THE DOCTOR'S CASEBOOK: WHEN LLMS REASON BY PATIENT ANALOGY VIA INFORMATION GAIN–GUIDED GRAPH SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical reasoning over electronic health records (EHRs) remains challenging for large language models (LLMs). Inspired by clinicians who combine internal medical knowledge with experience from prior patients, we investigate **Reasoning by Patient Analogy**, where a frozen LLM dynamically consults a clinical casebook to ground its reasoning in prior patient experience. Our pilot study reveals three core challenges. **(1) Perspective Limitation**: data-side similarity does not necessarily align with the target LLM's reasoning needs, while model-side signals are bounded by what the model itself can recognize. **(2) Cohort Awareness**: instance-level retrieval overlooks the population structure that organizes prior patient experience. **(3) Information Aggregation**: patient utility changes with the working context, causing overlapping cases to yield diminishing marginal gains as evidence accumulates. To address these challenges, we propose `GraphWalker`, a graph-guided framework for dynamic clinical casebook consultation. `GraphWalker` organizes historical patients into a clinically structured graph, uses cohort prototypes to localize relevant patient groups, and performs *Context-Conditioned Greedy Search with Frontier Expansion* to continually re-evaluate candidate patients by their marginal utility under the current working context, using the reduction in target-input cross-entropy as an information-gain proxy. Extensive experiments on real-world EHR benchmarks show that `GraphWalker` consistently outperforms strong patient-selection baselines, remains competitive with the evaluated supervised EHR predictors in-domain, and achieves stronger performance under cross-version and cross-institution transfer with only seconds-level additional inference overhead. `GraphWalker` further generalizes to black-box LLMs and remains compatible with downstream reasoning scaffolds. Our code is available at https://anonymous.4open.science/r/GraphWalker-4473.

1 INTRODUCTION

Clinical reasoning over electronic health records (EHRs) underpins critical applications such as diagnosis support, risk stratification, and treatment planning (Ma et al., 2023; Xu et al., 2023a;b; 2024b; Fang et al., 2023; Liao et al., 2025). In clinical practice, physicians progressively build and refine their understanding of a new patient through the interplay between *internal medical knowledge* and *experience from prior patients*, a process commonly described as *reasoning by analogy* (Norman, 2006; Ten Cate et al., 2017). Rather than matching the patient to a single similar case, clinical analogy unfolds through an iterative cycle of assessment and evidence seeking: physicians form an initial understanding, consult relevant prior cases, revise that understanding, and use it to guide subsequent case seeking (Lubarsky et al., 2015; Norman & Brooks, 1997; ten Cate & Durning, 2017). As LLMs are increasingly used for clinical reasoning (Zhang et al., 2025a; Ding et al., 2025a; Fang et al., 2025; Ding et al., 2025b), this clinician-style process suggests a simple but underexplored idea: equip the model not only with medical knowledge, but also with a clinical casebook it can consult as reasoning unfolds. Although prior work has explored demonstration selection, two fundamental questions remain open.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

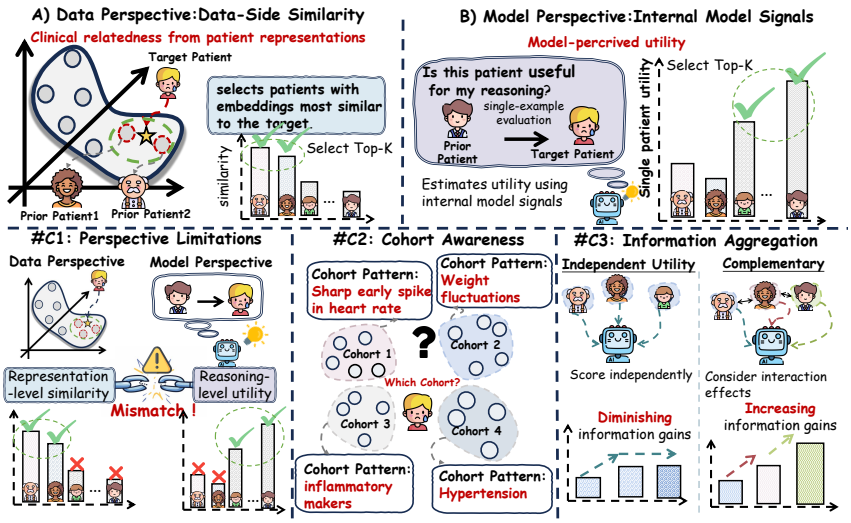


Figure 1: Overview of two existing perspectives for patient selection (top) and three challenges revealed by our pilot study (bottom).

1) *Fragmented Perspectives*. Data-driven approaches rely on *data-side similarity*, such as semantic or representation-level proximity in pretrained embedding spaces (Liu et al., 2022; Hongjin et al., 2022; Robertson et al., 2009; Yu et al., 2024) (Fig. 1(A)), whereas model-driven approaches rely on *model-side signals*, such as entropy, gradients, or influence (Gonen et al., 2023; Peng et al., 2024; Liu et al.; Li et al., 2025) (Fig. 1(B)). These views echo the external and internal sources of clinician reasoning, yet have largely been studied in isolation.

2) *Static Utility Assumption*. Existing methods largely treat example utility as fixed or independently determined. Clinical analogy, however, is iterative: each consulted case can refine the current understanding of the patient and thereby change which cases become informative next. This raises the question of how a clinical casebook should be dynamically consulted as reasoning proceeds.

Motivated by these questions, we conduct a systematic pilot study (Section 2) and identify three core challenges in using prior patient experience for LLM clinical reasoning, as summarized in Fig. 1(C).

- ***Perspective Limitation***: For the data perspective, high representation-level similarity does not consistently translate into utility for the target LLM: similar patient trajectories may still provide little benefit. For the model perspective, utility is estimated primarily from LLM-internal signals and is therefore bounded by what the model itself can recognize; clinically meaningful patterns may be missed when its clinical competence is insufficient. Thus, external patient evidence and internal model signals provide distinct but incomplete views of useful patient experience.
- ***Cohort Awareness***: Existing methods typically retrieve similar patients as isolated instances, overlooking the population structure that organizes prior patient experience. Clinically coherent cohorts instead provide population-level references, with cohort prototypes capturing shared characteristics beyond any single neighbor. Ignoring such structure can overemphasize idiosyncratic cases and lead to unstable reasoning contexts.
- ***Information Aggregation***: Patient utility changes with the working context. As evidence accumulates, overlapping patients may become redundant, yielding diminishing marginal gains, whereas complementary patients can provide additional value. Patient utility should therefore be evaluated by its marginal contribution relative to the current working context.

These challenges motivate `GraphWalker`, a graph-guided framework for **Reasoning by Patient Analogy** that operationalizes dynamic clinical casebook consultation without parameter updates. `GraphWalker` *integrates data and model perspectives* by organizing historical patients into a clinically structured graph and using an LLM-based information-gain proxy to guide traversal and patient selection. To capture *Cohort Awareness*, `GraphWalker` introduces *Cohort Discovery*, which uses cohort prototypes to localize relevant patient groups before selecting individual anchor patients. To address *Information Aggregation*, we introduce *Context-Conditioned Greedy Search with Frontier Expansion*, which continually re-evaluates candidate patients by their marginal information gain as the working context grows, while expanding the search frontier through graph neighbors.

In summary, our main contributions are as follows:

- **Systematic Analysis of Reasoning by Patient Analogy:** We systematically investigate how prior patient experience can support LLM clinical reasoning. Our pilot study reveals three key challenges, *Perspective Limitation*, *Cohort Awareness*, and *Information Aggregation*, showing that useful patient experience must be considered from both data- and model-side views, organized beyond isolated instances, and evaluated context-dependently as reasoning proceeds.
- **Cohort-Aware Marginal Information Gain Search:** We develop a graph-based search mechanism that localizes relevant patient groups through cohort prototypes and continually re-evaluates candidate patients using a marginal information-gain proxy conditioned on the current working context. *Context-Conditioned Greedy Search with Frontier Expansion* couples patient selection with graph exploration, allowing the search space to adapt as new cases are incorporated into the working context.
- **`GraphWalker` Framework:** Synthesizing these findings, we introduce `GraphWalker`, a graph-guided framework for dynamic clinical casebook consultation and **Reasoning by Patient Analogy**. Across multiple real-world EHR benchmarks, `GraphWalker` consistently outperforms strong patient-selection baselines, remains competitive with the evaluated supervised EHR predictors in-domain, and achieves stronger performance under cross-version and cross-institution transfer with only seconds-level additional inference overhead. It further generalizes to black-box LLMs and remains compatible with downstream reasoning scaffolds.

2 Observation & Motivation

In this section, we empirically investigate how prior patient experience should be evaluated, aggregated, and organized for LLM clinical reasoning. Detailed experimental settings are provided in Appendix C.

2.1 Perspective Limitation

We first ask whether useful patient experience can be reliably identified from a single perspective. For the *data perspective*, we consider Qwen3-Embedding-8B (Yang et al., 2025) and SMART (Yu et al., 2024), where patient utility is estimated from representation-level similarity. For the *model perspective*, we consider CONE (Peng et al., 2024) and LMS3 (Liu et al.), which estimate utility from LLM-internal signals. All methods are evaluated against Zero-shot and Random baselines under the same prompting setup. As shown in Fig. 2, SMART achieves the best performance in most settings, highlighting the value of clinically informed patient representations. Meanwhile, CONE and LMS3 generally outperform the generic Qwen3-Embedding-8B, showing that model-side signals can identify useful patients beyond representation similarity alone. Yet neither perspective consistently dominates across metrics. Data-side similarity may retrieve clinically related patients that offer limited utility to the target LLM, whereas model-side estimates remain bounded by the clinical patterns the LLM itself can recognize. These distinct strengths and failure modes motivate integrating the two perspectives for effective patient selection.

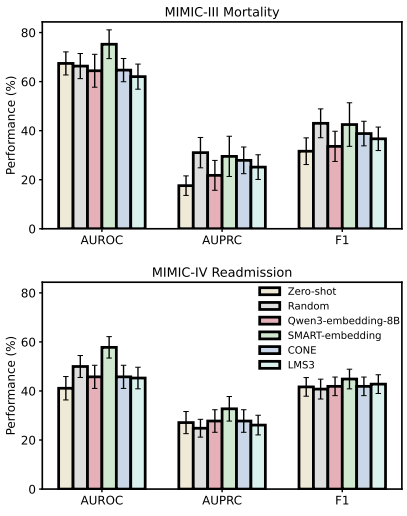


Figure 2: Analysis of *Perspective Limitation*: data- and model-side patient selection on MIMIC-III Mortality (top) and MIMIC-IV Readmission (bottom).

Observation I. *Data- and model-side signals exhibit complementary strengths and distinct failure modes, motivating their integration for patient selection.*

2.2 Information Aggregation

We next examine how patient utility changes as cases accumulate in the working context. Prior work observes that effective demonstration selection can reduce target-input conditional entropy,

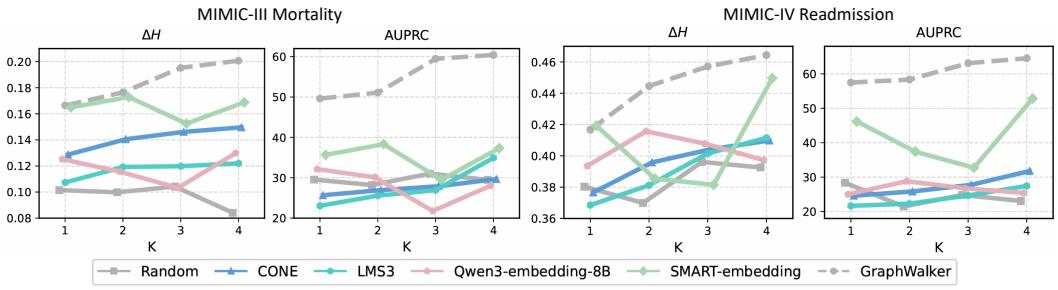


Figure 3: Analysis of *Information Aggregation*

suggesting that useful examples help the model better understand the input (Peng et al., 2024). Extending this intuition to patient sets, we use the composition-level information-gain proxy (ΔH; Appendix C), defined as the reduction in target-record conditional cross-entropy relative to zero-shot prompting. We interpret a larger ΔH as stronger contextual support from the selected cases for understanding the target input.

Figure 3 reveals two key findings. **(1)** Task performance closely follows the trend of ΔH, supporting information gain as an effective model-side signal for patient selection. **(2)** As more patients are incorporated, both similarity- and model-based methods exhibit clear aggregation effects. Similarity-based selection can degrade as overlapping or conflicting evidence accumulates, while model-side methods maintain performance improvements but show pronounced diminishing marginal gains. These observations suggest that evaluating patients independently is insufficient for multi-patient contexts: the value of a new patient should depend on the evidence already present in the working context.

Observation II. *Multi-patient contexts exhibit substantial redundancy and diminishing marginal gains, motivating patient utility to be evaluated by its marginal contribution relative to the current working context.*

2.3 Cohort Awareness

We finally examine whether prior patient experience should be retrieved as isolated instances or organized at the cohort level. Using SMART representations (Yu et al., 2024), we construct a patient graph. Instance Retrieval starts from the single most similar patient and expands locally to identify additional patients, whereas *Cohort Discovery* first localizes a relevant cohort through its prototype and then identifies anchor patients within that cohort. As shown in Fig. 4, Instance Retrieval is competitive when only a few patients are selected, but becomes less effective than *Cohort Discovery* as the number of selected patients increases. This suggests that anchoring subsequent search on a single nearest patient can be sensitive to idiosyncratic local structure. In contrast, cohort prototypes capture broader population-level characteristics and provide a more robust basis for discovering clinically coherent patient experience.

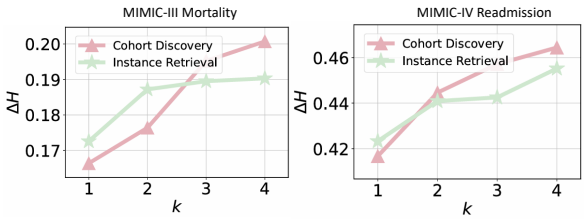


Figure 4: Analysis of *Cohort Awareness*: Cohort Discovery *vs.* Instance Retrieval.

Observation III. *Instance-level retrieval becomes less effective as more patients are selected, whereas cohort-level organization provides a more robust basis for multi-patient retrieval.*

Summary. Together, these observations motivate `GraphWalker` as an iterative casebook-consultation process rather than one-shot patient retrieval. `GraphWalker` combines data- and model-side guidance, continually re-evaluates candidate patients by their marginal information gain relative to the current working context, and uses *Cohort Discovery* to ground this search in clinically coherent patient groups. These principles are instantiated through *Context-Conditioned Greedy Search with Frontier Expansion*, where patient selection and graph exploration proceed jointly.

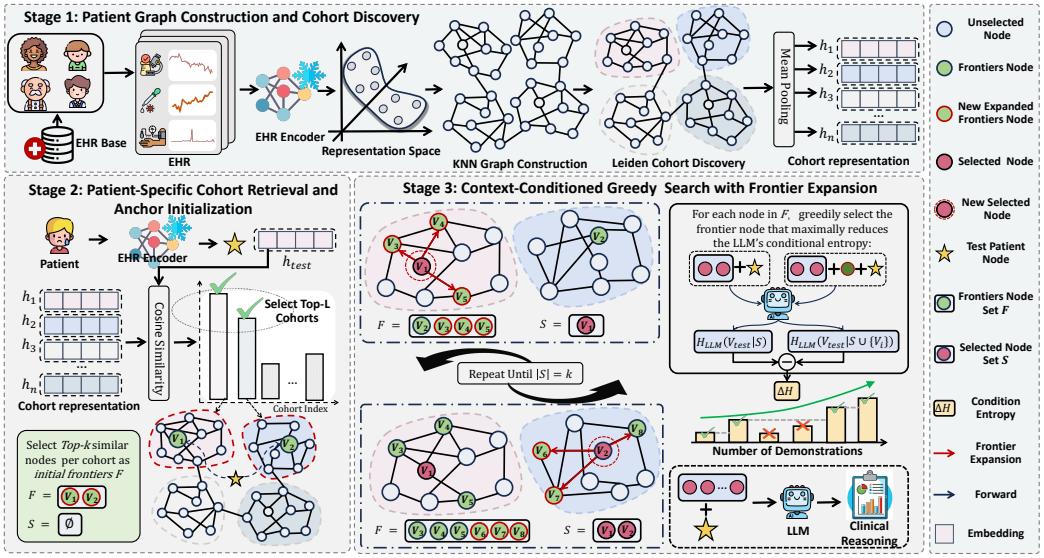


Figure 5: Overview of **GraphWalker**. Historical patients are organized into a graph with clinically coherent cohorts. Cohort retrieval and anchor initialization establish a target-specific search frontier. Context-conditioned greedy search then selects cases using a marginal information-gain proxy, expanding the frontier through graph neighbors and constructing the LLM's working context.

3 METHODOLOGY

GraphWalker operationalizes dynamic clinical casebook consultation by coupling population-level patient structure with context-conditioned model-side utility through three stages, as shown in Fig. 5. First, *Patient Graph Construction and Cohort Discovery* uses EHR representations to organize historical patients into a graph and identify clinically coherent cohorts. Second, *Patient-Specific Cohort Retrieval and Anchor Initialization* retrieves relevant cohorts through prototype similarity and selects anchor patients within each cohort to initialize a target-specific search frontier. Third, *Context-Conditioned Greedy Search with Frontier Expansion* iteratively builds the working context by re-evaluating candidates using a marginal information-gain proxy conditioned on the selected cases, while expanding the frontier through each newly selected patient's graph neighbors.

3.1 PATIENT GRAPH CONSTRUCTION AND COHORT DISCOVERY

Patient Representation. To instantiate data-side clinical relatedness, we encode each longitudinal EHR into a compact patient representation. Let a patient record be a sequence of T time-ordered visits $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. Given a historical patient repository $\{\mathbf{X}_i\}_{i=1}^{N}$, we use a pretrained EHR encoder $\mathcal{M}_{\text{exp}}$ to obtain $\mathbf{h}_i = \mathcal{M}_{\text{exp}}(\mathbf{X}_i) \in \mathbb{R}^d$,where $\mathbf{h}_i$ serves as the patient representation for subsequent graph construction.

Population-Level Graph Construction. We construct a k-nearest-neighbor patient graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to organize prior patient experience according to local clinical relatedness. Each node $v_i \in \mathcal{V}$ represents a patient, and an edge $(i, j) \in \mathcal{E}$ is established when patient j is among the k_g nearest neighbors of patient i under cosine similarity: $\mathcal{E} = \{(i, j) \mid \text{rank}_j[\text{sim}(\mathbf{h}_i, \mathbf{h}_j)] \leq k_g\}$, where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and k_g controls graph density.

Graph-Based Cohort Discovery. We partition the patient graph into clinically coherent cohorts using the Leiden algorithm (Traag et al., 2019) to obtain a set of cohort subgraphs $\{\mathcal{C}_m\}_{m=1}^{M}$ by optimizing graph modularity:

$$\mathcal{Q} = \frac{1}{2|\mathcal{E}|} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2|\mathcal{E}|} \right) \mathbb{I}(c_i = c_j), \tag{1}$$

where A_{ij} is an entry of the adjacency matrix of $\mathcal{G}$, d_i is the degree of node i, and $\mathbb{I}(\cdot)$ indicates whether two nodes belong to the same cohort. Each cohort $\mathcal{C}_m$ is summarized by a prototype obtained by mean-pooling its member representations: $\mathbf{z}_m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \mathbf{h}_i$. The cohort prototypes

provide compact population-level references for localizing clinically relevant regions of the patient graph.

3.2 Patient-Specific Cohort Retrieval and Anchor Initialization

Cohort Retrieval. Given a target patient with EHR record $\mathbf{X}^{(q)}$, we obtain its representation $\mathbf{h}^{(q)} = \mathcal{M}_{\text{exp}}(\mathbf{X}^{(q)})$. A patient may exhibit clinical patterns shared with several subpopulations. We therefore retrieve the top-K_c cohorts by prototype similarity, $\mathcal{C}_{\text{ret}}^{(q)} = \text{Top-}K_c\big(\text{sim}(\mathbf{h}^{(q)}, \mathbf{z}_m)\big)$, providing multiple population-level entry points for subsequent patient-level exploration.

Anchor Initialization. Given the retrieved cohorts $\mathcal{C}_{\text{ret}}^{(q)}$, we initialize the search frontier by selecting anchor nodes within each cohort. For each $\mathcal{C}_m \in \mathcal{C}_{\text{ret}}^{(q)}$, we select the top-$K_a$ patients whose embeddings are most similar to $\mathbf{h}^{(q)}$: $\mathcal{V}_m^{(q)} = \underset{v_i \in \mathcal{C}_m}{\text{Top-}K_a} \text{sim}(\mathbf{h}^{(q)}, \mathbf{h}_i)$. The initial search frontier is the union of these anchors: $\mathcal{F}^{(q)} = \bigcup_{\mathcal{C}_m \in \mathcal{C}_{\text{ret}}^{(q)}} \mathcal{V}_m^{(q)}$.

3.3 Context-Conditioned Greedy Search with Frontier Expansion

Context-Conditioned Marginal Information Gain. Starting from the initial frontier $\mathcal{F}^{(q)}$, `GraphWalker` iteratively constructs a selected patient set $\mathcal{S}^{(q)}$, which forms the patient-specific working context for the target LLM. Candidate utility is evaluated relative to the patients already selected. Let $H_\theta(x_{\text{test}} \mid \mathcal{S})$ denote the query-level conditional cross-entropy of the serialized target patient x_{test} given the selected patient set $\mathcal{S}$. Operationally, it is computed as the average teacher-forced negative log-probability of the target-patient tokens under the LLM; the target patient's ground-truth outcome is never used during selection. For a candidate patient $v_i \in \mathcal{F}^{(q)}$, we define its marginal information-gain proxy as

$$\Delta H_\theta\left(v_i \mid \mathcal{S}^{(q)}\right) = H_\theta\left(x_{\text{test}} \mid \mathcal{S}^{(q)}\right) - H_\theta\left(x_{\text{test}} \mid \mathcal{S}^{(q)} \cup \{v_i\}\right). \tag{2}$$

A larger ΔH_θ indicates that the candidate makes the target record more predictable given the selected cases. We interpret this reduction in conditional cross-entropy as stronger contextual support for the LLM to understand the target input. At each iteration, we select the candidate with the largest marginal information gain:

$$v^* = \arg\max_{v_i \in \mathcal{F}^{(q)}} \Delta H_\theta\left(v_i \mid \mathcal{S}^{(q)}\right), \qquad \mathcal{S}^{(q)} \leftarrow \mathcal{S}^{(q)} \cup \{v^*\}. \tag{3}$$

After each selection and frontier expansion, we recompute all frontier candidates' marginal information-gain proxies under the updated working context.

Graph-Based Frontier Expansion. After selecting v^*, we add its graph neighbors to the frontier and exclude patients already selected:

$$\mathcal{F}^{(q)} \leftarrow \left(\mathcal{F}^{(q)} \cup \mathcal{N}(v^*)\right) \setminus \mathcal{S}^{(q)}, \tag{4}$$

where $\mathcal{N}(v^*) = \{v_j \mid (v^*, v_j) \in \mathcal{E}\}$ denotes the graph neighbors of v^*. Each selected case thus contributes to the working context and exposes related patients for subsequent exploration. Re-evaluation under the updated context filters redundant candidates and favors those providing additional information.

Stopping Criterion. The search terminates when either (i) the patient budget K is reached, or (ii) no candidate in the current frontier yields a positive reduction in target-record conditional cross-entropy. Under our proxy, the latter indicates that no frontier candidate provides additional estimated support for understanding the target input.

LLM Inference with the Selected Working Context. Finally, the selected patients in $\mathcal{S}^{(q)}$ are formatted as few-shot demonstrations and combined with the target query to construct the final prompt. The target LLM then generates the clinical prediction without any task-specific parameter update. Prompt details and the full algorithm are provided in Appendix F and Appendix B.6, respectively.

Table 1: Performance comparison on MIMIC-III and MIMIC-IV datasets using *Qwen3-14B* and *LLaMA-3.1-8B-Instruct* as backbone LLMs. All metrics are higher-is-better (↑); values are means over three random seeds. **Bold**: best; underlined: second-best.

Method		MIMIC-III Mortality			MIMIC-III LOS		MIMIC-IV Readmission		
Paradigm	**Approach**	AUROC ↑	AUPRC ↑	F1 ↑	ma-ROC ↑	mi-ROC ↑	AUROC ↑	AUPRC ↑	F1 ↑
Qwen3-14B									
Vanilla	Zero-shot	64.09	17.58	31.62	65.22	37.49	41.11	27.11	41.66
	Random	66.33	31.04	43.00	58.12	61.28	49.98	24.81	40.77
Data-Side	Semantic-emb	64.43	21.78	33.60	57.06	66.60	45.12	26.91	42.01
	EHR-emb	75.24	29.52	42.49	62.03	66.98	57.79	32.73	44.87
	Time-series	62.28	29.62	39.34	58.03	55.15	53.46	26.97	40.55
Model-Side	SPELL	75.37	32.23	50.32	63.73	75.78	48.82	24.37	40.03
	Influence	60.54	27.20	36.10	50.03	62.70	52.85	28.55	43.19
	IDS	59.14	19.73	32.97	49.86	37.41	47.41	25.43	42.68
	CONE	64.69	27.90	38.82	57.84	67.91	45.76	27.75	41.89
	LMS3	66.54	26.92	39.07	59.44	31.89	44.34	24.68	42.28
	Delta-KNN	51.52	15.66	25.22	59.04	63.43	47.11	24.41	40.00
	GradSel	62.09	32.20	36.68	66.80	73.57	45.73	25.39	41.83
	`GraphWalker`	**84.19**	**59.46**	**58.24**	**69.51**	**84.64**	**75.91**	**60.61**	**58.01**
LLaMA-3.1-8B-Instruct									
Vanilla	Zero-shot	65.35	21.72	35.75	43.16	74.13	54.40	29.88	41.24
	Random	52.04	18.81	28.90	45.66	80.14	48.63	27.06	39.85
Data-Side	Semantic-emb	56.90	22.32	35.15	53.78	81.49	43.55	23.69	42.37
	EHR-emb	65.78	24.30	41.89	45.72	74.78	63.75	46.32	49.81
	Time-series	56.38	24.86	37.16	46.23	71.99	53.45	27.67	42.28
Model-Side	SPELL	55.43	20.92	34.51	49.48	82.02	48.11	25.03	40.09
	Influence	61.73	31.80	38.39	55.57	82.63	49.64	28.40	42.65
	IDS	59.89	28.40	36.11	42.12	71.14	45.25	24.86	42.65
	CONE	50.40	18.93	32.79	38.31	76.94	51.67	29.42	42.73
	LMS3	51.80	16.41	29.37	39.60	74.05	54.77	30.95	45.41
	Delta-KNN	57.42	21.48	33.61	47.69	75.71	51.23	26.50	41.13
	GradSel	47.26	17.32	31.23	53.92	80.83	52.33	28.70	42.17
	`GraphWalker`	**76.71**	**35.09**	**50.13**	**57.06**	**83.55**	**68.32**	**46.37**	**51.54**

4 Experiments

4.1 Experimental Setup

Datasets, Tasks, and Metrics. We evaluate `GraphWalker` on two widely used ICU EHR datasets, **MIMIC-III** (Johnson et al., 2016) and **MIMIC-IV** (Johnson et al., 2023), across three tasks: in-hospital mortality, length-of-stay (LOS), and ICU readmission. We report AUROC, AUPRC, and F1-score for mortality and readmission, and macro- and micro-averaged AUROC for LOS following prior work (Harutyunyan et al., 2019). We additionally use the **TJH** dataset (Yan et al., 2020) from Tongji Hospital to assess cross-institution transfer on mortality prediction. Dataset, preprocessing, and task details are provided in Appendices B.1 and B.2.

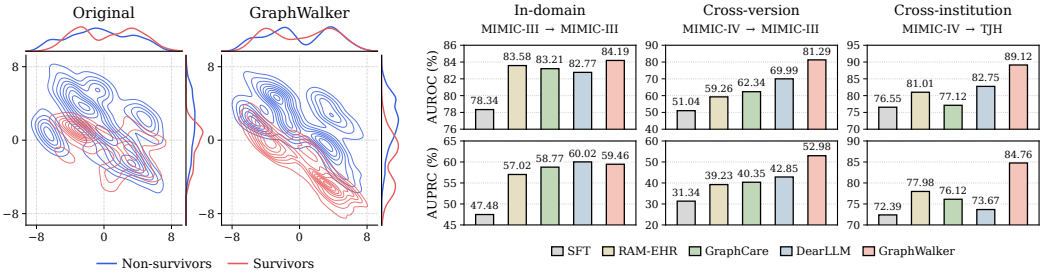
Baselines. We compare `GraphWalker` with representative patient-selection methods from the two perspectives studied in Section 2. From the *data perspective*, we include *Semantic-emb* (Qwen3-Embedding-8B-based) (Zhang et al., 2025b), *EHR-emb* (SMART-based) (Yu et al., 2024), and *Time-series* (Kawarada et al., 2024). From the *model perspective*, we include *SPELL* (Gonen et al., 2023), *Influence* (Nguyen & Wong, 2023), *IDS* (Qin et al., 2024), *CONE* (Peng et al., 2024), *LMS3* (Liu et al.), *GradSel* (Zhang et al., 2025c), and *Delta-KNN* (Li et al., 2025). We additionally include *Zero-shot* and *Random* as reference baselines. All LLM-based methods use the same serialized EHR inputs, backbone LLM, task instructions, output format, and patient budget. Implementation details are provided in Appendix B.4 and B.5.

4.2 Experimental Results

We report the main results with *Qwen3-14B* and *LLaMA-3.1-8B-Instruct* in Table 1; results with the larger *Qwen3-32B* backbone are provided in Appendix Table 4. **`GraphWalker` consistently achieves the strongest performance across backbones and tasks.** Across the evaluated settings, `GraphWalker` outperforms both data- and model-side baselines, consistent with the *Perspective Limitation* identified in our pilot study. Relative to the strongest baseline for each metric and setting, `GraphWalker` achieves average absolute gains of +9.93↑ in AUROC, +12.99↑ in AUPRC, +7.89↑ in F1, +2.86↑ in ma-ROC, and +4.31↑ in mi-ROC. To control for differences in candidate retrieval, we further equip model-side baselines with the same SMART-based retrieval. Although this improves

Table 2: Ablation study of `GraphWalker` using *Qwen3-14B* as the backbone LLM on MIMIC-III and MIMIC-IV. Values are means over three random seeds, and all metrics are higher-is-better (↑). **Bold**: best.

Method	MIMIC-III Mortality			MIMIC-III LOS		MIMIC-IV Readmission		
	AUROC ↑	AUPRC ↑	F1 ↑	ma-ROC ↑	mi-ROC ↑	AUROC ↑	AUPRC ↑	F1 ↑
`GraphWalker`	**84.19**	**59.46**	**58.24**	**69.51**	**84.64**	**75.91**	**60.61**	**58.01**
w/o EHR-emb	61.09	31.13	37.50	62.03	72.63	45.30	26.69	41.74
w/o Cohort	81.19	48.27	55.06	65.85	81.16	75.65	59.61	56.87
w/o Greedy & Frontier Expansion	69.43	46.72	55.34	60.93	53.33	62.43	43.69	47.24
w/o Early Stop	81.34	48.66	55.99	66.31	82.77	75.53	59.13	57.38
Leiden → Louvain	83.15	57.75	56.37	68.16	83.04	75.68	59.83	55.97



(a) KDE .

(b) Mortality prediction across settings.

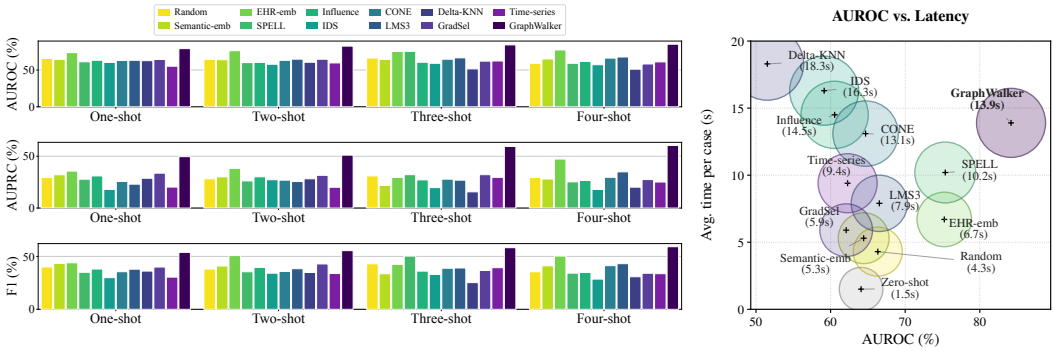
Figure 6: Left: class distributions for Original and `GraphWalker`; blue denotes non-survivors and red denotes survivors. Right: in-domain, cross-version, and cross-institution performance.

several baselines, `GraphWalker` remains stronger, suggesting that its gains cannot be explained by the EHR encoder alone (Appendix D.2.3).

Component Ablation. Table 2 reports ablations of `GraphWalker` on both EHR benchmarks using *Qwen3-14B*. *(i) w/o EHR-emb:* **EHR-informed patient structure matters.** Replacing SMART embeddings with *Qwen3-Embedding-8B* substantially degrades performance, indicating that generic semantic similarity is less effective for organizing patient relations. *(ii) w/o Context-Conditioned Greedy & Frontier Expansion:* **context-conditioned selection and graph exploration matter.** Using individual ΔH scores removes context-conditioned re-evaluation and iterative frontier expansion, substantially degrading performance, consistent with the redundancy and diminishing marginal gains in *Observation II*. A factorial ablation with identical initial candidates and a fixed demonstration count supports both components' contributions (Appendix D.2.8). Removing early stopping retains patients with non-positive marginal gains and further degrades performance. *(iii) w/o Cohort Discovery:* **population-level organization matters.** Starting from the top-K_a nearest patients removes cohort-level localization and reduces performance (*Observation III*). Matching both settings at nine initial candidates further supports the benefit of cohort-based initialization beyond candidate count (Appendix D.2.7). Replacing Leiden with Louvain (Blondel et al., 2008) also consistently reduces performance, suggesting that cohort organization affects subsequent patient search. Effectiveness across multiple pretrained EHR encoders further indicates that `GraphWalker` is not tied to a specific patient representation (Appendix D.2.2).

Understanding Patients through the Clinical Casebook. To examine how historical cases shape target-patient representations, we randomly sample 100 non-survivors and 100 survivors from the MIMIC-III Mortality validation split. Using Qwen3-14B, we extract each patient's final-layer hidden state at the last input token, without historical examples (*Original*) and with `GraphWalker`-selected cases. The t-SNE and kernel density visualizations (Fig. 6a) show less overlap between outcome groups with `GraphWalker`, indicating more distinct outcome-related structure. This qualitatively supports *Reasoning by Patient Analogy*: like clinicians consulting casebooks to interpret a patient's condition, the LLM appears to benefit from selected historical cases in contextualizing the target patient's clinical findings and forming a more outcome-relevant understanding.

Patient-Budget Scaling. We evaluate patient budgets $K \in \{1, 2, 3, 4\}$ on MIMIC-III Mortality and MIMIC-IV Readmission (Figs. 7a and 9). `GraphWalker` consistently outperforms all baselines and is the only method to improve consistently as K increases. CONE, LMS3, and SPELL show non-monotonic or dataset-dependent trends, indicating that more cases do not necessarily improve prediction. These findings suggest that cohort-aware, context-conditioned selection maintains effectiveness across budgets while benefiting from additional cases, underscoring the importance of assessing each case's contribution to the evolving working context. Details appear in Appendix D.2.6.



(a) Performance with different numbers of retrieved cases. (b) AUROC–latency tradeoff.

Figure 7: MIMIC-III mortality prediction. Left: AUROC, AUPRC, and F1 with one to four retrieved cases. Right: AUROC versus average inference time per case; shared methods use corresponding colors across panels.

Cross-Domain Transfer. We compare `GraphWalker` with four supervised baselines on mortality prediction within MIMIC-III and under cross-version (MIMIC-IV to MIMIC-III) and cross-institution (MIMIC-IV to TJH) transfer. `GraphWalker` uses a source-training casebook with both the EHR encoder and LLM frozen; neither transfer setting uses target-domain fine-tuning. As shown in Fig. 6b, `GraphWalker` remains competitive in-domain and leads both transfer settings, with AUROC/AUPRC gains of 11.30/10.13 percentage points on MIMIC-III and 6.37/6.78 on TJH over the strongest baseline for each metric. Given potential patient overlap between the source and target cohorts, we interpret the MIMIC-IV-to-MIMIC-III results as a supplementary cross-version comparison. The TJH results provide the primary evidence for cross-institution generalization without target-domain parameter updates. Protocols and controls appear in Appendices D.3 and D.2.4.

Black-Box Scoring and ReAct Integration. We evaluate `GraphWalker` as a **composable patient-analogy skill** supplying patient case context to downstream agentic workflows. On MIMIC-III Mortality, our black-box variant replaces token-probability-based ΔH with LLM self-evaluation while retaining the graph search. We integrate ReAct (Yao et al., 2022) by replacing direct prediction with ReAct-style reasoning over the same selected cases. The black-box variant substantially outperforms random and similarity-based baselines. ReAct further improves all three metrics for `GraphWalker`, but yields limited or inconsistent gains for weaker selection baselines. These results support `GraphWalker` as a pluggable patient-analogy skill that supplies useful case evidence across scoring interfaces and complements downstream reasoning. Results and scorer stability analyses appear in Appendices D.4 and D.4.2.

Extended Investigations and Key Insights. We further evaluate `GraphWalker` on additional medical reasoning benchmarks and task formats, observing consistent improvements over competing selection methods that support its applicability beyond EHR prediction (Table 16; Appendix E). We also examine the composition-level information-gain proxy and find that ΔH correlates positively with downstream performance, including at fixed patient budgets. The accompanying improvements in probability quality and calibration suggest that the predictive gains are not simply a consequence of increased confidence (Appendix D.2.1). By maintaining a local frontier rather than searching the full patient repository, `GraphWalker` bounds test-time search and achieves end-to-end latency comparable to other LLM-based patient-selection methods with Qwen3-14B (Fig. 7b; Appendix E.1). To guide hyperparameter selection in practice, we also conduct detailed sensitivity analyses, provided in Appendix D.2.5.

5 Conclusion

We presented `GraphWalker`, an inference-time framework for *Reasoning by Patient Analogy* that dynamically consults prior patient experience through cohort-aware graph search and context-conditioned marginal utility. Across EHR prediction and broader medical reasoning benchmarks, `GraphWalker` consistently improves over strong patient-selection baselines, exhibits stronger robustness under distribution transfer, and extends to black-box LLMs and downstream reasoning scaffolds. These results highlight dynamic clinical casebook consultation as a promising paradigm for LLM-based clinical reasoning.

AI USE STATEMENT

During the preparation of this manuscript, large language models (LLMs) assisted with manuscript drafting and language refinement, related-work literature search, coding, figure preparation, LaTeX formatting. All LLM-assisted content included in the manuscript was manually reviewed and validated by the authors. The core research design, experimental analysis, and all scientific conclusions were carried out independently by the authors without LLM involvement. The authors made all final decisions regarding the research design, experimental analysis, and scientific conclusions, and take full responsibility for the final content of this work.

ETHICS STATEMENT

Clinical decision-making is a high-stakes process that requires comprehensive medical evidence and professional judgment. The system studied in this work is intended solely for research and has not been deployed in real-world clinical settings. Its outputs should not be interpreted as definitive medical advice or used to replace qualified healthcare professionals. Our retrospective evaluation does not establish the safety or effectiveness of prospective clinical deployment.

This study uses existing, de-identified patient records from MIMIC-III, MIMIC-IV, and TJH, accessed under their applicable licenses and data-use agreements. No new patient recruitment or prospective clinical intervention was involved. Experiments involving proprietary models were conducted through the Azure OpenAI service with provider-side human review waived, while experiments involving open models were conducted locally. Data access conditions and external API usage are described in Appendices B.1 and B.7.

We acknowledge that demographic imbalances in the datasets, including differences in representation across age, race, ethnicity, and other population attributes, may affect model performance and generalizability. Any future clinical use would require further validation, evaluation of subgroup performance and calibration, secure data processing, and compliance with applicable regulatory and ethical requirements. Model outputs would also require clinician oversight and interpretation alongside clinical records, examinations, and other relevant medical evidence.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we document dataset access and preprocessing in Appendix B.1, task definitions in Appendix B.2, and EHR serialization and shared comparison settings in Appendix B.3. The patient-selection algorithm and prediction prompts are provided in Appendices B.6 and F. Baseline implementations are described in Appendix B.4, while the implementation settings, hyperparameters, and repeated-run protocol for **`GraphWalker`** are specified in Appendix B.5. Pilot-study protocols and the information-gain proxy are detailed in Appendix C. Additional experimental protocols and sensitivity analyses are provided alongside their corresponding appendix results. Our source code is available at `https://anonymous.4open.science/r/GraphWalker-4473`.

REFERENCES

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10):P10008, 2008.

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. Not the models you are looking for: Traditional ml outperforms llms in clinical prediction tasks. *medRxiv*, 2024.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024.

Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Weibin Liao, Xu Chu, Junfeng Zhao, and Yasha Wang. 3ds: Medical domain adaptation of llms via decomposed

difficulty-based data selection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19473–19495, 2025a.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Zheng Li, Junfeng Zhao, and Yasha Wang. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. *arXiv preprint arXiv:2508.13514*, 2025b.

Yue Fang, Yongxin Xu, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Hongyan Jin. A method and practice for menopausal disease prediction based on knowledge graph. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pp. 10–18. IEEE, 2023.

Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, Weibin Liao, Yongxin Xu, Yinghao Zhu, Zhibang Yang, Liantao Ma, et al. Toward better ehr reasoning in llms: Reinforcement learning with expert attention guidance. *arXiv preprint arXiv:2508.13579*, 2025.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.

Junyi Gao, Yinghao Zhu, Wenqing Wang, Zixiang Wang, Guiying Dong, Wen Tang, Hao Wang, Yasha Wang, Ewen M Harrison, and Liantao Ma. A comprehensive benchmark for covid-19 predictive modeling using electronic health records in intensive care. *Patterns*, 5(4), 2024.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10136–10148, 2023.

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

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Pengcheng Jiang, Cao Danica Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *International Conference on Learning Representations*, volume 2024, pp. 16839–16866, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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. Mimic-iii, a freely accessible critical care database. *Sci. Data*, 2016.

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

Masayuki Kawarada, Tatsuya Ishigaki, Goran Topić, and Hiroya Takamura. Demonstration selection strategies for numerical time series data-to-text. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7378–7392, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Chuyuan Li, Raymond Li, Thalia S Field, and Giuseppe Carenini. Delta-knn: Improving demonstration selection in in-context learning for alzheimer's disease detection. *arXiv preprint arXiv:2506.03476*, 2025.

Weibin Liao, Yinghao Zhu, Zhongji Zhang, Yuhang Wang, Zixiang Wang, Xu Chu, Yasha Wang, and Liantao Ma. Learnable prompt as pseudo-imputation: Rethinking the necessity of traditional ehr data imputation in downstream clinical prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 765–776, 2025.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

Jiayu Liu, Zhenya Huang, Chaokun Wang, Xunpeng Huang, ChengXiang Zhai, and Enhong Chen. What makes in-context learning effective for mathematical reasoning. In *Forty-second International Conference on Machine Learning*.

Stuart Lubarsky, Valérie Dory, Marie-Claude Audétat, Eugène Custers, and Bernard Charlin. Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian medical education journal*, 6(2):e61, 2015.

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

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

Liantao Ma, Chaohe Zhang, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Yinghao Zhu, Tianlong Wang, Xinyu Ma, Yasha Wang, Wen Tang, et al. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients. *Patterns*, 4(12), 2023.

Tai Nguyen and Eric Wong. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*, 2023.

Geoffrey Norman. Building on experience—the development of clinical reasoning, 2006.

Geoffrey R Norman and Lee R Brooks. The non-analytical basis of clinical reasoning. *Advances in health sciences education*, 2(2):173–184, 1997.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning. In *ACL (1)*, 2024.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455, 2024.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Olle ten Cate and Steven J Durning. Understanding clinical reasoning from multiple perspectives: a conceptual and theoretical overview. *Principles and practice of case-based clinical reasoning education: A method for preclinical students*, pp. 35–46, 2017.

Olle Ten Cate, Eugène JFM Custers, and Steven J Durning. Principles and practice of case-based clinical reasoning education: a method for preclinical students. 2017.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, et al. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6184–6205, 2024.

Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3):1035, 2013.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 754–765, 2024a.

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

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, volume 23, pp. 4921–4929, 2023b.

Yongxin Xu, Xinke Jiang, Xu Chu, Yuzhen Xiao, Chaohe Zhang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Protomix: Augmenting health status representation learning via prototype-based mixup. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3633–3644, 2024b.

Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Dearllm: Enhancing personalized healthcare via large language models-deduced feature correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 941–949, 2025.

Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature machine intelligence*, 2(5):283–288, 2020.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.

Zhihao Yu, Chu Xu, Yujie Jin, Yasha Wang, and Junfeng Zhao. Smart: Towards pre-trained missing-aware model for patient health status prediction. *Advances in Neural Information Processing Systems*, 37:63986–64009, 2024.

Jinyang Zhang, Yue Fang, Hongxin Ding, Weibin Liao, Muyang Ye, Xu Chu, Junfeng Zhao, and Yasha Wang. Adept: Continual pretraining via adaptive expansion and dynamic decoupled tuning. *arXiv preprint arXiv:2510.10071*, 2025a.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Ziniu Zhang, Zhenshuo Zhang, Dongyue Li, Lu Wang, Jennifer Dy, and Hongyang R Zhang. Linear-time demonstration selection for in-context learning via gradient estimation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16470–16488, 2025c.

Yinghao Zhu, Junyi Gao, Zixiang Wang, Weibin Liao, Xiaochen Zheng, Lifang Liang, Miguel O Bernabeu, Yasha Wang, Lequan Yu, Chengwei Pan, et al. Clinicrealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. *npj Digital Medicine*, 2026.

A Related Work

Recent years have witnessed growing interest in applying LLMs to clinical reasoning over EHRs. One line of work adapts LLMs to the medical domain through continual pre-training or post-training (Zhang et al., 2025a; Ding et al., 2025a; Fang et al., 2025; Ding et al., 2025b); a parallel line builds hybrid pipelines that incorporate LLM-derived knowledge into supervised EHR predictors (Jiang et al., 2024; Xu et al., 2025; 2024a). Both directions involve parameter updates, either to the LLM or to a downstream predictive model. Despite these advances, benchmark studies show that medical language competence does not consistently translate into strong predictive performance on longitudinal EHRs (Brown et al., 2024; Chen et al., 2024; Zhu et al., 2026). These challenges motivate inference-time strategies that support LLM predictions by drawing on relevant historical patient experience without additional parameter updates.

Demonstration Selection for LLM Reasoning. Existing demonstration selection methods use data-driven or model-driven criteria to identify useful examples. Data-driven approaches emphasize query relevance through lexical or embedding similarity, or the representativeness and diversity of example pools (Liu et al., 2022; Hongjin et al., 2022; Robertson et al., 2009). Model-driven approaches use model-dependent criteria and feedback, including perplexity, conditional entropy, gradients, and influence estimates, to assess prompt quality and guide demonstration selection (Gonen et al., 2023; Peng et al., 2024; Liu et al.; Li et al., 2025; Zhang et al., 2025c; Nguyen & Wong, 2023; Qin et al., 2024). For EHR-based clinical reasoning, our analysis highlights the need to connect patient relatedness with model-side utility (*Fragmented Perspectives*) and to account for changes in a case's marginal contribution as the working context evolves (*Static Utility Assumption*). These challenges motivate *Reasoning by Patient Analogy*, which organizes historical patient experience into a clinical casebook and couples context-dependent case selection with subsequent case exploration.

Positioning of GraphWalker. GraphWalker emphasizes complementary data- and model-side guidance for clinical case selection. On the data side, it uses population structure derived from historical EHRs as a structural prior for search. A pretrained EHR encoder provides patient representations, a patient graph captures local relationships, and graph-based cohort discovery identifies groups whose prototypes provide population-level entry points. The graph further guides candidate exploration through the neighbors of selected patients. On the model side, the target LLM evaluates each candidate's marginal utility under the currently selected cases. This formulation builds on, but differs from, existing demonstration selection approaches. TopK + ConE (Peng et al., 2024) combines relevance-based candidate filtering with model-dependent reranking; its TopK stage identifies individually relevant candidates without explicitly constructing patient cohorts or using their population structure to guide subsequent exploration. CEIL (Ye et al., 2023) models interactions among demonstrations through a learned conditional determinantal point process and performs incremental set selection. GraphWalker instead couples an EHR-derived population structure with direct, context-conditioned utility evaluation by the target LLM: patient cohorts initialize the search, model-side utility determines the next selected case, and that selection exposes graph neighbors for further evaluation.

B Experimental Setup

B.1 Datasets and Preprocessing

We conduct our main experiments on two widely used EHR benchmarks: **MIMIC-III v1.4** (Johnson et al., 2016) and **MIMIC-IV v2.2** (Johnson et al., 2023). Both are large-scale critical-care databases collected at the Beth Israel Deaconess Medical Center and made available through PhysioNet to credentialed researchers. They contain longitudinal, de-identified ICU records, including vital signs, laboratory measurements, and medication administrations. For out-of-distribution evaluation, we additionally use the **TJH** dataset (Yan et al., 2020), which contains de-identified structured EHRs from Tongji Hospital and serves as an independent cross-institution target domain.

Data Access and Research Scope. The datasets were accessed under their respective licenses and data-use agreements. Access to MIMIC-III and MIMIC-IV requires researcher credentialing, completion of the required training, and acceptance of the PhysioNet data-use agreement. Our study

Dataset	Split	Patients	Alive	Dead	Re.	No Re.
MIMIC-III	Train	8160	7355	805	–	–
	Validation	1020	919	101	–	–
	Test	1020	920	100	–	–
	Total	10200	9194	1006	–	–
MIMIC-IV	Train	8160	7486	674	1976	6184
	Validation	1020	936	84	247	773
	Test	1020	936	84	247	773
	Total	10200	9358	842	2470	7730
TJH	Total	361	195	166	–	–

Table 3: *Statistics of the MIMIC-III, MIMIC-IV, and TJH datasets.* "Re." denotes ICU readmission within 30 days after ICU discharge, while "No Re." denotes no ICU readmission within this window. TJH is reserved for external testing.

uses existing, de-identified records for retrospective evaluation and does not involve new patient recruitment or prospective clinical intervention.

Data Preprocessing. We preprocess the structured EHR data using dataset-specific procedures adapted from prior EHR benchmarks (Gao et al., 2024; Zhu et al., 2026). For long stays, we retain clinical measurements from the most recent seven days. Missing values are handled using the Last Observation Carried Forward (LOCF) strategy (Wells et al., 2013).

For each patient, we designate the last visit as the *target visit* and use all preceding visits as input. Records from the target visit are excluded from the input. For mortality and LOS prediction, the target visit refers to the hospital admission whose outcome is evaluated. Specifically, for **in-hospital mortality**, we predict whether the patient dies during the target hospital admission; for **length-of-stay (LOS)**, we predict the total LOS bin (Harutyunyan et al., 2019) of that admission. For **ICU readmission**, the outcome is defined relative to discharge from the target ICU stay: we predict whether the patient is readmitted to the ICU within 30 days after that discharge. The same prediction protocol is applied uniformly to all baselines and to `GraphWalker`.

To enable controlled comparisons across a broad set of LLM backbones and case-selection baselines, each requiring inference-time forward passes for every patient, we follow prior EHR studies (Zhu et al., 2026) and partition each MIMIC dataset into training, validation, and test sets at the patient level using an 8:1:1 ratio. All records, admissions, and ICU stays associated with the same patient are assigned exclusively to a single split, ensuring no patient overlap across the three splits. TJH is used exclusively as an external test set for cross-institution transfer and is not used for training, model selection, hyperparameter tuning, graph construction, or casebook construction. Detailed split sizes and outcome distributions are reported in Table 3. For the in-domain experiments, the patient graph and clinical casebook are constructed solely from the corresponding training split. For cross-institution transfer, they are constructed solely from the MIMIC-IV training split, while the entire TJH cohort is held out for evaluation.

B.2 Task Definitions

Definition 1 (EHR Dataset). A patient's historical EHR input is represented as a sequence of T time-ordered visits $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, where $\mathbf{x}_t$ denotes the clinical records associated with the t-th historical visit, including vital signs and laboratory measurements. The target visit is excluded from $\mathbf{X}$.

Definition 2 (Mortality Prediction). Given the historical EHR input $\mathbf{X}$, predict whether the patient dies during the target hospital admission. The label $y \in \{0, 1\}$ denotes death ($y = 1$) or survival ($y = 0$).

Definition 3 (ICU Readmission Prediction). Given the historical EHR input $\mathbf{X}$, predict whether the patient is readmitted to the ICU within 30 days after discharge from the target ICU stay. The label $y \in \{0, 1\}$ indicates ICU readmission within this window ($y = 1$) or no ICU readmission ($y = 0$).

Definition 4 (Length-of-Stay Prediction). Given the historical EHR input $\mathbf{X}$, predict the total length of the target hospital admission. Following prior work (Harutyunyan et al., 2019), we formu-

late this task as a classification problem. In our experiments, $y \in \{1, 2, 3, 4\}$ denotes one of four predefined length-of-stay categories, represented by the answer labels A, B, C, and D, respectively. The single-letter response specifies the predicted category, while AUROC is computed from scores for all four classes. Given the final prediction prompt, we compute the conditional log-likelihood of each candidate label at the answer position and apply a softmax over the four log-likelihoods to obtain normalized class scores. For labels spanning multiple tokens, we sum the conditional log-probabilities over the label tokens. Macro-AUROC is the unweighted mean of the four one-vs-rest AUROCs. Micro-AUROC is computed by pooling all patient–class pairs using the one-hot ground-truth labels and their corresponding class scores. The same label formatting and scoring procedure are used across all compared LLM-based LOS methods.

B.3 EHR Serialization and Shared Settings

Following (Fang et al., 2025), we convert each preprocessed longitudinal EHR into text using a shared template. This template is applied consistently to target patients and historical patient examples, ensuring that the same patient record has the same textual representation across selection methods. Within each matched comparison, we use the same backbone LLM, prediction instruction, output format, and maximum number of selected patient examples. Each method selects historical patients according to its own selection strategy and uses their textual records to construct the prediction context. Method-specific selection procedures are described below.

B.4 Baseline Implementation Details

We compare `GraphWalker` with representative baselines spanning multiple paradigms. In this subsection, we provide implementation details for each baseline. For all methods, we strictly constrain the LLM to directly output the final prediction in the task-specific format, without generating any intermediate reasoning or explanations, following the prompt templates described in Appendix F.

- **Vanilla in-context learning (ICL) baselines.** *Zero-shot* corresponds to a special case of ICL where no demonstration examples are provided, and the LLM directly produces predictions based on the task description and patient EHR data. *Random* samples a fixed number of patient examples uniformly at random as demonstrations. Both baselines adopt the same prompt structure as described in Appendix F, differing only in whether in-context examples are included.
- **Supervised.** *SFT* adapts LLMs to downstream tasks by fine-tuning model parameters on task-specific data. Specifically, we organize the training set into question–answer pairs using the same prompt templates described in Appendix F, and apply parameter-efficient fine-tuning via LoRA (Hu et al., 2022) to update the backbone LLM.
- **Data Perspective Selection.** *Semantic-emb* retrieves top-k demonstrations based on similarity in the embedding space of *Qwen3-Embedding-8B* (Zhang et al., 2025b), a widely used and strong general-purpose semantic model. For efficiency, we compute embeddings using vLLM (Kwon et al., 2023) to accelerate large-scale embedding inference. *EHR-emb* retrieves top-k demonstrations based on similarity computed from embeddings generated by *SMART* (Yu et al., 2024), a pre-trained model specifically designed for EHR data. We follow the hyperparameter settings recommended in the original work and use the `[CLS]` token embedding as the patient representation, as it encodes global semantic information of the longitudinal EHR sequence. *Time-series* (Kawarada et al., 2024) retrieves top-k demonstrations based on the similarity between the test patient and training examples' longitudinal EHR time series. We use a scale-invariant masked cosine similarity by aligning overlapping prefixes, truncating to the minimum length, and computing cosine similarity only over positions with valid values, ignoring missing entries. Demonstrations are ranked by similarity, and the top-k are selected as in-context examples.
- **Model Perspective Selection.** *SPELL* (Gonen et al., 2023) selects demonstrations by ranking candidates with LM perplexity (i.e., teacher-forced next-token cross-entropy) and choosing the lowest-perplexity top-k examples. In our implementation, we score all training examples with the same causal LM and select the top-k demonstrations by sorting the average loss in ascending order. The resulting few-shot set is shared across all test instances and inserted into the ICL prompt as demonstrations, without per-test-instance re-selection.
  Influence (Nguyen & Wong, 2023) estimates the importance of demonstrations by measuring their marginal contribution to the model's performance on a validation set. It calculates influence scores

based on the average performance difference between sampled subsets that include a specific example and those that exclude it.

IDS (Qin et al., 2024) selects demonstrations by calculating the semantic similarity between training examples and reasoning paths generated through Zero-shot CoT. The framework iteratively refines the demonstration set by utilizing the model's self-generated reasoning from the previous round as an updated query for subsequent retrieval.

CONE (Peng et al., 2024) selects demonstrations by minimizing the conditional entropy of the test input on a per-example basis, without modeling complementarity among selected examples. Following the original implementation, we first perform a semantic pre-filtering step using *Qwen3-Embedding-8B* to retrieve the top-10 semantically similar candidates. We then select the top-k demonstrations with the largest reduction in conditional entropy, where the entropy is computed independently for each candidate example.

LMS3 (Liu et al.) selects demonstrations by balancing LLM-oriented semantic similarity and inference stability, derived from a theoretical bound on reasoning efficacy. It uniquely incorporates a demonstration rejection mechanism that adaptively filters out unsuitable examples to prevent negative transfer.

Delta-KNN (Li et al., 2025) selects demonstrations by prioritizing candidates that were empirically observed to be most helpful for similar inputs, using a cached example-to-example gain matrix. In our implementation, we first take a fixed training subset of examples to construct the Delta matrix via paired zero-shot vs. one-shot inference, and cache the resulting matrix for each dataset/LLM setting. At test time, we use *Qwen3-Embedding-8B* to retrieve the top-10 semantically similar candidates, then re-rank them with the Delta-based scores and select the final top-k demonstrations. The selected three demonstrations are re-selected per test instance and inserted into the ICL prompt as demonstrations for final inference.

GradSel (Zhang et al., 2025c) estimates the utility of sampled demonstration subsets using a first-order approximation in the input-embedding space, based on the model output and gradients computed at anchor prompts. It then aggregates subset-level scores into point-wise demonstration scores and selects the top-ranked examples as final demonstrations.

B.5 Implementation Details of `GraphWalker`

For patient graph construction, we set the number of nearest neighbors to $k_g = 8$. For patient-specific cohort retrieval, the number of retrieved cohorts is set to $K_c = 3$, and within each retrieved cohort, we select $K_a = 3$ anchor nodes to initialize the search frontier. A detailed analysis and discussion of these key hyperparameters are provided in Appendix D.2.5. All experiments are repeated three times with different random seeds. Main-text tables report mean performance for readability, while variability is reported in the corresponding supplementary analyses where relevant. To accelerate query-level conditional cross-entropy computation during patient selection, we use `vLLM` (Kwon et al., 2023) for efficient inference.

B.6 Full Algorithm of `GraphWalker`

Algorithm 1 presents the full context-conditioned greedy search used by **`GraphWalker`**.

B.7 Data Privacy and External API Usage

Experiments involving proprietary models were conducted through the institutional Azure OpenAI service under a configuration that waived provider-side human review of submitted content. API requests contained de-identified target-patient and retrieved-case records, as described in the corresponding experimental sections. All experiments involving open models were conducted locally within the institutional computing environment.

Our evaluation was retrospective and does not establish the safety or effectiveness of prospective clinical deployment. Any future clinical use would require secure data processing, evaluation of subgroup performance and calibration, and clinician oversight.

Algorithm 1 `GraphWalker`

Require: Training EHR dataset $\{\mathbf{X}_i\}_{i=1}^{N}$; test patient record $\mathbf{X}^{(q)}$; pretrained EHR encoder $\mathcal{M}_{\text{exp}}$; LLM M_θ; graph neighbor size k_g; number of retrieved cohorts K_c; anchors per cohort K_a; patient budget K.

```
# Stage 1: Patient Graph Construction and Cohort Discovery
for i ← 1 to N do
    h_i ← M_exp(X_i)
end for
Construct a k_g-NN patient graph G = (V, E) over {h_i} using cosine similarity
{C_m}_{m=1}^{M} ← LEIDEN(G)
for m ← 1 to M do
    z_m ← (1/|C_m|) Σ_{i∈C_m} h_i
end for
# Stage 2: Patient-Specific Cohort Retrieval and Anchor Initialization
h^(q) ← M_exp(X^(q))
C_ret^(q) ← Top-K_c(sim(h^(q), z_m))
F^(q) ← ∅
for each cohort C_m ∈ C_ret^(q) do
    V_m^(q) ← Top-K_a(sim(h^(q), h_i)),  v_i ∈ C_m
    F^(q) ← F^(q) ∪ V_m^(q)
end for
# Stage 3: Context-Conditioned Greedy Search with Frontier Expansion
S^(q) ← ∅
H_0 ← H_θ(X^(q) | S^(q))
while |S^(q)| < K and F^(q) ≠ ∅ do
    v* ← None; Δ* ← 0; H* ← H_0
    for each v_i ∈ F^(q) do
        H_i ← H_θ(X^(q) | S^(q) ∪ {v_i})
        ΔH_θ(v_i | S^(q)) ← H_0 − H_i
        if ΔH_θ(v_i | S^(q)) > Δ* then
            v* ← v_i; Δ* ← ΔH_θ(v_i | S^(q)); H* ← H_i
        end if
    end for
    if v* = None or Δ* ≤ 0 then
        break
    end if
    S^(q) ← S^(q) ∪ {v*}
    H_0 ← H*
    F^(q) ← F^(q) \ {v*}
    F^(q) ← F^(q) ∪ N(v*)
    F^(q) ← F^(q) \ S^(q)
end while
return S^(q)
```

C Experimental Setup of the Pilot Study

C.1 Pilot Study Protocol

We use *Qwen3-14B* (Yang et al., 2025) as the backbone LLM for all pilot experiments and evaluate on MIMIC-III Mortality and MIMIC-IV Readmission. For each task, the working context is constructed by the corresponding patient-selection strategy. Dataset preprocessing, task definitions, and prompt templates follow Appendix B.1, Appendix B.2, and Appendix F, respectively.

C.2 Composition-Level Information-Gain Proxy

Prior work (Peng et al., 2024) observes that effective demonstration-selection methods can reduce the conditional entropy of the target input, and suggests that useful examples help the model better understand that input. Their practical selection procedure scores candidate examples individually, without explicitly accounting for interactions among the selected examples. In our clinical casebook setting, the contribution of a historical patient may depend on the other cases included in the working context. Cases that are individually useful may provide overlapping information, while others may offer complementary clinical context. This motivates evaluating the selected cases jointly.

To compare the contexts constructed by different patient-selection strategies, we measure the reduction in target-record conditional cross-entropy conditioned on the entire selected case set relative to the zero-shot context. We refer to this quantity as the *composition-level information-gain proxy*.

Formally, let x_{test} denote the serialized target-patient record and $\mathcal{S}_k$ a selected set of k historical patients. We define $H_\theta(x_{\text{test}} \mid \mathcal{S}_k)$ as the average teacher-forced negative log-probability of the target-record tokens conditioned on the task instruction and the selected patient examples. Only target-record tokens contribute to this loss; instruction and demonstration tokens serve as conditioning context. The target patient's ground-truth outcome is excluded from both the conditioning context and the scored tokens.

The composition-level information-gain proxy is

$$\Delta H_\theta(\mathcal{S}_k) = H_\theta(x_{\text{test}} \mid \emptyset) - H_\theta(x_{\text{test}} \mid \mathcal{S}_k). \tag{5}$$

Here, $\emptyset$ denotes the absence of historical patient examples, with the task instruction retained. A larger $\Delta H_\theta(\mathcal{S}_k)$ indicates a greater improvement in the predictability of the target record relative to the zero-shot context.

This composition-level quantity measures the benefit of the entire selected case set. The marginal information-gain proxy used during selection instead measures the additional contribution of a new case given those already included in the working context.

For each task, selection strategy, and $k \in \{1, 2, 3, 4\}$, we compute this quantity for every validation instance and report its mean over the corresponding validation set.

D Additional Experimental Results and Analyses

D.1 RQ1: Additional Results with Qwen3-32B

To evaluate whether the gains of **`GraphWalker`** persist with a larger target LLM, we additionally report results with *Qwen3-32B*. As shown in Table 4, **`GraphWalker`** maintains strong performance across MIMIC-III Mortality, MIMIC-III LOS, and MIMIC-IV Readmission, indicating that the benefits are not specific to a single backbone scale.

D.2 RQ2: Framework Analysis and Controls

D.2.1 Validation of the Information-Gain Proxy

Correlation with Downstream Performance. We first examine whether ΔH provides a useful information-gain proxy for patient selection. Using Qwen3-14B on the validation splits of MIMIC-III Mortality and MIMIC-IV Readmission, we evaluate each patient-selection method under budgets $K \in \{1, 2, 3, 4\}$. For each method–budget configuration, we average composition-level ΔH over validation patients and compute AUPRC from the corresponding predictions. For each dataset, we then compute Spearman rank correlations across all method–budget configurations and separately across methods at each fixed budget.

As shown in Table 5, ΔH exhibits consistent positive correlations with downstream AUPRC on both datasets. The correlations remain positive within every fixed-budget comparison (0.54–0.83), suggesting that the overall association is not solely attributable to differences in the prescribed patient budget. These results support ΔH as a practical information-gain proxy for patient selection.

Method		MIMIC-III Mortality			MIMIC-III LOS		MIMIC-IV Readmission		
Paradigm	**Approach**	AUROC ↑	AUPRC ↑	F1 ↑	ma-ROC ↑	mi-ROC ↑	AUROC ↑	AUPRC ↑	F1 ↑
Qwen3-32B									
Vanilla	Zero-shot	65.67	30.54	38.43	55.79	23.53	44.19	25.06	42.57
	Random	59.90	24.63	36.22	59.50	<u>80.44</u>	48.25	28.13	42.22
Data-side	Semantic-emb	60.62	19.87	32.08	57.65	74.16	47.86	26.07	42.47
	EHR-emb	<u>85.02</u>	<u>58.56</u>	<u>58.84</u>	<u>63.02</u>	67.00	<u>61.32</u>	<u>44.25</u>	<u>46.24</u>
	Time-series	63.71	28.74	37.58	51.95	38.06	52.34	27.02	40.79
Model-side	SPELL	63.26	27.26	38.33	53.87	75.06	48.56	25.90	40.17
	Influence	62.45	29.28	40.18	54.24	41.07	50.45	27.63	42.69
	IDS	56.72	28.37	35.23	52.47	45.82	46.93	27.71	43.16
	CONE	64.28	27.83	36.05	56.30	68.90	45.35	27.85	41.73
	LMS3	62.03	25.14	36.68	58.14	73.15	45.28	26.09	42.79
	Delta-KNN	61.08	29.41	39.36	56.55	64.86	51.75	26.74	40.37
	GradSel	60.77	24.53	39.30	53.87	44.22	47.08	25.53	41.93
	`GraphWalker`	**85.88**	**59.19**	**61.89**	**67.40**	**83.60**	**77.57**	**63.09**	**59.51**

Table 4: Additional results on MIMIC-III and MIMIC-IV datasets using *Qwen3-32B* as the backbone LLM. All metrics are higher-is-better (↑); values are means over three random seeds. **Bold**: best; <u>underlined</u>: second-best.

Dataset	Overall	1-shot	2-shot	3-shot	4-shot
MIMIC-III Mortality	**0.67**	0.71	0.60	0.54	0.66
MIMIC-IV Readmission	**0.72**	0.83	0.71	0.60	0.83

Table 5: Spearman rank correlation between composition-level ΔH and downstream AUPRC. Overall correlations include all method–budget configurations; fixed-budget correlations compare methods under the same prescribed patient budget.

Calibration Analysis. We further test whether the performance gains are accompanied merely by increased confidence rather than improved prediction quality. On MIMIC-III Mortality, we compare Zero-shot, EHR-emb, and `GraphWalker` using AUPRC, true-label negative log-likelihood (NLL), Brier score, and 10-bin expected calibration error (ECE).

As shown in Table 6, `GraphWalker` **achieves higher AUPRC while reducing true-label NLL, Brier score, and ECE. Thus, its performance gains are accompanied by improved probability quality and calibration, rather than simply increased model confidence.**

D.2.2 Generality across EHR Encoders

Setup. To assess the robustness of `GraphWalker` to the choice of EHR representation, we instantiate it with three pretrained EHR encoders of different architectures, all trained exclusively on the training split of the corresponding experiment: **SMART** (Yu et al., 2024), **ConCare** (Ma et al., 2020b), and **AdaCare** (Ma et al., 2020a). For each encoder, we compare two settings: (i) the encoder fine-tuned end-to-end as a task-specific supervised predictor (*encoder baseline*), and (ii) `GraphWalker` using the same pretrained encoder *without fine-tuning* for graph-based patient selection. Note that the encoder baseline here differs from *EHR-emb* in Table 1, which uses SMART embeddings for patient retrieval without fine-tuning.

Results. Table 7 shows that `GraphWalker` consistently improves over all encoder baselines across both MIMIC-III mortality and MIMIC-IV readmission, with consistent absolute gains across metrics and encoders. The improvements hold across encoders of distinct architectural styles, indicating that `GraphWalker`'s effectiveness is not tied to any specific EHR representation and benefits from its cohort-aware, context-conditioned patient search. Notably, `GraphWalker` **surpasses these encoders** ***even when they are fine-tuned end-to-end on the target task*****, underscoring that the gains**

Method	AUPRC ↑	NLL ↓	Brier ↓	ECE ↓
Zero-shot	17.58	0.431	0.119	0.097
EHR-emb	29.52	0.397	0.108	0.103
`GraphWalker`	**59.46**	**0.319**	**0.083**	**0.071**

Table 6: Prediction quality and calibration on MIMIC-III Mortality with Qwen3-14B. Values are means over three random seeds.

Method	MIMIC-III Mortality			MIMIC-IV Readmission		
	AUROC↑	AUPRC↑	F1↑	AUROC↑	AUPRC↑	F1↑
SMART (fine-tuned predictor)	80.64	45.40	47.18	74.41	54.62	57.38
`GraphWalker` (with SMART encoder)	**84.19**	**59.46**	**58.24**	**75.91**	**60.61**	**58.01**
Δ	+3.55	+14.06	+11.06	+1.50	+5.99	+0.63
ConCare (fine-tuned predictor)	69.19	36.01	40.65	70.30	41.87	50.26
`GraphWalker` (with ConCare encoder)	**70.46**	**36.93**	**45.90**	**77.01**	**59.49**	**59.44**
Δ	+1.27	+0.92	+5.25	+6.71	+17.62	+9.18
AdaCare (fine-tuned predictor)	60.67	19.37	26.81	78.40	59.37	54.87
`GraphWalker` (with AdaCare encoder)	**65.61**	**27.13**	**37.89**	**80.45**	**62.72**	**60.04**
Δ	+4.94	+7.76	+11.08	+2.05	+3.35	+5.17

Table 7: Performance of `GraphWalker` with different EHR encoders on MIMIC-III mortality and MIMIC-IV readmission, using *Qwen3-14B* as the backbone LLM. For each encoder, we compare two settings: (i) the encoder fine-tuned end-to-end as a task-specific supervised predictor, and (ii) `GraphWalker` using the same pretrained encoder (without fine-tuning) for graph-based demonstration selection. Δ denotes the absolute improvement (in percentage points) of `GraphWalker` over the supervised fine-tuned baseline. Note that this differs from *EHR-emb* in Table 1, which uses SMART embeddings for ICL retrieval without fine-tuning.

come from how patient experience is selected and composed, not merely from the underlying encoder.

D.2.3 Controlling for the EHR Encoder

We examine whether the advantage of `GraphWalker` can be explained solely by its use of domain-specific EHR representations. We construct SMART-enhanced variants of selected model-side baselines by using cosine similarity between SMART patient embeddings to retrieve candidate cases. Each variant then applies its method-specific scoring procedure to the retrieved candidates. For SPELL, our baseline implementation selects a case set shared across target patients. We additionally evaluate a SMART-restricted variant that introduces candidate filtering while retaining the same perplexity-based scoring criterion. All methods use the same *LLaMA-3.1-8B-Instruct* backbone, EHR serialization, task prompt, output format, and maximum patient budget of three. Random and EHR-emb are included as reference methods.

Equipping the five model-side baselines with SMART-based patient representations improves their performance relative to Table 1. Under this control, `GraphWalker` retains its advantage, outperforming the best SMART-enhanced baseline, Delta-KNN + SMART, by 9.40, 6.97, and 6.07 percentage points in AUROC, AUPRC, and F1, respectively. **These results suggest that the advantage of `GraphWalker` extends beyond the use of an EHR-specific encoder and support the additional contribution of its patient-selection strategy.**

D.2.4 Controlling for Direct SMART Prediction Signals

We examine whether the performance of `GraphWalker` benefits from direct access to prediction signals produced by the EHR encoder. In the default setting, SMART is used only to obtain frozen patient representations for graph construction, cohort discovery, cohort retrieval, and anchor initialization; its prediction logits are not provided to the target LLM.

As an additional control, we construct a logit-augmented variant that appends SMART prediction logits to the final LLM input while keeping all other components unchanged. We evaluate both variants with Qwen3-14B on MIMIC-III Mortality and MIMIC-IV Readmission.

Method	AUROC ↑	AUPRC ↑	F1 ↑
Random	52.04	18.81	28.90
EHR-emb (SMART-based)	65.78	24.30	41.89
SPELL + SMART	63.92	24.58	41.24
CONE + SMART	66.42	26.88	43.20
LMS3 + SMART	64.87	25.74	42.31
Delta-KNN + SMART	67.31	28.12	44.06
GradSel + SMART	62.84	24.96	40.87
`GraphWalker`	**76.71**	**35.09**	**50.13**

Table 8: Comparison of `GraphWalker`, SMART-enhanced model-side baselines, and reference methods on MIMIC-III Mortality using *LLaMA-3.1-8B-Instruct*. All metrics are higher-is-better (↑); values are means over three random seeds.

Dataset	Setting	AUROC ↑	AUPRC ↑	F1 ↑
MIMIC-III Mortality	`GraphWalker` (Embedding Only)	84.19	59.46	**58.24**
	`GraphWalker` + SMART Logits	**84.52**	**60.08**	57.91
MIMIC-IV Readmission	`GraphWalker` (Embedding Only)	**75.91**	60.61	58.01
	`GraphWalker` + SMART Logits	75.48	**60.96**	**58.36**

Table 9: Effect of providing SMART prediction logits directly to the target LLM. The default `GraphWalker` uses SMART only as a frozen patient representation encoder.

As shown in Table 9, adding SMART logits produces only small and inconsistent changes across metrics, with a maximum absolute difference of 0.62 points. The logit-augmented variant improves some metrics while degrading others, showing no consistent advantage over the embedding-only setting. These results indicate that **the gains of `GraphWalker` arise primarily from using EHR-informed representations to organize and dynamically consult prior patient experience, rather than from direct access to SMART predictions.**

D.2.5 Hyper-Parameter Sensitivity

We analyze three key hyper-parameters of `GraphWalker` on the validation set of *MIMIC-IV Readmission* with *Qwen3-14B* as the backbone LLM: the graph neighborhood size k_g, the number of retrieved cohorts K_c, and the number of anchor nodes per cohort K_a. All other settings are kept fixed.

Overall, performance generally favors moderate settings that balance candidate coverage and clinical coherence. We select the default configuration on the validation set and keep it fixed for all test evaluations; the following analyses characterize practical operating regions rather than statistically unique optima.

Graph Neighborhood Size k_g. We vary k_g from 4 to 10 (Fig. 8(a)). Performance steadily improves as k_g grows from 4 to 8, indicating that a moderate neighborhood captures richer population-level structure. When k_g further increases to 10, performance declines as the graph becomes overly dense and introduces weakly related neighbors, blurring local structure and propagating less informative signals during frontier expansion.

Number of Retrieved Cohorts K_c. Varying K_c from 1 to 4 (Fig. 8(b)), both AUPRC and F1 improve from 1 to 3, suggesting that retrieving multiple cohorts enriches the candidate pool with diverse but clinically related subpopulations and mitigates over-specialization to a single cohort. Performance declines at $K_c = 4$, where additional loosely related cohorts begin to dilute clinically meaningful signals.

Number of Anchor Patients K_a. We examine the sensitivity of `GraphWalker` to the number of anchor patients selected from each retrieved cohort. The cohort partition is kept fixed; K_a controls the size of the initial search frontier. A smaller K_a provides fewer starting points for patient search,

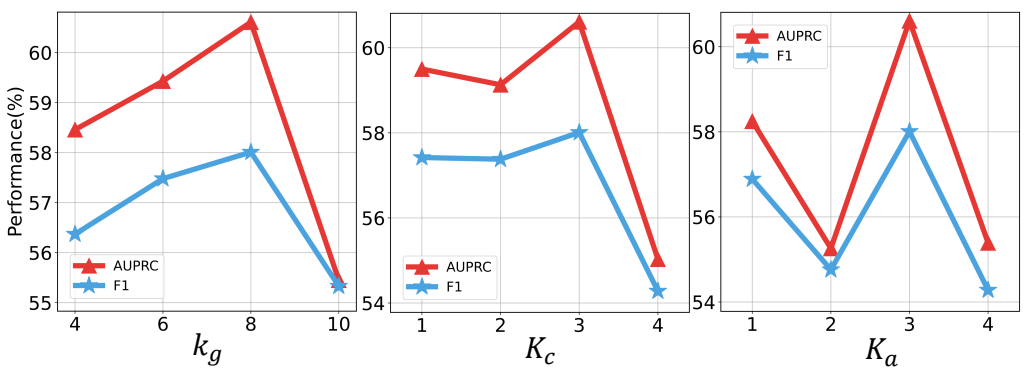


Figure 8: Performance of **GraphWalker** under different hyper-parameter settings on MIMIC-IV Readmission.

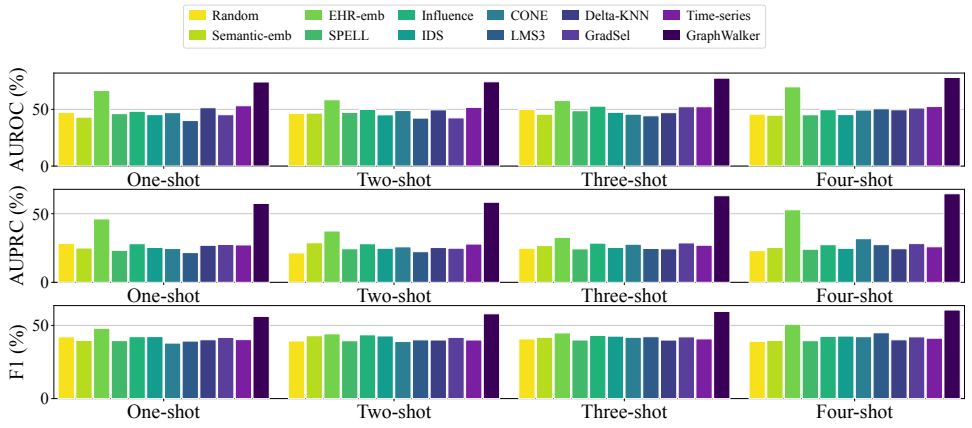


Figure 9: Patient-budget scaling performance on MIMIC-IV Readmission with *Qwen3-14B* as the backbone LLM.

whereas a larger K_a broadens the initial candidate pool. We evaluate $K_a \in \{1, 2, 3, 4\}$ on the MIMIC-IV Readmission validation split using *Qwen3-14B*, keeping all other hyper-parameters fixed and averaging results over three random seeds. As shown in Fig. 8(c), performance varies non-monotonically with K_a. Both AUPRC and F1 achieve their highest mean values at $K_a = 3$, while increasing K_a to 4 reduces both metrics. We therefore adopt $K_a = 3$ as the validation-selected default. These results suggest that expanding the initial search frontier does not necessarily yield better prediction performance, motivating a moderate anchor budget in the evaluated setting.

D.2.6 Patient-Budget Scaling

To examine the effect of patient budget on **GraphWalker**, we evaluate its performance under varying patient-budget settings ($k \in \{1, 2, 3, 4\}$) on *MIMIC-III Mortality* and *MIMIC-IV Readmission* with *Qwen3-14B* as the backbone LLM (Figures 7a and 9).

GraphWalker consistently outperforms all baselines across patient-budget settings. This demonstrates strong robustness to the choice of patient budget. In contrast, baseline methods exhibit non-monotonic and dataset-dependent trends, indicating that **simply adding more patients does not necessarily improve EHR reasoning performance**—irrelevant or misleading patients can disrupt rather than enhance coherent reasoning.

GraphWalker is the only method whose performance improves consistently with k. By grounding patient search in cohort-level clinical similarity and applying context-conditioned graph search, **GraphWalker** progressively identifies patient sets that maximize information gain rather than adding redundant or harmful cases.

Method	Initial candidates	MIMIC-III Mortality			MIMIC-IV Readmission		
		AUROC↑	AUPRC↑	F1↑	AUROC↑	AUPRC↑	F1↑
`GraphWalker` (cohort-based)	9	**84.19**	**59.46**	**58.24**	**75.91**	**60.61**	**58.01**
w/o Cohort (global top-3)	3	81.19	48.27	55.06	75.65	59.61	56.87
w/o Cohort (global top-9)	9	82.16	51.02	56.55	**75.91**	59.77	57.01

Table 10: Supplementary control for the main-text cohort ablation using Qwen3-14B and frozen SMART embeddings. Global top-9 retrieval matches the initial candidate count of `GraphWalker`. All settings retain the same subsequent search procedure and maximum demonstration budget of $K = 3$. Values are means over three random seeds. Bold indicates the best reported mean, including ties.

Model-side baselines exhibit variable performance across patient budgets. Methods such as CONE, LMS3, and SPELL show dataset-dependent performance and do not consistently benefit from larger patient budgets. These results motivate selection strategies that account for the contribution of additional cases to the current patient context.

D.2.7 Controlling for Initial Candidate Set Size

Motivation. We complement the main-text cohort ablation with a control for initial candidate set size. Under the default configuration, `GraphWalker` retrieves $K_c = 3$ cohorts and selects $K_a = 3$ anchors from each, initializing the search with nine candidates. The main-text *w/o Cohort* variant instead uses the global top-K_a nearest patients, initializing the search with three candidates. To distinguish the benefit of cohort-based initialization from that of a larger initial candidate set, we additionally evaluate a global top-9 variant.

Setup. We evaluate the three settings on MIMIC-III Mortality and MIMIC-IV Readmission using Qwen3-14B and frozen SMART embeddings. The full method selects three anchors from each of three retrieved cohorts. The two *w/o Cohort* variants select the global top-3 or top-9 nearest training patients under cosine similarity in the same embedding space. All settings share the training repository, patient graph, test split, EHR serialization, task prompts, and decoding settings. After initialization, they use identical context-conditioned scoring, greedy selection, graph-neighbor expansion, and stopping rules, with $k_g = 8$ and a maximum demonstration budget of $K = 3$. Thus, the global top-9 variant matches the full method's initial candidate count while changing how those candidates are selected. Subsequent frontier sizes and scoring workloads may differ as the search progresses. We report mean AUROC, AUPRC, and F1 over the same three random seeds.

Results. As shown in Table 10, expanding global retrieval from three to nine initial candidates improves mean performance across all reported metrics on both tasks, partially narrowing the gap observed in the main-text ablation. With the initial candidate count matched, however, `GraphWalker` retains advantages of 2.03, 8.44, and 1.69 percentage points in AUROC, AUPRC, and F1 on MIMIC-III Mortality. The additional benefit is smaller on MIMIC-IV Readmission: the two settings achieve the same reported AUROC, while `GraphWalker` improves AUPRC and F1 by 0.84 and 1.00 percentage points, respectively. These results strengthen the main-text ablation by suggesting that the benefits of cohort-based initialization extend beyond increasing the initial candidate count, particularly for mortality prediction.

D.2.8 Disentangling Context-Conditioned Re-scoring and Frontier Expansion

Motivation. The main-text joint ablation removes both context-conditioned re-scoring and graph-based frontier expansion. To disentangle their contributions, we independently enable or disable each mechanism while holding the initial candidate set and final demonstration count fixed.

Setup. We evaluate all variants on MIMIC-III Mortality and MIMIC-IV Readmission using Qwen3-14B and frozen SMART embeddings. Within each random seed, all variants share the same encoder checkpoint, training repository, patient graph, and test split. For each target patient, they start from the same nine candidates obtained by retrieving $K_c = 3$ cohorts and selecting $K_a = 3$ an-

Method	Context-conditioned re-scoring	Frontier expansion	MIMIC-III Mortality			MIMIC-IV Readmission		
			AUROC↑	AUPRC↑	F1↑	AUROC↑	AUPRC↑	F1↑
`GraphWalker` (w/o Early Stop)	Yes	Yes	**81.34**	**48.66**	**55.99**	**75.53**	**59.13**	**57.38**
w/o Re-scoring	No	Yes	77.01	46.99	55.25	70.01	52.37	52.89
w/o Expansion	Yes	No	77.13	46.91	54.90	71.08	52.86	54.03
w/o Both	No	No	68.99	46.89	51.08	63.97	43.98	49.56

Table 11: Factorial ablation of context-conditioned re-scoring and graph-based frontier expansion using Qwen3-14B and frozen SMART embeddings. All variants start from the same nine candidates and select exactly three demonstrations, with early stopping disabled. The reference row reuses the *w/o Early Stop* results from Table 2. Values are means over three random seeds; the best result in each column is bold.

chors from each. The graph neighborhood size is fixed at $k_g = 8$. EHR serialization, task prompts, decoding settings, and demonstration ordering by selection sequence are identical across variants.

All variants select exactly $K = 3$ demonstrations, with the stopping rule based on non-positive marginal utility disabled. This prevents changes in scoring from also changing the number of demonstrations. The reference configuration therefore corresponds to *w/o Early Stop* in the main-text ablation (Table 2), whose results we reuse here.

Ablation Variants. We compare four settings:

1. **`GraphWalker` (w/o Early Stop).** Both re-scoring and frontier expansion are enabled. At each iteration, we evaluate frontier candidates using their marginal information-gain proxy conditioned on the selected demonstrations, select the highest-scoring candidate, and expand the frontier through its graph neighbors.
2. **w/o Re-scoring.** Frontier expansion is retained, but each candidate v receives a fixed individual score independent of the selected demonstrations:
$$s(v) = H_\theta(x_{\text{test}} \mid \varnothing) - H_\theta(x_{\text{test}} \mid \{v\}).$$
We cache these scores and select the highest-scoring candidate in the current frontier at each iteration. Newly discovered candidates are scored using the same individual criterion.
3. **w/o Expansion.** Context-conditioned re-scoring is retained, but selection is restricted to the initial nine candidates. After each selection, we update the marginal scores of the remaining candidates without adding graph neighbors.
4. **w/o Both.** Both mechanisms are disabled. We rank the initial nine candidates by their fixed individual scores $s(v)$ and select the top three in descending score order.

Evaluation. We report mean AUROC, AUPRC, and F1 over the same three random seeds. Comparing the reference configuration with *w/o Re-scoring* measures the contribution of context-conditioned scoring when expansion is retained; comparing it with *w/o Expansion* measures the contribution of expansion when re-scoring is retained. The *w/o Both* variant provides a reference for evaluating either mechanism alone. This experiment controls the initial candidate set and final demonstration count, but not computational cost: candidate-scoring workloads may differ because re-scoring and expansion are the mechanisms being varied.

Results. Table 11 shows that enabling both mechanisms yields the highest mean performance across all six metrics. With expansion retained, removing re-scoring reduces AUROC/AUPRC by 4.33/1.67 percentage points on mortality prediction and 5.52/6.76 points on readmission prediction. With re-scoring retained, removing expansion reduces the corresponding metrics by 4.21/1.75 and 4.45/6.27 points, respectively. Neither mechanism alone consistently outperforms the other across tasks and metrics. These results support their complementary contributions under identical candidate initialization and a fixed demonstration count, further disentangling the joint ablation in Table 2.

D.3 Comparison with Supervised Predictors across Domains

Evaluation Settings. We evaluate in-hospital mortality prediction under three settings: **(i) in-domain evaluation**, from MIMIC-III to MIMIC-III; **(ii) cross-version transfer**, from MIMIC-

Method	Direct Answer (Few-shot)			+ ReAct		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1
Zero-shot	70.81	33.69	41.42	71.11	32.86	43.61
Random	70.41	31.32	39.12	68.46	29.08	39.08
Semantic-emb	69.91	25.78	36.30	70.91	26.01	36.92
EHR-emb	<u>81.59</u>	<u>48.45</u>	<u>46.07</u>	<u>82.09</u>	<u>47.98</u>	<u>48.51</u>
`GraphWalker`	**87.64**	**62.31**	**64.79**	**88.72**	**64.69**	**67.11**

Table 12: Results on MIMIC-III Mortality with GPT-5.2 as the black-box LLM, comparing two inference settings under the *same* few-shot demonstrations: **Direct Answer (Few-shot)** where the LLM directly outputs the prediction with no reasoning scaffold, and **+ ReAct** where the LLM applies ReAct-style intermediate reasoning before generating the prediction. Bold = best within column, underline = second-best. Higher is better (↑).

IV to MIMIC-III; and **(iii) cross-institution transfer**, from MIMIC-IV to TJH. MIMIC-III and MIMIC-IV originate from the same institution and have partially overlapping patient populations and collection periods (Johnson et al., 2023). We therefore treat MIMIC-IV-to-MIMIC-III evaluation as a supplementary cross-version comparison, subject to potential patient overlap between the source and target cohorts. To assess transfer to an independent institution, we additionally evaluate MIMIC-IV-to-TJH transfer and use this setting as the primary evidence for cross-institution generalization. Supervised baselines are trained on the corresponding source-training split and evaluated on the target test split. No target-domain fine-tuning is performed in either transfer setting.

Information Access and Fairness. Within each setting, all methods use the same preprocessing, aligned patient variables, task definitions, source/target splits, and evaluation metrics. They differ in how they use source-domain supervision: supervised baselines optimize task-specific parameters on the labeled source-training data, whereas `GraphWalker` uses the same source split to construct its patient graph, cohorts, and labeled clinical casebook. The retrieval base contains only source-training patients, and both the EHR encoder and target LLM remain frozen. SMART is used solely as a representation encoder.

Baselines. We compare with four supervised baselines:

- **SFT** fine-tunes the LLM backbone on source-training examples using next-token prediction loss to generate the binary mortality label.
- **RAM-EHR** (Xu et al., 2024a) combines knowledge retrieval from multiple sources with consistency regularization.
- **GraphCare** (Jiang et al., 2024) incorporates LLM-generated personalized knowledge triples into a graph neural network predictor.
- **DearLLM** (Xu et al., 2025) augments supervised prediction with LLM-derived feature relations and graph pooling.

Implementations follow the original papers and use validation-based model selection.

Results. Fig. 6b summarizes performance under the three evaluation settings. In-domain, `GraphWalker` achieves the highest AUROC (84.19) and the second-highest AUPRC (59.46), close to DearLLM's 60.02. Under MIMIC-IV-to-MIMIC-III transfer, it achieves 81.29 AUROC and 52.98 AUPRC, exceeding the strongest baseline for each metric by 11.30 and 10.13 percentage points, respectively. Under MIMIC-IV-to-TJH transfer, it achieves 89.12 AUROC and 84.76 AUPRC, with corresponding gains of 6.37 and 6.78 percentage points. These within-setting comparisons support competitive in-domain performance and stronger cross-domain transfer performance without target-domain parameter updates.

D.4 RQ4: Black-Box and Reasoning-Scaffold Compatibility

We investigate two practical questions beyond the main experiments: (i) **whether `GraphWalker` can be extended to closed-source or black-box LLMs** where token-level probabilities are not accessible, and (ii) **whether `GraphWalker` can be combined with reasoning scaffolds** such as ReAct to yield further gains under the same few-shot setting. We implement a black-box variant of `GraphWalker` (Sec D.4.1) and integrate it with a ReAct-style scaffold (Sec D.4.3).

D.4.1 BLACK-BOX `GraphWalker`

The white-box version of `GraphWalker` estimates candidate utility through the reduction in query-level conditional cross-entropy computed from token probabilities. While effective for open-source LLMs with accessible logits, this design is not directly applicable to closed-source frontier models exposed only through API-based text generation. To address this, we develop a black-box variant that replaces entropy-based scoring with LLM self-evaluation.

Pipeline. The graph construction and search structure remain unchanged. We build a patient KNN graph from the training set using pretrained EHR embeddings, then apply Leiden clustering to obtain patient cohorts. For each test patient, we retrieve the top-L most relevant cohorts and initialize anchor nodes by selecting the top-K most similar patients within each cohort, forming the initial frontier for graph-guided patient search. The key difference lies in the greedy selection step. Instead of computing ΔH from token-level probabilities, we score each frontier candidate by querying the black-box LLM with a self-evaluation prompt. Given the target patient, the already selected demonstrations, and one candidate demonstration, the model rates how much *additional useful information* the candidate provides for predicting the target patient's outcome on an integer scale from 0 to 10.

Greedy selection with self-evaluation. At each iteration, all frontier candidates are scored in parallel through API calls. We select the highest-scoring candidate, add it to the demonstration set, and expand the frontier with its graph neighbors. The process repeats until either the maximum number of demonstrations is reached or the highest candidate score becomes zero. This preserves the cohort-guided frontier expansion and greedy search structure of `GraphWalker`, while replacing white-box entropy computation with a black-box scoring proxy.

Scoring prompt. The scoring prompt (Table D.4.1) approximates the same intuition as the original information-gain objective: a useful candidate should be clinically relevant to the test patient, provide complementary evidence beyond the already selected examples, and reduce uncertainty about the target prediction. Concretely, the black-box scorer assesses each candidate from three aspects: (1) clinical similarity or complementarity to the target patient, (2) whether it contributes information not already covered by selected examples, and (3) how much it helps improve confidence in the target prediction. The model outputs only a single integer score from 0 to 10.

Black-box GraphWalker Scoring Prompt

You are evaluating whether a candidate patient example would be helpful for predicting a test patient's clinical outcome.
Task:
{TASK_DESCRIPTION}
Already Selected Examples for In-Context Learning:
{SELECTED_EXAMPLES}
Candidate Example to Evaluate:
{CANDIDATE_EXAMPLE}
Test Patient to Predict:
{TEST_PATIENT_DETAIL}
Instruction:
Rate how much *additional useful information* the candidate example provides for predicting the test patient's outcome, considering:

- Clinical similarity or complementarity to the test patient
- Whether it provides new information not already covered by selected examples
- How much it would help reduce uncertainty about the test patient's outcome

Score Definition (0–10):

- 0: completely redundant or irrelevant
- 5: moderately helpful, provides some new perspective
- 10: extremely informative, highly relevant and complementary

Output Format:
Output **ONLY** a single integer score from 0 to 10, and nothing else.

Prompt	AUROC ↑	AUPRC ↑	F1 ↑	Rank ρ ↑	Jaccard ↑
Original	**87.64**	**62.31**	**64.79**	–	–
Paraphrase 1	86.38	59.72	62.41	0.68	0.69
Paraphrase 2	87.02	61.08	63.57	0.75	0.76

Table 13: Sensitivity of black-box utility estimation to scoring-prompt wording. Rank ρ denotes Spearman correlation with the candidate ranking produced by the original prompt, and Jaccard measures overlap between the final selected patient sets. Values are averaged over three random seeds.

Discussion. This black-box implementation does not claim to exactly reproduce the white-box entropy objective. Instead, it provides a practical approximation that preserves the core design principle of `GraphWalker`: selecting demonstrations based on *target-conditioned marginal utility* rather than static similarity alone. As a result, `GraphWalker` extends to stronger closed-source LLMs without requiring access to internal logits, broadening its applicability in realistic deployment settings.

Experiments. We evaluate the black-box variant under a standard few-shot setting on *MIMIC-III Mortality*, where the downstream LLM directly outputs the final prediction without any additional reasoning scaffold.

Results. As shown in Table 12, we summarize three observations:

(i) Generic semantic similarity is actively harmful under black-box settings. *Semantic-emb* not only substantially underperforms the zero-shot baseline on AUPRC and F1 (e.g., AUPRC 25.78 vs. 33.69), but is also worse than *Random* retrieval (25.78 vs. 31.32 on AUPRC). This indicates that surface-level semantic similarity, when divorced from clinical structure, can mislead demonstration selection rather than help it.

(ii) Clinically informed similarity helps but remains single-perspective. *EHR-emb* provides a stronger retrieval signal by capturing clinically meaningful patient similarity, clearly outperforming both zero-shot and random retrieval. Yet as a purely similarity-based method, it still leaves substantial headroom relative to `GraphWalker`.

(iii) Black-box `GraphWalker` substantially outperforms all single-perspective baselines. Compared with the strongest baseline *EHR-emb*, black-box `GraphWalker` improves AUPRC from 48.45 to 62.31. This confirms that its dual-view advantage, which bridges cohort-level clinical similarity and model-side utility estimation, generalizes beyond the original entropy-based implementation: the core strength lies in cohort-guided, target-conditioned demonstration construction rather than any specific scoring interface.

D.4.2 Stability of Black-Box Utility Estimation

We evaluate the stability of the black-box utility estimator from two aspects: sensitivity to scoring-prompt wording and consistency with the white-box utility signal.

Prompt Sensitivity. We construct two semantic paraphrases of the original self-evaluation prompt and evaluate all three variants under the same target patients, candidate pools, decoding configuration, and three random seeds. The variants differ only in prompt wording. We report downstream performance together with the agreement of candidate rankings and final selected patient sets relative to the original prompt.

As shown in Table 13, **paraphrasing the scoring prompt leads to only modest performance changes.** Candidate rankings remain positively aligned with the original prompt ($\rho = 0.68$–0.75), while the final selected patient sets retain substantial overlap (Jaccard $= 0.69$–0.76). These results suggest that black-box utility estimation is reasonably stable to prompt wording.

Consistency with White-Box Scoring. We further compare the white-box ΔH criterion and black-box self-evaluation using the same Qwen3-14B model on MIMIC-III Mortality. Both inter-

Comparison	Rank $\rho \uparrow$	Top-1 Agreement $\uparrow$	Set Jaccard $\uparrow$
White-box vs. Black-box	0.64	66.8%	0.65

Table 14: Candidate-selection consistency between white-box ΔH and black-box self-evaluation on MIMIC-III Mortality.

faces use identical target patients, candidate frontiers, working contexts, patient budgets, and input formatting; only the utility-scoring interface differs.

The two scoring interfaces exhibit meaningful agreement in candidate ranking and selected patient sets (Table 14). The black-box variant also retains performance close to the white-box implementation, with differences of 1.23 AUROC, 2.55 AUPRC, and 1.77 F1 points. Together, these results indicate that self-evaluation provides a practical approximation to the white-box utility signal when token-level likelihoods are unavailable.

D.4.3 Compatibility with ReAct-Style Reasoning

Recent LLM agents show strong reasoning and decision-making across many domains (Ferrag et al., 2025). Rather than positioning `GraphWalker` as an alternative to ReAct-style reasoning, we view it as a **complementary upstream plugin**: even powerful reasoning-scaffold inference depends on the quality of the provided case context, and `GraphWalker`'s cohort-guided, target-conditioned demonstration selection can supply more informative examples to better unlock downstream ReAct-style reasoning.

Experiments. We choose *ReAct* (Yao et al., 2022) as a representative ReAct-style reasoning framework: it tightly couples intermediate reasoning with final decision making and is a widely used backbone for dynamic agent-style inference. We keep the few-shot demonstrations selected by each method unchanged and only replace the downstream inference style from direct answer generation to ReAct-style reasoning. This design isolates the effect of combining `GraphWalker` with ReAct-style reasoning, directly testing whether better case selection can further enhance a more dynamic reasoning process under the same few-shot setting. Experiments are conducted on *MIMIC-III Mortality* using the same black-box LLM and demonstration selection baselines as in the previous subsection.

Results. As shown in Table 12, we summarize three observations:

(i) ***`GraphWalker` + ReAct further improves over standalone `GraphWalker` across all metrics.*** Compared with the standard black-box setting, `GraphWalker` + ReAct improves AUROC from 87.64 to 88.72, AUPRC from 62.31 to 64.69, and F1 from 64.79 to 67.11.

(ii) ReAct does not consistently help, and sometimes hurts, weaker baselines. For *Random*, ReAct *actively degrades* performance across all three metrics (e.g., AUROC 70.41 $\rightarrow$ 68.46, AUPRC 31.32 $\rightarrow$ 29.08). For *Semantic-emb*, changes are marginal. For the stronger *EHR-emb*, gains are mixed: AUROC and F1 improve slightly while AUPRC drops (48.45 $\rightarrow$ 47.98).

(iii) Demonstration quality is a prerequisite for ReAct-style reasoning to be effective. ReAct-style reasoning compounds positively with high-quality demonstrations (consistent improvements for `GraphWalker`) but compounds neutrally or even negatively with low-quality ones. By constructing more informative and clinically coherent demonstrations, `GraphWalker` provides a stronger evidence base on which ReAct-style reasoning can build, positioning it as a complementary upstream module for downstream reasoning.

E Applicability to Additional Medical Reasoning Tasks

To further examine the generality of `GraphWalker` beyond EHR-centric prediction tasks, we extend our evaluation to three additional medical reasoning benchmarks: **MedQA** (Jin et al., 2021), **CMB** (Wang et al., 2024), and **CMB-Clin** (Wang et al., 2024).

MedQA consists of medical examination questions designed to assess diagnostic reasoning and clinical decision-making, with each question grounded in a realistic patient case description. Although the full benchmark is multilingual, we focus on the English subset, which contains approximately 12.7K case-based questions and has been widely adopted for evaluating LLM-based medical reasoning.

CMB is a large-scale Chinese medical benchmark covering a broad range of clinical specialties and question types. Following (Ding et al., 2025b), we filter the dataset to retain only case-based questions involving explicit patient scenarios, ensuring its suitability for in-context reasoning and in-context evaluation.

CMB-Clin (Wang et al., 2024) is a medical question-answering benchmark containing 208 questions associated with 74 clinical cases. We partition the dataset at the clinical case level, assigning all questions associated with the same case to the same training, validation, or test split. The resulting splits contain 83, 42, and 83 questions, respectively, with no clinical case shared across splits. The CMB-Clin counts in Table 15 therefore refer to questions rather than distinct cases.

Dataset statistics are summarized in Table 15.

Split	MedQA	CMB	CMB-Clin
Train	10,178	15,465	83
Val	1,272	1,940	42
Test	1,273	1,935	83
Total	12,723	19,340	208

Table 15: Dataset statistics for additional medical reasoning benchmarks: MedQA, CMB, and CMB-Clin.

E.0.1 Experimental Setup

In this experimental setting, both `GraphWalker` and the Semantic-embedding baseline use *Qwen3-Embedding-8B* as the retrieval encoder. For fair comparison, all methods share the same backbone LLM, namely *Qwen3-14B*.

E.0.2 Results Analysis

Overall, the results in Table 16 demonstrate that `GraphWalker` generalizes well beyond EHR-centric prediction settings and remains consistently effective across diverse medical reasoning tasks. On the two multiple-choice benchmarks, `GraphWalker` achieves the best EM on both CMB (84.05%) and MedQA (61.10%), outperforming all compared baselines, including strong retrieval-based and influence-based selection methods. This indicates that **`GraphWalker` is not limited to structured EHR classification, but can also transfer effectively to broader medical reasoning scenarios requiring diagnostic understanding and clinical decision-making.**

More importantly, `GraphWalker` also shows strong generalization to the open-ended generation benchmark CMB-clin, where it achieves the best performance on all three generation metrics, including BLEU-1, BLEU-4, and ROUGE-L. The gains on CMB-clin suggest that **the benefits of `GraphWalker` are not confined to answer selection or label prediction, but also extend to multi-round, free-form medical response generation**, where models must produce more comprehensive and coherent answers grounded in complex clinical contexts.

Taken together, these results verify that **`GraphWalker` is a generally applicable patient-analogy framework for medical reasoning, rather than a task-specific solution tailored only to EHR analysis**. Its consistent improvements across both discriminative and generative settings highlight its robustness and suggest that the underlying cohort-guided retrieval and example composition mechanism captures transferable reasoning patterns that remain beneficial across different task formats and evaluation protocols.

Method	CMB	MedQA	CMB-clin		
	EM	EM	BLEU-4	BLEU-1	ROUGE-L
Zero-shot	80.21	56.56	21.06	59.85	44.97
Random	81.50	52.96	26.85	66.05	38.63
Semantic-emb	80.52	<u>59.58</u>	26.86	64.90	41.26
Time-series	78.05	52.11	<u>29.05</u>	66.56	40.12
SPELL	81.98	54.10	28.41	66.75	36.82
Influence	<u>83.50</u>	53.02	23.88	67.81	<u>47.05</u>
IDS	79.57	52.59	22.63	43.79	25.58
CONE	80.41	59.17	29.04	<u>68.10</u>	41.10
LMS3	82.02	54.09	26.78	62.58	38.02
Delta-KNN	83.02	53.47	23.09	63.73	46.50
GradSel	81.48	52.93	24.83	64.98	45.13
Ours	**84.05**	**61.10**	**29.68**	**71.51**	**47.48**

Table 16: Performance comparison on CMB, MedQA, and CMB-clin using *Qwen3-14B* as the backbone LLM. CMB and MedQA are evaluated by exact match (EM), while CMB-clin is evaluated by BLEU-4, BLEU-1, and ROUGE-L. Higher values indicate better performance (↑). We highlight the **best** and <u>second best</u> results.

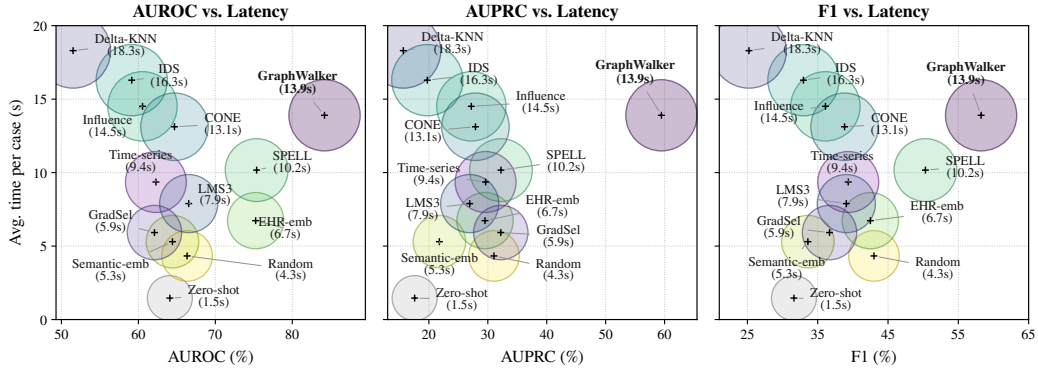


Figure 10: Performance–latency tradeoff on MIMIC-III mortality across AUROC, AUPRC, and F1 (left to right). Each bubble represents one case retrieval method, with position showing mean performance (x-axis) and average inference time per case (y-axis); bubble color identifies the method. **`GraphWalker`** achieves the strongest performance across all three metrics while keeping latency comparable to prior LLM-based case selection methods.

E.1 Computational Cost and Efficiency

The central design objective of **`GraphWalker`** is to improve clinical reasoning accuracy by deliberately allocating a small additional inference budget at test time. In clinical decision-support scenarios, predictive accuracy often carries far greater practical value than marginal reductions in latency, since higher-quality predictions can directly affect the quality of downstream clinical decisions. Motivated by this, **`GraphWalker`** invests extra inference cost to construct working contexts that are clinically coherent, complementary, and informative for each target patient—following the same high-level principle as test-time scaling in reasoning models, where additional computation is intentionally exchanged for better reasoning outcomes.

Crucially, this extra cost is tightly bounded rather than spent on unconstrained global search. Cohort retrieval, anchor initialization, frontier expansion, context-conditioned greedy selection, and early stopping jointly restrict the inference budget to a small set of clinically relevant candidates, making **`GraphWalker`** a targeted form of test-time patient-context search.

Experiments. We conduct runtime analysis on *MIMIC-III Mortality* using *Qwen3-14B* as the backbone LLM. For each patient-selection method, we measure the average wall-clock evaluation

Cost Item	Value
Offline graph-construction time	13.6 min
Offline peak memory	10.8 GB
Average final context length	8.4K tokens
Average candidate evaluations per patient	22.3
Average cumulative input tokens per patient	154K
Peak online GPU memory	46.2 GB
End-to-end latency per patient	13.9 s

Table 17: Computational-resource breakdown of `GraphWalker` under the white-box setting on MIMIC-III Mortality with Qwen3-14B.

Cost Item	Value
Average candidate-scoring calls per patient	39.6
Sequential selection rounds	3
Average concurrent calls per round	13.2
End-to-end latency per patient	19.0 s
Average API cost per patient	$0.50

Table 18: Per-patient inference profile of the black-box variant.

time per test case—including both patient selection and final prediction generation—and compare the resulting performance–latency tradeoff across methods.

Results. As shown in Fig. 10, `GraphWalker` achieves the strongest performance across all three metrics while keeping inference cost on par with existing LLM-based selection methods. The search overhead is effectively bounded by the cohort-guided frontier design rather than expanded into exhaustive global exploration. Two observations make this concrete:

(i) The absolute overhead is modest. Compared with the strongest similarity-based baseline *EHR-emb*, `GraphWalker` incurs only 7.2s additional inference time per case (13.9s vs. 6.7s), yet yields substantial gains of $+9.0$ AUROC, $+29.9$ AUPRC, and $+15.8$ F1.

(ii) The overhead is competitive among LLM-based selection methods. Among approaches that similarly invoke LLM-side scoring, `GraphWalker` (13.9s) is on par with *CONE* (13.1s), and notably faster than *Influence* (14.5s), *IDS* (16.3s), and *Delta-KNN* (18.3s), while delivering substantially larger accuracy gains across all three metrics.

Together, these results show that **`GraphWalker` converts a limited test-time cost into significant improvements in clinical prediction quality.**

Detailed Resource Profile. *White-box setting.* We evaluate computational cost on MIMIC-III Mortality with Qwen3-14B. Table 17 reports graph-construction cost, context length, candidate-scoring workload, cumulative token consumption, memory usage, and end-to-end latency.

The 154K input tokens denote the cumulative token volume processed across candidate-scoring passes and final prediction, rather than the length of a single prompt; the final prediction context averages only 8.4K tokens. Moreover, `GraphWalker` searches over a dynamically maintained local frontier instead of repeatedly evaluating the entire patient repository. Under this setting, the complete patient-specific search and final prediction require 13.9 seconds on average.

Black-box setting. We further report the inference profile when candidate utility is estimated through black-box LLM self-evaluation. Candidate evaluations within the same selection round are independent given the current working context and can therefore be issued in parallel, while successive rounds remain sequential.

Although the black-box variant requires more candidate-scoring calls, these calls are parallelized within each selection round. Thus, the raw API-call count primarily affects total request volume and monetary cost rather than translating directly into sequential latency. Overall, `GraphWalker` introduces a bounded test-time search cost in exchange for avoiding task-specific parameter updates.

F PROMPT DETAILS

We provide the task-specific instructions and shared prompt templates used for patient-outcome prediction. For each patient, the final visit is designated as the *target visit*, whose outcome is to be predicted. The input contains EHR records from preceding visits and excludes records from the target visit.

We use two prompting schemes. The *zero-shot* template provides the task instruction and the patient's observed EHR history. The *few-shot* template additionally includes selected patient examples and their known task-specific outcomes. Each example follows the same construction: its input contains records preceding its own target visit, and its label specifies the outcome associated with that visit. Both templates use the same task-specific instructions and EHR representation.

Prompt for Mortality Prediction

You are tasked with predicting whether the patient will die during the target visit, using EHR records from preceding visits. The target visit is the patient's final visit, and its records are excluded from the input.
Provide only a floating-point number between 0 and 1 representing the predicted probability of in-hospital mortality during the target visit. A higher value indicates a higher probability of death.
Do **not** provide any reasoning, explanation, or additional text. Output **only** the numerical value.
Example: `0.XX`

Prompt for ICU Readmission Prediction

You are tasked with predicting whether the patient will be readmitted to the ICU within 30 days after discharge from the target ICU stay, using EHR records from preceding visits. The target visit is the patient's final visit, and its records are excluded from the input.
Provide only a floating-point number between 0 and 1 representing the predicted probability of a new ICU admission within this 30-day window.
Do **not** provide any reasoning, explanation, or additional text. Output **only** the numerical value.
Example: `0.XX`

Prompt for Length-of-Stay (LOS) Prediction

You are tasked with predicting the patient's total hospital length of stay during the target visit, using EHR records from preceding visits. The target visit is the patient's final visit, and its records are excluded from the input.
Provide only a single letter (**A**, **B**, **C**, or **D**) representing the predicted length-of-stay category:

- **A**: Less than 3 days (< 3 days)
- **B**: 3 to 7 days (3–7 days)
- **C**: 8 to 14 days (8–14 days)
- **D**: More than 14 days (> 14 days)

Do **not** provide any reasoning, explanation, or additional text. Output **only** the letter (**A**, **B**, **C**, or **D**).
Example: `B`

Zero-shot Prompt Template

You will be provided with longitudinal electronic health record (EHR) data from a patient's visits preceding the final visit. We refer to the final visit as the **target visit**. Records from the target visit and its outcome are not provided.
Each clinical feature is represented as a time-ordered sequence of measurements from the preceding visits. Missing values are denoted as `NaN`. Units and reference ranges are provided where applicable. Use the provided historical records to predict the specified outcome associated with the target visit.
Task Description:
{`TASK_DESCRIPTION`}
Instructions & Output Format:
{`RESPONSE_FORMAT`}
Target Patient's Historical Records:

- Number of measurements: {`LENGTH`}
- Measurement times in the historical records: [{`RECORD_TIME_LIST`}]

Clinical Features Over Time:

- **Heart Rate** (Unit: bpm. Reference range: 60–100):
  `[82, 86, 90, 95, NaN, 88, 84, ...]`
- **Systolic Blood Pressure** (Unit: mmHg. Reference range: $<$ 120):
  `[110, 118, 125, 132, 120, NaN, 115, ...]`
- **Oxygen Saturation** (Unit: %. Reference range: 95–100):
  `[100, 99, 98, NaN, 96, 95, 97, ...]`
- **Glasgow Coma Scale (Total)** (Unit: /. Reference range: /.):
  `[15, 14, NaN, 12, 10, 8, ...]`
- ⋮

Your Answer:
{`MODEL_OUTPUT`}

Few-shot Prompt Template

You will be provided with longitudinal electronic health record (EHR) data from a patient's visits preceding the final visit. We refer to the final visit as the **target visit**. Records from the target visit and its outcome are not provided.
Each clinical feature is represented as a time-ordered sequence of measurements from the preceding visits. Missing values are denoted as `NaN`. Units and reference ranges are provided where applicable.
You will also receive selected examples from other patients. Each example contains that patient's historical records preceding their own target visit, together with the known outcome associated with that visit. Use these examples and the target patient's historical records to predict the specified outcome for the target patient.
Task Description:
{`TASK_DESCRIPTION`}
Instructions & Output Format:
{`RESPONSE_FORMAT`}
Selected Patient Examples:

- **Example 1:**
  Historical EHR Records: `[...]`
  Outcome Label: {`LABEL_1`}
- **Example 2:**
  Historical EHR Records: `[...]`
  Outcome Label: {`LABEL_2`}
- ⋮

Target Patient's Historical Records:

- Number of measurements: {`LENGTH`}
- Measurement times in the historical records: `[`{`RECORD_TIME_LIST`}`]`

Clinical Features Over Time:

- **Heart Rate** (Unit: bpm. Reference range: 60–100):
  `[82, 86, 90, 95, NaN, 88, 84, ...]`
- **Systolic Blood Pressure** (Unit: mmHg. Reference range: < 120):
  `[110, 118, 125, 132, 120, NaN, 115, ...]`
- **Oxygen Saturation** (Unit: %. Reference range: 95–100):
  `[100, 99, 98, NaN, 96, 95, 97, ...]`
- **Glasgow Coma Scale (Total)** (Unit: /. Reference range: /.):
  `[15, 14, NaN, 12, 10, 8, ...]`
- ⋮

Your Answer:
{`MODEL_OUTPUT`}

G Reproducibility and Transparency

G.1 Computational Resources and Software Environment

All experiments were conducted on a server equipped with two NVIDIA H20 GPUs (96 GB memory each) and 503 GB system RAM, running Ubuntu 22.04.5 LTS. The software environment was based on Python 3.11.11 with Conda 23.5.2, and experiments were implemented using PyTorch 2.6.0, HuggingFace Transformers 4.51.3, and SpaCy 3.8.4, all with default settings unless otherwise specified.