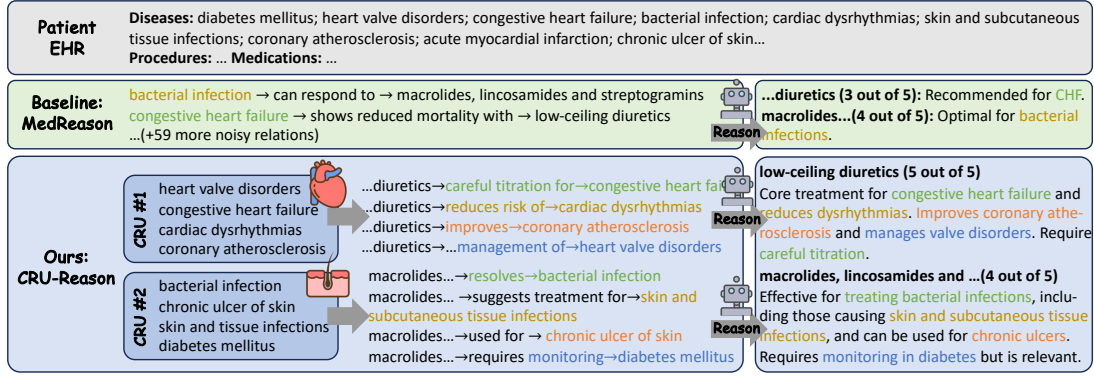


Figure 6: A case for drug recommendation. Text colors show the provenance of knowledge in rationales.

5 Conclusion

We introduce CRU-Reason, a framework that organizes structured EHR concepts and retrieved KG relations into disease-centered Clinical Rationale Units before LLM-based prediction. CRU-Reason achieved the strongest overall results among the evaluated EHR-prediction baselines on drug recommendation and diagnosis prediction, with improved cross-dataset robustness and rationale quality. This work advances clinical decision support by enabling structure-aware reasoning through unified knowledge and data-driven integration.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2025YFC2708700), Zhejiang Province Major Science and Technology Plan Project (2026C01021), the Central Government-Guided Fund for Local Science and Technology Development (2025ZY01034), the Provincial Key Laboratory Program (2025E10048), and Zhejiang Province Science and Technology Plan Project (2025C01128).

Limitations

Our study has several limitations. First, CRU-Reason is evaluated on structured, standardized concepts from MIMIC-III and MIMIC-IV. It does not address entity extraction from raw clinical notes or robustness to extraction errors. In addition, these ICU-centered datasets may not generalize to other institutions, populations, or care settings, and subgroup performance and calibration require further evaluation.

Second, we adopt the KARE knowledge graph without independently validating its factual completeness or accuracy. Errors in the KG or LLM-

based relevance filtering may propagate into predictions and rationales. In particular, a rationale may be plausible and knowledge-linked while still supporting an incorrect prediction.

Finally, CRU-Reason provides discrete scores rather than calibrated uncertainty estimates. Future work should study calibration, uncertainty quantification, rationale-faithfulness evaluation, and error detection.

Ethical Considerations

We use the de-identified MIMIC-III and MIMIC-IV datasets in accordance with the applicable PhysioNet credentialing and data-use requirements. No attempt is made to re-identify any individual. Three board-certified physicians volunteer to participate in the evaluation. Before participating, they are informed of the study purpose, evaluation procedure, expected workload, and intended use of their annotations. Model outputs are evaluated for research purposes only and are not intended to provide autonomous diagnosis, prescribing, or medical advice.

References

- Moumita Bhattacharya, Claudine Jurkowitz, and Hagit Shatkay. 2018. Co-occurrence of medical conditions: Exposing patterns through probabilistic topic modeling of snomed codes. *Journal of Biomedical Informatics*, 82:31–40.
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.
- Katherine E. Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X. Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A. Malin. 2025. Large language models are less effective at clinical prediction tasks than

- locally trained machine learning models. *J. Am. Medical Informatics Assoc.*, 32(5):811–822.
- Laura Cabello, Carmen Martin-Turrero, Uchenna Akujobi, Anders Søgaard, and Carlos Bobed. 2024. MEG: medical knowledge-augmented large language models for question answering. *CoRR*, abs/2411.03883.
- Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).
- Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. 2024. Predictive modeling with temporal graphical representation on electronic health records. In *IJCAI*, pages 5763–5771. ijcai.org.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025a. Towards medical complex reasoning with llms through medical verifiable problems. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pages 14552–14573. Association for Computational Linguistics.
- Wanyi Chen, Zihua Zhao, Jiangchao Yao, Ya Zhang, Jiajun Bu, and Haishuai Wang. 2025b. Multi-modal medical diagnosis via large-small model collaboration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30763–30773. IEEE.
- Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. ACL.
- Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, pages 3504–3512.
- Hejie Cui, Xinyu Fang, Ran Xu, Xuan Kan, Joyce C. Ho, and Carl Yang. 2025. Multimodal fusion of EHR in structures and semantics: Integrating clinical records and notes with hypergraph and LLM. In *MedInfo*, volume 329 of *Studies in Health Technology and Informatics*, pages 753–757. IOS Press.
- DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.
- Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M. Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *WWW*, pages 530–540. ACM / IW3C2.
- Yang Gao, Xiang Zhang, Zhongquan Sun, Payal Chandak, Jiajun Bu, and Haishuai Wang. 2025. Precision adverse drug reactions prediction with heterogeneous graph neural network. *Advanced Science*, 12(4):2404671.
- Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. 2024. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.
- Jiawei Guo, Yang Gao, Zihang Liu, Yihan Rao, Jiang Hu, and Haishuai Wang. 2025. Eaglecare: Augmenting icu monitoring with llms and knowledge graph reasoning. In *2025 IEEE International Conference on Knowledge Graph (ICKG)*, pages 130–137. IEEE.
- Pengfei Hu, Chang Lu, Fei Wang, and Yue Ning. 2024. Dualmar: Medical-augmented representation from dual-expertise perspectives. *CoRR*, abs/2410.19955.
- Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *CoRR*, abs/2504.00869.
- Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. In *COLING*, pages 9278–9298. Association for Computational Linguistics.
- Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2024. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *ICLR*. OpenReview.net.
- Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha A. Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *ICLR*. OpenReview.net.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2022. MIMIC-IV. *PhysioNet*. Version 2.0.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.
- Yan Kang, Mingjian Yang, Yue Peng, Jingwen Cai, Lei Zhao, Zhan Gao, Ningshu Li, and Bin Pu. 2025. LLM-DG: leveraging large language model for enhanced disease prediction via inter-patient and intra-patient modeling. *Inf. Fusion*, 121:103145.
- Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework

- with prompt-generated rationales. In *AAAI*, pages 18417–18425. AAAI Press.
- Feiyang Li, Yingjian Chen, Haoran Liu, Rui Yang, Han Yuan, Yuang Jiang, Tianxiao Li, Edison Marrese-Taylor, Hossein Rouhizadeh, Yusuke Iwasawa, Douglas Teodoro, Yutaka Matsuo, and Irene Li. 2025a. Mkg-rank: Enhancing large language models with knowledge graph for multilingual medical question answering. *CoRR*, abs/2503.16131.
- Mengxuan Li, Ke Liu, Jialong Guo, Jiajun Bu, Hongwei Wang, and Haishuai Wang. 2025b. Imputeinr: Time series imputation via implicit neural representations for disease diagnosis with missing data. *arXiv preprint arXiv:2505.10856*.
- Shiyu Li, Ziqi Yan, Zhihao Wu, Jielong Lu, Weiran Liao, Jiajun Yu, Genjie Li, Zeyu Chu, Jiajun Bu, and Haishuai Wang. 2026. Deepbd: A grounded agentic workflow for variant prioritization and diagnosis of genetic birth defects. *arXiv preprint arXiv:2606.24779*.
- Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *CoRR*, abs/2406.03712.
- Moonshot-AI. 2025. *Kimi-k2-instruct*.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and others. 2022. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005.
- Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. Deepr: A convolutional net for medical records. *IEEE J. Biomed. Health Informatics*, 21(1):22–30.
- OpenAI. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.
- OpenAI. 2025. *Introducing gpt-4.1 in the api*.
- Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, Yifei Shen, and Rex Ying. 2026. Experience retrieval-augmentation with electronic health records enables accurate discharge QA. In *ACL (1)*, pages 23406–23430. Association for Computational Linguistics.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *CoNLL*, pages 314–334. Association for Computational Linguistics.
- Bin Pu, Mingjian Yang, Yan Kang, Guanyuan Chen, and Pengchen Liang. 2026. Hierarchical knowledge fusion for enhanced health event prediction: Discriminating between frequent and new diseases. *Inf. Fusion*, 127:103741.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and others. 2026. *Medgemma technical report*. Preprint, arXiv:2507.05201.
- Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer. 2018. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(6):1968–1978.
- Qipeng Wang, Rui Sheng, Yafei Li, Huamin Qu, Yushi Sun, and Min Zhu. 2025a. Medkgi: Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. *CoRR*, abs/2512.24181.
- Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, and Yixuan Yuan. 2025b. Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications. *CoRR*, abs/2508.00669.
- Zixiang Wang, Yinghao Zhu, Junyi Gao, Xiaochen Zheng, Yuhui Zeng, Wen Tang, Yasha Wang, Ewen M Harrison, Chengwei Pan, Liantao Ma, and Ling Wang. 2024. *Retcare: Towards interpretable clinical decision making through LLM-driven medical knowledge retrieving*. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.
- Zixiang Wang, Yinghao Zhu, Huiyi Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen M. Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. 2025c. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *WWW*, pages 2250–2261. ACM.
- Chenrui Wu, Haishuai Wang, Xiang Zhang, Chengqi Zhang, and Jiajun Bu. 2025a. Efficient personalized adaptation for physiological signal foundation model. In *Forty-Second International Conference on Machine Learning*.
- Jialun Wu, Xin Mei, Rui Mao, Kai He, and Erik Cambria. 2026a. Takecare: A temporal-hierarchical framework with knowledge fusion for personalized clinical predictive modeling. *Inf. Fusion*, 126:103620.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin

- Cho, Changin Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025b. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *CoRR*, abs/2504.00993.
- Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025c. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *CoRR*, abs/2505.11733.
- Lei Wu, Hui Lin, Fengling Hu, Kaiying Shen, Wenjie Liang, Liwei Hu, Weilin Wang, Jiajun Bu, and Haishuai Wang. 2026b. Federated orthogonal learning for detection of liver lesions from multi-phase contrast-enhanced ct images. *npj Digital Medicine*.
- Muhao Xu, Zhenfeng Zhu, Youru Li, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. 2024a. Flexcare: Leveraging cross-task synergy for flexible multimodal healthcare prediction. In *KDD*, pages 3610–3620. ACM.
- Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce C. Ho, and Carl Yang. 2024b. RAM-EHR: retrieval augmentation meets clinical predictions on electronic health records. In *ACL (Short Papers)*, pages 754–765. Association for Computational Linguistics.
- Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, pages 4921–4929. ijcai.org.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and others. 2025. *Qwen3 technical report*. Preprint, arXiv:2505.09388.
- Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. 2023. Pyhealth: A deep learning toolkit for healthcare applications. In *KDD*, pages 5788–5789. ACM.
- Zhihao Yu, Yujie Jin, Yongxin Xu, Xu Chu, Yasha Wang, and Junfeng Zhao. 2024. Intellicare: Improving healthcare analysis with variance-controlled patient-level knowledge from large language models. *CoRR*, abs/2408.13073.
- Kai Zhang, Rui Zhu, Shutian Ma, Jingwei Xiong, Yejin Kim, Fabricio Murai, and Xiaozhong Liu. 2025. Kedrec-lm: A knowledge-distilled explainable drug recommendation large language model. *CoRR*, abs/2502.20350.
- Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025a. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *WWW*, pages 4442–4457. ACM.
- Yibowen Zhao, Yinan Zhang, Zhixiang Su, Lizhen Cui, and Chunyan Miao. 2025b. Beyond traditional diagnostics: Transforming patient-side information into predictive insights with knowledge graphs and prototypes. *CoRR*, abs/2512.08261.
- Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024a. EMERGE: enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *CIKM*, pages 3549–3559. ACM.
- Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M. Harrison, Liantao Ma, and Chengwei Pan. 2024b. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *CoRR*, abs/2402.01713.
- Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. Kg4diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program*, volume 281 of *Proceedings of Machine Learning Research*, pages 195–204. PMLR.

A Related Works

A.1 Machine Learning-Based EHR Prediction

Electronic Health Records (EHRs) demonstrate significant potential in enhancing clinical decision-making through AI-assisted support systems. Machine learning models have been proposed to learn statistical patterns from large-scale EHRs from scratch and provide data-driven predictions for clinical support (Choi et al., 2016; Chen et al., 2024; Pu et al., 2026; Wu et al., 2026a; Xu et al., 2023, 2024a). Despite these methods providing direct alignment between input EHR features and downstream outcomes, they often suffer from limited semantic understanding and struggle to leverage external medical knowledge.

A.2 LLM-Augmented EHR Prediction

To address these limitations, recent studies integrate LLMs as knowledge bases to enrich medical knowledge (Jiang et al., 2024) or as encoders to refine EHR representations (Xu et al., 2023, 2024b; Hu et al., 2024; Cui et al., 2025; Zhu et al., 2024a; Yu et al., 2024). However, these methods still largely adhere to a data-driven paradigm. They typically lack the ability to provide human-readable reasoning chains, limiting their interpretability and trustworthiness in high-stakes applications.

Recent advances in LLMs (DeepSeek-AI, 2024; Yang et al., 2025; OpenAI, 2024) have substantially advanced reasoning-based clinical applications. Studies (Li et al., 2026; Wu et al., 2025a; Chen et al., 2025b) prompt LLMs to directly make clinical decisions and generate human-readable rationales for their predictions. A complementary line of work improves KG-free medical reasoning through verifiable medical problems, test-time scaling, and reasoning-oriented prompting or training (Chen et al., 2025a; Huang et al., 2025; Wang et al., 2025b). These approaches primarily strengthen the internal reasoning capabilities of LLMs. In contrast, our work investigates how external, structured medical knowledge can be selectively organized and grounded in structured EHR prediction. Thus, CRU-Reason is complementary to, rather than a direct replacement for, KG-free medical reasoning approaches. Nevertheless, LLM-generated rationales can be plausible but incorrect (Pal et al., 2023), motivating explicit grounding and careful clinical oversight.

A.3 Structural Knowledge-Enhanced EHR Prediction

To ground LLM reasoning in reliable and up-to-date medical expertise, retrieval-augmented generation (RAG) has been used to enhance reasoning by retrieving relevant knowledge from various sources (Zhu et al., 2024b; Wang et al., 2024, 2025c; Zhao et al., 2025b; Wang et al., 2025a). Other methods utilize medical knowledge graphs (KGs) to facilitate structured reasoning by retrieving triplets (Zhang et al., 2025), paths (Jia et al., 2025; Wu et al., 2025b), or subgraphs (Zhao et al., 2025a; Jiang et al., 2025; Zuo et al., 2025) to better understand complex relationships between medical concepts. However, a remaining challenge is that generic LLMs often lack task-specific optimization. To bridge this gap, fine-tuning enables LLMs to refine their understanding of inputs and develop task-appropriate reasoning strategies (Brown et al., 2025; Kwon et al., 2024; Wu et al., 2025b,c). Nevertheless, these fine-tuning-based methods are computationally expensive and can reduce the model’s generalizability to out-of-distribution scenarios. This motivates our work to develop a novel reasoning framework that is aware of the clinical structures of EHRs. Compared with representative methods with holistic patient-level graph aggregation (Jiang et al., 2024; Wu et al., 2025b; Jiang et al., 2025), our method enables fine-grained clinical reasoning with disease-centric communities and rationales. More specifically, compared with KARE (Jiang et al., 2025), which retrieves standalone communities directly from KGs with statistical features, our CRUs originate from the patient’s records with minimal noise and stronger relevance.

B Implementation Details

Datasets. We conduct experiments on two publicly available clinical EHR datasets, MIMIC-III (v1.4, Johnson et al., 2016) and MIMIC-IV (v2.0, Johnson et al., 2022, and follow Jiang et al. (2024) for preprocessing. Clinical Classification Software (CCS) is used for disease and procedure mappings. Anatomical Therapeutic Chemical classification system (ATC-3) is used for medication mappings. For the drug recommendation task, to avoid information leakage, we only use historical diseases and procedures to predict medications. On the MIMIC-IV dataset, we randomly downsample (seed=42) to 40,000 samples for drug recommendation and

Split	MIMIC-III						MIMIC-IV					
	#Patient	#Sample	#Visit	#D	#P	#M	#Patient	#Sample	#Visit	#D	#P	#M
Drug Recommendation												
Train	28,565	35,578	1.38	12.36	4.41	29.85	26,895	31,907	2.10	15.66	3.97	28.84
Validation	3,571	4,346	1.32	12.07	4.30	29.90	3,407	4,025	2.10	15.48	3.92	28.85
Test	783	1,000	1.39	10.77	4.56	29.11	968	1,000	2.08	16.03	4.04	29.00
Diagnosis Prediction												
Train	4,931	7,814	1.92	16.54	5.11	37.12	7,245	8056	2.41	17.90	4.34	35.02
Validation	616	896	1.63	15.06	4.84	34.90	845	934	2.49	18.28	4.49	34.67
Test	617	983	1.77	15.49	4.93	37.35	908	1,000	2.43	17.88	4.42	35.47

Table 4: Dataset statistics. #Visit: average number of visits per sample. #D, #P, #M: average number of diseases, procedures, and medications per sample.

10,000 samples for diagnosis prediction. Then we split the selected samples into three sets: 80% as the training set, 10% as the validation set, and at most 1,000 samples from the remaining as the test set, and ensure that samples from the same patient are allocated to the same subset to prevent data leakage. The demographic features are removed for all methods for fairness, following Jiang et al. (2025). Table 4 shows the statistics of the datasets.

Knowledge Graphs. The knowledge graph is directly adopted from the public official code of Jiang et al. (2025). We obtain the vectorized representations of all entities, relations, and triplets with text-embedding-3-large (Neelakantan et al., 2022) with a dimension of 1,024. We do not claim an independent validation of the KG’s factual completeness or accuracy. Assessing how KG errors propagate through clinical reasoning remains important future work.

Baselines. For ML models, we use off-the-shelf implementations from the `pyhealth` package (Yang et al., 2023). All models are trained on the full training set with a hidden state of 128 for 50 epochs and an early-stop patience of 10. The learning rate and weight decay are both $1e-5$. Results of ML models are averaged over 10 runs with different random seeds.

For LLM-only baselines, we adapt prompts from Jiang et al. (2025) to drug recommendation and diagnosis prediction tasks with the following structure: task description, key considerations, patient’s record, chain-of-thought instruction, and output format instruction. For all KG+LLM baselines and RetCare, we insert retrieved knowledge from KGs into the same prompt. Results of LLM-related methods are obtained from a single run with temperature=0 for cost considerations.

Some related works are excluded due to a lack of publicly available codes (Kang et al., 2025; Wu

et al., 2026a; Zhang et al., 2025), involving complex agent systems (Wang et al., 2025c), or designed for multimodal data (Xu et al., 2023, 2024a).

LLMs. We employ DeepSeek-V3 (DeepSeek-AI, 2024) as our primary backbone model. For LLM-only methods, we also use Qwen3-8B, Qwen3-235B-Instruct, Qwen3-235B-Thinking (Yang et al., 2025), Kimi-K2-Instruct (Moonshot-AI, 2025), GPT4.1-mini (OpenAI, 2025), DeepSeek-R1-0528 (DeepSeek-AI, 2025) and include fine-tuned Qwen3-8B models distilled from DeepSeek-V3. For LLMs, we deploy them (except GPT) on local servers with full precision and greedy decoding with a temperature of 0. Fine-tuned Qwen3-8B models are obtained by 1) distilling a training corpus with 8,000 samples from DeepSeek-V3 given ground truth labels, 2) fine-tuning full parameters with DeepSpeed-Zero3 for 1,000 steps, and 3) selecting the best per-200-step checkpoint with the lowest loss on a 500-sample validation set.

Software and Hardware. Experiments are conducted on Linux machines with Python 3.10.15, PyTorch 2.4.0, Transformers 4.51.3, TRL 0.20.0, DeepSpeed 0.16.2, and CUDA 12.4, equipped with Intel Xeon Platinum 8369B CPUs and NVIDIA A100 80G GPUs.

Hyper-parameters. During the target discovery process, we use a threshold of 0.4 to obtain ML candidates, and set $L = 1$ to obtain KG candidates. During the CRU construction process, we set the maximum exploration steps of the BFS strategy to 3 for the drug recommendation task and 2 for the diagnosis prediction task. During the CRU-wise reasoning process, we rank the KG-derived candidates by their scores given by ML models, and select only the top 60 candidates to evaluate. The top 2 ML candidates are used as few-shot examples for each KG candidate. During the LLM

Query for Neighbor Constraint Scoring with LLMs
Instruction Determine whether the concepts are related to the same disease or body part and should be treated as a single disease. Please provide a brief reasoning. Then, provide your prediction with "YES" or "NO".
Example Concept: Hypertension Candidate: Coronary artery disease Relation: Coronary artery disease $\rightarrow$ can occur with $\rightarrow$ Hypertension Answer: Hypertension and coronary artery disease are both cardiovascular conditions, affecting the same body system (heart and blood vessels). YES. ...
Input Concept: {current node u } Candidate: {neighbor v } Relation: {triplet in " $u \rightarrow r \rightarrow v$ " format}
Output Answer: [Reasoning then YES or NO]

Table 5: Prompt for Neighbor Constraint.

reasoning processes of both LLM-based baseline methods and CRU-Reason, we use batched inference for the candidates with a batch size of 20, i.e., simultaneously sending 20 candidate targets to the model and obtaining their rationales in a single generation. During the aggregation process, for a candidate that appears in multiple CRUs, we average its CRU-specific scores to obtain the final prediction score.

Threshold Decision. Since F1 and Jaccard metrics are threshold-sensitive, we individually search for the optimal threshold for each method for a fair comparison. For LLM-based methods with discrete outputs, the search range is $\{0, 1, 2, 3, 4\}$ out of 5. For ML methods with continuous outputs, the search range is $\{0.0, 0.2, 0.4, 0.6, 0.8\}$ out of 1. Candidate targets with scores equal to or higher than the threshold are regarded as positive predictions, while scores lower than the threshold are negative predictions.

Impact of Locality To compare the discriminative ability between isolated and localized concepts for drug recommendation in Figure 2(c), we select the top-10,000 four-concept subgraphs appearing most frequently in the training set, denoted as $\mathbb{G}$.

For each subgraph $G \in \mathbb{G}$, we count for its probability w.r.t. a candidate c as $P_{joint}(G, c) = \frac{N(c \in M)}{N(G \subset X)}$, where $N(\text{condition}) = \sum_{(X, M) \in \mathbf{X}_{\text{train}}} \mathbb{I}(\text{condition})$ counts for the frequency when the condition is true within the training set. X is the patient’s EHR, and M is the ground-truth. We also count for the average probability of individual concepts within

subgraph G w.r.t. a candidate c as $P_{id}(G, c) = \frac{1}{\|\mathbb{G}\|} \sum_{e \in G} \frac{N(c \in M)}{N(e \in X)}$. Finally, we compare the average probability of individual concepts $P_{id}(G, c)$ and the probability of joint concepts $P_{joint}(G, c)$. Specifically, the number of improved selections of positive candidates is defined as $N(c \in M, P_{joint}(G, c) > P_{id}(G, c), G \in \mathbb{G}, G \subset X)$ and the number of harmed selections is defined as $N(c \in M, P_{joint}(G, c) < P_{id}(G, c), G \in \mathbb{G}, G \subset X)$. The number of improved/harmed rejections of negative candidates is defined respectively as $N(c \notin M, P_{joint}(G, c) < P_{id}(G, c), G \in \mathbb{G}, G \subset X)$ and $N(c \notin M, P_{joint}(G, c) > P_{id}(G, c), G \in \mathbb{G}, G \subset X)$. In experiments, the subgraphs’ probabilities are higher in 48.2M/71.6M comparisons when the candidate is included in ground-truths, indicating its better capability in predicting true-positives. When the candidate is excluded in ground-truths, the subgraphs’ probabilities are lower in 294.4M/400.4M comparisons, indicating its better capability in predicting true-negatives.

B.1 CRU Construction

Disease Ranking The disease ranking of the CRU construction process (Section 3.3) includes three steps. (1) Pair Frequency Analysis: for each disease in the training data, we count how frequently it co-occurs with every possible output target (medication or disease). (2) Entropy Calculation: for each disease, we compute the entropy of its associated target distribution. Lower entropy indicates that the disease predominantly predicts a limited set of targets, demonstrating stronger indicative power. (3) Sorting based on Entropy: we sort all diseases according to the entropy in ascending order.

Prompts for Constraint We use the prompt in Table 5 to utilize LLMs to evaluate the clinical relevance in Section 3.3. Notably, we cache the results to avoid excessive LLM calls.

C Supplementary Results

C.1 Inference Cost

We compare the inference cost of LLM-involved methods using the average number of LLM calls, prompt tokens, and completion tokens per patient case. Table 8 shows that CRU-Reason incurs higher overhead than lightweight prompting and several retrieval-based baselines because it evaluates candidates within multiple CRUs. This effect is more

Query for Clinical Rationale Unit Summarization
Task: Patient Overview Go over the key information in the patient’s context, and consider the relationship among conditions and procedures in the context. Summarize the patient’s context in one paragraph. ## Patient’s Context ### Conditions chronic kidney disease fluid and electrolyte disorders mycoses ### Procedures hemodialysis ### Medications anxiolytics drugs for constipation antiinfectives and antiseptics, excl. combinations with corticosteroids ... ### Relations among Concepts mycoses → treatedwith → intestinal antiinfectives mycoses → are treatable with → antiinfectives and antiseptics, excl. combinations with corticosteroids mycoses → treatedby → antiepileptics mycoses → could improve prognosis for → antithrombotic agents ... # Output Format Patient Overview: [Your output]
Response
A patient with chronic kidney disease on maintenance hemodialysis is being managed for fluid/electrolyte disorders and mycoses. The antifungal treatment plan includes intestinal anti-infectives, systemic anti-infectives/antiseptics, and possibly antiepileptic agents, while additional medications cover dialysis-related needs (antithrombotics), gastrointestinal protection (PUD/GORD drugs ...

Table 6: An example of CRU summarization.

Variants	F1	Jaccard	Rank
MIMIC-III: Drug Recommendation			
Full*	0.657	0.506	1.639
w/o rel	0.650	0.481	1.649
w/o meta	0.651	0.486	1.649
w/ BFS-1	0.629	0.467	1.630
w/ BFS-2	0.640	0.489	1.642
w/o CNST*	0.648	0.494	1.679
MIMIC-III: Diagnosis Prediction			
Full*	0.397	0.260	2.047
w/o rel	0.390	0.254	2.098
w/o meta	0.394	0.255	2.051
w/ BFS-1	0.392	0.254	2.053
w/o CNST*	0.385	0.252	2.016

Table 7: Extended ablation results on MIMIC-III. Asterisk (*) refers to results copied from Table 3.

pronounced for drug recommendation, where each CRU contains more candidates on average (51.29) than for diagnosis prediction (18.85).

However, CRU-Reason is substantially less expensive than candidate-wise methods such as MK-GRank and MedIKAL in several settings. On an 8×NVIDIA A100 80GB system with vLLM acceleration, our implementation processes a patient case in under 10 seconds on average. These costs remain material for deployment.

C.2 Additional Ablation Studies

We expand our ablation study to systematically evaluate the CRU-construction design choices, in-

cluding: *w/o rel* (removing LLM-based edge-relevance filtering), *w/o meta* (removing metapath restrictions), *w/ BFS-1* and *w/ BFS-2* (restrict the maximum BFS depth to one and two hops, respectively).

The results in Table 7 show that on MIMIC-III drug recommendation, removing relevance scoring or metapath constraints reduces F1 by 1.1% and 0.9%, respectively, and limiting exploration to one hop reduces F1 by 3.8%. This indicates that over-fragmented CRUs fail to preserve coherent disease-centric context. The extended results also characterize sensitivity to CRU granularity: overly shallow exploration fragments evidence, whereas removing constraints weakens clinical coherence.

D Data Availability

The MIMIC-III dataset is available in <https://physionet.org/content/mimiciii/1.4/>. The MIMIC-IV dataset is available in <https://physionet.org/content/mimiciv/2.0/>. The source codes are available in <https://github.com/lzhmarkk/CRU-Reason>.

E Examples

Table 6 shows an example for CRU summarization. Table 9 shows an example for ML-guided reasoning on the diagnosis prediction task.

Method	MIMIC-III					MIMIC-IV				
	# Calls	# Prompt	# Completion	Δ_{prompt}	$\Delta_{\text{completion}}$	# Calls	# Prompt	# Completion	Δ_{prompt}	$\Delta_{\text{completion}}$
Drug Recommendation										
Prompt	10.0	16,468	12,606	-57.0%	-29.0%	11.0	25,308	13,223	-59.3%	-57.2%
KARE	7.1	42,742	9,478	11.5%	-46.6%	7.5	51,507	9,506	-17.2%	-69.2%
MedRAG	8.5	22,069	10,233	-42.4%	-42.4%	8.8	28,922	10,469	-53.5%	-66.1%
MKGRank	177.3	58,924	97,130	53.7%	446.9%	182.9	61,883	101,154	-0.5%	227.4%
MedReason	8.5	15,234	9,803	-60.3%	-44.8%	8.8	17,499	10,097	-71.9%	-67.3%
MedIKAL	134.7	544,200	11,421	1,319.8%	-35.7%	143.4	608,120	13,934	877.7%	-54.9%
RetCare	17.0	33,718	45,209	-12.0%	154.5%	17.6	48,644	60,354	-21.8%	95.4%
CRU-Reason	31.1	38,328	17,761	—	—	57.0	62,200	30,895	—	—
Diagnosis Prediction										
Prompt	8.7	15,751	9,810	-65.8%	-13.0%	8.7	23,342	9,989	-63.3%	-36.8%
KARE	8.7	66,042	11,589	43.4%	2.8%	8.7	70,540	11,492	10.8%	-27.3%
MedRAG	12.7	53,246	15,538	15.6%	37.8%	12.8	60,488	15,518	-5.0%	-1.9%
MKGRank	269.2	95,560	152,472	107.6%	1,252.2%	273.2	99,754	157,138	56.6%	893.7%
MedReason	12.7	111,926	15,088	143.1%	33.8%	12.8	98,510	15,197	54.7%	-3.9%
MedIKAL	164.3	550,280	19,659	1,095.2%	74.3%	166.3	615,153	19,311	865.9%	22.1%
RetCare	25.4	94,910	112,156	106.1%	894.6%	25.5	108,251	125,613	70.0%	694.4%
CRU-Reason	25.4	46,041	11,276	—	—	37.4	63,689	15,813	—	—

Table 8: Inference cost of LLM-involved methods. # Calls, # Prompt, and # Completion denote the average number of LLM calls, prompt tokens, and completion tokens per sample, respectively. Δ is computed relative to CRU-Reason within the same dataset and task.

Query for ML-guided LLM Reasoning on Diagnosis Prediction	
# Task: Next-visit Diseases Prediction Predict the likelihood of specific diseases or health conditions that may manifest or progress in the patient’s next medical visit based on their current health status. Use available patient data to assess risk factors and early warning signs. Scores: 0 (Very Unlikely), 1 (Unlikely), 2 (Possible), 3 (Likely), 4 (Very Likely) ## Key Considerations Conditions: - Existing Diagnoses: Current diseases/conditions that may worsen or trigger complications (e.g., diabetes, increasing infection risk)... thyroid dysfunction). Procedures: - Recent Interventions: Surgeries or treatments that may lead to complications (e.g., post-op infection risk, stent thrombosis).. Medications: - Adverse Effects: Drugs known to cause conditions (e.g., steroids → hyperglycemia, PPIs → C. diff infection).. ## Patient’s Context A patient with chronic kidney disease on maintenance hemodialysis is being managed for fluid/electrolyte disorders and mycoses.. ## Recommended Diseases {.. "fluid and electrolyte disorders": ["antiinfectives and antiseptics, excl. combinations with corticosteroids → can cause → fluid and electrolyte disorders", "fluid and electrolyte disorders → may require → iron antianemic preparations",],.. }] # Instructions Justify Recommended Diseases (the keys). These diseases are generated by a trustworthy model and are precisely recommended to the patient. Consider the relevance of each disease to the patient’s medical conditions, and predict scores from 0 to 4 in JSON format.. # Output Format Justify Each Recommended Disease: [Your output in JSON]	
Response	
{.. "fluid and electrolyte disorders": { "reason": "End-stage kidney disease eliminates renal water/solute clearance, making hyperkalemia, hyponatremia, hypocalcemia, hyperphosphatemia, and volume overload are daily risks. Current antifungals (amphotericin B, azoles) and repeated dialysis sessions provoke shifts in K^+ , Mg^{2+} and Na^+ ; constipation agents and opioids can further elevate potassium. Inter-current infection or missed dialysis days can tip the balance. These disorders are therefore the single most common reason for unscheduled visits or admission before the next routine appointment.", "predict": 4 },.. } }	

Table 9: An example of ML-guided reasoning on the diagnosis prediction task.