EviCare: Enhancing Diagnosis Prediction with Deep Model-Guided Evidence for In-Context Reasoning

Hengyu Zhang
hengyu.zhang3@hdr.mq.edu.au
Macquarie University
Sydney, Australia

Xuyun Zhang*
xuyun.zhang@mq.edu.au
Macquarie University
Sydney, Australia

Pengxiang Zhan
yyytdms@gmail.com
Fuzhou University
Fuzhou, China

Linhao Luo
Linhao.Luo@monash.edu
Monash University
Melbourne, Australia

Hang Lv
lvhangkenn@gmail.com
Fuzhou University
Fuzhou, China

Yanchao Tan
yctan@fzu.edu.com
Fuzhou University
Fuzhou, China

Shirui Pan
s.pan@griffith.edu.au
Griffith University
Brisbane, Australia

Carl Yang
j.carlyang@emory.edu
Emory University
Atlanta, USA

Abstract

Recent advances in large language models (LLMs) have enabled promising progress in diagnosis prediction from electronic health records (EHRs). However, existing LLM-based approaches tend to overfit to historically observed diagnoses, often overlooking novel yet clinically important conditions that are critical for early intervention. To address this, we propose EviCare, an in-context reasoning framework that integrates deep model guidance into LLM-based diagnosis prediction. Rather than prompting LLMs directly with raw EHR inputs, EviCare performs (1) deep model inference for candidate selection, (2) evidential prioritization for set-based EHRs, and (3) relational evidence construction for novel diagnosis prediction. These signals are then composed into an adaptive in-context prompt to guide LLM reasoning in an accurate and interpretable manner. Extensive experiments on two real-world EHR benchmarks (MIMIC-III and MIMIC-IV) demonstrate that EviCare [1] achieves significant performance gains, which consistently outperforms both LLM-only and deep model-only baselines by an average of 20.65% across precision and accuracy metrics. The improvements are particularly notable in challenging novel diagnosis prediction, yielding average improvements of 30.97%.

CCS Concepts

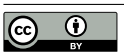
• **Applied computing** → **Life and medical sciences**;

Keywords

Diagnosis Prediction, Large Language Model, In-context Reasoning, Deep Model-Guided Evidence

*Xuyun Zhang is the corresponding author.

[1] https://github.com/zhyccc/EviCare



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3780257

ACM Reference Format:
Hengyu Zhang, Xuyun Zhang, Pengxiang Zhan, Linhao Luo, Hang Lv, Yanchao Tan, Shirui Pan, and Carl Yang. 2026. EviCare: Enhancing Diagnosis Prediction with Deep Model-Guided Evidence for In-Context Reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770854.3780257

1 Introduction

Accurate diagnosis prediction from electronic health records (EHRs) plays a vital role in clinical decision-making, enabling early detection of complications and personalized treatment planning [2, 5]. With growing access to longitudinal EHR data, recent studies have begun exploring large language models (LLMs) for diagnosis prediction [4, 18, 29, 34]. These approaches leverage LLMs' biomedical knowledge and reasoning capabilities to support medical decision-making. For example, Kwon et al. [18] propose a reasoning-aware prompting strategy to generate interpretable diagnostic rationales. DearLLM [34] enhances diagnosis prediction by extracting feature correlations deduced by LLMs to support personalized care.

Despite promising results for LLM-based diagnosis prediction, a critical unmet challenge remains: accurately predicting novel diagnoses, namely those not previously recorded in a patient's history but clinically relevant for proactive care. These diagnoses often indicate disease progression, new comorbidities, or acute deterioration, making them highly valuable for early intervention. However, the task requires temporal reasoning over patient-specific visit sequences that vary widely across individuals.

As illustrated in Figure 1(a), LLMs often repeat previously observed diagnoses. For instance, predicting CHD (Coagulation and Hemorrhagic Disorders) merely because it appeared in earlier visits, even though it is typically an acute condition unlikely to persist. This behavior arises not from hallucination, but from the LLM's next-token prediction objective [26], which favors contextually familiar outputs. While LLMs generate a mix of historical and novel predictions (Figure 1(d)), the majority are historical. The further demo experiment shows that they are accurate primarily on those

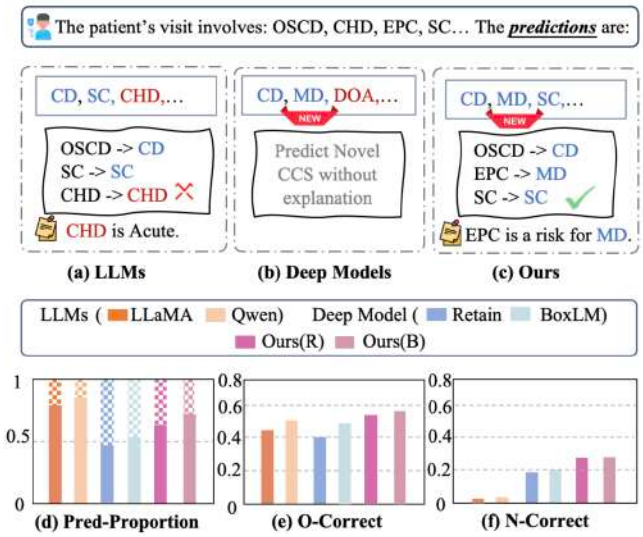


Figure 1: Diagnosis Prediction Patterns Across Models. (a) LLMs tend to predict repeating historical diagnoses (e.g., CHD); (b) Deep models are effective at predicting novel diagnoses (e.g., MD) but lack explanation; (c) Our method integrates deep model-guided evidence into LLMs to achieve accurate and explainable predictions (e.g., EPC → MD). (d) Proportion of predictions: historical (solid bars) vs. novel (hatched bars); (e) Overall prediction performance (Precision@10, O-Correct); (f) Performance on novel diagnoses (Precision@10, N-Correct).

repeated diagnoses (Figure 1(e)), and fail on novel ones (Figure 1(f)), exposing a clear limitation in discovery-oriented reasoning.

In contrast, deep models (e.g., Retain [7] and BoxLM [28]) trained on EHR data better capture temporal and structured patterns, enabling detection of previously unseen conditions and outperforming LLMs on novel diagnosis prediction. As shown in Figure 1(b), these models successfully surface MD (Mood Disorders) and DOA (Deficiency and Other Anemia) given relevant comorbid signals like EPC (Epilepsy and Convulsions). However, these deep models are fundamentally restricted to categorical outputs. While they can effectively identify what the likely diagnosis is based on latent patterns, they lack the generative capability to articulate the clinical reasoning process (e.g, explaining why lead to the prediction), which is essential for verifiable clinical decision-making.

To address these complementary limitations, we propose `EviCare`, a hybrid reasoning framework that transforms numerical signals from deep models into structured clinical reasoning. Rather than relying solely on LLMs to reason over raw EHR inputs, we extract structured evidence from a deep model and embed it into an adaptive in-context prompt to guide LLM reasoning.

Specifically, `EviCare` incorporates four core evidential components: (1) Deep model inference for candidate selection, which identifies top-K diagnoses covering both historical and novel using prediction logits from a deep model; (2) Evidential prioritization for set-based EHRs, which ranks historical diagnoses based on their contribution to the deep model's prediction, transforming unordered EHR codes into weighted organized inputs; (3) Relational evidence construction for novel prediction, which builds symbolic links between historical and candidate diagnoses by extracting co-occurrence and ontology-based relations from large-scale EHRs, helping justify novel predictions; (4) LLM-based diagnosis via evidence-driven in-context learning, which composes the above components into a structured prompt that guides the LLM to perform accurate and interpretable predictions. As shown in Figure 1(c), (e), and (f), `EviCare` improves both novel and overall diagnosis accuracy, and provides interpretable reasoning.

We evaluate `EviCare` on two real-world benchmarks (MIMIC-III and MIMIC-IV) and under both overall and novel prediction settings. Experiments demonstrate that `EviCare` consistently outperforms LLM-only and deep-only baselines, particularly in challenging novel diagnosis scenarios, while offering transparent and evidence-grounded outputs.

Our contributions are summarized as follows:

- *Identification of LLM Limitations in Diagnosis Prediction.* We reveal key limitations of LLMs in EHR-based diagnosis, including a tendency toward repeating historical conditions and poor performance on clinically important novel diagnoses.
- *A Deep Model-Guided Reasoning Paradigm.* We introduce `EviCare`, a novel hybrid reasoning framework that bridges deep EHR models with LLMs via structured evidence-guided prompting. This paradigm enables LLMs to reason beyond surface patterns by leveraging deep model-generated signals through three aligned mechanisms: Candidate Selection, Evidential Prioritization, and Relational Evidence Construction.
- *Comprehensive Evaluation on Real-World EHRs.* We validate the proposed `EviCare` on two large-scale clinical datasets (MIMIC-III and MIMIC-IV), showing consistent improvements over LLM-only and deep-only baselines. Our approach significantly enhances novel diagnosis prediction while maintaining interpretability and generalizability under limited supervision.

2 Related Work

2.1 Diagnosis Prediction with Deep Models

Diagnosis prediction from electronic health records (EHRs) has become a central problem in clinical informatics, essential for early detection and personalized care [1, 11, 15, 24]. Existing deep learning approaches can be broadly categorized into two categories:

The first line of work focuses on capturing temporal and contextual dependencies across patient visits using neural sequence models. For instance, RETAIN [7] and Dipole [23] encode visit sequences to learn time-aware patient embeddings, while Transformer-based methods [3] enhance the modeling of long-range dependencies and clinical state transitions via self-attention mechanisms.

The second line of work augments patient representations using structured medical knowledge. For example, GRAM [6] incorporated hierarchical relations from clinical ontologies via an attention mechanism. CGL [20] jointly modeled ontology and co-occurrence signals through graph learning. SeqCare [33] leveraged personalized knowledge graphs and label dependencies to suppress task-irrelevant noise. More recently, geometry-inspired models such as BoxCare [22] and BoxLM [28] encoded medical concepts as high-dimensional boxes to capture inclusion hierarchies and semantic

proximity. Despite their accuracy, these models operated in latent spaces without offering explicit reasoning, hindering their integration with LLMs for interpretable clinical decision-making.

2.2 Large Language Models for Healthcare

Recent years have witnessed the rapid progress of large language models (LLMs) in the medical domain [25, 27]. Foundation models like GPT-4 have demonstrated impressive performance in medical question answering, radiology interpretation, and differential diagnosis [37, 42], with some specialized LLMs (e.g., AMIE [39]) even surpassing the diagnostic accuracy of primary care physicians. These models not only extract clinical evidence from text but also support clinical reasoning by generating fluent and coherent diagnostic rationales [38].

To enhance LLMs' clinical utility, two major research directions have emerged. The first is tool-augmented prompting, where external medical knowledge or similar patient records are retrieved and provided as context for LLMs [14]. In this way, LLMs are empowered by retrieval [41], search APIs [37], or agent-based systems [40] to incorporate up-to-date clinical knowledge. Recent work such as KARE [13] further extends this line by organizing large-scale medical knowledge graphs into community-level structures, enabling more focused retrieval to support LLM-based healthcare prediction and mitigate hallucinations. The second focuses on instruction tuning, where LLMs are adapted to the medical domain using collected datasets from biomedical literature, knowledge bases, or self-generated rationales [27, 30]. These approaches significantly improve LLMs' generalization in medical tasks, including diagnosis, prognosis, and treatment planning.

However, existing LLM-based methods often overlook the temporal structure and symbolic relationships embedded in EHRs. While recent approaches like DearLLM [34] and CKLE [36] have attempted to incorporate code frequency, co-occurrence patterns, and ontology-based signals into LLM reasoning, they still fall short in effectively capturing patient-specific relational context. Related efforts also explore integrating structured knowledge with EHRs, such as GraphCare [12], which constructs personalized patient knowledge graphs using LLMs and external biomedical KGs for downstream prediction, and InKrat [19], which leverages cross-modal semantic retrieval and LLM-generated explanations to enhance interpretability. As a result, LLMs tend to treat diagnosis prediction as a sequence generation task, leading to over-repetition of historical codes and limited recognition of novel conditions. These limitations underscore the need for a framework that grounds LLM reasoning in structured, clinically meaningful evidence derived from the EHR itself.

3 THE EviCare FRAMEWORK

3.1 Problem Setup and Framework Overview

We consider the task of diagnosis prediction from longitudinal electronic health records (EHRs). Each patient p is represented by a sequence of historical visits $\mathcal{V}_p = \{v_1, v_2, \ldots, v_t\}$, Where each visit v_i is associated with a set of diagnosis codes drawn from either the ICD (*International Classification of Diseases*) or CCS (*Clinical Classifications Software*) vocabulary. Since ICD represents specific diagnoses, while CCS further groups ICD codes, we involve both of them to encode hierarchical and complementary semantics.

Given this visit history $\mathcal{V}_p$, the goal is to predict the relevant CCS codes $\hat{C}_{t+1}$ for the patient's next visit v_{t+1}. We formulate two sub-tasks:

(1) **Overall diagnosis prediction**, which includes all clinically relevant CCS codes regardless of whether they have appeared in history; (2) **Novel diagnosis prediction**, which filters out previously observed diagnoses to assess the model's capacity for discovering unseen yet clinically consistent outcomes. Both tasks share the same EHR input, but differ in label construction.

We summarize the main modules of the EviCare framework in Figure 2 to provide an overview. The first component, *Deep Model-Guided Candidate Selection*, uses a deep model (e.g., BoxLM or RETAIN) to identify the top-K most probable diagnosis codes, grounding the LLM's reasoning within a clinically relevant candidate space. The second component, *Historical Evidence Prioritization*, ranks historical diagnoses based on their predictive contribution as determined by the deep model, transforming unordered EHR codes into a weighted and structured representation. The third component, *Relational Evidence Construction*, builds symbolic links between historical and candidate diagnoses using co-occurrence statistics and ontology hierarchies, offering interpretable justification for novel predictions. The fourth component, *LLM-based Diagnosis via Evidence-driven In-Context Learning*, composes all evidence into a structured prompt that guides the LLM toward accurate and explainable diagnostic reasoning.

3.2 Deep Model Inference for Candidate Selection

As shown in Figure 1, while LLMs exhibit strong reasoning capabilities, they often struggle to make clinically grounded predictions when faced with structured EHR inputs, frequently overemphasizing previously seen diagnoses. In contrast, deep models trained on EHR data provide a more reliable estimation of diagnosis probabilities by learning task-specific relational patterns. However, such models typically lack interpretability.

To combine the strengths of both paradigms, we propose to leverage deep model prediction logits as a compact and model-agnostic representation of diagnosis likelihood. These logits enable us to construct a candidate diagnosis set C_p^{cand} by selecting the top-K most probable CCS codes for each patient visit. Rather than being a heuristic filter, this design is theoretically motivated by the principle that the ranked logits list from deep models encodes a structured approximation of the underlying information distribution. Let $y_{p,c}$ be the deep model logit for patient p and candidate diagnosis c. We formalize the theoretical guarantees as follows:

Lemma 3.1 (Ranked-Logit Information Prior and Denoising). *The ranked-logit prior establishes the following properties:*

1. *Conditioning on C_p^{cand} reduces the conditional entropy $H(C \mid X_p)$ and increases mutual information, thereby effectively denoising the reasoning space;*
2. *The Top-K selection by posterior probability is Bayes-optimal under a K-budget constraint [10];*
3. *Using rank-based weights as priors does not increase Bayes risk compared to flat prompting.*

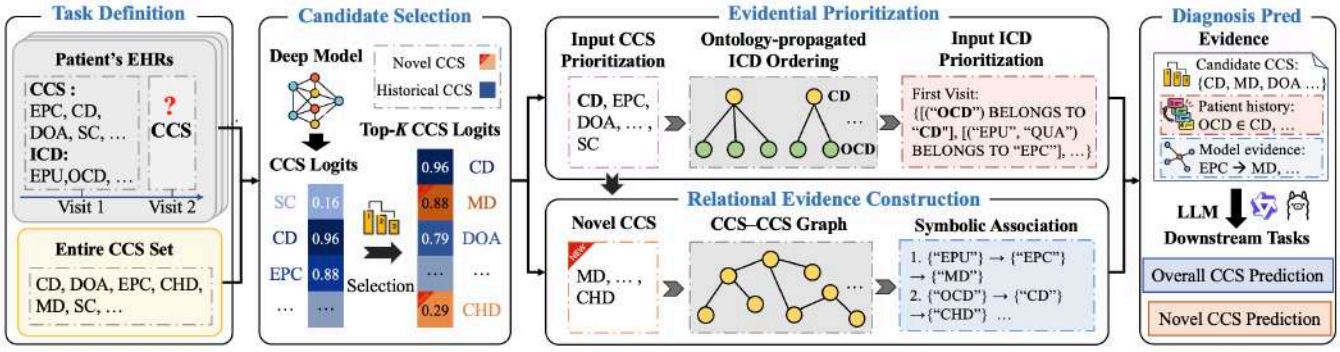


Figure 2: The overall framework of our proposed EviCare, which consists of (1) deep model inference for Candidate Selection, (2) Evidential Prioritization for set-based EHRs, (3) Relational Evidence Construction for novel diagnosis prediction, and (4) LLM-based Diagnosis Prediction via evidence-driven in-context learning.

This analysis provides a theoretical justification for using ranked logits as an information prior, explaining why constraining the candidate space improves robustness without sacrificing flexibility.

In practive, this design grounds the LLM's inference within a clinically relevant scope while avoiding open-ended generation, which often leads to noisy or redundant outputs. Importantly, using logits ensures compatibility across different deep architectures. By extracting only the final probabilistic output, our framework remains lightweight and flexible, supporting plug-and-play integration with diverse backbone models.

We instantiate this component with two representative deep architectures (BoxLM [28] due to its effective performance and RETAIN [7] due to its wide adoption) to illustrate the framework's generality, where both output clinically grounded prediction logits can be directly used in downstream reasoning.

BoxLM (Structure-Aware Logits). BoxLM encodes each diagnosis (ICD or CCS) as a high-dimensional box $\mathbf{b}_i = (\mathbf{b}_i^{\text{Cen}}, \mathbf{b}_i^{\text{Off}})$, where the center $\mathbf{b}_i^{\text{Cen}}$ encodes the semantic core (initialized from BioBERT) and the offset $\mathbf{b}_i^{\text{Off}}$ represents conceptual uncertainty (learned via ontology-aware GCN). Visit-level and patient-level representations are aggregated via attention and temporal pooling:

$$\mathbf{b}_{v_t}^{\text{Cen}} = \sum_{i \in v_t} \alpha_i \mathbf{b}_i^{\text{Cen}}, \quad \mathbf{b}_{v_t}^{\text{Off}} = \max_{i \in v_t} \mathbf{b}_i^{\text{Off}}, \tag{1}$$

$$\mathbf{b}_p^{\text{Cen}} = \sum_{v_i \in \mathcal{V}_p} w_i \mathbf{b}_{v_i}^{\text{Cen}}, \quad \mathbf{b}_p^{\text{Off}} = \max_{v_i \in \mathcal{V}_p} \mathbf{b}_{v_i}^{\text{Off}}, \tag{2}$$

where i indexes diagnosis codes within visit v_t, and v_t denotes the t-th visit of patient p.

We calculate the relevance of a candidate CCS c by measuring the intersection volume between patient and candidate boxes:

$$\hat{y}_{p,c} = \log\left(\text{Vol}(\mathbf{b}_p \cap \mathbf{b}_c)\right), \tag{3}$$

where the intersection volume is computed using a differentiable approximation. Specifically, we adopt a *Gumbel-softplus approximation* to estimate the volume of intersection between two boxes in a smooth and numerically stable manner:

$$\text{Vol}(\mathbf{b}_p \cap \mathbf{b}_c) \approx \prod_{k=1}^{d} \beta \log\left(1 + \exp\left(-\frac{\mu_k^{\max} - \mu_k^{\min}}{\beta} - 2\gamma\right)\right), \tag{4}$$

where $\mu_k^{\min}$ and $\mu_k^{\max}$ are minimum and maximum corners corresponding to the intersection box between $\mathbf{b}_p$ and $\mathbf{b}_c$. β is the scale of the Gumbel distribution and γ is the Euler-Mascheroni constant.

RETAIN (Sequence-Aware Logits). RETAIN computes diagnosis prediction scores by attending to historical visit representations in reverse chronological order. Each visit $x_t \in \mathbb{R}^r$ is embedded as $\mathbf{v}_t$. Two RNNs are applied over the reversed visit sequence $\{\mathbf{v}_T, \ldots, \mathbf{v}_1\}$ to generate visit-level attention weights α_t and variable-level attention vectors β_t.

Then, the patient representation is aggregated via:

$$\mathbf{c}_p = \sum_{t=1}^{T} \alpha_t \cdot (\beta_t \odot \mathbf{v}_t). \tag{5}$$

The prediction logits for each candidate CCS are computed via:

$$\hat{y}_{p,c} = \left(W_o \cdot \mathbf{c}_p + b_o\right)_c. \tag{6}$$

Unified Candidate Selection. Regardless of architecture, we extract the top-K CCS codes based on logits $\hat{y}_{p,c}$, and define these as the candidate set C_p^{cand} for downstream LLM-based reasoning. This modular design ensures that any diagnosis model producing logits—regardless of its internal structure—can be integrated into our framework, facilitating broad applicability across settings.

3.3 Evidential Prioritization for Set-based EHRs

Raw EHR inputs are typically represented as unordered sets of ICD/CCS codes, lacking clinical prioritization or semantic hierarchy. However, not all diagnoses contribute equally to a patient's current state. When such flat, unstructured code sets are provided to LLMs, their ability to focus on the most relevant clinical cues is diminished, leading to noisy or generic reasoning.

To address this, we introduce an evidential prioritization strategy that enhances both the informativeness and structure of the patient history used in prompting. Specifically, we first rank each historical CCS code by computing its marginal contribution to the deep model's prediction logits. Then, using the CCS–ICD ontology[2], we propagate these priorities to all corresponding ICD codes. The

[2]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

intuition is to quantify each diagnosis's contribution to the deep model's prediction and reflect that in the prompt structure.

Input CCS Prioritization. To highlight clinically relevant conditions within the unordered historical CCS codes $\mathcal{H}_p$, we rank them by their predictive influence using the deep model's output logits. This prioritization helps guide the LLM to focus on diagnoses that most impact the model's predictions. This choice is both model-agnostic and task-aligned: unlike attention weights or hidden states that vary significantly across architectures, logits provide a consistent and interpretable signal that reflects each input code's contribution to the model's prediction. Clinically, this prioritization helps emphasize persistent or influential conditions (e.g., chronic comorbidities or early indicators of disease onset), which are known to shape diagnostic outcomes over time [7, 23]. Then, we convert the unordered set $\mathcal{H}_p$ into an ordered list $[\tilde{C}_1, \tilde{C}_2, \ldots]$ by sorting according to:

$$[\tilde{C}_1, \tilde{C}_2, \ldots] = \text{Sort}_{h \in \mathcal{H}_p}(\hat{y}_{p,h}). \tag{7}$$

This ordering reflects the model's estimation of how well each CCS code aligns with the patient's current condition, enabling downstream prompt construction to emphasize more relevant diagnoses.

Ontology-Propagated ICD Ordering. In most diagnostic models, prediction is performed on either CCS or ICD code space—but not both. As a result, we typically obtain prediction logits only for one level of the ICD-CCS ontology, while the other lacks direct supervision. In our setting, the deep model outputs logits for CCS codes, leaving the ICD-level relevance unspecified.

To incorporate fine-grained ICD-level detail while maintaining the diagnostic relevance provided by CCS-level prediction, we propagate CCS priorities to their associated ICD codes based on a standardized ontology.

Formally, for each CCS code C_h with prediction score $\hat{y}_{p,h}$, we collect the corresponding ICD codes from the historical visit as:

$$\mathcal{S}_h = \{\text{ICD}_i \in \mathcal{V}_p \mid \text{ICD}_i \text{ belongs to } C_h\}, \tag{8}$$

where "belongs to" follows the CCS–ICD ontology mapping.

Each ICD group $\mathcal{S}_h$ inherits the priority of its parent CCS C_h, allowing us to present historical diagnoses in a relevance-aware, semantically structured format. A corresponding prompt segment is constructed as:

Evidential Prioritization for Set-based EHRs

Patient Historical Diagnoses (Prioritized):
`[{"Essential Hypertension", "Hypertensive Heart Disease"} BELONG TO "Hypertension"], [{"Type 2 Diabetes without Complications"} BELONG TO "Diabetes Mellitus"], ...]`

This formulation enhances the prompt with both concept-level relevance and fine-grained clinical specificity, enabling the LLM to reason over patient history in a hierarchically informed and interpretable manner.

3.4 Relational Evidence Construction for Novel Diagnosis Prediction

As shown in Figure 1, LLMs often struggle to infer novel diagnoses absent from a patient's history due to the lack of explicit supporting evidence. Without structured cues, they tend to repeat previously seen conditions or overlook emerging comorbidities. In contrast, deep models trained on large-scale EHR datasets can implicitly capture rich statistical relationships between diagnosis codes, enabling them to assign high scores even to novel diagnoses through learned co-occurrence patterns.

However, such reasoning remains hidden within the deep model's black-box architecture. To enhance interpretability and support LLM-based inference, we extract symbolic relational evidence between historical and candidate CCS codes. Specifically, we mine co-occurrence statistics from the training data to discover potential associative links between past diagnoses and each top-ranked novel candidate identified by the deep model. While the construction process itself is model-agnostic, we selectively extract symbolic relations only for the novel candidate diagnoses predicted by the deep model. By bridging global co-occurrence with instance-specific deep model predictions, we enable the LLM to reference relevant medical associations that may reflect latent patterns implicitly captured by the deep model. Specifically, we extract such relational evidence in two steps: (1) constructing a CCS–CCS co-occurrence matrix from patient-level diagnosis records, and (2) identifying, for novel candidate CCS generated by deep models, the most statistically related historical CCS from the patient's record.

CCS Co-occurrence Matrix Construction. Let $\mathcal{A} \in \mathbb{R}^{N \times C}$ denote the patient-to-CCS binary adjacency matrix, where N is the number of patients and C is the number of CCS codes. Each entry $\mathcal{A}(n, c) = 1$ indicates that patient n was diagnosed with CCS code c. The CCS co-occurrence matrix is then defined as:

$$\mathbf{G}_{\text{ccs}} = \mathcal{A}^\top \cdot \mathcal{A}, \tag{9}$$

where $\mathbf{G}_{\text{ccs}}(i, j)$ denotes the total number of patients who have been diagnosed with both C_i and C_j. This matrix encodes diagnosis associations that are implicitly used by deep models during learning.

Historical-to-Candidate Relation Extraction. Let C_p^{cand} be the set of deep model-generated candidate CCS codes for patient p, and $\mathcal{H}_p$ the set of their historical CCS codes. For each candidate $C_c \in C_p^{\text{cand}} \setminus \mathcal{H}_p$ that is not previously observed, we identify the most related historical diagnosis C_h^* using the co-occurrence matrix:

$$C_h^* = \arg \max_{C_h \in \mathcal{H}_p} \mathbf{G}_{\text{ccs}}(C_h, C_c). \tag{10}$$

This relation $C_h^* \Rightarrow C_c$ suggests that the novel diagnosis C_c is statistically supported by its co-occurrence with C_h^* in historical patient records. It can be interpreted as an externalized rationale for why the deep model considers C_c clinically plausible, and subsequently passed to the LLM to enhance its reasoning process.

Relational Evidence for Novel Diagnosis Inference

Relational Evidence Support:
`"Cardiac dysrhythmias"` ⇒ `"Conduction disorders"`
`"Hypertension"` ⇒ `"Pulmonary heart disease"` ...

These historical-to-candidate relations offer interpretable, model-aligned support for novel predictions. By explicitly encoding statistical justifications into the LLM prompt, we enable the model to go beyond memorized history and generalize toward clinically coherent yet previously unobserved diagnoses.

3.5 LLM-based Diagnosis Prediction via Evidence-driven In-Context Learning

Building on the structured evidential signals derived from deep model outputs, our framework enables large language models (LLMs) to perform clinical diagnosis prediction through evidence-guided in-context reasoning. Instead of relying on model fine-tuning, we inject task-relevant information, including filtered candidate sets, prioritized historical diagnoses, and relational associations to novel conditions, into the prompt in a structured and interpretable format.

This framework accommodates both *overall diagnosis prediction*, where the LLM observes all historical visits and may output both recurring and new conditions, and *novel diagnosis prediction*, where only the most recent visit is retained and previously seen diagnoses are excluded from the candidate list. In both settings, the prompt is composed from three evidence types extracted by prior modules: prioritized diagnosis history, relational support for novel conditions, and a task-specific candidate list.

An example of the structured prompt for the *novel diagnosis prediction* task is shown below:

Prompt for Novel Diagnosis Prediction

Last Diagnostic Visit (8 days ago): [...]
Evidential Prioritization for Set-based EHRs: [...]
Relational Evidence for Novel Diagnoses: [...]
Candidate CCS Codes (Novel Only): [...]
Instruction: - Re-rank the candidate CCS categories from most to least likely.
- Output format: `Answer: <CCS 1>, <CCS 2>, ...`

This unified prompting strategy allows LLMs to flexibly perform both general and novel diagnosis prediction in an interpretable, evidence-grounded manner—without any additional model training or adaptation.

4 Experiments

We conduct comprehensive experiments to evaluate the predictive performance, reasoning ability, and generalization capability of our proposed `EviCare` framework. Our evaluation seeks to answer the following research questions:

- **RQ1:** How does `EviCare` perform compared with existing state-of-the-art methods in diagnosis prediction?
- **RQ2:** What is the contribution of each component in `EviCare` to overall and novel diagnosis prediction?
- **RQ3:** How does `EviCare` compare to existing reasoning approaches in predicting novel diagnoses?
- **RQ4:** How do factors such as candidate size and LLM backbone affect prediction performance?
- **RQ5:** Does `EviCare` provide clinically valid and interpretable justifications for its predictions?

Table 1: Statistics of the datasets used in our experiments.

Dataset	MIMIC-III	MIMIC-IV
# of patients	5,449	79,393
# of visits	14,141	329,605
Avg. # visits per patient	2.60	4.15
Avg. # CCS per visit	12.08	11.30
Avg. # Novel CCS per visit	6.15	4.82
Max. # visits per patient	29	169
# of unique diagnoses	3,874	37,917
# of CCS codes	285	842

4.1 Experiment Settings

4.1.1 Datasets and Evaluation Protocols. We evaluate `EviCare` on two widely-used EHR datasets: MIMIC-III [16] and MIMIC-IV [17], following standard benchmarks [3, 28]. Each dataset is preprocessed by mapping all ICD-9/10 codes to CCS categories, resulting in 285 and 842 unique CCS codes in MIMIC-III and MIMIC-IV, respectively. The statistics are summarized in Table 1.

To reflect realistic diagnostic workflows, we focus on next-visit prediction using patients with at least 2 visits. Given historical visits, the model predicts CCS codes for the next one. We consider two evaluation settings: (1) *overall prediction,* where all ground-truth CCS codes are evaluated; and (2) *novel prediction,* where only codes not present in the patient's history are considered, evaluating the model's ability to identify clinically emerging diagnoses.

We report both visit-level and code-level metrics using standard multi-label measures: Precision@k (P@k) and Accuracy@k (Acc@k). Details and metric formulas are included in Appendix A.

4.1.2 Baselines. We compare `EviCare` with a comprehensive set of baseline methods from three categories: (1) LLM-based methods: We include LLaMA3.1-8B [8] and Qwen3-8B [35], which perform diagnosis prediction via direct prompting over EHR inputs. These models are tested with the same prompt template for fair comparison; (2) Sequential modeling methods: RETAIN [7], StageNet [9], and TRANS [3], which encode visit sequences with RNNs, LSTM or Transformer mechanisms for diagnosis prediction, respectively; (3) Structure-aware methods: CGL [20], HiTANet [21], BoxCare [22], and BoxLM [28], which incorporate EHR graph or hierarchical ontology to enhance the representation of medical concepts. The details of the baselines are provided in Appendix B.

4.1.3 Implementation Details. We follow the same experimental setup as used in TRANS [3] and BoxLM [28]. Both datasets are split into training, validation, and test sets using a 7:1:2 ratio at patient level. All compared methods are trained using Adam optimizer, and their hyperparameters are tuned as recommended in their original papers. We set the embedding dimension to 16 for all models. For the in-context reasoning, we employ Qwen3-8B as the foundation model. All experiments are conducted with 5-fold cross-validation.

4.2 Main Results and Analysis (RQ1)

To assess the overall effectiveness of `EviCare` in diagnosis prediction, we compare its performance against a range of baselines on both MIMIC-III and MIMIC-IV datasets under 5% training settings

Table 2: Results of novel and overall diagnosis prediction performance on the MIMIC-III dataset with 5% training data. The best results are highlighted in bold while the second best are underlined.

Task	Novel				Overall			
Metric	Visit-Level		Code-Level		Visit-Level		Code-Level	
	P@5	P@10	Acc@5	Acc@10	P@10	P@20	Acc@10	Acc@20
LLaMA3.1	$1.01_{\pm 0.09}$	$2.22_{\pm 0.47}$	$0.43_{\pm 0.19}$	$2.19_{\pm 0.45}$	$41.19_{\pm 2.11}$	$42.06_{\pm 1.03}$	$30.46_{\pm 0.84}$	$39.88_{\pm 0.71}$
Qwen3	$0.65_{\pm 0.05}$	$2.76_{\pm 0.51}$	$0.68_{\pm 0.16}$	$2.54_{\pm 0.28}$	$47.03_{\pm 1.05}$	$48.61_{\pm 1.12}$	$32.73_{\pm 0.66}$	$46.10_{\pm 1.02}$
RETAIN	$10.14_{\pm 0.17}$	$16.41_{\pm 0.24}$	$6.95_{\pm 0.18}$	$15.50_{\pm 0.21}$	$38.69_{\pm 0.18}$	$45.62_{\pm 0.26}$	$27.68_{\pm 0.24}$	$44.15_{\pm 0.23}$
StageNet	$12.91_{\pm 0.12}$	$20.39_{\pm 0.23}$	$8.76_{\pm 0.18}$	$19.27_{\pm 0.23}$	$35.70_{\pm 0.17}$	$43.59_{\pm 0.24}$	$25.71_{\pm 0.23}$	$42.83_{\pm 0.19}$
TRANS	$12.14_{\pm 0.16}$	$18.39_{\pm 0.21}$	$8.17_{\pm 0.13}$	$17.03_{\pm 0.17}$	$38.60_{\pm 0.18}$	$45.77_{\pm 0.23}$	$27.56_{\pm 0.22}$	$44.29_{\pm 0.23}$
CGL	$14.24_{\pm 0.17}$	$20.39_{\pm 0.18}$	$9.64_{\pm 0.12}$	$19.10_{\pm 0.22}$	$38.98_{\pm 0.20}$	$45.98_{\pm 0.23}$	$28.06_{\pm 0.23}$	$44.89_{\pm 0.28}$
HiTANet	$12.90_{\pm 0.15}$	$19.61_{\pm 0.18}$	$8.71_{\pm 0.13}$	$18.79_{\pm 0.24}$	$34.94_{\pm 0.18}$	$43.88_{\pm 0.24}$	$25.25_{\pm 0.20}$	$43.10_{\pm 0.26}$
BoxCare	$11.57_{\pm 0.14}$	$17.56_{\pm 0.17}$	$7.80_{\pm 0.12}$	$16.68_{\pm 0.25}$	$39.91_{\pm 0.16}$	$46.08_{\pm 0.25}$	$28.53_{\pm 0.18}$	$44.80_{\pm 0.20}$
BoxLM	$9.79_{\pm 0.17}$	$16.51_{\pm 0.19}$	$6.47_{\pm 0.15}$	$15.31_{\pm 0.24}$	$45.44_{\pm 0.23}$	$51.46_{\pm 0.28}$	$32.21_{\pm 0.27}$	$49.21_{\pm 0.24}$
EviCare (Retain)	$\underline{21.32}_{\pm 1.63}$	$\underline{23.59}_{\pm 1.17}$	$\underline{14.26}_{\pm 0.72}$	$\underline{21.83}_{\pm 0.41}$	$\underline{49.66}_{\pm 0.58}$	$\underline{52.40}_{\pm 0.56}$	$\underline{35.11}_{\pm 0.73}$	$\underline{50.10}_{\pm 1.07}$
EviCare (BoxLM)	$\mathbf{21.54}_{\pm 1.60}$	$\mathbf{24.41}_{\pm 1.16}$	$\mathbf{14.48}_{\pm 0.64}$	$\mathbf{22.05}_{\pm 0.57}$	$\mathbf{51.73}_{\pm 0.72}$	$\mathbf{55.05}_{\pm 0.95}$	$\mathbf{36.44}_{\pm 0.47}$	$\mathbf{52.26}_{\pm 0.74}$

Table 3: Results of novel diagnosis prediction performance on the MIMIC-IV dataset with 5% training data. The best results are highlighted in bold while the second best are underlined.

Metric	Visit-Level		Code-Level	
	P@5	P@10	Acc@5	Acc@10
LLaMA3.1	$5.29_{\pm 0.76}$	$6.13_{\pm 0.66}$	$4.46_{\pm 0.56}$	$7.31_{\pm 0.70}$
Qwen3	$5.14_{\pm 0.46}$	$6.36_{\pm 0.61}$	$4.36_{\pm 0.28}$	$7.53_{\pm 0.74}$
RETAIN	$8.12_{\pm 0.06}$	$14.03_{\pm 0.08}$	$5.71_{\pm 0.11}$	$12.14_{\pm 0.15}$
StageNet	$7.81_{\pm 0.06}$	$13.10_{\pm 0.07}$	$5.38_{\pm 0.10}$	$11.20_{\pm 0.11}$
TRANS	$9.03_{\pm 0.07}$	$14.35_{\pm 0.11}$	$6.31_{\pm 0.14}$	$12.32_{\pm 0.13}$
CGL	$8.02_{\pm 0.04}$	$14.54_{\pm 0.05}$	$5.08_{\pm 0.07}$	$11.71_{\pm 0.09}$
HiTANet	$7.43_{\pm 0.07}$	$10.88_{\pm 0.07}$	$5.09_{\pm 0.05}$	$9.24_{\pm 0.10}$
BoxCare	$8.09_{\pm 0.05}$	$14.21_{\pm 0.08}$	$5.21_{\pm 0.10}$	$11.81_{\pm 0.10}$
BoxLM	$8.35_{\pm 0.09}$	$15.08_{\pm 0.07}$	$5.57_{\pm 0.12}$	$12.12_{\pm 0.11}$
EviCare (R)	$\underline{11.87}_{\pm 0.98}$	$\underline{16.24}_{\pm 0.85}$	$\underline{8.82}_{\pm 0.46}$	$\underline{15.41}_{\pm 0.72}$
EviCare (B)	$\mathbf{12.17}_{\pm 0.76}$	$\mathbf{16.88}_{\pm 0.28}$	$\mathbf{8.84}_{\pm 0.45}$	$\mathbf{15.44}_{\pm 0.26}$

(shown in Tables 2, 3 and 8). The performance comparison under varying ratios of training data scenarios (e.g., 1%, 5%, 10%, 50%, and 100%) is provided in Appendix E.

In general, EviCare consistently outperforms all baseline methods, with performance gains ranging from reasonably large (6.2% achieved with Acc@20 on MIMIC-III) to significantly large (51.26% achieved with P@5 on MIMIC-III). These gains answer RQ1, showing that our in-context reasoning strategy can leverage deep model guidance to enhance LLM prediction accuracy and interpretability.

Performance on Novel Diagnosis Prediction. We observe the most pronounced improvements in the novel diagnosis setting, which evaluates the model's ability to identify CCS codes that have not appeared in a patient's history. This task is particularly challenging for LLM-only approaches that tend to rely on shallow textual associations or the repetition of previously seen diagnoses. For example, Qwen3 achieves only 2.76 P@10 and 2.54 Acc@10 on MIMIC-III, while EviCare (BoxLM) improves these metrics to 24.41 and 22.05 respectively, achieving a relative improvement of more than 700%. These results highlight the necessity of relational evidence construction via deep models.

Compared to deep learning baselines, EviCare also yields substantial gains across both temporal (e.g., RETAIN) and structure-aware models (e.g., BoxLM). On MIMIC-III, RETAIN achieves 16.41 P@10 and 15.50 Acc@10, while our EviCare (RETAIN) variant achieves 23.59 and 21.83, with relative improvements of 43.8% and 40.8%. Similarly, BoxLM's performance improves from 16.51 to 24.41 in P@10 and from 15.31 to 22.05 in Acc@10 when integrated into EviCare. These results indicate that our framework successfully enhances the generalization ability of deep models by shifting the LLM's focus to novel but informative diagnoses.

Performance on Overall Diagnosis Prediction. The overall diagnosis prediction results on MIMIC-III are presented in Table 2, while the corresponding results on MIMIC-IV are shown in Table 8, with the complete analysis provided in Appendix D.

EviCare consistently outperforms both LLM-only and deep learning baselines. On MIMIC-III, EviCare (BoxLM) achieves 51.73 P@10 and 36.44 Acc@10, improving upon the BoxLM backbone by +6.29 and +4.23 respectively. Compared to Qwen3 (47.03 P@10, 32.73 Acc@10), our model yields notable absolute gains while maintaining interpretability. We further observe consistent gains on MIMIC-IV, where EviCare (Retain) achieves 55.42 P@10 and outperforms all baselines. These improvements reflect the benefit of guiding LLMs with deep model-derived evidence and structured prompts.

4.3 Ablation Study (RQ2)

To examine the contribution of each component in EviCare, we conduct ablation experiments on MIMIC-III, incrementally adding components: *Candidate Selection (Candidate)*, *Evidential Prioritization (Prioritization)*, and *Relational Evidence (Relational)*. Table 4

Table 4: Ablation study of EviCare components on MIMIC-III.

Task	Novel				Overall			
Metric	Visit-Level		Code-Level		Visit-Level		Code-Level	
	P@5	P@10	Acc@5	Acc@10	P@10	P@20	Acc@10	Acc@20
Qwen3-8B (Base)	$0.65_{\pm 0.05}$	$2.76_{\pm 0.51}$	$0.68_{\pm 0.16}$	$2.54_{\pm 0.28}$	$47.03_{\pm 1.05}$	$48.61_{\pm 1.12}$	$32.73_{\pm 0.66}$	$46.10_{\pm 1.02}$
+ Candidate	$9.61_{\pm 0.42}$	$12.11_{\pm 0.57}$	$6.45_{\pm 0.43}$	$11.04_{\pm 0.62}$	$50.81_{\pm 0.82}$	$52.50_{\pm 0.74}$	$35.89_{\pm 0.79}$	$49.05_{\pm 0.69}$
+ Prioritization	$10.71_{\pm 0.77}$	$14.34_{\pm 0.53}$	$7.53_{\pm 0.68}$	$13.95_{\pm 1.03}$	$52.61_{\pm 0.76}$	$54.19_{\pm 0.78}$	$37.17_{\pm 0.55}$	$51.52_{\pm 0.70}$
+ Relational (Full)	$21.54_{\pm 1.60}$	$24.41_{\pm 1.16}$	$14.48_{\pm 0.64}$	$22.05_{\pm 0.57}$	$51.73_{\pm 0.72}$	$55.05_{\pm 0.95}$	$36.44_{\pm 0.47}$	$52.26_{\pm 0.74}$

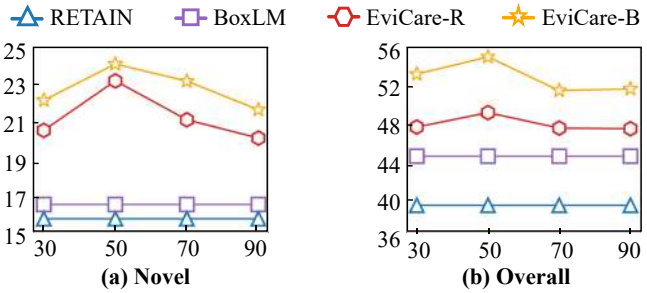


Figure 3: P@10 performance on novel and overall diagnosis prediction under varying candidate CCS sizes (MIMIC-III).

Table 5: Comparison with different reasoning paradigm.

Method	P@10	Reasoning Example Highlights
Qwen (CoT)	3.17	"*Atrial flutter and lung edema* are often seen in *heart failure* and *pneumonia*"
Qwen (SC)	4.16	"*Cardiomyopathies* and *dysrhythmias* are strong indicators of heart failure"
EviCare (Ours)	**24.58**	"AICD adjustment suggests *conduction issues*, which links to nonhypertensive heart failure"

reports the results on both novel and overall diagnosis tasks. We observe general performance gains as more components are added.

The most substantial improvements appear in the novel setting. Starting from the base configuration, which achieves only 2.76 P@10 and 2.54 Acc@10, the inclusion of candidate selection boosts performance to 12.11 and 11.04, respectively. The results show that narrowing the reasoning space significantly improves precision. Adding prioritization further improves P@10 to 14.34 and Acc@10 to 13.95. Finally, relational brings the largest jump, reaching 24.41 P@10 and 22.05 Acc@10, demonstrating the critical role of structured evidence in uncovering clinically novel diagnoses.

While overall gains are more moderate, each component provides measurable improvements. Although overall P@10 slightly drops from 52.50 to 51.73 after adding symbolic relational evidence, novel prediction improves significantly (e.g., visit-level P@10 increase by 70.22%). This trade-off reflects a shift toward broader clinical coverage and generalization, as top-20 metrics continue to improve despite minor top-10 fluctuations. For example, from the base to the full model, P@20 continuously increases from 48.61 to 55.05.

4.4 Comparison with Generic Reasoning (RQ3)

We compare EviCare with two representative reasoning paradigms: Chain-of-Thought (CoT) and Self-Consistency (SC). All methods use Qwen3-8B as the base model and operate over the same candidate CCS set for fairness. As shown in Table 5, EviCare achieves a P@10 of 24.58 under the novel diagnosis setting, significantly outperforming Qwen (CoT) (3.17) and Qwen (SC) (4.16).

Although CoT encourages step-by-step reasoning and SC promotes output diversity through sampling, both approaches lack EHR-derived structure and clinical grounding. As a result, their reasoning often relies on superficial semantic correlations. For example, CoT may associate *lung edema* with *pneumonia*, leading to plausible yet clinically misaligned conclusions. SC, on the other hand, tends to produce generic statements (e.g., cardiomyopathies suggest heart failure), which offers little patient-specific insight.

In contrast, EviCare focuses on novel candidates and explicitly links them to the patient's history using symbolic relational evidence guided by deep model predictions. This strategy enables more targeted and clinically informed reasoning. Additional performance metrics and representative examples are provided in Appendix F.

4.5 Generalization Analysis (RQ4)

To assess the robustness and generalizability of EviCare, we analyze its performance under varying candidate set sizes and across different LLM backbones, focusing on the MIMIC-III dataset.

Candidate Size. Figure 3 illustrates the impact of varying candidate set sizes on P@10. We observe that performance on both diagnosis prediction peaks when $K = 50$, after which performance slightly declines. This trend suggests that smaller candidate sets may omit true positives, while overly large sets introduce noise, weakening the focus of in-context reasoning. Thus, moderate pruning (e.g., $K = 50$) best facilitates the LLM's evidence-driven prediction process.

LLM Backbones. To assess whether EviCare generalizes across different LLM backbones, we fix the underlying deep model (RETAIN or BoxLM) and compare the performance between two representative LLMs (LLaMA3.1 and Qwen3). As shown in Table 7, both LLMs achieve substantial gains over their standalone baselines, indicating that EviCare consistently improves reasoning capabilities across model types.

Notably, while Qwen3 shows stronger standalone performance (e.g., 2.76 vs. 2.22 P@10 on novel diagnoses), LLaMA3.1 combined with BoxLM (50.09) surpasses outperforms Qwen3 with RETAIN (49.66) in overall precision. This indicates that EviCare 's structured backbone design can compensate for weaker LLM capacity. These

Table 6: Case study on a MIMIC-III patient comparing base model (Qwen3-8B), deep model (BoxLM) and EviCare under overall CCS prediction settings. Blue = Novel CCS in ground truth; Red = Incorrect prediction. Arrows (→) indicate evidence chains inferred from patient history to support predicted diagnoses.

Method	Diagnosis Prediction Set
Input EHR Data	**ICD** :Cardiac complications, not elsewhere classified (**CNC**); Acute pancreatitis (**AP**); Peritoneal adhesions postoperative postinfection (**Papp**); Congestive heart failure, unspecified (**CHFu**); ... **CCS** : Complications of surgical procedures or medical care (**CSPM**); Pancreatic disorders (not diabetes) (**PD**); Other gastrointestinal disorders (**OGD**); Congestive heart failure; nonhypertensive (**CHFn**); ...
Ground-Truth Num:10	Coronary atherosclerosis and other heart disease(**CAHD**); **CHFn**, Peripheral and visceral atherosclerosis(**PVA**); **CSPM**; Deficiency and other anemia(**DOA**); Acute posthemorrhagic anemia(**APA**); Essential hypertension(**EH**);...
Qwen3-8B Correct: 3	**CHFn** (...directly listed in the patient's history...is a complication of **AP**..can lead respiratory failure...also in history.) **CSPM** (...listed in the patient's history...linked to **Papp** and **AP**...) **PVA** (...patient's history of heart disease and **CNC** supports this possibility...) Pulmonary heart disease (...patient history has **CHFn**,...lead pulmonary hypertension an pulmonary heart disease.) Nonspecific chest pain (...a symptom of various cardiac and non-cardiac conditions.) ...
BoxLM (3)	**CSPM**, **CAHD**, **CHFn**, Cardiac dysrhythmias, Septicemia (except in labor) ...
EviCare Correct: 5	**CSPM** (...supported by history of **CNC** and **Papp**,... -> **CSPM**) **CHFn** (...patient's history has **CHFn** and **CHFu**, ... -> **CHFn**...) **DOA** (...has history of **Papp**...-> **DOA**, **Papp** -> nutritional deficiencies and anemia.) **EH** (...has history of **AP**...-> **EH**, **AP** -> hypertension due to systemic inflammation...) **CAHD** (...has history of Subendocardial infarction, initial episode of care -> **CAHD**,...due to the pathophysiology of myocardial infarction.) ...

Table 7: Generalization of EviCare across different LLMs on MIMIC-III. The metric used is Precision@10.

LLM	Deep	Novel	Overall	Avg. Imp.
LLaMA3.1	None	2.22	41.19	–
	RETAIN	20.73	48.21	425.4%
	BoxLM	22.15	50.09	459.4%
Qwen3	None	2.76	47.03	–
	RETAIN	23.59	49.66	380.1%
	BoxLM	24.41	51.73	397.4%

trends demonstrate that EviCare generalizes well across diverse LLM architectures, and its design effectively activates reasoning regardless of the LLM backbone.

4.6 Interpretable Case Study (RQ5)

We present a case study to highlight how EviCare enhances both accuracy and interpretability in complex diagnostic scenarios. Compared to Qwen3-8B and BoxLM, EviCare better captures clinically relevant yet previously unseen diagnoses by integrating deep model predictions with symbolic co-occurrence evidence.

As shown in Table 6, Qwen3-8B correctly predicts 3 CCS codes, mainly relying on history repetition. Its explanation often overgeneralizes, such as linking *Pulmonary heart disease* to a vague mention of heart failure, despite no evidence of pulmonary hypertension. It also exhibits hallucinations, e.g., *Nonspecific chest pain*. Although BoxLM achieves similar accuracy with two historical and one novel diagnosis, it fails to provide explanation, limiting its clinical interpretability.

In contrast, EviCare correctly predicts 5 ground-truth CCS codes, including 3 novel ones, and provides clear, clinically grounded explanations. For example, *Essential hypertension* is associated with *acute pancreatitis* based on known inflammatory mechanisms; *Deficiency and other anemia* is explained by the presence of *peritoneal adhesions (postinfection)*. This demonstrates EviCare 's ability to move beyond surface-level pattern matching toward evidence-guided, interpretable reasoning.

5 Conclusion

In this paper, we propose EviCare, a deep model-guided in-context reasoning framework for LLM-based diagnosis prediction. By incorporating candidate selection, evidential prioritization, and relational evidence construction, EviCare enables LLMs to make targeted and interpretable predictions over complex clinical histories. Experiments on MIMIC-III and MIMIC-IV show that EviCare consistently outperforms both LLM-only and deep model-only competitors, particularly in predicting novel diagnoses that are unseen previously but clinically important. This task is clinically valuable for discovering emerging pathological patterns and comorbidities. Extensions to other applications, such as drug recommendation, will be explored in future work.

Acknowledgements

This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship, the State Key Laboratory of Novel Software Technology (KFKT2024A03), the Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grant (2025H0042), and Fujian Provincial Natural Science Foundation of China under Grants (2025J01540). Carl Yang was not supported by any funds from China.

References

[1] Tanisha Aggarwal et al. 2025. Harnessing AI Algorithms for Accurate Medical Diagnosis from Electronic Health Record. In *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Vol. 3. IEEE, 1–5.

[2] Emmi Antikainen, Joonas Linnosmaa, Adil Umer, Niku Oksala, Markku Eskola, Mark van Gils, Jussi Hernesniemi, and Moncef Gabbouj. 2023. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. *Scientific Reports* 13, 1 (2023), 3517.

[3] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. 2024. Predictive modeling with temporal graphical representation on electronic health records. In *IJCAI: proceedings of the conference*, Vol. 2024. 5763.

[4] Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. RareBench: Can LLMs Serve as Rare Diseases Specialists?. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4850–4861.

[5] Chih-Chou Chiu, Chung-Min Wu, Te-Nien Chien, Ling-Jing Kao, Chengcheng Li, and Chuan-Mei Chu. 2023. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent Dirichlet allocation method. *International journal of environmental research and public health* 20, 5 (2023), 4340.

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

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

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[9] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of the web conference 2020*. 530–540.

[10] Ritwik Gupta, Rodolfo Corona, Jiaxin Ge, Eric Wang, Dan Klein, Trevor Darrell, and David M Chan. 2025. Enough Coin Flips Can Make LLMs Act Bayesian. *arXiv preprint arXiv:2503.04722* (2025).

[11] Janneke MT Hendriksen, Geert-Jan Geersing, Karel GM Moons, and Joris AH de Groot. 2013. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis* 11 (2013), 129–141.

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

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

[14] Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, et al. 2024. Health-LLM: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746* (2024).

[15] Li Jiu, Junfeng Wang, Francisco Javier Somolinos-Simón, Jose Tapia-Galisteo, Gema García-Sáez, Mariaelena Hernando, Xinyu Li, Rick A Vreman, Aukje K Mantel-Teeuwisse, and Wim G Goettsch. 2024. A literature review of quality assessment and applicability to HTA of risk prediction models of coronary heart disease in patients with diabetes. *Diabetes research and clinical practice* 209 (2024), 111574.

[16] 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 (2016), 1–9.

[17] 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 (2018), 32–39.

[18] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*.

[19] Qing Li, Zehao Li, Jingjing Song, Jianshuo Bao, Jin Yang, and Zhuhong You. 2025. InKrat: Interpretable diagnosis prediction models based on cross-modal knowledge graph semantic retrieval fusion. *Information Fusion* (2025), 103546.

[20] 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 *International Joint Conference on Artificial Intelligence*.

[21] 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*. 647–656.

[22] Hang Lv, Zehai Chen, Yacong Yang, Guofang Ma, Tan Yanchao, and Carl Yang. 2024. BoxCare: a box embedding model for disease representation and diagnosis prediction in healthcare data. In *Companion Proceedings of the ACM Web Conference 2024*. 1130–1133.

[23] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1903–1911.

[24] Sidra Nasir, Rizwan Ahmed Khan, and Samita Bai. 2024. Ethical framework for harnessing the power of AI in healthcare and beyond. *IEEE Access* 12 (2024), 31014–31035.

[25] Harsha Nori et al. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375* (2023).

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[27] Karan Singhal et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[28] Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang. 2025. BoxLM: Unifying Structures and Semantics of Medical Concepts for Diagnosis Prediction in Healthcare. In *Forty-second International Conference on Machine Learning*.

[29] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *ACL Findings*.

[30] Jinqiang Wang, Huansheng Ning, Yi Peng, Qikai Wei, Daniel Tesfai, Wenwei Mao, Tao Zhu, and Runhe Huang. 2024. A survey on large language models from general purpose to medical applications: Datasets, methodologies, and evaluations. *arXiv preprint arXiv:2406.10303* (2024).

[31] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

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

[34] Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2025. DearLLM: Enhancing Personalized Healthcare via Large Language Models-Deduced Feature Correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[35] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[36] Sirui Ding, Jiancheng Ye, Xia Hu, and Na Zou. 2024. Distilling the knowledge from large-language model for health event prediction. *Scientific Reports* 14, 1 (2024), 30675.

[37] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* 40, 2 (2024), btae075.

[38] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine* 7, 1 (2024), 20.

[39] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature* 642, 8067 (2025), 442–450.

[40] Hao Wu, Yinghao Zhu, Zixiang Wang, Xiaochen Zheng, Ling Wang, Wen Tang, Yasha Wang, Chengwei Pan, Ewen M Harrison, Junyi Gao, et al. 2024. Ehrflow: A large language model-driven iterative multi-agent electronic health record data analysis workflow. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

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

[42] Sunyi Zheng, Nannan Zhao, Jing Wang, Tao Yu, Dongsheng Yue, Wenjia Zhang, Shuxuan Fan, Xiaolei Wang, Guilin Tang, Yuxuan Sun, et al. 2025. Comparison of a specialized large language model with GPT-4o for CT and MRI radiology report summarization. *Radiology* 316, 2 (2025), e243774.

APPENDIX

A Detail of Evaluation Metrics

We evaluate diagnosis prediction at visit and code levels using Precision@k (P@k) and Accuracy@k (Acc@k).

Visit-Level Precision@k. This metric computes the proportion of correctly predicted codes within the top-k predictions for each individual visit. Formally, for a visit t, let $\hat{Y}_t$ be the top-k predicted codes, and Y_t the ground-truth codes for the next visit. The visit-level precision is defined as:

$$\text{P@k}_{\text{visit}} = \frac{1}{|T|} \sum_{t=1}^{|T|} \frac{|\hat{Y}_t \cap Y_t|}{\min(k, |Y_t|)}, \tag{11}$$

where T is the set of all visits in the test set. The normalization ensures fairness when $|Y_t| < k$.

Code-Level Accuracy@k. This metric measures the overall proportion of correctly predicted codes across all patients, weighted by total ground-truth codes. Let P denote the set of test patients, and for each patient t, Y_t and $\hat{Y}_t$ are defined similarly. The code-level accuracy is given by:

$$\text{Acc@k}_{\text{code}} = \frac{\sum_{t \in P} |\hat{Y}_t \cap Y_t|}{\sum_{t \in P} |Y_t|}. \tag{12}$$

Novel Diagnosis Evaluation. For the *novel diagnosis prediction* setting, a predicted CCS code is counted as correct only if it satisfies:

(i) It appears in the ground-truth codes Y_t, and
(ii) It does not appear in any of the patient's previous visits.

This stricter evaluation reflects the model's ability to identify unseen or emerging diagnoses.

B Baseline Details

We provide a summary of all baseline models used in our experiments. The models are organized into four categories based on their core design characteristics.

B.1 LLM-based Methods

- **LLaMA3.1-8B** [8]: An open-source LLM for efficient language understanding and generation.
- **Qwen3-8B** [35]: supports multilingual instruction following and complex reasoning with strong alignment performance.
- **Chain-of-Thought (CoT)** [32]: prompts LLMs to generate intermediate reasoning steps for better decision-making.
- **Self-Consistency (SC)** [31]: selects the most consistent answer from multiple reasoning paths to enhance reliability.

B.2 Sequential Modeling Methods

- **RETAIN** [7]: applies RNNs with a reverse-time attention mechanism to predict patient diagnoses.
- **StageNet** [9]: models disease progression using stage-aware LSTMs and stage-adaptive convolutions.
- **TRANS** [3]: constructs temporal heterogeneous graphs to capture dynamics and structure in EHRs.

B.3 Ontology-aware Methods

- **CGL** [20]: models patient-disease interactions and external knowledge via collaborative graph learning.
- **HiTANet** [21]: models temporal patterns in EHRs using a hierarchical time-aware transformer.
- **BoxCare** [22]: represents inclusion and exclusion relations among diseases via box embeddings in a structured space.
- **BoxLM** [28]: aligns medical concepts through unified structure- and semantics-aware box embeddings.

C Ontology Graph Construction and Data Sources

We construct the ontology graph by mapping ICD-10-CM and ICD-10-PCS codes to multi-level clinical categories defined in the Clinical Classifications Software Refined (CCSR)[3].

D Full Overall Diagnosis Prediction Analysis

In the overall diagnosis setting, which evaluates a model's ability to predict all ground-truth CCS codes. EviCare achieves consistent improvements over both LLM-only and deep learning baselines.

LLM-only models such as Qwen3 and LLaMA3.1 demonstrate strong performance due to their memorization of historical patterns and semantic associations in clinical texts. For instance, Qwen3 achieves 50.99 P@20 and 46.06 Acc@20 on MIMIC-IV, outperforming several traditional deep models (e.g., RETAIN and StageNet), which primarily rely on sequential visit modeling. However, their predictions are often biased toward frequent diagnoses and lack fine-grained reasoning over complex patient histories. By contrast, EviCare supplements LLMs with candidate selection and evidential scaffolding derived from deep models, enabling more accurate and context-aware reasoning with improvement.

Table 8: Results of overall diagnosis prediction performance on the MIMIC-IV dataset with 5% training data. The best results are highlighted in bold while the second best are underlined.

Metric	Visit-Level		Code-Level	
	P@10	P@20	Acc@10	Acc@20
LLaMA3.1	$38.32_{\pm 1.02}$	$43.62_{\pm 1.36}$	$31.22_{\pm 0.63}$	$33.82_{\pm 0.72}$
Qwen3	$50.42_{\pm 0.90}$	$50.99_{\pm 1.20}$	$36.79_{\pm 0.69}$	$46.06_{\pm 0.78}$
RETAIN	$43.19_{\pm 0.08}$	$49.05_{\pm 0.07}$	$31.49_{\pm 0.07}$	$45.60_{\pm 0.09}$
StageNet	$37.69_{\pm 0.08}$	$43.45_{\pm 0.07}$	$27.69_{\pm 0.06}$	$40.89_{\pm 0.07}$
TRANS	$36.00_{\pm 0.09}$	$41.95_{\pm 0.08}$	$26.52_{\pm 0.07}$	$39.83_{\pm 0.08}$
CGL	$32.48_{\pm 0.05}$	$38.01_{\pm 0.07}$	$22.42_{\pm 0.06}$	$36.72_{\pm 0.10}$
HiTANet	$27.34_{\pm 0.07}$	$34.00_{\pm 0.11}$	$20.07_{\pm 0.09}$	$32.54_{\pm 0.12}$
BoxCare	$34.61_{\pm 0.09}$	$40.33_{\pm 0.12}$	$24.29_{\pm 0.08}$	$36.50_{\pm 0.10}$
BoxLM	$43.02_{\pm 0.07}$	$48.12_{\pm 0.08}$	$31.20_{\pm 0.06}$	$44.44_{\pm 0.09}$
EviCare (R)	$\mathbf{55.42}_{\pm 0.74}$	$\underline{57.12_{\pm 0.61}}$	$\mathbf{40.41}_{\pm 0.51}$	$\underline{53.27_{\pm 0.65}}$
EviCare (B)	$\underline{55.04_{\pm 0.76}}$	$\mathbf{57.63}_{\pm 0.61}$	$\underline{40.17_{\pm 0.65}}$	$\mathbf{53.80}_{\pm 0.67}$

In contrast, deep models like BoxLM offer strong structural priors by encoding hierarchical and co-occurrence-based relationships between diagnoses, but they may lack interpretability and flexibility in complex clinical scenarios. EviCare addresses these limitations

[3] https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp

by using the deep model to generate a semantically meaningful candidate space and relational evidence, which are then synthesized through LLM-based reasoning. For instance, EviCare (BoxLM) improves upon its backbone by achieving 57.63 P@20 and 53.80 Acc@20, marking absolute gains of +9.51 and +9.36, respectively. This demonstrates that our hybrid design effectively combines the structural precision of deep models with the semantic generalization of LLMs. These benefits generalize well to MIMIC-IV, further validating the scalability and robustness of our framework across datasets with varying label spaces and patient distributions.

E Impact of Training Size

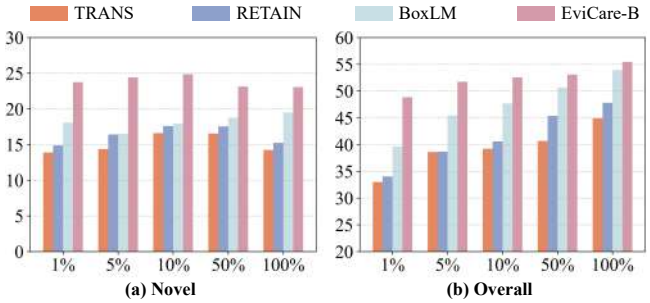


Figure 4: P@10 performance on novel and overall diagnosis prediction under different training data ratios (MIMIC-III).

Figure 4 reports P@10 performance of EviCare and three base models (TRANS, RETAIN, BoxLM) under varying training data ratios (1%, 5%, 10%, 50%, 100%) on MIMIC-III. Across all settings, EviCare-B consistently achieves higher accuracy on both novel and overall diagnosis tasks. Even with only 1% of training data, EviCare-B outperforms base models trained on significantly more data, highlighting its robustness under limited supervision.

F Full Reasoning Strategy Study

Table 9: Performance Comparison of Different Reasoning Strategies.

Novel	P@5	P@10	Acc@5	Acc@10
Qwen (CoT)	1.25	3.17	1.63	3.07
Qwen (SC)	2.09	4.26	2.71	4.02
EviCare	22.03	24.58	14.56	22.12
Overall	P@10	P@20	Acc@10	Acc@20
Qwen (CoT)	41.91	45.66	29.12	43.80
Qwen (SC)	42.23	46.09	30.15	44.26
EviCare	52.01	55.23	36.67	52.11

Table 9 shows the prediction performance under different reasoning strategies. Table 10 presents an example of predicting a novel diagnosis *Congestive heart failure; nonhypertensive* using different reasoning methods. Although the target condition was correctly predicted by all methods, the generated explanations showed differences.

We retain only the clinically relevant entries from the patient's historical EHR for clarity. All models were given the same candidate set and input. Among the three structured prompting (SC) trials, two generated plausible explanations, while one failed to associate the history with the target diagnosis.

In contrast, our method (EviCare) generates a more comprehensive explanation based on relational evidence. Specifically, it links multiple historical conditions—*Cardiac dysrhythmias*, *Other primary cardiomyopathies*, and the *Fitting and adjustment of an automatic implantable cardiac defibrillator (AICD)*—into a coherent causal chain.

This explanation is clinically supported. The implantation of an AICD is commonly indicated in patients with severe conduction disorders or ventricular arrhythmias, which are well-known precursors of sudden cardiac death and nonhypertensive heart failure. Such conduction abnormalities impair cardiac synchronization and function, directly contributing to heart failure.

Thus, the path from AICD adjustment, through conduction disorders, to nonhypertensive heart failure, reflects a medically established progression supported by evidence from clinical trials such as MADIT-II and SCD-HeFT. This validates the relational evidence used by our method and illustrates its advantage in generating faithful, interpretable justifications for novel diagnosis prediction.

Table 10: Explanation comparison for a novel diagnosis prediction case.

Patient History	Other primary cardiomyopathies; Paroxysmal ventricular tachycardia; Fitting and adjustment of AICD; Cardiac dysrhythmias; Conduction disorders; ...
Target Diagnosis	Congestive heart failure; nonhypertensive
Qwen (CoT)	The patient has a history of "Other primary cardiomyopathies" and "Paroxysmal ventricular tachycardia" which are both associated with heart failure.
Qwen (SC)	1. The patient has "Other primary cardiomyopathies" and "Cardiac dysrhythmias", both of which are precursors or components of nonhypertensive heart failure. Although not explicitly stated, the progression is clinically plausible. 2. The patient's history includes "Other primary cardiomyopathies" (ICD) and "Cardiac dysrhythmias" (CCS), which are strong indicators of heart failure. Cardiomyopathies can progress to heart failure, and dysrhythmias like paroxysmal ventricular tachycardia are common in heart failure. 3. (No relevant prediction)
EviCare (Ours)	This is supported by the patient's history of cardiac dysrhythmias, other primary cardiomyopathies, and AICD adjustment. These conditions are often associated with heart failure, especially in the context of chronic disease and device management.
Relational Evidence	"Fitting and adjustment of automatic implantable cardiac defibrillator" → "Conduction disorders" → "Congestive heart failure; nonhypertensive"