

EHR-R1: A Reasoning-Enhanced Foundational Language Model for Electronic Health Record Analysis

Yusheng Liao^{*,1}, Chaoyi Wu^{*,1}, Junwei Liu^{*,3,4}, Shuyang Jiang^{2,5}, Pengcheng Qiu^{1,2}, Haowen Wang³, Yun Yue³, Shuai Zhen³, Jian Wang³, Qianrui Fan³, Jinjie Gu³, Ya Zhang^{1,2}, Yanfeng Wang^{1,2}, Yu Wang^{1,2,†} and Weidi Xie^{1,2,†}

¹Shanghai Jiao Tong University, Shanghai, China

²Shanghai Artificial Intelligence Laboratory, Shanghai, China

³Intelligence Healthcare Department, AntGroup, Hangzhou, China

⁴Intelligence Computing and Sensing Laboratory, Peking University, Beijing, China

⁵Fudan University, Shanghai, China

^{*}Equal contributions [†]Corresponding author

Yu Wang: yuwang@sjtu.edu.cn; Weidi Xie: weidi@sjtu.edu.cn

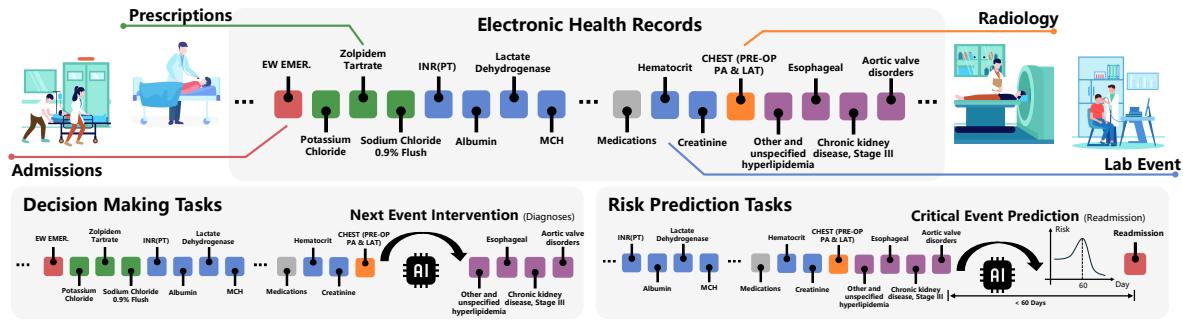
Electronic Health Records (EHRs) contain rich yet complex information, and their automated analysis is critical for clinical decision-making. Despite recent advances of large language models (LLMs) in clinical workflows, their ability to analyze EHRs remains limited due to narrow task coverage and lack of EHR-oriented reasoning capabilities. This paper aims to bridge the gap, specifically, we present **EHR-Ins**, a large-scale, comprehensive EHR reasoning instruction dataset, comprising 300k high-quality reasoning cases and 4M non-reasoning cases across 42 distinct EHR tasks. Its core innovation is a **thinking-graph-driven framework** that enables to generate high-quality reasoning data at scale. Based on it, we develop **EHR-R1**, a series of reasoning-enhanced LLMs with up to 72B parameters tailored for EHR analysis. Through a multi-stage training paradigm, including **domain adaptation**, **reasoning enhancement**, and **reinforcement learning**, EHR-R1 systematically acquires domain knowledge and diverse reasoning capabilities, enabling accurate and robust EHR analysis. Lastly, we introduce **EHR-Bench**, a new benchmark curated from MIMIC-IV, spanning 42 tasks, to comprehensively assess reasoning and prediction across EHR scenarios. In experiments, we show that the resulting EHR-R1 consistently outperforms state-of-the-art commercial and open-source LLMs (including DeepSeek-V3 and GPT-4o), surpassing GPT-4o by over 30 points on EHR-Bench and achieving a 10% higher zero-shot AUROC on EHRSHOT. Collectively, EHR-Ins, EHR-R1, and EHR-Bench have significantly advanced the development for more reliable and clinically relevant EHR analysis.

1 INTRODUCTION

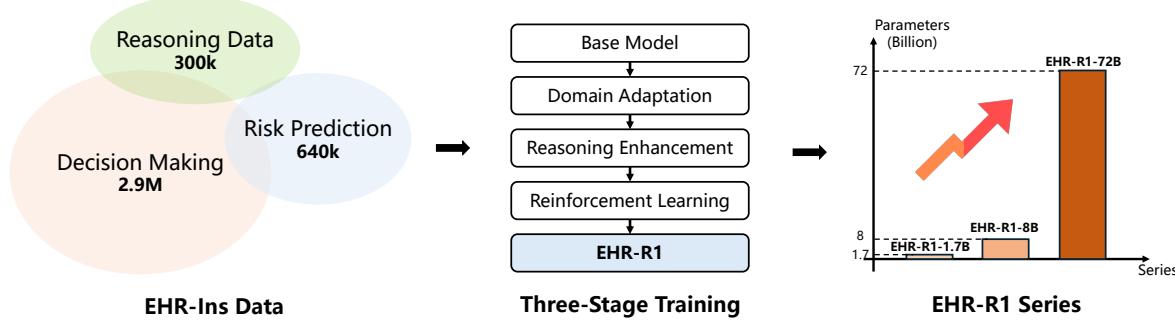
Electronic Health Records (EHRs) are comprehensive digital repositories of patient information, encompassing laboratory tests, medications, diagnoses, procedures, clinical notes, *etc.* [1, 2, 3, 4, 5]. Systematic analysis of EHRs is essential for modern healthcare, as accurate interpretation of patient histories enables early disease detection, personalized treatment planning, and improved clinical outcomes. In everyday clinical practice, physicians rely on electronic health records (EHRs) to address a wide spectrum of analytical needs, spanning decision-support queries, for example, “*what is the most probable next diagnosis for this patient?*”, to prognostic assessments, such as “*what is the risk of hospital readmission within 60 days?*” Advanced EHR analysis systems can substantially accelerate clinical workflows and augment medical decision-making, ultimately shaping the quality and timeliness of patient care.

Large language models (LLMs) have recently transformed biomedical natural language processing (BioNLP), achieving impressive results across a range of medical tasks [6, 7, 8, 9, 10, 11, 12, 13], yet their performance on EHR-related tasks remains limited [14, 15, 16, 17]. Even state-of-the-art commercial models struggle to extract, integrate, and reason over EHR data, which markedly limits their interoperability with daily used hospital information systems [18] and, consequently, their wide adoption in clinical practice.

a. EHR Analysis Tasks



b. Methods Overview



c. Results

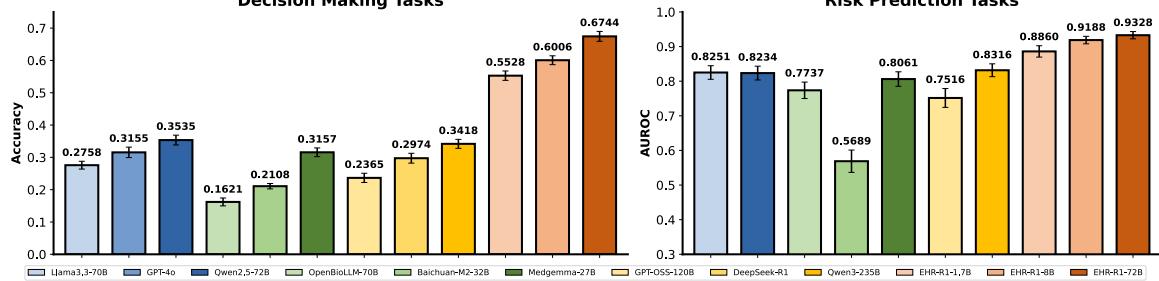


Figure 1 | Overview of the EHR tasks and the proposed method. **a. EHR Analysis Tasks.** EHR analysis tasks are defined as consisting of two types of tasks: decision-making and risk-prediction. **b. Methods Overview.** Our approach addresses these challenges with a three-stage training pipeline. First, a large volume of non-reasoning data is used for continual pre-training. This is followed by an instruction-tuning phase that leverages reasoning data. Finally, reinforcement learning with Group Reward Policy Optimization (GRPO) is applied to further refine the model. **c. Results.** This figure compares the performance of our model against several baseline LLMs on both decision-making and risk-prediction tasks, showcasing its superior performance.

Existing work on EHR analysis with LLMs is fundamentally constrained by two major challenges. First, in terms of **task coverage**, prior research has focused on narrow, task-specific objectives (*e.g.*, risk prediction for a specific condition or outcome), typically restricted to particular diseases or event types, and has yet to deliver the holistic capabilities required to support evolving clinical workflows [19, 20, 21]. Second, in terms of **reasoning ability**, existing models struggle to construct reliable, EHR-oriented reasoning chains that demand both selective information filtering and multi-source integration. As a result, they are highly susceptible to the redundancy inherent in EHR data [22, 23] and struggle to integrate isolated findings into a coherent longitudinal understanding of disease progression [22, 24].

In this paper, to address these, we propose a holistic framework for EHR analysis with two primary contributions: (i) a large-scale super-instruction dataset that unifies diverse EHR analysis tasks into a generative format and incorporates reasoning supervision, and (ii) a reasoning-enhanced EHR analysis LLM trained with synthetic reasoning data via a three-stage training paradigm. The overview of the EHR tasks

and our proposed method are shown in Figure 1.

We present **EHR-Ins**, a large-scale dataset capturing broad task diversity and explicit medical reasoning for EHR analysis. The corpus comprises **300K** high-quality reasoning cases and **3.5 million** non-reasoning cases spanning 42 EHR tasks. The tasks generally fall into two categories: **decision-making** (*e.g.*, diagnosis and treatment recommendations) and **risk-prediction** (*e.g.*, mortality and readmission). To ensure clinical fidelity and relevance, we develop a ‘thinking-graph’ driven reasoning data synthesis pipeline that (i) applies statistical analysis of entity co-occurrence ratios to identify key related entities, (ii) links them via knowledge from UMLS [25], and (iii) prompts GPT-4o to produce structured, step-by-step clinical reasoning. The resulting dataset is extensive and clinically grounded, enabling models to acquire diverse, context-rich reasoning capabilities.

Second, we introduce **EHR-R1**, a family of reasoning-enhanced LLMs (up to **72B** parameters) tailored for EHR analysis and trained on EHR-Ins. The training curriculum is designed to enhance domain knowledge and diverse reasoning patterns through three stages: (i) large-scale domain adaptation on extensive non-reasoning data, (ii) reasoning enhancement on high-quality reasoning cases, and (iii) reinforcement learning with Group Relative Policy Optimization (GRPO) on a smaller, curated set. This multi-stage regime improves the models’ ability to handle complex EHR tasks and to produce accurate, clinically meaningful outputs across diverse scenarios.

We evaluate **EHR-R1** on EHRSHOT [19] and MIMIC-IV-CDM [26], as well as on a new benchmark, **EHR-Bench**, derived from MIMIC-IV. EHR-Bench spans 42 tasks across decision-making and risk-prediction settings, providing a balanced, comprehensive assessment of both reasoning and task-specific performance. Together, these benchmarks cover two distinct clinical centers—Stanford Medicine and Beth Israel Deaconess Medical Center—and encompass a broad range of EHR analysis tasks.

Experimental results demonstrate that our model consistently outperforms leading commercial and open-sourced LLMs. In particular, EHR-R1-72B achieves an average performance improvement of over **30 points** compared to GPT-4o across all 42 tasks on EHR-Bench, highlighting the effectiveness of our method in enhancing LLMs to tackle the full spectrum of EHR-related challenges. On the out-of-distribution EHRSHOT dataset, which features significantly different structures and medical concepts, EHR-R1-72B achieves a zero-shot AUROC score that is 10% higher than baseline models. These results highlight not only the superior task adaptability of EHR-R1, but also its robustness and generalizability to unseen datasets, underscoring its potential as a transformative tool for clinical decision support.

2 Results

In the following, we first give an overview on the data quality of our constructed training data, EHR-Ins, then describe the benchmarks and evaluation metrics. Afterwards, we report the performance of our final reasoning-enhanced model, EHR-R1, highlighting its consistent gains over leading LLMs.

2.1 Training Data Overview

To enhance the LLM’s understanding of EHR data and its ability to identify relationships between medical entities within redundant EHR records, we propose a novel ‘thinking-graph’ pipeline to curate a new large-scale reasoning-enhanced EHR-analysis instruction data, **EHR-Ins**. This pipeline is designed to augment GPT-4o [27], enabling it to generate high-quality EHR reasoning chains. The details of this methodology can be found in Section 4.2.2. The synthesized reasoning data distribution for each analysis task in EHR-Ins is illustrated in Figure 2a, widely covering diverse decision-making tasks.

To demonstrate the effectiveness of our thinking-graph pipeline in synthesizing reasoning chains for EHR analysis, we hired medical experts to manually evaluate 100 synthesized reasoning chains across eight decision-making tasks to check the data quality. As shown in Figure 2b, the experts are tasked to rate each generated reasoning samples on a 5-point scale. The scores from highest to lowest represent how completely the reasoning process supports the predicted label set. A score of 5 indicates that the reasoning process fully supports the predicted result, while a score of 1 indicates the reasoning process is completely irrelevant to the predicted result. Alongside our *thinking-graph* pipeline, we adopt a naïve data distillation strategy with GPT-4o as

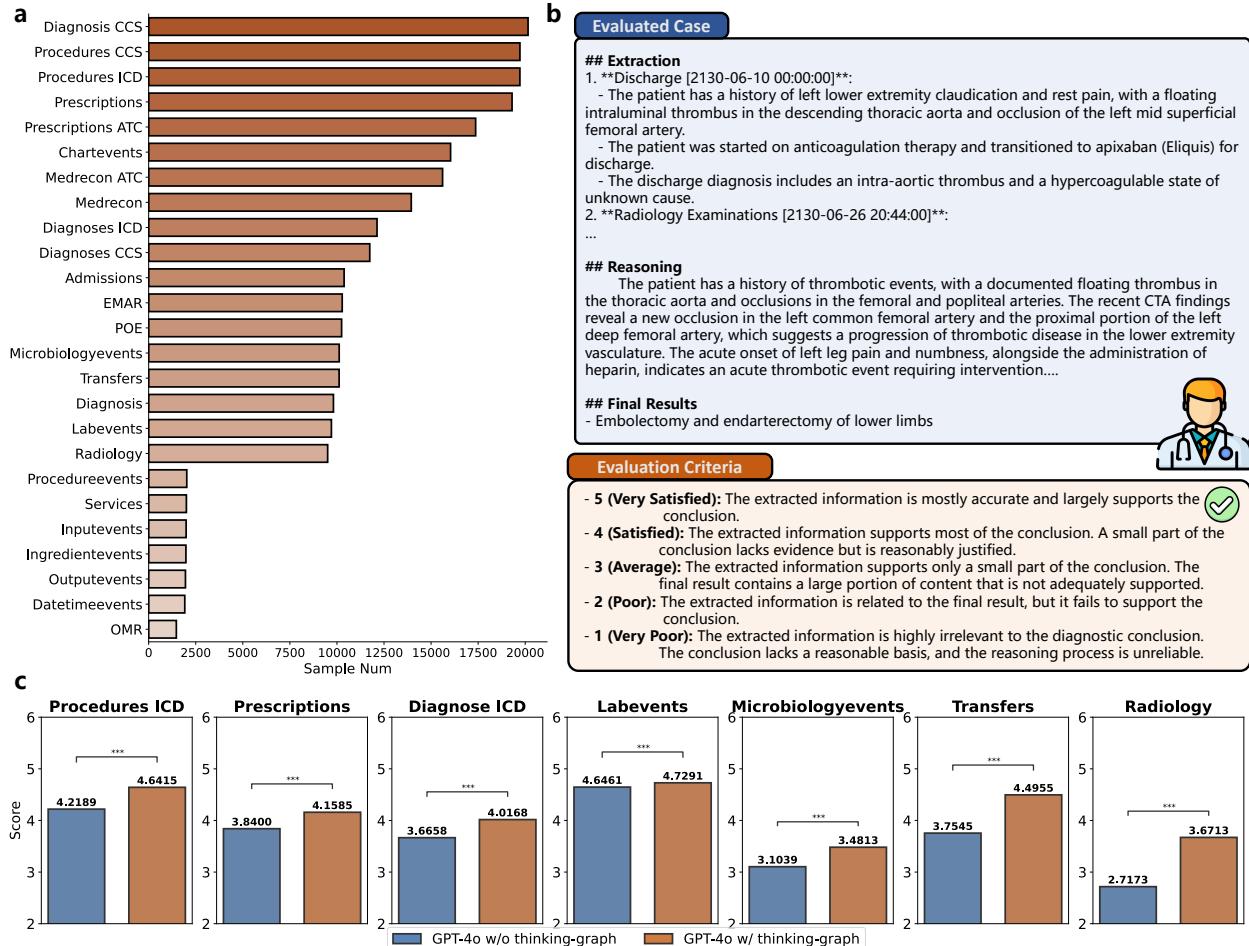


Figure 2 | Overview of reasoning data in EHR-Ins. **a** The sample size of each task in the reasoning data of EHR-Ins. **b** Example of human evaluation on the reasoning data. **c** Manual evaluation results on EHR reasoning data across eight decision-making tasks, each associated with a distinct type of decision-making event. We compared the quality of synthetic reasoning data with and without thinking-graph enhancement, where ‘***’ represents a significance level of $p < 0.001$.

a baseline, *i.e.*, directly prompting it (using Prompt 1) to derive reasoning chains from the original EHR analysis cases.

The corresponding results are shown in Figure 2c. First, **in relative comparison**, the *thinking-graph* pipeline produced reasoning chains with more adequate EHR evidence across all eight tasks than the naïve data distillation approach. These findings demonstrate that our extracted *thinking-graph* yields more reliable references for GPT-4o, effectively mitigating its knowledge limitations and enhancing curated data quality. Second, **in terms of absolute scores**, the medical experts expressed satisfaction with most of the synthesized reasoning data, assigning it an average rating above 4 on a five-point scale. This evaluation underscores the ability of the thinking-graph pipeline to effectively extract and utilize sufficient medical evidence from EHR data to support target entities, thereby substantially enhancing the capabilities of the LLM.

2.2 Evaluation Setting

2.2.1 Benchmarks

We evaluate on three benchmarks to comprehensively assess capabilities. EHR-Bench is our primary in-distribution benchmark, constructed from the same MIMIC-IV source as training and providing a holistic assessment across diverse EHR tasks. MIMIC-IV-CDM is also based on MIMIC-IV but uses different task

Category	EHR-Bench	EHR-Ins
Reassignment	1,000	250,000
Service	1,000	170,000
Procedure	1,000	316,299
Test & Exam	2,000	320,000
Diagnose	2,000	700,000
Treatment	2,500	773,644
ICU Event	2,500	200,000
Mortality	3,500	160,000
Readmission	2,000	160,000
LengthOfStay	2,000	160,000
Transfer	1,000	80,000
Outcome	500	40,000
Total	21,000	3,429,943

Item	Number
Task Types	2
Categories	12
Tasks	42

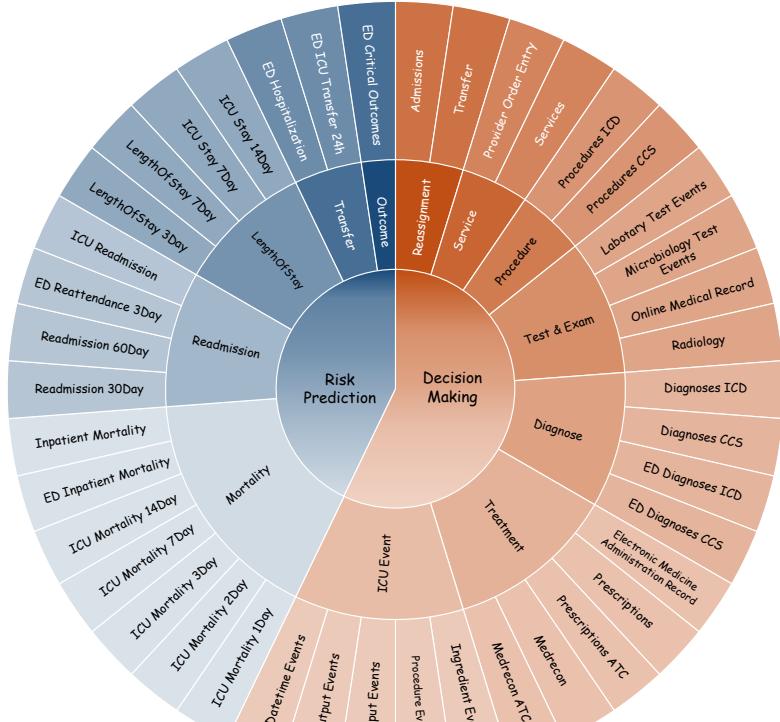


Figure 3 | Overview of EHR-Ins and EHR-Bench. The hierarchical ring chart displays the distribution of both datasets. The inner ring partitions tasks into two types: risk prediction and decision making. The middle ring shows 12 task categories (subtypes). The outer ring details all 42 specific tasks.

formulations and preprocessing—including filtered, denoised patient histories—and emphasizes diagnostic tasks, testing generalization to task formulation. EHRSHOT is derived from a different healthcare system with distinct event types and medical entities, serving as a benchmark to evaluate generalization on a entirely new patient ditribution. More detailed desription and case demonstration for each dataset can be found in Section §4.4.1.

EHR-Bench. To comprehensively evaluate LLM performance on EHR analysis, we introduce EHR-Bench, a benchmark derived from MIMIC-IV [28]. As summarized in Figure 3, EHR-Bench spans **12 subtypes** and **42 tasks**, organized into two major groups—decision making and risk prediction—covering a broad spectrum of clinically relevant settings [29, 30, 31].

The decision-making tasks are generative, requiring the model to recommend the next appropriate intervention given a specific medical event. We organize these into seven subtypes—*reassignment*, *service*, *procedure*, *test & exam*, *diagnosis*, *treatment*, and *ICU event*—covering, for example, where a patient should be transferred (transfer), which tests to order (test & exam), or the likely disease (diagnosis). These subtypes comprise **24 tasks** that assess the model’s ability to map a patient’s current state to clear, actionable decisions.

In contrast, risk-prediction tasks are binary classification problems in which the model forecasts whether a significant medical event will occur within a specified horizon. We group these into five subtypes—*mortality*, *readmission*, *length of stay*, *transfer*, and *outcome*. Specifically, *transfer* refers to events such as a patient’s admission or transfer to another department, and outcome aggregates severe events such as death or ICU transfer. These account for **18 tasks** that probe the model’s capacity to identify risks from longitudinal patient data.

By spanning both generative and predictive settings, EHR-Bench offers a comprehensive framework that mirrors real-world EHR challenges and rigorously tests LLM reasoning and adaptability in clinical contexts.

MIMIC-IV-CDM. The MIMIC-IV Clinical Decision Making (MIMIC-IV-CDM) [26] benchmark is also derived from MIMIC-IV. Compared with EHR-Bench, its distinct patient preprocessing pipeline and task design focus enable the evaluation of models’ generalization across domain shifts in EHR analysis tasks. Specifically, it targets diagnostic accuracy for four diseases—appendicitis, cholecystitis, diverticulitis, and pancreatitis—framed as decision-making tasks. We assess two diagnostic granularities: main disease diagnosis and ICD-level full diagnosis, enabling evaluation across coarse and fine levels of clinical specificity.

EHRSHOT. EHRSHOT, a public dataset from Stanford Medicine, is used to evaluate models’ generalization to an entirely new healthcare system, which exhibits domain shifts not only in task formulations but also in population demographics, event type distributions, and recording practices. The benchmark comprises 14 risk-prediction tasks across three subtypes: *operational outcomes* (long length of stay, 30-day readmission, ICU transfer), *anticipating lab test results* (thrombocytopenia, hyperkalemia, hypoglycemia, hyponatremia, anemia), and *assignment of new diagnoses* (hypertension, hyperlipidemia, pancreatic cancer, celiac disease, lupus, acute MI). Following the original protocol, we assess both zero-shot performance (directly on unseen data) and few-shot adaptation given 1 to 128 examples, probing robustness and sample-efficient transfer.

2.2.2 Metrics

We evaluate decision-making and risk-prediction tasks with metrics aligned to their outputs, more detailed formulation can be found in Section §4.4.2.

Decision-making tasks are formulated as a multi-label prediction problem, where the model outputs a set of medical entities per sample. We report **F1 score**, balancing **precision** (fraction of predicted entities that are correct) and **recall** (fraction of ground-truth entities recovered). Only exact entity matches are counted as correct. To accommodate general-purpose LLMs without EHR-specific fine-tuning (*e.g.*, GPT-4o, DeepSeek-R1), we provide a candidate pool: 100 randomly sampled labels from the task’s full label space combined with the true labels for models to select, avoiding penalization for unfamiliar output spaces.

Risk-prediction tasks are formulated as binary classification, which can be evaluated with **Area Under the Receiver Operating Characteristic curve (AUROC)**. It is a robust metric that measures a model’s ability to distinguish between positive and negative outcomes across all possible classification thresholds, making it ideal for medical risk prediction where datasets can often be imbalanced.

Since large language models don’t have a traditional classification head, we get our probability estimates by using the tokens `yes` and `no` as our positive and negative classes, under a yes/no question prompt. At inference, we apply a technique called `logit_biases` to isolate the model’s scores for `yes` and `no` [6]. We then normalize these scores using a softmax function to get a probability for the positive class, which we use to calculate AUROC. Further metric details are provided in Section §4.4.2.

2.2.3 Baselines

To provide a comprehensive performance comparison, we select a diverse type of LLMs as our baselines. More details can be found in Section §4.4.3.

- **Llama3.3-70B** [32]: Developed by Meta, this series is a family of powerful open-source models that serve as a robust foundation for a wide range of applications.
- **Qwen2.5-72B** [33]: As a versatile, general-purpose multilingual model from Alibaba Cloud, Qwen2.5 is recognized as one of the strongest open-source models available.
- **GPT-4o** [27]: As OpenAI’s flagship closed-source model, GPT-4o is a powerful multimodal model known for its advanced reasoning and real-time responsiveness.
- **Qwen3-235B** [34]: The **reasoning model** offers a unique ‘thinking’ and ‘non-thinking’ hybrid reasoning engine for enhanced problem-solving.
- **DeepSeek R1** [35]: This is the largest open-source **reasoning model** we evaluate, primarily focused on pushing the limits of reinforcement learning to achieve state-of-the-art results in complex reasoning.
- **GPT-OSS-120B** [36]: Released by OpenAI, this open-source **reasoning model** provides strong

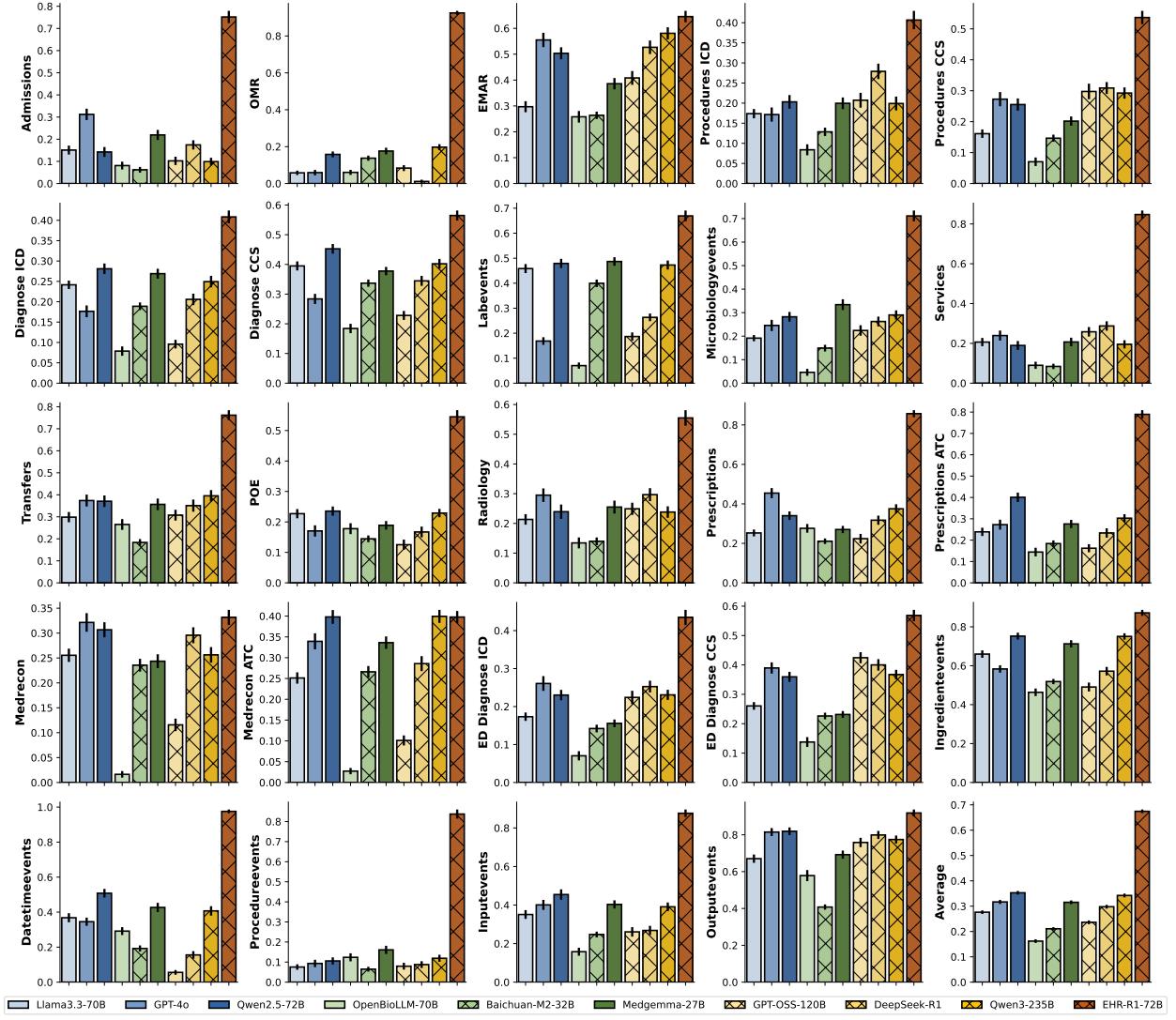


Figure 4 | Performance comparison of EHR-R1 and nine baseline LLMs across 24 decision-making tasks on EHR-Bench. The performance is measured with F1 score. Cross-hatched bars denote reasoning-enhanced models, highlighting the effect of explicit reasoning. In each subplot, our EHR-R1 (rightmost bar) achieves a clear performance advantage on nearly all tasks.

reasoning and agentic capabilities to the open community.

- **Medgemma-27B [37]:** Created by Google, this is a specialized **medical model** optimized for medical text and image comprehension, with a key multimodal capability.
- **OpenBioLLM-70B [38]:** An open-source **medical model** meticulously fine-tuned for the biomedical field based on the Llama3-70B, consistently outperforming larger general-purpose models on biomedical benchmarks.
- **Baichuan-M2-32B [39]:** This is a powerful open-source **medical reasoning model** based on Qwen3-32B. It has been specifically optimized for medical scenarios and trained from scratch on an unprecedented amount of high-quality medical data, enabling it to achieve deep medical expertise.

2.3 Results on EHR-Bench

As shown in Figure 4 and 5, we report the performance of EHR-R1 on both decision-making and risk-prediction tasks, benchmarking against strong baselines to demonstrate the effectiveness across diverse EHR analyses.

2.3.1 Decision-making Tasks

We show the results of evaluated LLMs with bar graph in Figure 4 and the accurate numerical performance in the Supplementary Table 1.

EHR-R1-72B demonstrates superior results. It achieves an average F1 of 0.6744, outperforming the next-best model, Qwen2.5-72B (0.3535), by over 30 points. The advantage is consistent across all decision-making tasks, validating the effectiveness and scalability of our domain-specific, reasoning-driven approach. EHR-R1-72B also surpasses strong closed-source models (GPT-4o: 0.3155) and specialized medical LLMs (Med-Gemma-27B: 0.3157), highlighting superior clinical reasoning and EHR-specific knowledge.

Strong models perform inconsistently. A notable observation is that closed-source commercial models do not reliably outperform open-source ones. For example, GPT-4o (0.3155) performs the second best on average, but is often surpassed by MedGemma-27B (0.3157) and Qwen2.5-72B (0.3535) in inpatient diagnostic settings. Similarly, the general-purpose DeepSeek-R1 model (0.2974) underperformed both MedGemma-27B and Qwen2.5-72B. This variability underscores limitations in generalization for prior LLMs. In contrast, EHR-R1-72B delivers consistently strong F1 across diverse decision-making tasks, suggesting it can construct task-tailored reasoning pathways absent in existing models.

General reasoning shows limited benefit. Despite the common consensus that reasoning capabilities improve model performance, existing reasoning models that are not tailored to EHR tasks actually fail to demonstrate a clear advantage over non-reasoning ones. For instance, the top-performing non-reasoning model, Qwen2.5-72B (0.3535), outperforms the best-performing reasoning baseline model, Qwen3-235B (0.3418). Similarly, the non-reasoning MedGemma-27B (0.3157) achieves a better score than the reasoning-enhanced DeepSeek-R1 (0.2977). This indicates that effective reasoning for EHR requires tight integration of medical knowledge with case-structured analysis rather than general chain-of-thought scaling.

In summary, EHR-R1-72B not only sets a new state-of-the-art for decision-making tasks in EHR analysis but also addresses critical limitations of prior models by seamlessly integrating domain-specific reasoning with medical expertise.

2.3.2 Risk-prediction Tasks

EHR-R1-72B consistently leads. Figure 5 illustrates the performance of our proposed model, against several open-source baselines on 18 risk-prediction tasks, measured by ROC curves. EHR-R1-72B attains an average AUROC of 0.9523, significantly outperforming the second-best baseline model, Qwen3-235B (0.8245). The performance gains are especially pronounced in challenging Emergency Department (ED) tasks, where rapid and accurate risk assessment is critical. For instance, on the ED reattendance 3day task, EHR-R1-72B reaches 0.9007 versus Qwen2.5-72B's 0.5540. For core clinical predictions, the model also excels—for example, inpatient mortality at 0.9787 versus Qwen3-235B's 0.9028.

Baseline LLMs lack task versatility. The baselines' performance varies significantly across task types, indicating limited generalization. Notably, Qwen2.5-72B—strong in decision-making—trails Qwen3-235B (0.8245) on risk prediction despite similar or smaller parameter counts, and both Qwen2.5-72B and Med-Gemma-27B fluctuate across tasks. These inconsistencies suggest current general-purpose and medical LLMs struggle to deliver uniformly reliable performance across the full spectrum of EHR analysis.

Prediction accuracy decreases with longer horizons. All models exhibit decreasing AUROC as the prediction horizon extends, reflecting the complexity of forecasting long-term outcomes from longitudinal, multi-factor trajectories. For ICU Mortality, EHR-R1-72B drops from 0.9898 (1-day) to 0.9610 (14-day). While our model consistently outperforms baselines even on these difficult, long-range predictions—for example, scoring 0.7562 on the Readmission 30Day task against Qwen2.5-72B's 0.6735—it's clear that predicting long-term patient outcomes remains a significant challenge.

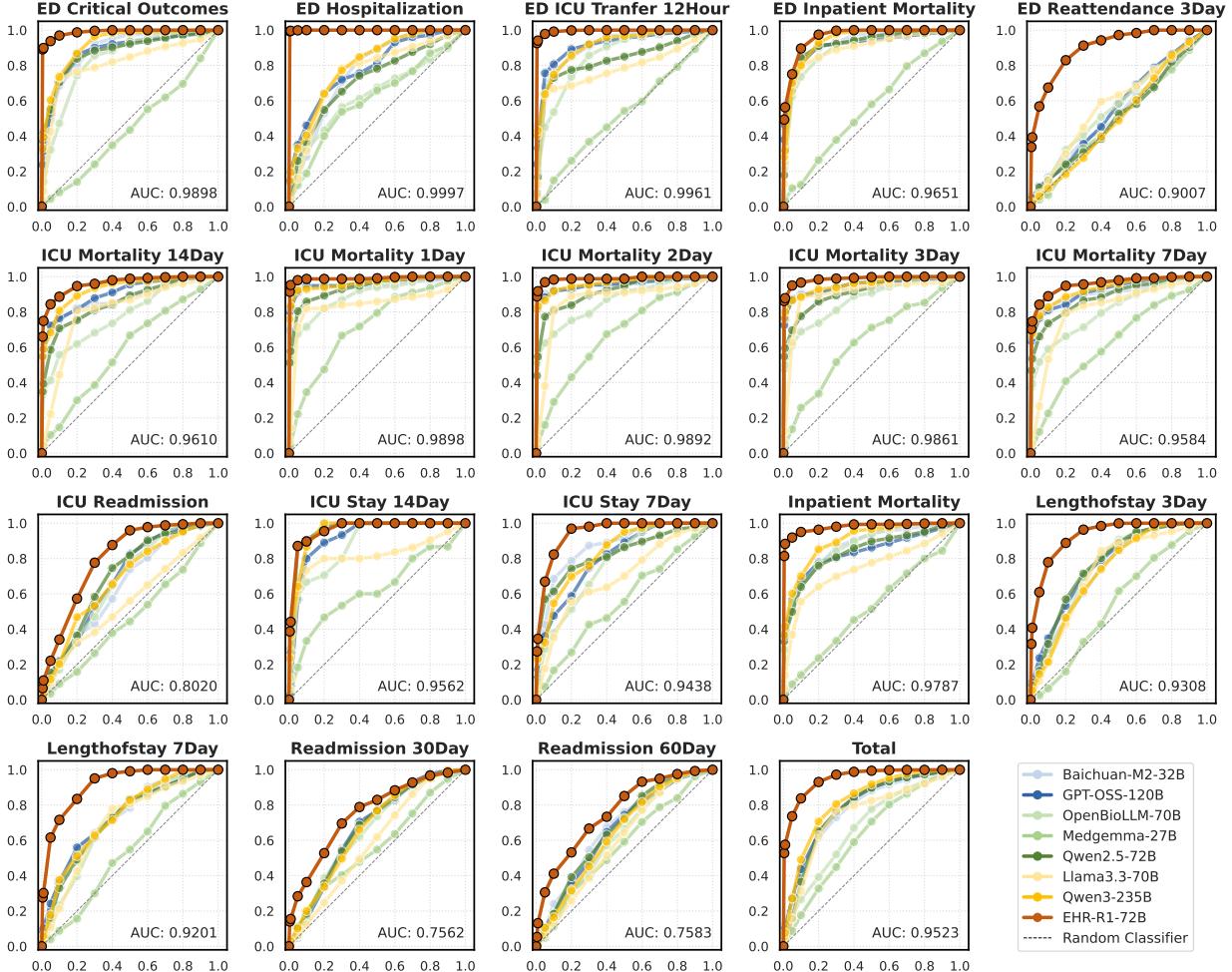


Figure 5 | Performance comparison of EHR-R1 and seven baseline LLMs on 18 risk-prediction tasks on EHR-Bench. Each subplot shows the ROC curves per task, with EHR-R1 highlighted in orange; the bottom-right corner of each plot reports EHR-R1-72B’s AUROC. The final ‘Total’ subplot summarizes aggregated performance across all 18 tasks.

2.4 Generalization Evaluation on MIMIC-IV-CDM

The results of our generalization evaluation on the MIMIC-IV-CDM tasks are shown in Figure 6a. We focus on the zero-shot setting, where models are directly prompted [40] to perform EHR analysis tasks in MIMIC-IV-CDM without any task-specific training. The prompts used for evaluation are shown in Supplementary Table 5. We evaluate LLMs’ performance on two levels of diagnosis tasks: the main disease diagnoses and ICD coding level diagnoses. It can be observed that our model, EHR-R1-72B, achieves the highest performance on both types of tasks.

EHR-R1-72B excels on main disease prediction. For the task of predicting the main disease, which provides a less noisy and more idealized scenario, our model sets a new state-of-the-art with a performance of 0.8913. While other powerful models like DeepSeek-R1 (0.8841) and GPT-OSS-120B (0.8793) also perform well, our model maintains a slight lead. This demonstrates that our model’s diagnostic capabilities are highly effective even in a simplified and refined clinical context.

EHR-R1-72B leads in multi-level diagnoses. Most models struggle to perform well on both main disease and ICD-level coding simultaneously. For example, DeepSeek-R1 (0.8841 Main Disease) and GPT-OSS-120B (0.8793) lag on ICD Code prediction (0.2597 and 0.2422), whereas Med-Gemma-27B attains higher ICD accuracy (0.2860) but lower Main Disease (0.7939). In contrast, EHR-R1-72B, is the only one that achieves

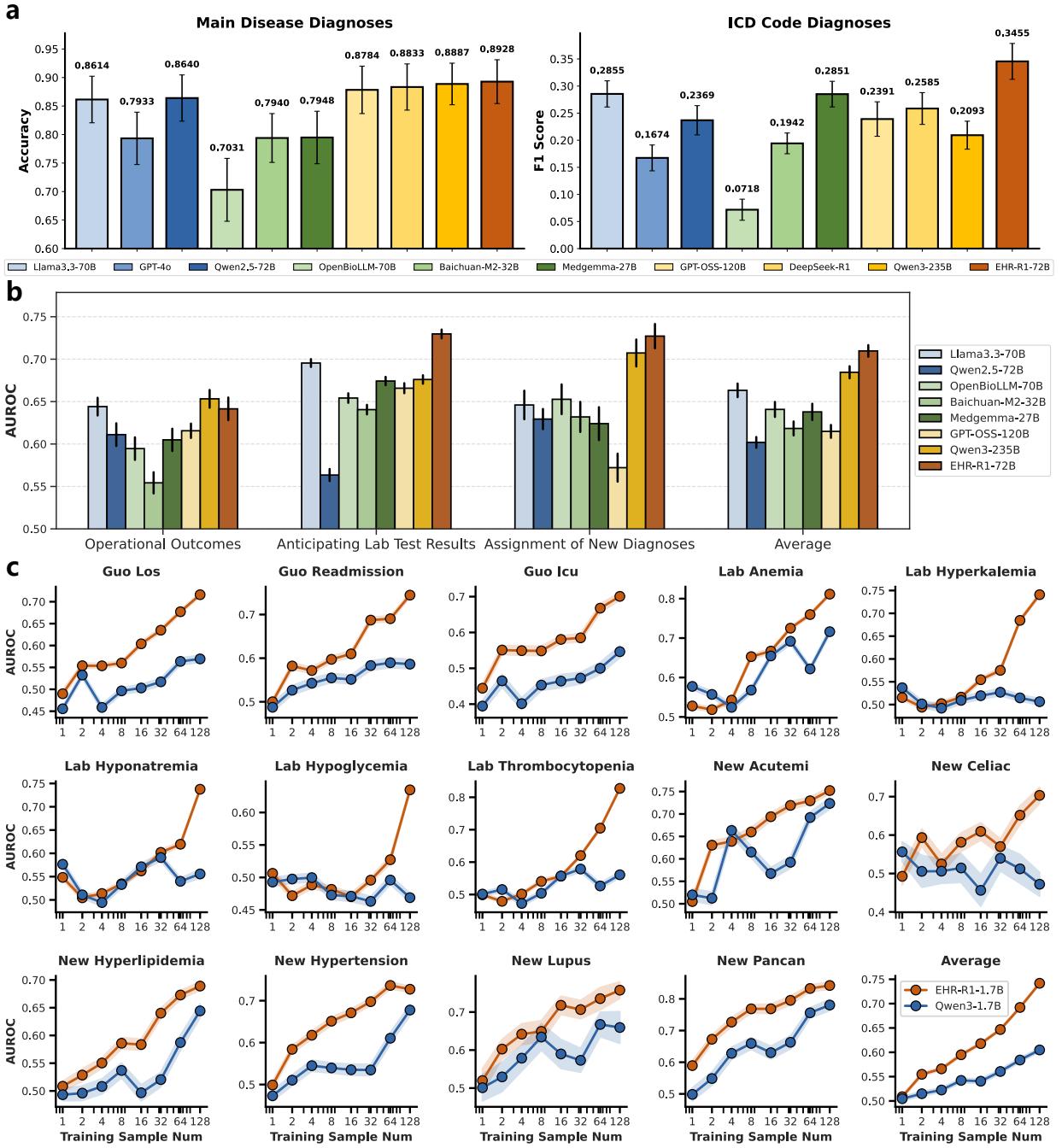


Figure 6 | Performance comparison of EHR-R1 and baseline LLMs in generalization evaluation. The score metric for MIMIC-IV-CDM and EHRSHOT is F1 and AUROC score, respectively. **a** Zero-shot on MIMIC-IV-CDM: performance on Main Disease and ICD-level diagnoses using the same samples. **b** Zero-shot on EHRSHOT for 70B-parameters LLMs: aggregated AUROC across all subtasks within each of the three category groups, plus the overall average across all tasks. **c** Few-shot on EHRSHOT for small-scale language models (Qwen3-1.7B vs. EHR-R1-1.7B): per-task performance across $\{1, 2, 4, 8, 16, 32, 64, 128\}$ shots.

the highest performance on both tasks, with scores of 0.8913 for main disease and 0.3501 for ICD code. These results underscore its ability to generalize zero-shot to a new system while retaining granular coding precision.

2.5 Generalization Evaluation on EHRSHOT

We follow the EHRSHOT protocol to assess generalization in both zero-shot and few-shot scenarios: (1) zero-shot, where we directly evaluate on an unseen dataset with prompts to measure out-of-the-box generalization similar as n MIMIC-IV-CDM, and (2) few-shot, where we **train** on a small number of labeled examples to gauge how quickly LLMs adapt with minimal supervision. The prompts used for evaluation are shown in Supplementary Table 5.

2.5.1 Zero-shot Setting

In the zero-shot setting (Figure 6b), we compare EHR-R1-72B to similarly sized (70B) baselines across three categories—operational outcomes, lab test forecasting, and new-diagnosis assignment—and the overall average. EHR-R1-72B attains the highest AUROC in every category and the overall average, outperforming strong general models (*e.g.*, Qwen2.5-72B) and specialized medical models (*e.g.*, Med-Gemma-27B). Given that EHRSHOT differs from EHR-Bench in both data format and task types, these results underscore robust cross-dataset generalization, which is a critical feature for practical clinical applications.

2.5.2 Few-shot Setting

Considering that model adaptation in clinical settings is often constrained by limited computational resources. Therefore, in the few-shot setting, we use a small-parameter 1.7B LLMs with a minimal number of training samples (at most 128) to evaluate the performance of the LLMs in rapid adaptation scenarios.

EHR-R1 shows better learning efficiency. Figure 6c and the accompanying table present a detailed performance comparison between our small model, EHR-R1-1.7B, and its base model, Qwen3-1.7B, on a variety of EHRSHOT tasks under a few-shot fine-tuning setting. Following EHRSHOT, the models are fine-tuned on a limited number of examples (k) for each task, ranging from 1 to 128. The results show the profound superiority of our model, which achieves an average AUROC of 0.7465 at $k = 128$, far exceeding Qwen3-1.7B’s average of 0.5998 at the same shot setting.

Performance gap grows with more data. Our model’s performance advantages are primarily manifested in two types of tasks: operational outcomes and lab test forecasting. When $k < 16$, the gap is modest—likely because extremely limited data under-utilizes large models. For $k \geq 16$, the gap widens markedly, suggesting that EHR-focused training equips EHR-R1-1.7B to leverage small-to-moderate datasets more efficiently.

General models narrow the gap on diagnosis tasks. For new-diagnosis prediction, the gap peaks around $k = 16$ (*e.g.*, new hypertension: 0.6593 vs. 0.5108 AUROC for EHR-R1-1.7B vs. Qwen3-1.7B) but narrows as k increases. This likely reflects that disease diagnosis is a core medical task extensively represented in general-model pretraining; with sufficient fine-tuning data, general models can catch up.

In summary, our model not only achieves higher performance but also improves faster as data scales, indicating superior learning efficiency and a stronger EHR-specific knowledge base.

2.6 Ablation Study

We conduct ablation experiments to validate the rationale and effectiveness of our proposed framework from **three key perspectives**: the impact of incorporating reasoning data into model training, the effectiveness of reasoning during inference compared to direct answer output, and the generality of our approach across model scales.

Experiments design. We conducted an ablation analysis on five configurations, progressively adding our training and inference components (Figure 7): (i) **BaseModel**: the original base model with direct answers (no reasoning at inference, no training enhancements). (ii) **BaseModel (w/ reasoning inference)**: the base model evaluated with test-time reasoning to guide predictions. (iii) **EHR-R1 (w/o reasoning training)**: the model has undergone our continual pre-training to specialize in the EHR domain, but without reasoning data; inference uses direct answers. (iv) **EHR-R1 (full training)**: our three-stage model—continual pre-training, reasoning-data training, and task-specific fine-tuning—with direct-answer inference. (v) **EHR-R1 (full training, w/ reasoning inference)**: the full model augmented with test-time reasoning. Note that,

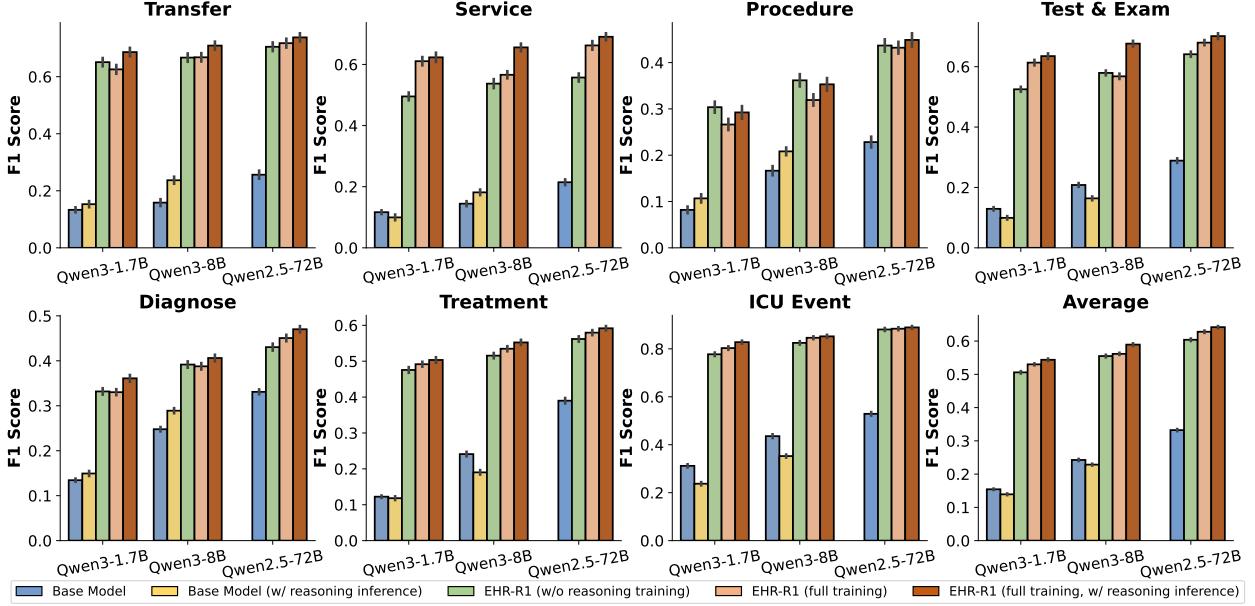


Figure 7 | Ablation experiments results on decision making tasks of EHR-Bench. The figure show 7 categories of sub-type decision making tasks and the average performance of 5 variant of our methods, including Base Model, Base Model (with reasoning inference), EHR-R1 (w/o reasoning training), EHR-R1 (full training), and our final model EHR-R1 (full training, w/ reasoning inference), which showcase the incremental performance gains from each stage of our training pipeline.

Qwen2.5-72B is not applied with **BaseModel (w/ reasoning inference)** because it is not a reasoning-enabled model. The accurate experimental results can be found in the Supplementary Table 2.

Effectiveness of reasoning data. Incorporating reasoning data into training consistently improves performance: for EHR-R1-1.7B, direct-answer F1 rises from 0.5060 (without reasoning data) to 0.5300 (with reasoning data), and for EHR-R1-72B from 0.6039 to 0.6281, indicating that synthesized reasoning injects useful EHR-specific knowledge. The gains are amplified when reasoning is also used at inference, with the full EHR-R1-1.7B reaching 0.5438 F1 compared to 0.5060 without reasoning data, demonstrating that training-time reasoning and test-time reasoning are complementary—training equips the model with reasoning primitives, and inference-time reasoning leverages them to further refine predictions—yielding a more knowledgeable and robust foundation across inference strategies.

Effectiveness of reasoning inference. Reasoning at inference only helps when the model has been explicitly equipped with EHR-specific reasoning. Applying general reasoning prompts to base models offers little-to-negative benefit (Qwen3-1.7B drops from 0.1624 to 0.1456 F1; Qwen3-8B also drops from 0.2425 to 0.2286), underscoring that general reasoning alone is insufficient for clinical tasks. In contrast, our fully trained EHR-R1 models gain substantially from reasoning inference: EHR-R1-1.7B improves from 0.5060 to 0.5438 (± 0.0035), EHR-R1-8B from 0.5549 to 0.5894, and EHR-R1-72B from 0.6039 to 0.6418. These results validate the effectiveness of our ‘thinking-graph’ synthesis pipeline in instilling domain-specific reasoning pathways that inference-time reasoning can reliably exploit.

Effectiveness across model scales. Our approach generalizes robustly with scale: applying the same pipeline from 1.7B to 72B parameters yields monotonic gains, with average F1 rising from 0.5438 (Qwen3-1.7B-based) to 0.5894 (Qwen3-8B-based) and 0.6481 (Qwen2.5-72B-based). This consistent improvement indicates that the benefits of reasoning-data training and reasoning inference are not scale-specific but instead compound as model capacity increases, confirming that our framework scales effectively across parameter regimes.

3 Discussion

Despite the remarkable progress of LLMs across diverse medical tasks, their application to clinical EHR analysis remains a significant challenge [14, 15, 16, 17]. Existing approaches fall short in two critical dimensions. First, task coverage remains narrow, as most efforts are confined to specific objectives or disease cohorts rather than enabling holistic analytical support for evolving clinical workflows. Second, reasoning ability is underdeveloped, with current LLMs struggling to generate reliable EHR-focused reasoning processes that progressively highlight key records, integrate fragmented evidence, and construct longitudinal, synthesized understanding of patient health.

Overview of Our Approach

We introduce a thinking-graph-driven, auto-generation pipeline for EHR reasoning that systematically converts raw, longitudinal records into structured, query-ready insights for generative LLM training. The pipeline: (1) extracts salient medical entities from heterogeneous EHR sources to create clinical focused insights; (2) links these entities into a thinking graph that encodes temporal relations and causal hypotheses, transforming disjoint timesteps into a coherent longitudinal narrative; and (3) synthesizes explicit, stepwise reasoning over this graph to support adaptive solutions across diverse clinical queries. This pipeline enables to curate **EHR-Ins**, a super-instruction corpus with 300K reasoning cases and 3.5M non-reasoning cases spanning 42 EHR tasks. We then train our reasoning-enhanced model, **EHR-R1**, via three-stage training, including large-scale domain adaptation, reasoning enhancement followed by reinforcement learning with Group Relative Policy Optimization (GRPO) to further strengthen analytical fidelity, temporal reasoning, and clinical robustness.

Main Contribution

A super-instruction dataset for EHR analysis with reliable reasoning. We release EHR-Ins, a large-scale super-instruction corpus that pairs 300K high-quality reasoning cases with 3.5M non-reasoning cases across 42 EHR tasks. Reasoning traces are auto-generated via our thinking-graph pipeline, which extracts entities, links temporal/causal relations, and synthesizes clear stepwise rationales—transforming structured EHR data into LLM-ready, actionable insights. This design enables models to both answer diverse clinical queries and explain their decisions.

First comprehensive benchmark for EHR analysis. This paper introduces EHR-Bench, a new benchmark based on the MIMIC-IV dataset, which provides a comprehensive evaluation of LLMs on 42 diverse EHR tasks. This benchmark is designed to holistically assess a model’s ability to handle the full range of diverse clinical queries that exist in real-world EHRs, addressing coverage gaps in existing evaluations.

A robust and generalizable foundation model. We introduce EHR-R1, a state-of-the-art model that demonstrates superior performance on EHR analysis. By learning from the explicit reasoning paths in our data, our model not only sets new performance benchmarks on the proposed EHR-Bench but also shows exceptional zero-shot generalization capabilities on the out-of-distribution EHRSHOT dataset. This highlights EHR-R1’s robust ability to translate EHR content into clinical insight and understand temporal data to form a longitudinal narrative, which is critical for real-world clinical applications.

Key Findings

Effective handling of diverse tasks. Trained on our EHR-Ins dataset, EHR-R1 performs a broad range of EHR tasks with high accuracy. Across 42 tasks in EHR-Bench—including clinical coding and other less-common or challenging tasks—it delivers strong, adaptive reasoning over diverse clinical queries. In addition, its zero-shot performance on unseen EHR data and tasks from the EHRSHOT dataset confirms that the model captures transferable structure rather than memorizing training data.

Effectiveness of reasoning for EHR tasks. The gains derive from an auto-synthesis pipeline that leverages relationships among medical entities to teach explicit reasoning over structured EHR data. This enables the model to denoise heterogeneous inputs, analyze temporal relations, and construct longitudinal clinical narratives. Applying reasoning chains at inference yields substantial improvements, supporting the value of explicit, knowledge-informed reasoning.

Generalizability of our framework. The effectiveness of our method is not limited to a specific model scale. Improvements from our training framework and reasoning data are consistent for models from 1.7B to

72B parameters. This finding demonstrates the generalizability and scalability of our approach, confirming that it provides a robust and widely applicable solution for the EHR domain.

Limitations and Future Work

Scope of reasoning data. Our reasoning-data synthesis was applied only to decision-making tasks, for which target labels include the medical entities necessary for building the thinking graph. This constrains coverage of diverse clinical queries. Although stage-three reinforcement learning transferred some of these capabilities to risk-prediction tasks, future work should explore procedures to construct explicit reasoning data for these binary classification.

Data attrition during graph construction. We excluded many samples for which the thinking graph could not be built—either due to insufficiently related entity pairs or unsuccessful retrieval of medical relations from UMLS. This limits our ability to convert temporal data into a longitudinal narrative for every case. Refining graph-building heuristics, incorporating additional or domain-specific knowledge bases, and improving entity linking could reduce data loss and broaden case coverage.

Specialization versus breadth. The approach yields a specialized EHR model and may attenuate general-purpose capabilities. Emphasizing structured temporal narratives and prescribed reasoning steps risks overfitting to clinical settings. Future work should explore strategies to balance specialization and breadth—for example, multi-domain continued pretraining, alternating-task curricula, modular adapters, or mixture-of-experts routing—to preserve general capabilities while deepening clinical expertise.

4 Methods

In this section, we present the details of our method, starting with the problem formulation, followed by the reasoning data curation pipeline, model training, and finally, evaluation.

4.1 Problem Formulation

We formulate the EHR analysis problems in an instruction-tuning-based [41] generative framework.

An EHR can be represented as a chronologically ordered sequence of longitudinal records (also referred to as events), including laboratory results, medication administrations, transfers, *etc.*:

$$\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_K\}. \quad (1)$$

The k -th record is defined as $\mathbf{r}_k = (c_k, E_k, t_k)$, where $c_k \in \mathcal{C}$ denotes the clinical event category and \mathcal{C} represents the set of all possible event types, such as *medication*, *diagnosis*, *laboratory test*, *etc.*. The term E_k represents the set of medical entities contained in the event (e.g., test items and results in a laboratory event, or diagnoses in a diagnostic event). Finally, t_k is the timestamp of the event.

At any given prediction time step t , a training sample for EHR analysis is expressed as

$$S = (\mathcal{I}, \mathcal{R}_{\leq t}, \mathcal{A}),$$

where \mathcal{I} is a free-text instruction specifying the analysis task, $\mathcal{R}_{\leq t} = \{\mathbf{r}_k \mid t_k \leq t\}$ is the observable history of the EHR up to time t , and \mathcal{A} is the ground-truth free-text answer.

Our objective is to train an LLM, Φ_{LLM} , to perform the following conditional text generation task:

$$\Phi_{\text{LLM}}(\mathcal{I}, \mathcal{R}_{\leq t}) \rightarrow \mathcal{A}, \quad (2)$$

that is, to output an appropriate answer \mathcal{A} given the task instructions \mathcal{I} and the observed EHR history $\mathcal{R}_{\leq t}$.

4.2 Data Curation

In this section, we introduce the curation pipeline of **EHR-Ins** and **EHR-Bench**. We begin by processing the raw MIMIC-IV EHR cases, filtering and reorganizing them into an instruction-tuning generative format.

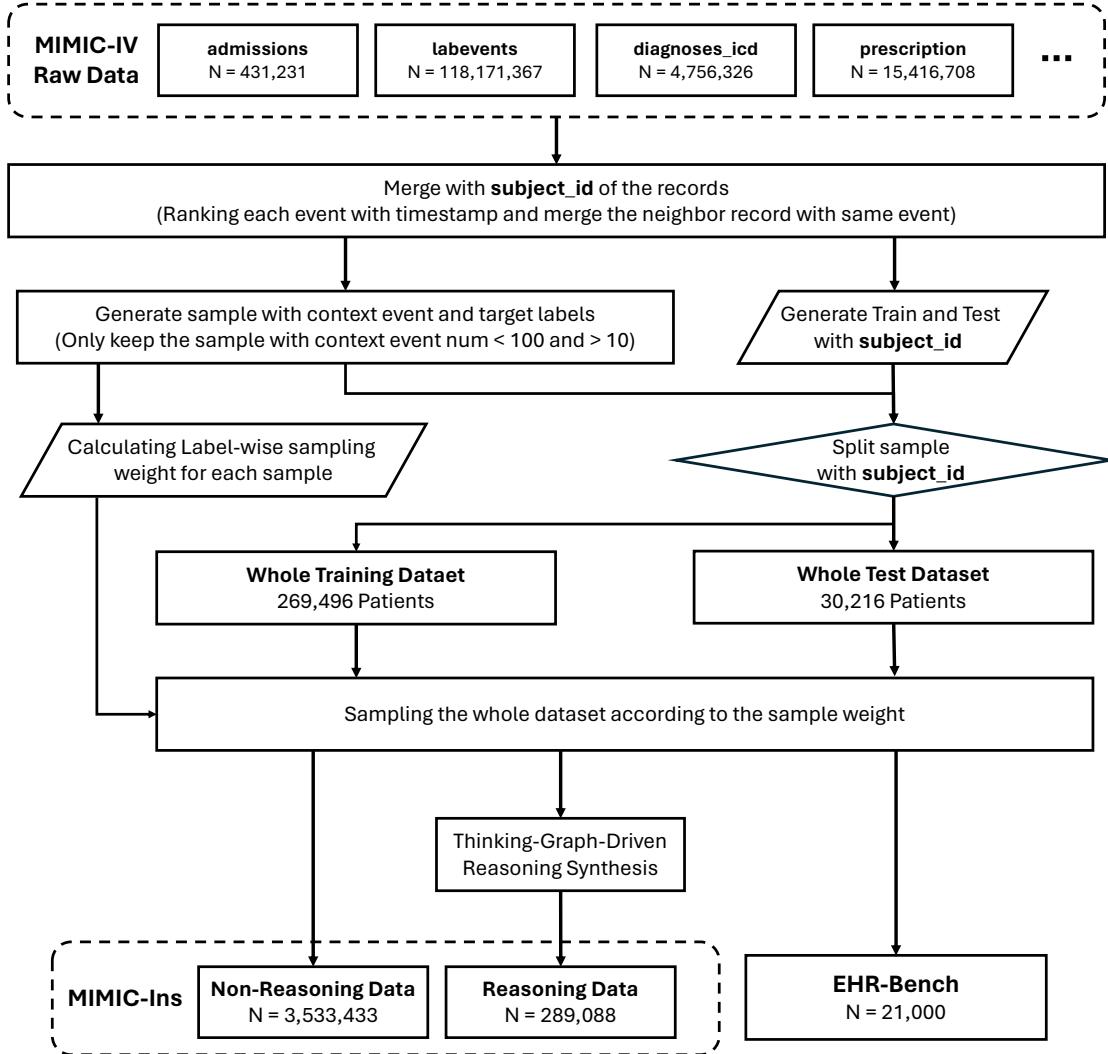


Figure 8 | Overview of Data curation pipeline. The pipeline for our data curation process begins with the original MIMIC-IV dataset. From there, we establish EHR-Ins with reasoning-enhanced EHR analysis instructions through thinking-graph–driven reasoning synthesis, along with a new comprehensive EHR-Bench.

The data is then carefully split into training and testing sets, where the test set forms EHR-Bench, and the training set is further enhanced with reasoning, resulting in the final EHR-Ins. The overview of the preprocessing pipeline is shown in Figure 8.

4.2.1 MIMIC-IV Processing

Here, we describe the construction process of both EHR-Ins and EHR-Bench. First, we introduce the processing details of MIMIC-IV. Then, we describe the free-text formatting of the EHR data, which transforms the long-horizon EHRs into detailed generative training samples. Finally, we describe the approach to achieve a balanced distribution of the data with label-wise weighted sampling methods and how to split the whole dataset into a training and a test set.

Information Enrichment. To enable chronological construction from EHR data, we reformat and sort the original MIMIC-IV dataset at patient level. We first extract all events for each patient and reordered them by timestamp with second precision. Recognizing that *diagnoses_icd*, *procedure_icd*, and *diagnosis* event types in MIMIC-IV lack second-level timestamps, we manually added them to enrich the information. In particular,

for *diagnoses_icd* and *diagnosis*, we associate them with their respective admission events via *hadm_id*. We then set *diagnoses_icd* to 1 minute before the *discharge* event and *diagnosis* to 1 minute before the emergency department discharge event. For *procedure_icd* events, which only provide day-level timestamps, we default their time to 23:59:59 on that day.

To further enrich the EHR information, we map the International Classification of Diseases (ICD) codes in the *diagnoses_icd* and *procedure_icd* tables to Clinical Classifications Software (CCS) categories using the ‘ICD-to-CCS’ script¹. Likewise, we translated National Drug Codes (NDCs) in the *prescriptions* and *pharmacy* tables into Anatomical Therapeutic Chemical (ATC) classification codes adopting the off-the-shelf mapping script². These enrichment consolidate the originally vast and heterogeneous sets of disease, procedure, and medication codes into clinically coherent categories. We also identify that information within discharge events, such as social history and chief complaint, as being observable upon admission. We move these details to the admission event to provide more comprehensive input for LLMs. Additionally, to mitigate potential data leakage in EHR data, we mask *pharmacy* event information within *prescription* events.

After the above processing and enrichment, all MIMIC-IV data are organized into a chronological sequence of records for each patient.

Training Sample Formatting Using the structured EHRs, we construct an instruction dataset on EHR analysis by defining detailed instruction–answer pairs for each training sample, enabling generative training, regarding the two EHR analysis task categories: **decision-making** and **risk-prediction**.

Decision-making tasks involve predicting the next-step event based on the history of events. Therefore, naturally, for a given EHR sequence $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_K\}$, we can generate training samples by sampling an arbitrary timestep t and predicting the next event. Assuming the next-step record as $\mathbf{r}_{t+1} = (c_{t+1}, E_{t+1}, t_{t+1})$, we formulate the instruction \mathcal{I} for decision-making task according to the type of next-step event: c_{k+1} (detailed prompts can be found in Supplementary Table 4), and the target answer set is composed of the entities contained in E_{k+1} :

$$\mathcal{A} = \{y_1, y_2, \dots, y_N\}, \text{ where } y_i \in E_{k+1}, i \in \{1, \dots, N\}, \quad (3)$$

For example, in a case where the next-step event is diagnoses, E_{k+1} contains all the diagnosis results of the patients and y_i represents the name of a disease. In the implementation, \mathcal{A} is formatted as free-text by concatenating the entity list with `str.join` function in Python.

In addition, at the time step t , we can also construct risk-prediction samples. We organize the instruction of the risk-prediction tasks by filling $\{c_{\text{critic}}, T\}$ into a predefined template (details can be found in Supplementary Table 4), where the objective of the model is to determine whether a critical outcome event of type c_{critic} will occur within a specified time frame T . Therefore, the binary ground truth answer of the task can be formulated as:

$$\mathcal{A} = \begin{cases} \text{yes,} & \text{if there exists } i \in \{t + 1, \dots, K\} \text{ such that } c_i = c_{\text{critic}} \text{ and } t_i < T, \\ \text{no,} & \text{otherwise.} \end{cases} \quad (4)$$

Thus far, we have organized training samples in the form $S = (\mathcal{I}, R_{\leq t}, \mathcal{A})$, encompassing EHR analysis tasks that fall into two main categories: decision-making and risk-prediction. **Notably**, since our instructions are generated through prompt filling, all training samples associated with a specific task (e.g., diagnostic decision-making) share the same instruction. Hence, in the following, \mathcal{I} can also be used to denote a specific task.

Lastly, to convert the tabular observed EHR data $R_{\leq t}$ into a format that LLMs can process, we serialize the structured data[42, 43] into Markdown format[21], which is widely present in LLM pre-training corpora, facilitating better model comprehension. Specifically, each event is represented as plain text with a title (event name and standardized start time) and content. The content uses bullet-point syntax for events with single records and table syntax for others with multiple records, as shown in Case 1 and Case 2, respectively. The

¹ICD9 to CCS: <https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>
²ICD10 to CCSR: <https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp>

²NDC to ATC: <https://github.com/sunlabuiuc/PyHealth>

free-text case of EHR is shown in Case 5.

Label-wise Weighted Sampling. Because the constructed dataset over all valid EHR time steps contains a large number of samples and the target candidate sets in MIMIC-IV are highly imbalanced, we apply sampling to form the final training and test sets. We apply weighted sampling to rebalance the target label distribution in the training set. In practice, for decision-making tasks, given a certain task prompted by instruction \mathcal{I} , we first count the frequency of all possible label entities y appeared in the target answer \mathcal{A} and use the reciprocal of the frequency as the sampling weight. Due to that, each training sample’s answer label is a list of entities, and we assign a weight based on the average of its constituent labels. Formally, for a sample S with ground-truth answer as $A = \{y_1, y_2, \dots, y_N\}$, where N is the size of the label set, its weight w_S is given by:

$$w_S = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{Count}_{\mathcal{I}}(y_i)}, \quad (5)$$

where $\text{Count}_{\mathcal{I}}(y_i)$ represents the frequency of label entity y_i in a same type of task. This strategy can also be applied to risk-prediction tasks, where the entities in answering contain only two binary items ('yes' or 'no', thus $N=1$). The weighting formula, therefore, simplifies to the reciprocal of the frequency of the sample’s single label, rebalancing the positive and negative classes.

The label-wise weighted sampling strategy is adopted to generate the train and test sets simultaneously. This approach ensures that the EHR-Ins maintains a balanced distribution of target labels while addressing the issue of class imbalance. Besides, the test set can also cover diverse types of target labels, which makes the EHR-Bench evaluating the performance of the models more comprehensive.

Data Split. Prior studies commonly treat each hospital admission (which may come from the same patient) as an independent sample for data splitting. However, in our setting, to prevent potential information leakage between the training and test sets, we adopted a rigorous patient-level data split. This approach ensures that no patient, identified by their unique `subject_id`, appears in both the training and test sets.

Furthermore, due to context length constraints and computational resources, we limit the observable historical time window to 24 hours. To ensure an appropriate length for historical event trajectories, we filter out samples with fewer than 10 or more than 100 historical events. This process prevents issues of insufficient information or excessively long inputs.

From the entire training dataset, we sample cases using a label-wise weighting scheme to ensure balanced representation across all tasks. This process yielded 3.5M samples designated as non-reasoning data and an additional 300K samples specifically for synthesizing reasoning data. These two sets collectively form our proposed super-instruction dataset, EHR-Ins. Similarly, a test set of 21K samples was created from the whole test set using the same label-wise sampling approach to form our benchmark, EHR-Bench.

4.2.2 Thinking-Graph-Driven Reasoning Synthesis

In this section, to enable the model to provide convincing reasoning in clinical scenarios, we further enhance the answers in EHR-Ins with detailed rationales by incorporating the concept distribution in the EHR and external medical knowledge sources.

We mainly consider the decision-making tasks. We first explain how to identify medical entities that are potentially related to the task labels. We then filter these entities and adopt bidirectional graph searching on external medical knowledge bases to gather the thinking graph, which links the context entities and target entities with medical relations. Finally, we describe how to synthesize the reasoning subset for decision-making tasks in EHR-Ins. The overview of the proposed reasoning synthesis pipeline is shown in Figure 9.

Entity Relevance Identification. Compared with regular medical data, a distinctive feature of EHRs is their tabular format, where most records are clearly indexed by a list of well-defined medical entities. Therefore, by analyzing the relationships among these entities across the training set, we can uncover potential reasoning pathways that trace key observed entities from noisy raw EHRs to task labels.

In particular, we first process some unstructured exception free-text records, such as discharge summaries and

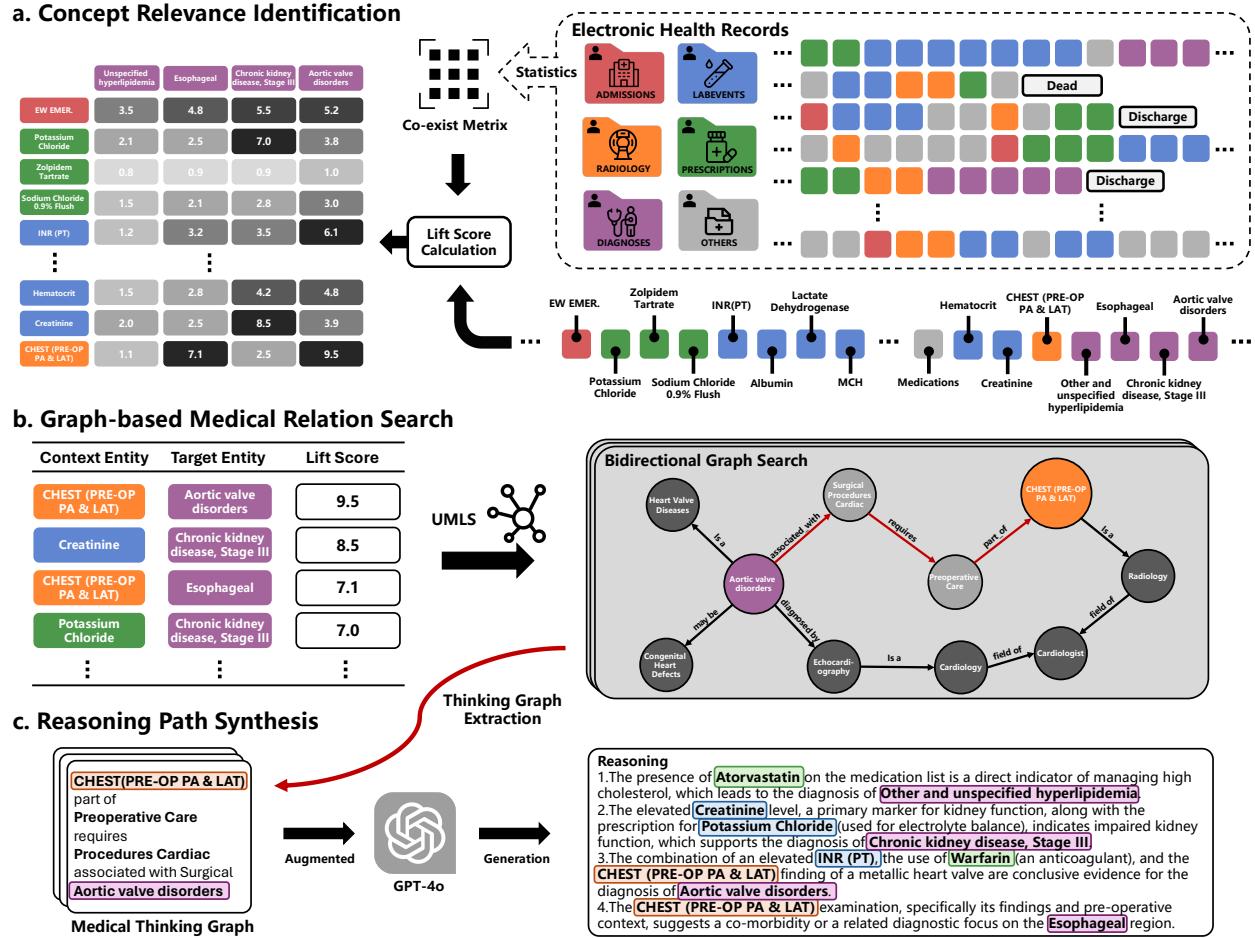


Figure 9 | Overview of Thinking-Graph-Driven Reasoning Synthesis. The whole pipeline comprises three parts: **a** Entity Relevance Identification analyzes electronic health records to calculate a co-exist matrix and Lift Score for medical entities. **b** Graph-based Medical Relation Search filters the entity pair with a higher lift score and conducts a bidirectional graph search on a large-scale medical knowledge graph, UMLS; The red lines indicate the search paths **c** Reasoning Path Synthesis arguments GPT-4o with retrieved medical relations to generate a coherent reasoning path. The entity with the color box in the reasoning chain indicates the context entity or the target labels.

examination reports. We adopt QuickUMLS³ to extract the meaningful medical entities from the raw texts. Then, all the medical entities in the observable events can be expressed as $\mathcal{E}_{\leq t} = \bigcup_{t_i < t} E_i$, $E_i = \{e_1^i, e_2^i, \dots, e_k^i\}$ is the entity set of the i -th events.

Then we count the co-occurrence frequencies of the context entity $e_i \in \mathcal{E}_{\leq t}$ with the target label entities $y_j \in \mathcal{A}$ for a specific decision-making task denoted by its instruction \mathcal{I} and calculate the Lift [44] between them. The Lift metric can discover strong associations within the data by measuring how much more frequently two entities appear together than expected by chance, allowing us to pinpoint the most relevant entities for the reasoning pathways. In our method, the Lift metric is calculated as:

$$\text{Lift}(e_i, y_j, \mathcal{I}) = \frac{P_{\mathcal{I}}(e_i, y_j)}{P_{\mathcal{I}}(e_i)P_{\mathcal{I}}(y_j)} = \frac{\text{Count}_{\mathcal{I}}(e_i, y_j)\text{Count}_{\mathcal{I}}}{\text{Count}_{\mathcal{I}}(e_i)\text{Count}_{\mathcal{I}}(y_j)}, e_i \in \mathcal{E}_{\leq t}, y_j \in \mathcal{A} \quad (6)$$

where $\text{Count}_{\mathcal{I}}(e_i, y_j)$ represents the co-occurrence frequency of e_i and y_j in the training set for task \mathcal{I} and $\text{Count}_{\mathcal{I}}$ represents the total number of samples in task \mathcal{I} . Since rare occurrences of e_i and y_j (for example,

³<https://github.com/Georgetown-IR-Lab/QuickUMLS>

occurring only once) can disproportionately inflate the **Lift** score, we only retain co-occurrence pairs where both the frequency of e_i and y_j are greater than 5. Finally, we retain only the co-occurrence pairs where the **Lift** value is greater than 5 for each sample:

$$\mathcal{K}_{\text{Lift}} = \{(e_i, y_j) \mid \text{Count}_{\mathcal{I}}(e_i) > 5, \text{Count}_{\mathcal{I}}(y_j) > 5, \text{Lift}(e_i, y_j, \mathcal{I}) > 5\}, e_i \in \mathcal{E}_{\leq t}, y_j \in \mathcal{A} \quad (7)$$

This filtering process helps to reduce noise while ensuring that the remaining entity-label pairs have a strong potential association, thus providing candidate references for evidence in the subsequent reasoning process.

Graph-based Entity Relation Search. Although Lift can uncover potentially related medical entities from the historical EHR events, the medical relationships between these entities (including both the context entities in input and the label entities) and the labels remain uncovered and thus fail to capture the underlying clinical associations. To address this issue, we leverage the Unified Medical Language System (UMLS) [25] knowledge graph, which provides a structured representation of medical concepts and their relationships. UMLS integrates a wide range of medical terminologies and ontologies, including SNOMED CT, ICD, and MeSH, among others, making it an ideal resource to validate and strengthen the medical relationships between entities and labels in EHR data. Specifically, we map all medical entities identified in the EHR to their corresponding UMLS concepts. To increase the accuracy and completeness of this mapping, we use fuzzy matching provided in UMLS [25], allowing us to match medical terms in the EHR with UMLS concepts even when there are slight variations in spelling, abbreviations, or synonyms. This approach ensures a broader and more robust set of entity-to-concept mappings, improving the coverage of medical entities in the EHR data. For entities that cannot be mapped to UMLS concepts, we discard the related co-occurrence pairs, as these entities would not contribute valid medical relationships and could introduce noise into the analysis.

After mapping entities to UMLS, we search for medical relations between entity pairs, progressively linking context input entities towards the final label entities, compensating the missed concept nodes, forming a instance specific ***thinking-graph***. However, many medical entities in UMLS do not have direct associations, which makes it challenging to acquire their clinical relationships. To overcome this, we employ a bidirectional graph search algorithm to find connections between entity pairs by traversing across multiple nodes and links. This process allows us to uncover implicit medical relationships between entities that are not directly linked in the UMLS knowledge graph.

The final discovered relationships between the entities are then extracted to form the ***thinking-graph*** to enhance the subsequent synthesis of reasoning paths for this training case:

$$\mathcal{K}_{\text{Think-Graph}} = \{\text{Path}(e_i, y_j) \mid \text{Path}(e_i, y_j) \in \text{UMLS}, (e_i, y_j) \in \mathcal{K}_{\text{Lift}}\} \quad (8)$$

where the function $\text{Path}(e_i, y_j)$ is the connection between e_i and y_j extracted from UMLS.

Reasoning Path Synthesis With the relationships between label entities and historical event entities for each sample, we leverage a powerful large language model, GPT-4o, to generate a detailed reasoning chain based on the historical events, entity relationships, and label entities. The prompt used for the reasoning synthesis are shown in Prompt 2.

Specifically, the reasoning chain consists of three stages: extraction, reasoning, and final result. In the extraction stage, the model is required to extract relevant information from the historical events based on the task's objective. This extraction process must specify the events and their corresponding timestamps to clearly identify the source of the information, ensuring the accuracy and traceability of the extracted data. The reasoning stage requires the model to integrate the extracted information and, using the entity relationship data, link the extracted entities to the corresponding label entities. The model then constructs a coherent reasoning path that logically connects the historical events to the predicted label, providing an explanation for the clinical decision. In the final result stage, the model outputs the predicted result based on the reasoning path constructed in the previous stage. This output represents the model's interpretation of the task based on the historical data and reasoning process.

It is important to note that each sample may contain multiple label entities. For those label entities for which the model cannot construct a valid reasoning chain, the synthesized data may introduce hallucinations.

To address this issue, we instruct GPT-4o to generate reasoning chains only for the label entities that are logically inferable from the available data. For those label entities that cannot be reasoned through, we discard the associated reasoning chains and results. Finally, we retain only those samples where the number of inferable label entities constitutes at least 70% of the original label entities. This threshold ensures that the data distribution remains close to the original raw data and minimizes the potential bias introduced by the discarded entities. Figure 1 shows an example of the reasoning synthesis process.

4.3 Model Training

In this section, we detail the training procedure. The proposed model follows a three-stage training process: large-scale domain adaptation to learn the data distribution and task knowledge in the EHR, reasoning enhancement to extract key information from observable historical events and infer prediction results, and reinforcement learning to enhance the model’s reasoning capabilities on EHR data.

We first serialize the training sample $S = (\mathcal{I}, \mathcal{R}_{\leq t}, \mathcal{A})$ into the final input-output generative format, which can be represented as $(\mathcal{X}, \mathcal{A})$. The input \mathcal{X} is constructed by concatenating the markdown format of EHR historical events with the task-specific instruction \mathcal{I} :

$$\mathcal{X} = \text{Concat}(\text{Markdown}(\mathcal{R}_{\leq t}), \mathcal{I}) \quad (9)$$

Concurrently, the target output \mathcal{A} is formed by joining all items from the corresponding label set, with each item separated by a newline character.

Large-scale Domain Adaptation. The initial step in our training pipeline is to adapt the base LLM to the clinical domain of EHR data by instruction-tuning on naïve question-answering pairs without reasoning. In our experiments, we found that supervising the model on both the input \mathcal{X} and the target labels \mathcal{A} during this phase yields markedly better performance on downstream EHR tasks than training on \mathcal{A} alone. We believe this improvement stems from the fact that \mathcal{X} encodes rich distributional and temporal patterns of patient trajectories, which guide the model to learn co-occurrence and sequence dependencies more effectively. Formally, we minimize the combined loss:

$$\mathcal{L}_{\text{SFT}}(\Phi_{\text{LLM}}) = - \sum_{i=1}^{|\mathcal{A}|} \log \Phi_{\text{LLM}}(\mathcal{A}_i | \mathcal{A}_{<i}, \mathcal{X}) - \sum_{j=1}^{|\mathcal{X}|} \log \Phi_{\text{LLM}}(\mathcal{X}_j | \mathcal{X}_{<j}) \quad (10)$$

Reasoning Enhancement. Following the domain adaptation phase, we conduct instruction-tuning on the reasoning samples aiming at teaching the model how to perform explicit reasoning on EHR analysis tasks. To achieve this, we adopt the reasoning data generated by our automatic synthesis pipeline (as detailed in Section 4.2.2).

While the input \mathcal{X} remains the patient’s observable events and the task instruction, the output \mathcal{A} is an augmented sequence that contains both a detailed reasoning chain encapsulated within the ‘<think>’ and ‘</think>’ tags and the final result. The training objective for this stage remains the same as that of large-scale domain adaptation to ensure consistency and reinforce the model’s ability to learn from both the input sequence and the structured reasoning output.

Reinforcement Learning. Despite the gains achieved through instruction-tuning, the model’s reasoning capacity remains limited by the quality of the generated data. To further unlock its reasoning potential, we apply Group Reward Policy Optimization (GRPO) on top of the SFT checkpoint, with the expectation that end-to-end reinforcement learning will guide the model to self-explore improved reasoning trajectories.

GRPO frames each reasoning chain as a sequential decision process, rewarding trajectories that yield correct predictions and coherent, evidence-grounded explanations. By exploring multiple candidate chains and reinforcing the most effective ones, the model learns to prioritize high-quality inference paths. Specifically, we design GRPO’s reward function $R_{\text{GRPO}}(\tau)$ for the inference reasoning trajectory τ as a sum of two components: format reward and accuracy reward:

Table 1 | Training hyperparameters for three base models. In Instruction Tuning, all models use a high learning rate and large batch sizes for efficient pre-training. For Reasoning Tuning and GRPO, models train for more epochs (3 and 2 respectively) to compensate for smaller, specialized datasets. During GRPO, larger models (Qwen3-8B and Qwen2.5-72B) adopt smaller learning rates (5e-7) to ensure training stability given their extensive parameters. Finally, Qwen2.5-72B’s shorter maximum sequence length (8192) across Reasoning Tuning and GRPO is a necessary compromise due to computational resource limitations.

Training Stage	Model	LR	Batch Size	Epoch	Max Seq Len	GPU Num
Instruction Tuning	Qwen3-1.7B	2e-5	128	1	8192	64
	Qwen3-8B	2e-5	256	1	8192	128
	Qwen2.5-72B	2e-5	512	1	8192	256
Reasoning Tuning	Qwen3-1.7B	2e-5	256	3	12288	128
	Qwen3-8B	2e-5	256	3	12288	128
	Qwen2.5-72B	2e-5	256	3	8192	128
GRPO	Qwen3-1.7B	2e-6	128	2	12288	128
	Qwen3-8B	5e-7	32	2	12288	128
	Qwen2.5-72B	5e-7	32	2	8192	128

$$R(\tau, \hat{\mathcal{A}}, \mathcal{A}, \mathcal{I}) = \lambda_{\text{fmt}} \cdot R_{\text{fmt}}(\tau) + \lambda_{\text{acc}} \cdot R_{\text{acc}}(\hat{\mathcal{A}}, \mathcal{A}, \mathcal{I}) \quad (11)$$

where $\hat{\mathcal{A}}$ is the predicted entity set of LLMs, \mathcal{A} is the target entity set, and \mathcal{I} is the type of EHR analysis task. The format reward $R_{\text{fmt}}(\tau)$ checks that whether the model wraps its chain of thought between the tags ‘<think>’ and ‘</think>’ and respects the three prescribed stages: Extraction, Reasoning, and Final Result:

$$R_{\text{fmt}}(\tau) = \begin{cases} 1, & \text{if the format of } \tau \text{ is right} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The accuracy reward $R_{\text{acc}}(\tau)$ uses the task-specific metric. ACC for the risk-prediction task and F1 for the decision-making task are used to score the final prediction against the ground truth:

$$R_{\text{acc}}(\hat{\mathcal{A}}, \mathcal{A}, \mathcal{I}) = \begin{cases} \text{ACC}(\hat{\mathcal{A}}, \mathcal{A}), & \text{if } \mathcal{I} \text{ is decision-making Task} \\ \text{F1}(\hat{\mathcal{A}}, \mathcal{A}), & \text{if } \mathcal{I} \text{ is Risk Prediction Task} \end{cases} \quad (13)$$

To guard against runaway drift and hallucination, we restrict GRPO updates to just 500 examples per task, ensuring that the model refines its reasoning without straying too far from the validated chains generated during instruction tuning. This targeted reinforcement step produces a model that not only reasons more flexibly over EHR data but also maintains the reliability and interpretability of its clinical explanations.

Implementation Details For our experiments, we adopt Qwen3-1.7B, Qwen3-8B, and Qwen2.5-72B [34] as our base models. The same three stage training strategy us applied to all three models. All the training are conduct based on the code of `Transformers`⁴ and `TRL`⁵ framework. For GRPO stage, The weighted hyperparameters of rewards λ_{fmt} and λ_{acc} are both set to 1 and the number of rollout is 8 per sample. More detail training hyperparameters are shown in Table 1.

4.4 Evaluation

Lastly, we introduce the evaluation settings, including used benchmarks, considered baselines, and metrics.

⁴<https://github.com/huggingface/transformers>

⁵<https://github.com/huggingface/trl>

4.4.1 Benchmarks

In this section, we present the benchmarks considered in this paper and describe how we reformatted them to align with our generative evaluation pipeline, accompanied by detailed case demonstrations.

EHR-Bench. In EHR-Bench, we’ve included two categories of tasks. One of these, decision-making tasks, requires the model to predict multiple possible target answers. For evaluation, a prediction is only considered correct if the model’s predicted medical entities precisely match those in the target answer. However, since most baseline models aren’t fine-tuned on MIMIC-IV data, we generate a candidate set by randomly sampling from the task’s target label set. Baseline models then simply select medical entities from this candidate set. It’s worth noting that our proposed model, EHR-R1, can directly generate answers without needing this candidate set, demonstrating its practical usability. To ensure the difficulty for baseline models selecting from options is comparable to EHR-R1 directly generating answers, we set the candidate set size for each sample to 100 (including the number of target labels). If a task has fewer than 100 target label types, we provide all available target labels for the model to choose from. The example is shown in Case 5.

MIMIC-IV-CDM Benchmark. MIMIC-IV-CDM meticulously selected four diseases—appendicitis, cholecystitis, diverticulitis, and pancreatitis—to test model diagnostic accuracy. Unlike EHR-Bench and EHRSHOT, the historical event information in MIMIC-IV-CDM is not arranged chronologically. Instead, it extracts the most recent information for each event to serve as context. Consequently, the context in MIMIC-IV-CDM is less noisy, and the task is simpler. To maintain consistency with the other two benchmarks, we also reorganized the MIMIC-IV-CDM data into a Markdown-formatted free text. Given that the information within it lacks timestamps, the title for each event only includes the event name without the time it occurred. The example case of the free-text input is shown in Case 4.

EHRSHOT Benchmark. EHRSHOT, a public dataset collected by Stanford University, comprises electronic health records from 7,000 patients. Its publicly available portion includes 14 risk-prediction tasks categorized into three subtypes: Operational Outcomes, Anticipating Lab Test Results, and Assignment of New Diagnoses. EHRSHOT is originally designed for traditional machine learning models, presents a challenge for Large Language Models (LLMs) due to its direct use of various medical codes, which are inherently difficult for LLMs to interpret. To address this, we’ve reformatted the EHRSHOT data into a Markdown-like free-text, similar to EHR-Bench, enabling direct testing with LLMs. Specifically, we utilized the descriptive mapping lexicon provided within EHRSHOT to convert item codes into natural language text. However, certain codes, such as CPT4 and ICD10PCS, could not be directly transformed. Following previous work on evaluating LLMs with EHRSHOT [21], we manually added extra mapping files to resolve these inconsistencies⁶. Furthermore, for EHRSHOT’s measurement and observation events, we semantically clustered the 24 most frequently occurring items. We then enriched these items with additional information, including their units, normal value ranges, and anomaly indicators, significantly enhancing the informational content of EHRSHOT. The example case of the free-text input is shown in Case 3.

4.4.2 Metrics

Based on the two high-level EHR analysis task types, *i.e.*, decision-making and risk-prediction, different evaluation metrics are adopted to quantitatively reflect the performance of LLMs, as formulated below.

Decision-Making Tasks. For each decision-making sample, let the ground-truth target set be \mathcal{A} and the model’s predicted set be $\hat{\mathcal{A}}$. We adopt the precision, recall, F1 score to evaluate prediction accuracy in this multi-label setting:

$$\text{Precision} = \frac{|\mathcal{A} \cap \hat{\mathcal{A}}|}{|\hat{\mathcal{A}}|}, \quad \text{Recall} = \frac{|\mathcal{A} \cap \hat{\mathcal{A}}|}{|\mathcal{A}|}, \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (14)$$

⁶CPT4 mapping: <https://gist.github.com/lieldulev/439793dc3c5a6613b661c33d71fdd185>

ICD10PCS mapping: https://hcup-us.ahrq.gov/toolssoftware/ccsr/PRCCSR_v2025-1.zip

Risk-Prediction Tasks. For risk-prediction tasks, we follow prior work [19] and adopt the area under the receiver operating characteristic curve (AUROC) as the evaluation metric. AUROC measures a model's ability to distinguish between positive and negative outcomes across all decision thresholds, and can be written as

$$\text{AUROC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t), \quad (15)$$

where TPR and FPR denote the true positive and false positive rates, respectively.

4.4.3 Baselines

The following LLMs are used as baselines for comparison:

Qwen2.5/3 Series. Developed by Alibaba Cloud, the Qwen2.5/3 series [34] offers versatile, general-purpose multilingual models with a unique ‘thinking’ and ‘non-thinking’ hybrid reasoning engine and Mixture-of-Experts (MoE) architecture for efficient inference. Trained on 36 trillion tokens across over 100 languages, Qwen3 excels in general reasoning, coding, and agent capabilities. While not specifically medical, its robust multilingual support and advanced reasoning could serve as a powerful general foundation for processing diverse medical literature or assisting in global health information dissemination.

Medgemma 4B/27B. Created by Google, MedGemma [37] is a collection of Gemma 3 variants specifically optimized for medical text and image comprehension. Its key innovation lies in its multimodal capability, seamlessly integrating natural language processing with computer vision to analyze medical images (e.g., X-rays, histopathology) alongside textual patient data. Trained on extensive de-identified medical datasets, MedGemma demonstrates superior performance on clinical benchmarks like MedQA and is designed to accelerate healthcare AI applications while supporting local deployment for data privacy.

Llama3 Series. The Llama3 Series [32], developed by Meta, comprises a family of powerful open-source general-purpose LLMs ranging from 8 billion to 405 billion parameters, with evolving multimodal capabilities in Llama 3.2. Trained on over 15 trillion tokens, these models excel in broad applications, including complex reasoning, coding, and multilingual understanding. While not inherently specialized for medicine, the Llama3 architecture serves as a robust foundation, with fine-tuned variants like Me-LLAMA demonstrating superior performance in various medical question-answering and summarization tasks, highlighting its adaptability for domain-specific healthcare solutions.

OpenBioLLM. Developed by Saama AI Labs, OpenBioLLM [38] is an 8-billion-parameter open-source language model specifically designed for the biomedical field, built upon the Llama 3 architecture. It is meticulously fine-tuned on high-quality biomedical data, enabling it to accurately understand and generate text for tasks like clinical entity recognition, medical question answering, and patient data de-identification. OpenBioLLM consistently outperforms larger general-purpose models on diverse biomedical benchmarks, making it a highly effective and transparent choice for healthcare and life sciences applications.

Baichuan M2 32B. Baichuan-M2 [39], from Baichuan Intelligent, stands out as the first open-source LLM specifically optimized for medical scenarios, trained “from scratch” on an unprecedented 20 trillion tokens of high-quality medical and general data. This dedicated training, coupled with a multi-stage curriculum learning approach, enables it to achieve deep medical expertise, outperforming models five times its size in clinical practice benchmarks while maintaining strong general capabilities. Its specialized focus makes it highly relevant for tasks such as clinical decision support, medical education, and rare disease diagnosis.

GPT-4o. OpenAI’s GPT-4o (omni) [27] is a flagship multimodal model designed for natural human-computer interaction, accepting and generating any combination of text, audio, image, and video through a single, end-to-end neural network . It offers real-time responsiveness and enhanced tokenization for non-English languages, making it faster and more cost-effective. While a general-purpose model, its advanced multimodal understanding and reasoning capabilities could be applied to complex medical cases involving diverse data types, from patient consultations (audio/text) to diagnostic images.

DeepSeek R1. Developed by DeepSeek, R1 [35] is an open-source AI model released in January 2025, primarily focused on pushing the limits of reinforcement learning as a post-training technique for complex reasoning. Utilizing a Mixture-of-Experts (MoE) architecture, it achieves high accuracy in mathematics, programming, and general reasoning tasks at a significantly lower operating cost. Notably, DeepSeek R1 has shown strong performance on medical benchmarks like MedQA [45], demonstrating its potential for efficient and accurate reasoning in healthcare contexts.

GPT-OSS. The GPT-OSS (Open-Source Series) [36] is OpenAI’s first open-source GPT-class model since GPT-2, released under an Apache 2.0 license to provide powerful reasoning and agentic capabilities to the open community. The models, gpt-oss-120b and gpt-oss-20b, utilize a Mixture-of-Experts (MoE) architecture for efficient inference and support a long 128k token context window. While a general-purpose reasoning model, its performance on complex, health-related queries is noteworthy; the gpt-oss-120b model achieved a score of 30% on the HealthBench benchmark, outperforming OpenAI’s own o4-mini and o3-mini models on this specific task. This demonstrates its potential as a strong, customizable foundation for a variety of applications, including those with intricate medical reasoning requirements.

5 Data Availability

The data source for EHR-Bench and EHR-Ins derives from MIMIC-IV, a public EHR resource. Both datasets are under review of PhysioNet⁷.

6 Code Availability

All source codes of this paper have been released in <https://github.com/MAGIC-AI4Med/EHR-R1>.

⁷<https://physionet.org/>

References

- [1] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.
- [2] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark - A comprehensive machine learning benchmark on high-resolution ICU data. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.
- [3] José Roberto Ayala Solares, Francesca Elisa Dilettia Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir Hossein Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, and Gholamreza Salimi Khorshidi. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Informatics*, 101:103337, 2020.
- [4] Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department triage prediction models with machine learning and large public electronic health records. In *AMIA 2022, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 5-9, 2022*. AMIA, 2022.
- [5] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In Marzyeh Ghassemi, editor, *ACM CHIL '20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 222–235. ACM, 2020.
- [6] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- [7] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [8] Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*, 2024.
- [9] Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*, 2025.
- [10] Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards evaluating and building versatile large language models for medicine. *npj Digit. Medicine*, 8(1), 2025.
- [11] Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. Reflectool: Towards reflection-aware tool-augmented clinical agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13507–13531. Association for Computational Linguistics, 2025.
- [12] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

- [13] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.
- [14] Hossein Rouhizadeh, Anthony Yazdani, Boya Zhang, David Vicente Alvarez, Matthias Hüser, Alexandre Vanobberghen, Rui Yang, Irene Li, Andreas Walter, and Douglas Teodoro. Large language models struggle to encode medical concepts—a multilingual benchmarking and comparative analysis. *medRxiv*, pages 2025–01, 2025.
- [15] Yixiang Qu, Yifan Dai, Shilin Yu, Pradham Tanikella, Travis Schrank, Trevor Hackman, Didong Li, and Di Wu. Enhancing llms with smart preprocessing for ehr analysis. *arXiv preprint arXiv:2412.02868*, 2024.
- [16] Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):A1dbp2300040, 2024.
- [17] Weijieying Ren, Jingxi Zhu, Zehao Liu, Tianxiang Zhao, and Vasant Honavar. A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models. *arXiv preprint arXiv:2507.12774*, 2025.
- [18] Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Liantao Ma, and Lequan Yu. Healthflow: A self-evolving ai agent with meta planning for autonomous healthcare research. *arXiv preprint arXiv:2508.02621*, 2025.
- [19] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 67125–67137. Curran Associates, Inc., 2023.
- [20] Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training llms for ehr-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*, 2025.
- [21] Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *arXiv preprint arXiv:2502.17403*, 2025.
- [22] Zhenbang Wu, Anant Dadu, Mike Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. *Advances in Neural Information Processing Systems*, 37:54772–54786, 2024.
- [23] Ziyi Chen, Mengyuan Zhang, Mustafa Mohammed Ahmed, Yi Guo, Thomas J George, Jiang Bian, and Yonghui Wu. Narrative feature or structured feature? a study of large language models to identify cancer patients at risk of heart failure. In *AMIA Annual Symposium Proceedings*, volume 2024, page 242, 2025.
- [24] Samuel Kim, In Gu Sean Lee, Mijeong Irene Ban, and Jane Chiang. Detection and prediction of clopidogrel treatment failures using longitudinal structured electronic health records. *arXiv preprint arXiv:2310.08757*, 2023.
- [25] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.
- [26] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaassis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making.
- [27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

- [28] 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.
- [29] Paweł Renc, Michał K. Grzeszczyk, Nassim Oufattolle, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew B. A. McDermott, Jarosław Was, Anthony E. Samir, Jonathan W. Cunningham, David W. Bates, and Arkadiusz Sitek. Foundation model of electronic medical records for adaptive risk estimation. *CoRR*, abs/2502.06124, 2025.
- [30] Paweł Renc, Yugang Jia, Anthony E. Samir, Jarosław Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digit. Medicine*, 7(1), 2024.
- [31] Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for MIMIC-IV. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Machine Learning for Health, ML4H 2022, 28 November 2022, New Orleans, Louisiana, USA & Virtual*, volume 193 of *Proceedings of Machine Learning Research*, pages 311–325. PMLR, 2022.
- [32] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- [33] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [35] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.
- [36] OpenAI. Introducing gpt-oss. <https://openai.com/index/introducing-gpt-oss/>. Accessed: 2025-08-08.

- [37] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [38] Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*, 2024.
- [39] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.
- [40] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [42] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor. In *International Conference on Learning Representations*, 2022.
- [43] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3):1–28, 2024.
- [44] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.
- [45] 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.

7 Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), and the Scientific Research Innovation Capability Support Project for Young Faculty (ZY-GXQNJSKYCXNLZCXM-I22).

8 Author Contributions

All listed authors clearly meet the ICMJE 4 criteria. Y.L., C.W., and J.L. contribute equally to this work. Y.W. and W.X. are the corresponding authors. Specifically, Y.L., C.W., J.L., S.J., P.Q., H.W., Y.Y., S.Z., J.W., Q.F., J.G., Y.Z., Y.W., Y.W., and W.X. all make contributions to the conception or design of the work, and Y.L., C.W., J.L. further perform acquisition, analysis, or interpretation of data for the work. In writing, Y.L., C.W., J.L. draft the work. S.J., P.Q., H.W., Y.Y., S.Z., J.W., Q.F., J.G., Y.Z., Y.W., Y.W., and W.X. review it critically for important intellectual content. All authors approve of the version to be published and agree to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

9 Supplementary

9.1 Details of Experiments

Supplementary Table 1 | Accurate results of decision making tasks in EHR-Bench. The performance are measured with the metric F1 score and shown in in the format ‘mean \pm std’.

Tasks	Llama3.3-70B	GPT-4o	Qwen2.5-72B	OpenBioLLM-70B	Baichuan-M2-32B	Medgemma-27B	GPT-OSS-120B	DeepSeek-R1	Qwen3-235B	EHR-R1-72B
Admissions	0.1438 \pm 0.0169	0.3151 \pm 0.0208	0.1452 \pm 0.0172	0.0787 \pm 0.0122	0.0633 \pm 0.0074	0.2172 \pm 0.0164	0.1029 \pm 0.0141	0.1728 \pm 0.0166	0.0974 \pm 0.0134	0.7512 \pm 0.0186
OMR	0.0569 \pm 0.0059	0.0582 \pm 0.0080	0.1591 \pm 0.0092	0.0598 \pm 0.0077	0.1373 \pm 0.0088	0.1739 \pm 0.0099	0.0835 \pm 0.0092	0.0115 \pm 0.0032	0.1989 \pm 0.0107	0.9242 \pm 0.0059
EMAR	0.2941 \pm 0.0156	0.5543 \pm 0.0213	0.5087 \pm 0.0227	0.2558 \pm 0.0166	0.2639 \pm 0.0114	0.3858 \pm 0.0183	0.4118 \pm 0.0198	0.5290 \pm 0.0225	0.5801 \pm 0.0179	0.6503 \pm 0.0192
Procedures ICD	0.1737 \pm 0.0085	0.1693 \pm 0.0138	0.2049 \pm 0.0156	0.0848 \pm 0.0118	0.1272 \pm 0.0073	0.2001 \pm 0.0103	0.2074 \pm 0.0172	0.2770 \pm 0.0173	0.1976 \pm 0.0139	0.4079 \pm 0.0189
Procedures CCS	0.1634 \pm 0.0109	0.2718 \pm 0.0196	0.2557 \pm 0.0158	0.0689 \pm 0.0106	0.1466 \pm 0.0086	0.2047 \pm 0.0116	0.2971 \pm 0.0186	0.3064 \pm 0.0174	0.2907 \pm 0.0160	0.5340 \pm 0.0189
Diagnoses ICD	0.2432 \pm 0.0088	0.1767 \pm 0.0116	0.2805 \pm 0.0129	0.0786 \pm 0.0074	0.1880 \pm 0.0067	0.2681 \pm 0.0088	0.0972 \pm 0.0078	0.2058 \pm 0.0119	0.2470 \pm 0.0111	0.4088 \pm 0.0123
Diagnoses CCS	0.3972 \pm 0.0107	0.2847 \pm 0.0121	0.4529 \pm 0.0122	0.1880 \pm 0.0119	0.3366 \pm 0.0080	0.3760 \pm 0.0099	0.2293 \pm 0.0137	0.3451 \pm 0.0131	0.3992 \pm 0.0123	0.5654 \pm 0.0134
Labevents	0.4558 \pm 0.0152	0.1667 \pm 0.0108	0.4788 \pm 0.0123	0.0702 \pm 0.0070	0.4001 \pm 0.0126	0.4901 \pm 0.0138	0.1878 \pm 0.0096	0.2625 \pm 0.0107	0.4740 \pm 0.0135	0.6681 \pm 0.0178
Microbiologevents	0.1915 \pm 0.0091	0.2437 \pm 0.0183	0.2826 \pm 0.0174	0.0458 \pm 0.0093	0.1494 \pm 0.0078	0.3325 \pm 0.0159	0.2250 \pm 0.0200	0.2627 \pm 0.0158	0.2888 \pm 0.0140	0.7101 \pm 0.0215
Services	0.2055 \pm 0.0185	0.2421 \pm 0.0188	0.1888 \pm 0.0150	0.0902 \pm 0.0129	0.0860 \pm 0.0083	0.2060 \pm 0.0174	0.2599 \pm 0.0199	0.2834 \pm 0.0202	0.1938 \pm 0.0191	0.8451 \pm 0.0169
Transfers	0.2997 \pm 0.0182	0.3686 \pm 0.0217	0.3697 \pm 0.0227	0.2668 \pm 0.0203	0.1831 \pm 0.0092	0.3626 \pm 0.0221	0.3035 \pm 0.0211	0.3500 \pm 0.0219	0.3968 \pm 0.0195	0.7634 \pm 0.0153
POE	0.2276 \pm 0.0101	0.1716 \pm 0.0147	0.2401 \pm 0.0121	0.1793 \pm 0.0144	0.1457 \pm 0.0083	0.1880 \pm 0.0091	0.1220 \pm 0.0139	0.1699 \pm 0.0126	0.2280 \pm 0.0123	0.5482 \pm 0.0200
Radiology	0.2115 \pm 0.0136	0.2943 \pm 0.0197	0.2418 \pm 0.0192	0.1364 \pm 0.0164	0.1408 \pm 0.0090	0.2583 \pm 0.0169	0.2540 \pm 0.0199	0.2982 \pm 0.0205	0.2383 \pm 0.0163	0.5542 \pm 0.0210
Prescriptions	0.2512 \pm 0.0110	0.4533 \pm 0.0218	0.3390 \pm 0.0141	0.2727 \pm 0.0192	0.2124 \pm 0.0086	0.2705 \pm 0.0124	0.2271 \pm 0.0154	0.3157 \pm 0.0194	0.3761 \pm 0.0162	0.8560 \pm 0.0116
Prescriptions ATC	0.2375 \pm 0.0128	0.2686 \pm 0.0183	0.3996 \pm 0.0183	0.1443 \pm 0.0162	0.1848 \pm 0.0091	0.2739 \pm 0.0137	0.1617 \pm 0.0149	0.2316 \pm 0.0174	0.2974 \pm 0.0134	0.7891 \pm 0.0138
Medrecon	0.2551 \pm 0.0114	0.3210 \pm 0.0159	0.3030 \pm 0.0133	0.0171 \pm 0.0048	0.2343 \pm 0.0102	0.2436 \pm 0.0113	0.1139 \pm 0.0102	0.2961 \pm 0.0149	0.2545 \pm 0.0133	0.3319 \pm 0.0131
Medrecon ATC	0.2524 \pm 0.0099	0.3383 \pm 0.0137	0.4006 \pm 0.0128	0.0271 \pm 0.0049	0.2647 \pm 0.0093	0.3345 \pm 0.0133	0.1002 \pm 0.0100	0.2842 \pm 0.0159	0.4003 \pm 0.0128	0.3981 \pm 0.0142
ED Diagnoses ICD	0.1740 \pm 0.0076	0.2583 \pm 0.0150	0.2320 \pm 0.0131	0.0691 \pm 0.0087	0.1421 \pm 0.0064	0.1500 \pm 0.0076	0.2231 \pm 0.0144	0.2514 \pm 0.0154	0.2310 \pm 0.0110	0.4319 \pm 0.0168
ED Diagnoses CCS	0.2593 \pm 0.0098	0.3877 \pm 0.0173	0.3595 \pm 0.0127	0.1365 \pm 0.0127	0.2271 \pm 0.0079	0.2321 \pm 0.0076	0.4240 \pm 0.0151	0.3982 \pm 0.0153	0.3669 \pm 0.0118	0.5674 \pm 0.0149
Ingredientevents	0.6613 \pm 0.0124	0.5826 \pm 0.0146	0.7537 \pm 0.0101	0.4662 \pm 0.0117	0.5176 \pm 0.0077	0.7131 \pm 0.0119	0.4914 \pm 0.0165	0.5704 \pm 0.0166	0.7489 \pm 0.0110	0.8739 \pm 0.0080
Datetimeevents	0.3659 \pm 0.0164	0.3438 \pm 0.0196	0.5096 \pm 0.0206	0.2956 \pm 0.0192	0.1931 \pm 0.0099	0.4280 \pm 0.0181	0.0572 \pm 0.0082	0.1557 \pm 0.0131	0.4089 \pm 0.0167	0.9758 \pm 0.0043
Procedureevents	0.0745 \pm 0.0077	0.0938 \pm 0.0122	0.1030 \pm 0.0100	0.1231 \pm 0.0146	0.0655 \pm 0.0059	0.1628 \pm 0.0150	0.0787 \pm 0.0122	0.0884 \pm 0.0115	0.1192 \pm 0.0109	0.8377 \pm 0.0179
Inputevents	0.3504 \pm 0.0158	0.3979 \pm 0.0204	0.4516 \pm 0.0192	0.1565 \pm 0.0163	0.2457 \pm 0.0085	0.4030 \pm 0.0181	0.2621 \pm 0.0159	0.2705 \pm 0.0193	0.3911 \pm 0.0152	0.8763 \pm 0.0129
Outputevents	0.6730 \pm 0.0180	0.8107 \pm 0.0170	0.8231 \pm 0.0154	0.5745 \pm 0.0200	0.4049 \pm 0.0087	0.6949 \pm 0.0184	0.7556 \pm 0.0190	0.8012 \pm 0.0176	0.7794 \pm 0.0168	0.9176 \pm 0.0116
Average	0.2758 \pm 0.0025	0.3155 \pm 0.0037	0.3535 \pm 0.0031	0.1621 \pm 0.0027	0.2108 \pm 0.0018	0.3157 \pm 0.0027	0.2365 \pm 0.0035	0.2974 \pm 0.0033	0.3418 \pm 0.0025	0.6744 \pm 0.0029

Supplementary Table 2 | Accurate results of ablation experiments. The table presents the performance of different model configurations across various decision-making tasks. We evaluate our method on base models of different sizes, including Qwen3-1.7B, Qwen3-8B, Qwen2.5-72B, and Qwen3-235B. The performances are measured with the metric F1 score and shown in the format ‘mean \pm std’, with the average performance across all tasks provided in the final column.

Methods	Decision Making Tasks							Average
	Transfer	Service	Procedure	Test&Exam	Diagnose	Treatment	ICU Event	
Qwen3-1.7B								
Base Model	0.1332 \pm 0.0087	0.1166 \pm 0.0059	0.0820 \pm 0.0071	0.1292 \pm 0.0053	0.1344 \pm 0.0035	0.1225 \pm 0.0032	0.3119 \pm 0.0064	0.1544 \pm 0.0020
Base Model (w/ reasoning inference)	0.1530 \pm 0.0107	0.0996 \pm 0.0093	0.1068 \pm 0.0087	0.0993 \pm 0.0055	0.1493 \pm 0.0055	0.1182 \pm 0.0051	0.2374 \pm 0.0066	0.1393 \pm 0.0024
EHR-R1 (w/o reasoning training)	0.6510 \pm 0.0144	0.4953 \pm 0.0127	0.3037 \pm 0.0119	0.0525 \pm 0.0081	0.3319 \pm 0.0171	0.4756 \pm 0.0078	0.7773 \pm 0.0071	0.5060 \pm 0.0034
EHR-R1 (full training)	0.6257 \pm 0.0153	0.6106 \pm 0.0134	0.2663 \pm 0.0121	0.6135 \pm 0.0089	0.3302 \pm 0.0066	0.4916 \pm 0.0067	0.8031 \pm 0.0072	0.5300 \pm 0.0029
EHR-R1 (full training, w/ reasoning inference)	0.6863 \pm 0.0148	0.6232 \pm 0.0150	0.2924 \pm 0.0136	0.6350 \pm 0.0089	0.3610 \pm 0.0075	0.5035 \pm 0.0075	0.8280 \pm 0.0060	0.5438 \pm 0.0035
Qwen3-8B								
Base Model	0.1585 \pm 0.0119	0.1445 \pm 0.0080	0.1665 \pm 0.0098	0.2082 \pm 0.0062	0.2478 \pm 0.0048	0.2411 \pm 0.0062	0.4358 \pm 0.0077	0.2425 \pm 0.0029
Base Model (w/ reasoning inference)	0.2370 \pm 0.0123	0.1815 \pm 0.0091	0.2083 \pm 0.0083	0.1639 \pm 0.0062	0.2890 \pm 0.0056	0.1903 \pm 0.0062	0.3527 \pm 0.0061	0.2286 \pm 0.0026
EHR-R1 (w/o reasoning training)	0.6670 \pm 0.0144	0.5372 \pm 0.0145	0.3617 \pm 0.0132	0.5793 \pm 0.0078	0.3917 \pm 0.0070	0.5155 \pm 0.0073	0.8249 \pm 0.0069	0.5549 \pm 0.0038
EHR-R1 (full training)	0.6681 \pm 0.0145	0.5661 \pm 0.0117	0.3192 \pm 0.0122	0.5681 \pm 0.0088	0.3877 \pm 0.0071	0.5348 \pm 0.0070	0.8462 \pm 0.0057	0.5616 \pm 0.0031
EHR-R1 (full training, w/ reasoning inference)	0.7091 \pm 0.0138	0.6558 \pm 0.0124	0.3531 \pm 0.0135	0.6764 \pm 0.0092	0.4062 \pm 0.0072	0.5523 \pm 0.0071	0.8521 \pm 0.0063	0.5894 \pm 0.0037
Qwen2.5-72B								
Base Model	0.2564 \pm 0.0142	0.2146 \pm 0.0092	0.2283 \pm 0.0116	0.2888 \pm 0.0074	0.3310 \pm 0.0053	0.3900 \pm 0.0074	0.5289 \pm 0.0072	0.3323 \pm 0.0031
EHR-R1 (w/o reasoning training)	0.7049 \pm 0.0151	0.5572 \pm 0.0131	0.4368 \pm 0.0133	0.6412 \pm 0.0085	0.4304 \pm 0.0077	0.5619 \pm 0.0072	0.8808 \pm 0.0061	0.6039 \pm 0.0037
EHR-R1 (full training)	0.7176 \pm 0.0155	0.6623 \pm 0.0140	0.4322 \pm 0.0126	0.6797 \pm 0.0085	0.4506 \pm 0.0071	0.5796 \pm 0.0070	0.8840 \pm 0.0060	0.6281 \pm 0.0034
EHR-R1 (full training, w/ reasoning inference)	0.7379 \pm 0.0143	0.6905 \pm 0.0113	0.4489 \pm 0.0143	0.7013 \pm 0.0092	0.4702 \pm 0.0070	0.5918 \pm 0.0060	0.8894 \pm 0.0054	0.6418 \pm 0.0031

9.2 Details of EHR-Bench

Supplementary Table 3 | Details of task information in EHR-Bench. This table presents a comprehensive overview of the 42 tasks, categorized into 2 main Task Types and further divided into 12 SubTypes. The Target Event column specifies the type of event associated with the target label for each task. For decision making tasks, the Item Name indicates the column name in MIMIC-IV where the target label is located. It's important to note that Risk Prediction tasks do not have an Item Name as their target labels are not specific entities. Columns marked with '*' represent additional columns created through manual mapping. The Candidates column shows the size of the target candidate set for each task.

Task Type	SubType	Task	Target Event	Item Name	Metric	Candidates
	Reassignment	Admissions Transfer	<i>admissions</i> <i>transfer</i>	<i>admission_type</i> <i>eventtype</i>	F1 F1	8 39
	Test & Exam	OMR Labevents Microbiologyevents Radiology	<i>omr</i> <i>labevents</i> <i>microbiologyevents</i> <i>radiology</i>	<i>result_name</i> <i>item_name</i> <i>test_name</i> <i>exam_name</i>	F1 F1 F1 F1	11 698 165 961
	Diagnoses	Diagnose ICD Diagnose CCS Diagnosis ICD Diagnosis CCS	<i>diagnose_icd</i> <i>diagnose_icd</i> <i>diagnosis_icd</i> <i>diagnosis_icd</i>	<i>diagnoses</i> <i>CCT Type*</i> <i>icd_title</i> <i>CCS Type*</i>	F1 F1 F1 F1	24467 279 13171 271
Decision Making	Procedures	Procedures ICD Procedures CCS	<i>procedures_icd</i> <i>procedures_icd</i>	<i>procedures</i> <i>CCS Type*</i>	F1 F1	11098 230
	Services	Services POE	<i>services</i> <i>poe</i>	<i>curr_service</i> <i>order_type</i>	F1 F1	18 15
	Treatments	EMAR Prescriptions Prescriptions ATC Medrecon Medrecon ATC	<i>emr</i> <i>prescriptions</i> <i>prescriptions</i> <i>medrecon</i> <i>medrecon</i>	<i>medication</i> <i>drug</i> <i>ATC Type*</i> <i>name</i> <i>ATC Type*</i>	F1 F1 F1 F1 F1	4153 9233 913 18641 899
	ICU Event	Ingredientevents Datetimenevents Procedureevents Inputevents Outputevents	<i>ingredientevents</i> <i>datetimenevents</i> <i>procedureevents</i> <i>inputevents</i> <i>outputevents</i>	<i>item_name</i> <i>item_name</i> <i>item_name</i> <i>item_name</i> <i>item_name</i>	F1 F1 F1 F1 F1	15 137 138 222 63
	Transfer	ED Hospitalization ED ICU Tranfer 12hour	<i>edstays</i> <i>edstays</i>	-	AUROC	2 2
	Critical Outcomes	ED Critical Outcomes	<i>edstays</i>	-	AUROC	2
	Readmission	Readmission 30day Readmission 60day ICU Readmission ED Reattendance 3day	<i>discharge</i> <i>discharge</i> <i>icustays</i> <i>edstays</i>	-	AUROC	2 2 2 2
	LengthOfStay	LengthOfStay 3day LengthOfStay 7day ICU Stay 7day ICU Stay 14day	<i>admissions</i> <i>admissions</i> <i>icustays</i> <i>icustays</i>	-	AUROC	2 2 2 2
	Mortality	ED Inpatient Mortality Inpatient Mortality ICU Mortality 1day ICU Mortality 2day ICU Mortality 3day ICU Mortality 7day ICU Mortality 14day	<i>edstays</i> <i>admissions</i> <i>icustays</i> <i>icustays</i> <i>icustays</i> <i>icustays</i> <i>icustays</i>	-	AUROC	2 2 2 2 2 2 2

9.3 Prompt Collection

Prompt
<p>Prompt 1. Reasoning Chain Evaluation Instruction for Clinicians</p> <p>Task Description: Evaluation of Clinical Reasoning Chain Validity</p> <p>I. Background</p> <p>We have developed a large language model (LLM) designed to extract key information from electronic health records (EHRs) and perform clinical reasoning based on that information. For evaluation purposes, we have hidden the full EHR context, presenting only the critical information extracted by the model and the final diagnostic conclusion. Your task is to assess the relationship and logical consistency between the extracted information and the final conclusion.</p> <p>II. Data Description</p> <p>The model's response consists of three main sections: ##Extraction, ##Reasoning, and ##Final Result. You need to evaluate whether the content in ##Extraction and ##Reasoning adequately supports the ##Final Result and provide your final score in the Score column.</p> <p>III. Evaluation Objective</p> <p>Using the provided extracted information, you must determine if it is sufficient and accurate to support the model's final diagnostic conclusion. Specifically, you need to confirm:</p> <ul style="list-style-type: none">- Whether the key information extracted by the model is relevant to the diagnosis and accurately expressed.- Whether the final diagnostic conclusion is fully supported by the extracted information, without any missing essential evidence. <p>Note: The model's reasoning result is considered correct and does not need to be validated for correctness.</p> <p>IV. Evaluation Process</p> <p>Please read the entire reasoning chain, including the extracted information and the diagnostic conclusion. Based on your professional judgment, evaluate the degree of correspondence between the information and the conclusion. The scoring criteria are as follows (five levels):</p> <ul style="list-style-type: none">- 5 points (Very Satisfied): The extracted information is mostly accurate and largely supports the conclusion.- 4 points (Satisfied): The extracted information supports most of the conclusion. A small part of the conclusion lacks evidence but is reasonably justified.- 3 points (Average): The extracted information supports only a small part of the conclusion. The final result contains a large portion of content that is not adequately supported.- 2 points (Poor): The extracted information is related to the final result, but it fails to support the conclusion.- 1 point (Very Poor): The extracted information is highly irrelevant to the diagnostic conclusion. The conclusion lacks a reasonable basis, and the reasoning process is unreliable. <p>Scoring Differences Explained:</p> <ul style="list-style-type: none">- The main difference between 5 points and 4 points is whether the information extracted by the model fully supports the final result.- The difference between 5 points and 4 points is whether the final conclusion is fully supported by the extracted information.- The difference between 4 points and 3 points is whether a large part or a small part of the final conclusion is supported.- The difference between 3 points and 2 points is whether at least one conclusion is supported. If so, it's 3 points; if not, but the information is related, it's 2 points.- The difference between 2 points and 1 point is whether the extracted information is relevant and accurate enough to support the final conclusion.

2 points and 1 point is whether the extracted and reasoned information is relevant to the final conclusion.

V. Evaluation Examples

Example 1 (5 points, Lab Test)

1. Task Type:

Predict the lab tests the patient should undergo.

2. Model Reasoning Chain:

##Extraction:

Source: Outpatient Record [2178-09-23 11:20:00]

- Key Information: Patient complains of fatigue and pale skin/mucous membranes.

- Analysis: Typical signs of anemia, requiring clarification of anemia type.

Source: Lab Results [2178-09-22 15:45:00]

- Key Information: Hemoglobin 8.2 g/dL (normal 13-17), Creatinine 2.1 mg/dL (normal 0.7-1.3).

- Analysis: Moderate to severe anemia combined with abnormal kidney function suggests possible renal anemia.

Source: Past Medical History [2177-05-12 09:00:00]

- Key Information: History of chronic kidney disease (CKD stage 3), not on dialysis.

- Analysis: Chronic kidney disease often leads to renal anemia. Erythropoietin levels need to be assessed.

Source: Past Surgical History [2176-08-05 10:00:00]

- Key Information: Appendectomy, good recovery.

- Analysis: Not directly related to the current anemia evaluation. This is irrelevant information.

##Reasoning:

Combining "anemia symptoms + significantly low hemoglobin + abnormal kidney function + chronic kidney disease," the following lab tests are recommended to clarify the anemia mechanism:

Serum iron

Ferritin

Transferrin saturation

Erythropoietin levels

##Final Result:

Recommended Tests: Serum iron, Ferritin, Transferrin saturation, Erythropoietin levels.

3. Score Explanation:

Score: 5 points

Reasoning:

The first three pieces of extracted information ("fatigue + pale skin," "Hemoglobin 8.2 + Creatinine 2.1," "chronic kidney disease history") fully support the necessity of performing tests for iron metabolism and erythropoietin levels.

The "past appendectomy" is extraneous information unrelated to the conclusion, but it does not affect the correctness of the conclusion.

All key points of the conclusion are fully justified by the first three essential pieces of information, thus a score of 5 points.

Example 2 (4 points, Medication Prescription)

1. Task Type:

Predict the patient's medication plan.

2. Model Reasoning Chain:

##Extraction:

Source: Inpatient Record [2180-02-15 08:10:00]

- Key Information: Type 2 diabetes mellitus, fasting blood glucose 12.5 mmol/L, postprandial blood glucose 16.0 mmol/L.

- Analysis: Blood glucose is significantly high; the current oral medication regimen is insufficient.

Source: Lab Results [2180-02-14 07:50:00]

- Key Information: HbA1c 8.5%, Creatinine 60 µmol/L (normal).

- Analysis: High HbA1c indicates the need for improved blood glucose control. Normal kidney function means GLP-1 agonists are a viable option.

Source: Past Medication History [2179-10-10 09:00:00]

- Key Information: Taking oral Metformin 1000 mg twice daily and Gliclazide 80 mg once daily.

- Analysis: The current dual-drug therapy has failed to achieve the target, requiring an upgrade.

##Reasoning:

Poor glycemic control (fasting 12.5, postprandial 16.0, HbA1c 8.5%) → Needs escalation of antidiabetic treatment.

Normal kidney function → GLP-1 agonists are a safe option.

Has a history of hypertension and is taking Amlodipine → Needs to continue antihypertensive therapy.

##Final Result:

Prescription: Add a GLP-1 receptor agonist, continue Amlodipine treatment.

3. Score Explanation:

Score: 4 points

Reasoning:

Support for adding GLP-1: The information "fasting 12.5 mmol/L, HbA1c 8.5%" clearly suggests that the current oral medications are insufficient, and an escalation to a GLP-1 is reasonable. Additionally, "normal kidney function" proves this drug can be used safely. This part of the conclusion aligns perfectly with the extracted information.

Support for continuing Amlodipine: The extracted information "history of hypertension and is taking Amlodipine" indicates the patient is already on antihypertensive treatment. This sub-conclusion is also directly supported by the extracted information.

Overall: In this case, both parts of the conclusion—"add GLP-1" and "continue Amlodipine"—are supported by the corresponding information. The model's mention of "continuing Amlodipine" is a reasonable inference (most patients with T2DM have concurrent hypertension) but is not directly proven by the extracted information.

Therefore, the majority of the conclusion (the GLP-1 part) is fully supported by the information, while a small part of the conclusion (continuing Amlodipine) is a reasonable but unsupported inference. This fits the criteria for a "4-point" scenario.

Example 3 (3 points, Diagnosis)

1. Task Type:

Predict the patient's diagnosis.

2. Model Reasoning Chain:

##Extraction:

Source: Outpatient Record [2179-05-10 10:30:00]

- Key Information: Patient complains of sore throat, cough, fever, and fatigue for 3 days.

- Analysis: Suggests symptoms of an upper respiratory tract infection (URTI).

Source: Physical Examination [2179-05-10 10:45:00]

- Key Information: Temperature 38.5°C, pharyngeal congestion, bilateral tonsillar enlargement (Grade II).

- Analysis: Supports acute tonsillitis or acute pharyngitis.

Source: Lab Results [2179-05-10 11:00:00]

- Key Information: White blood cell count $12.5 \times 10^9/L$ (normal 4-10), Neutrophils 80% (normal 50-75%).
- Analysis: Both white blood cells and neutrophils are elevated, suggesting a bacterial infection.

##Reasoning:

Sore throat, fever, cough, temperature 38.5°C , pharyngeal congestion, tonsillar enlargement → Suggests URTI.

Elevated white blood cells and neutrophils → Suggests bacterial infection.

Combining the above information, the diagnosis is "acute upper respiratory tract infection with bacterial infection."

Given the symptoms and signs, acute bronchitis is also a possibility.

##Final Result:

Diagnosis: Acute upper respiratory tract infection with bacterial infection, acute bronchitis.

3. Score Explanation:

Score: 3 points

Reasoning:

The extracted information (sore throat, fever, elevated temperature, pharyngeal congestion, elevated white blood cells and neutrophils) fully supports the diagnosis of "acute upper respiratory tract infection with bacterial infection." This part of the conclusion is highly consistent with the information and is correct.

The diagnostic conclusion also includes "acute bronchitis," but the extracted information does not contain direct evidence to support this diagnosis. For example, there is no mention of worsening cough, productive cough, or abnormal findings on lung auscultation (such as rales or rhonchi).

While the majority of the conclusion ("acute upper respiratory tract infection with bacterial infection") is supported by the information, the diagnosis of "acute bronchitis" lacks a basis. According to the scoring criteria, "The extracted information only supports a small part of the conclusion, and the final result contains a large portion of content that is not adequately supported," therefore, the score is 3 points.

Example 4 (2 points, Medication Prescription)

1. Task Type:

Predict the patient's medication plan.

II. Model Reasoning Chain:

##Extraction:

Source: Inpatient Record [2180-03-05 09:00:00]

- Key Information: Patient admitted due to palpitations and chest tightness. ECG shows atrial fibrillation.

- Analysis: The patient has a cardiac arrhythmia.

Source: Past Medical History [2178-01-15 14:00:00]

- Key Information: History of hypertension, long-term use of Irbesartan.

- Analysis: Blood pressure control needs to be continued.

Source: Lab Results [2180-03-05 10:30:00]

- Key Information: Normal kidney function, normal liver function.

- Analysis: Facilitates drug selection.

##Reasoning:

The patient has atrial fibrillation and needs heart rate control.

Long-term hypertension requires continued blood pressure management.

Considering all factors, Metoprolol is recommended.

##Final Result:
Prescription: Metoprolol.

3. Score Explanation:

Score: 2 points

Reasoning:

The extracted information—"palpitations, chest tightness, atrial fibrillation"—does indicate a cardiac arrhythmia, and Metoprolol is a commonly used first-line drug for treating atrial fibrillation. From a pharmacological perspective, this conclusion is related to the extracted information.

However, the extracted information does not provide sufficient evidence to support Metoprolol as the sole or preferred drug. The treatment plan for atrial fibrillation requires a comprehensive consideration of the patient's heart rate control needs (type and rate of A-fib), stroke risk (CHA₂DS₂-VASc score), presence of comorbidities (like heart failure), and blood pressure levels. Based solely on "atrial fibrillation" and "hypertension," one cannot directly conclude that Metoprolol should be prescribed, as there are various other drug options (e.g., Diltiazem, Verapamil).

The conclusion "Metoprolol" lacks sufficient evidence to be fully supported. It might be a plausible treatment, but it is not the only or best choice that the extracted information can definitively determine. This aligns with the scoring criterion: "The extracted information is related to the final result, but it fails to support the conclusion."

Example 5 (1 point, Lab Test)

1. Task Type:

Predict the lab tests the patient should undergo.

2. Model Reasoning Chain:

##Extraction:

Source: Outpatient Record [2179-11-20 15:00:00]

- Key Information: Patient complains of abdominal pain and diarrhea.
- Analysis: Suggests a gastrointestinal issue.

Source: Physical Examination [2179-11-20 15:20:00]

- Key Information: Abdominal palpation shows no significant tenderness, bowel sounds are active.
- Analysis: Suggests no organic lesion or inflammation in the abdomen.

Source: Past Medication History [2179-11-19 18:00:00]

- Key Information: Recently took Aspirin.
- Analysis: No direct correlation with diarrhea.

##Reasoning:

Patient with abdominal pain and diarrhea → Gastrointestinal symptoms.

No tenderness on abdominal palpation → Rules out conditions like acute peritonitis.

Patient took Aspirin → May affect the gastrointestinal tract.

Based on the above information, an autoimmune disease is suspected.

##Final Result:

Recommended Tests: Antinuclear antibody (ANA) profile, Complement, Rheumatoid factor.

3. Score Explanation:

Score: 1 point

Reasoning:

The extracted information is "abdominal pain, diarrhea," "no tenderness on abdominal palpation," and "took Aspirin." This information points to common gastrointestinal symptoms and potential drug-related GI discomfort.

However, the final conclusion, "Recommended Tests: Antinuclear antibody profile, Complement,

"Rheumatoid factor," is used to diagnose autoimmune diseases (e.g., Systemic Lupus Erythematosus, Rheumatoid Arthritis).

There are no symptoms or signs in the extracted information (e.g., joint pain, rash, oral ulcers) that are related to autoimmune diseases. Linking "abdominal pain and diarrhea" to tests for autoimmune diseases lacks any reasonable basis.

The conclusion is severely irrelevant to the extracted information, and the reasoning process is highly unreliable. Therefore, the score is 1 point.

Prompt

Prompt 2. Reasoning Chain Synthesis Prompt for GPT-4o

Patient EHR Context #
{context}

Retrieved Medical Knowledge #
{medical_knowledge}

Ground Truth #
{ground_truth}

Task #
{task}

Data Description

- # Patient EHR Context #: Contains all medical events and content from the patient's hospitalization journey.
- # Retrieved Medical Knowledge #: Contains Patient EHR Context elements and their relationships potentially relevant to the Ground Truth entity.
- # Ground Truth #: Contains the correct answers for this task; your predictions must exactly match these.
- # Task #: Contains the specific task description; you need to complete the task based on the information in # Patient EHR Context #.

Instructions

Please provide a logically rigorous medical reasoning process so that the # Ground Truth # can be derived from the content in # Patient EHR Context # and # Task #.

Requirements

The reasoning process should include three stages: Extraction, Reasoning, and Final Results.

Extraction

- In this stage, extract and identify each piece of "key information" from the # Patient EHR Context # according to the provided # Retrieved Medical Knowledge #.
- Don't pay attention to information that you think is not helpful for the reasoning.
- Each step in the Extraction stage should follow the format below, **You need to specify the event name and time for each extracted information**:

Event Name [Event Time]: list the information extracted from the event and analyze the potential relationship between the key information and the ground truth.

Reasoning

- Analyze the relationship between the context information and the item in # Ground Truth # in a very specific and professional manner, providing detailed reasoning steps.
- Your analysis should include the item in # Ground Truth # as much as possible. Items that cannot be inferred from the context can be omitted.
- Do not use the word "maybe", "possible" or "though" in the generated reasoning. You should do your best to find all the supporting information you can to ensure the correctness of your reasoning.
- The reasoning process should be concise and rigorous, and each step should explain the specific medical knowledge involved, making the reasoning process more credible.
- All reasoning must be based on the context information to infer the items in the ground truth and no reverse inference can be performed.

Final Results - Provide the final result for the task. **Note that the final result should only contain the items contained in # Gound Truth # that have been correctly inferred in.

Reasoning stage.**

- Each item in the ## Final Results should be contained in the # Gound Truth # with extactly same string.

Important Notes!!!

- **For each piece of # Retrieved Medical Knowledge # that is relevant to completing the # Task #, locate the exact position of its first item within the # Patient EHR Context # and explicitly annotate it during the Extraction phase to ensure a more thorough analysis.**
- **During the ## Reasoning stage, remember to analyze very carefully how each item in # Ground Truth # is inferred**
- **Most importantly, integrate references to # Ground Truth # and # Retrieved Medical Knowledge # in an implicit manner. At any point in the reasoning process, do not use phrases such as “according to the medical knowledge above”, “as shown in ground truth” or any wording that reveals you are aware of the underlying medical knowledge or the ground truth.**

Output Format

Extraction

[YOUR OUTPUT]

Reasoning

[YOUR OUTPUT]

Final Results

[YOUR OUTPUT]

Tasks	Instruction
Admissions	Given the sequence of events that have occurred in a hospital, please give the next Admissions suggestion for the patient.
OMR	Given the sequence of events that have occurred in a hospital, please give the next Online Medical Record suggestion for the patient.
EMAR	Given the sequence of events that have occurred in a hospital, please give the next Electronic Medicine Administration Record suggestion for the patient.
Procedures ICD	Given the sequence of events that have occurred in a hospital, please give the next Procedures International Classification of Diseases Item suggestion for the patient.
Procedures CCS	Given the sequence of events that have occurred in a hospital, please give the next Procedures Clinical Classifications Software Item suggestion for the patient.
Diagnoses ICD	Given the sequence of events that have occurred in a hospital, please give the next Diagnoses International Classification of Diseases Item suggestion for the patient.
Diagnoses CCS	Given the sequence of events that have occurred in a hospital, please give the next Diagnoses Clinical Classifications Software Item suggestion for the patient.
Labevents	Given the sequence of events that have occurred in a hospital, please give the next Laboratory Test Events suggestion for the patient.
Microbiologyevents	Given the sequence of events that have occurred in a hospital, please give the next Microbiology Test Events suggestion for the patient.
Services	Given the sequence of events that have occurred in a hospital, please give the next Services suggestion for the patient.
Transfers	Given the sequence of events that have occurred in a hospital, please give the next Transfers suggestion for the patient.
POE	Given the sequence of events that have occurred in a hospital, please give the next Provider Order Entry suggestion for the patient.
Radiology	Given the sequence of events that have occurred in a hospital, please give the next Radiology Examinations suggestion for the patient.
Prescriptions	Given the sequence of events that have occurred in a hospital, please give the next Prescriptions suggestion for the patient.
Prescriptions ATC	Given the sequence of events that have occurred in a hospital, please give the next Anatomical Therapeutic Chemical Classification Prescriptions suggestion for the patient.
Medrecon	Given the sequence of events that have occurred in a hospital, please give the next ED Medrecon suggestion for the patient.
Medrecon ATC	Given the sequence of events that have occurred in a hospital, please give the next ED Medrecon on Anatomical Therapeutic Chemical (ATC) Classification suggestion for the patient.
Diagnosis	Given the sequence of events that have occurred in a hospital, please give the next ED Diagnoses on International Classification of Diseases suggestion for the patient.
Diagnosis CCS	Given the sequence of events that have occurred in a hospital, please give the next ED Diagnoses on Clinical Classifications Software Item suggestion for the patient.
Ingredientevents	Given the sequence of events that have occurred in a hospital, please give the next Ingredient Events suggestion for the patient.
Datetimeevents	Given the sequence of events that have occurred in a hospital, please give the next Datetime Events suggestion for the patient.
Procedureevents	Given the sequence of events that have occurred in a hospital, please give the next Procedure Events suggestion for the patient.
Inputevents	Given the sequence of events that have occurred in a hospital, please give the next Input Events suggestion for the patient.
Outputevents	Given the sequence of events that have occurred in a hospital, please give the next Output Events suggestion for the patient.
ED Hospitalization	Given the sequence of events that have occurred in a hospital, please predict whether the patient will be hospitalized after the emergency room visit.
ED Inpatient Mortality	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die during hospitalization.
ED ICU Tranfer 12Hour	Given the sequence of events that have occurred in a hospital, please predict whether the patient will be transferred to the ICU within 12 hours after the emergency room.
ED Reattendance 3DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will return to the emergency department within 72 hours after the emergency visit.
ED Critical Outcomes	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die during hospitalization or will be transferred to the ICU within 12 hours after the emergency room.
Readmission 30Day	Given the sequence of events that have occurred in a hospital, please predict whether the patient will be readmitted to the hospital within 30 days
Readmission 60Day	Given the sequence of events that have occurred in a hospital, please predict whether the patient will be readmitted to the hospital within 60 days
Inpatient Mortality	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die during hospitalization.
Lengthofstay 3DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient's hospital stay will exceed 3 days.
Lengthofstay 7DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient's hospital stay will exceed 7 days
ICU Mortality 1DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die within 1 day.
ICU Mortality 2DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die within 2 day.
ICU Mortality 3DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die within 3 day.
ICU Mortality 7DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die within 7 day.
ICU Mortality 14Day	Given the sequence of events that have occurred in a hospital, please predict whether the patient will die within 14 day.
ICU STAY 7DAY	Given the sequence of events that have occurred in a hospital, please predict whether the patient will stay in the ICU for more than 7 days.
ICU STAY 14Day	Given the sequence of events that have occurred in a hospital, please predict whether the patient will stay in the ICU for more than 14 days.
ICU Readmission	Given the sequence of events that have occurred in a hospital, please predict whether the patient will be admitted to the ICU again during this hospitalization

Supplementary Table 4 | Instruction for each task in EHR-Bench.

Dataset	Tasks	Instruction
MIMIC-IV-CDM	Main Disease Diagnoses	Given the sequence of events that have occurred in a hospital, please give the main Diagnoses of Diseases Item suggestion for the patients.
	ICD Code Diagnoses	Given the sequence of events that have occurred in a hospital, please give the next Diagnoses in International Classification of Diseases Item suggestion for the patients.
EHRSHOT	Length of Stay	Given the sequence of events that have occurred in a hospital, please predict whether a patient's total length of stay during a visit to the hospital will be at least 7 days.
	Readmission	Given the sequence of events that have occurred in a hospital, please predict whether a patient will be re-admitted to the hospital within 30 days after being discharged from a visit.
	ICU Transfer	Given the sequence of events that have occurred in a hospital, please predict whether a patient will be transferred to the ICU during a visit to the hospital.
	Anemia	Given the sequence of events that have occurred in a hospital, please predict whether an anemia lab comes back as normal (≥ 120 g/L).
	Hyperkalemia	Given the sequence of events that have occurred in a hospital, please predict whether a hyperkalemia lab comes back as normal (≤ 5.5 mmol/L).
	Hyponatremia	Given the sequence of events that have occurred in a hospital, please predict whether a hyponatremia lab comes back as normal (≥ 135 mmol/L).
	Hypoglycemia	Given the sequence of events that have occurred in a hospital, please predict whether a hypoglycemia lab comes back as normal (≥ 3.9 mmol/L).
	Thrombocytopenia	Given the sequence of events that have occurred in a hospital, please predict whether a thrombocytopenia lab comes back as normal (≥ 150 109/L).
	Acute MI	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of an acute myocardial infarction within the next year.
	Celiac	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of celiac disease within the next year.
	Hyperlipidemia	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of hyperlipidemia within the next year.
	Hypertension	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of essential hypertension within the next year.
	Lupus	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of lupus within the next year.
	Pancan	Given the sequence of events that have occurred in a hospital, please predict whether the patient will have his first diagnosis of pancreatic cancer within the next year.

Supplementary Table 5 | Instruction for each task in MIMIC-IV-CDM and EHRSHOT.

9.4 Case Collection

Case
Case 1. Markdown Format for Event with Single Item ## Event Name [Event Time] - Info Key 1: Info Value 1 - Info Key 2: Info Value 2 ... - Info Key N: Info Value N
Case 2. Markdown Format for Event with Multiple Item ## Event Name [Event Time] Info Key 1 Info Key 2 ... Info Key N ----- ----- --- ----- Info Value Info Value ... Info Value Info Value Info Value ... Info Value ... Info Value Info Value ... Info Value
Case 3. Free-text Input Example of EHRSHOT ## Person [1991-04-12 00:00:00] - Birth - White - Not Hispanic or Latino - MALE

```

## Measurement [2015-02-24 02:25:00]
| Item_Name | Valuenum | Valueuom | Ref_Range_Lower | Ref_Range_Upper | Flag |
| — | — | — | — | — | — |
| Lactate [Moles/volume] in Blood | 0.6000000238418579 | mmol/L | nan | nan | nan |
| Lactate [Mass/volume] in Blood | 0.6000000238418579 | mmol/L | nan | nan | nan |
| Glasgow coma scale | 15.0 | nan | nan | nan | nan |
| Mean blood pressure | 103.0 | mmHg | nan | nan | nan |
| Pain severity [Score] Visual analog score | 6.0 | nan | nan | nan | nan |
| Respiratory rate | 18.0 | breaths/min | 12 | 18 | normal |
| Systolic blood pressure | 152.0 | mmHg | 90 | 140 | abnormal |
| Body temperature | 98.4000015258789 | F | 95 | 100.4 | normal |
| Diastolic blood pressure | 79.0 | mmHg | 60 | 90 | normal |
| Heart rate | 97.0 | bpm | 60 | 100 | normal |

```

Observation [2015-02-24 02:25:00]

```

| Item_Name | Valuenum | Valueuom | Ref_Range_Lower | Ref_Range_Upper | Flag |
| — | — | — | — | — | — |
| Body temperature measurement site | 1.0 | nan | nan | nan | nan |
| Oxygen saturation | 100.0 | % | 95 | 100 | normal |

```

Drug_Exposure [2015-02-24 04:22:00]

- oxycodone hydrochloride 1 MG/ML Oral Solution
- ibuprofen 400 MG Oral Tablet
- gabapentin 300 MG Oral Capsule
- levetiracetam 500 MG Oral Tablet
- 1 ML hydromorphone hydrochloride 2 MG/ML Prefilled Syringe
- acetaminophen 325 MG Oral Tablet
- 0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe
- zolpidem tartrate 5 MG Oral Tablet
- triamcinolone acetonide 1 MG/ML Topical Cream
- oxycodone hydrochloride 5 MG Oral Tablet

Procedure_Occurrence [2015-02-24 23:59:00]

- Routine venipuncture
- Taking patient vital signs assessment
- Ambulating patient

Condition_Occurrence [2015-02-25 11:07:00]

- Tobacco dependence syndrome
- Benign neoplasm of colon
- Inflammatory dermatosis
- Epilepsy

Case

Case 4. Free-text Input Example of MIMIC-IV-CDM

Patient Demographics

- Patient History: ____ s/p emergency tissue AVR and type A aortic dissection repair ____ and sternal washout ____, c/b tamponade requiring clot evacuation ____, trach ____, PEG _____. Overall, other complications included seizures, renal failure requiring CRRT, shock liver, HITT, sepsis, afib, left colonic ischemia. Discharged at end of ____ and returned to ____ in ____, at which time he underwent a bronch. Has been at rehab. Was on tube feeds until several weeks ago and now only on

regular diet. He has been eating, but decreased appetite and intake due to post-prandial epigastric abdominal pain. This has been occurring for approximately three weeks, and he initially attributed it to indigestion. He had routine follow up in thoracic clinic with Dr _____, at which time his PEG tube was removed. Then he was then sent to the ED for his abdominal pain. PEG site was clean prior to removal.

- Past Medical History: Hypertension Appendectomy Right facial cyst drainage
- Social History: _____ Family History: None
- Physical Examination: Physical exam VS: 98.6, 80, 135/72, 18, 96% RA Gen: NAD CV: RRR Pulm: CTA b/l Abd: soft, nondistended. slightly tender in epigastric area and right upper quadrant without rebound/guarding/rigidity Ext: left hand with ischemic digits, b/l _____ wrapped with gauze PE

Labotary Test Events

Item_name	Valuenum	Value uom	Ref_range_lower	Ref_range_upper
Alanine Aminotransferase (ALT)	18.0	IU/L	0.0	40.0
PTT	32.8	sec	25.0	36.5
PT	14.9	sec	9.4	12.5
INR(PT)	1.4	None	0.9	1.1
White Blood Cells	25.8	K/uL	4.0	11.0
Red Blood Cells	3.84	m/uL	4.6	6.2
RDW	14.1	%	10.5	15.5
Platelet Count	177.0	K/uL	150.0	440.0
Neutrophils	91.5	%	50.0	70.0
Monocytes	3.6	%	2.0	11.0
MCV	92.0	fL	82.0	98.0
MCHC	33.0	%	31.0	35.0
Lymphocytes	4.8	%	18.0	42.0
Hemoglobin	11.6	g/dL	14.0	18.0
Hematocrit	35.2	%	40.0	52.0
Eosinophils	0.1	%	0.0	4.0
MCH	30.2	pg	27.0	32.0
Light Green Top Hold	HOLD.	None	None	None
Albumin	3.3	g/dL	3.5	5.2
Alkaline Phosphatase	117.0	IU/L	40.0	130.0
Anion Gap	15.0	mEq/L	8.0	20.0
Basophils	0.1	%	0.0	2.0
Bicarbonate	28.0	mEq/L	22.0	32.0
Bilirubin, Total	1.2	mg/dL	0.0	1.5
Chloride	94.0	mEq/L	96.0	108.0
Aspartate Aminotransferase (AST)	20.0	IU/L	0.0	40.0
Estimated GFR (MDRD equation)	Using this patient's age, gender, and serum creatinine value of 1.2., Estimated GFR = 61 if non African-American (mL/min/1.73 m ²). Estimated GFR = 74 if African-American (mL/min/1.73 m ²). For comparison, mean GFR for age group 60-69 is 85 (mL/min/1.73 m ²). GFR<60 = Chronic Kidney Disease, GFR<15 = Kidney Failure.	None	None	None
Glucose	141.0	mg/dL	70.0	100.0
Lipase	15.0	IU/L	0.0	60.0
Creatinine	1.2	mg/dL	0.5	1.2
Potassium	4.6	mEq/L	3.3	5.1
Sodium	132.0	mEq/L	133.0	145.0
Urea Nitrogen	27.0	mg/dL	6.0	20.0

Specimen Type VEN. None None None
Lactate 1.3 mmol/L 0.5 2.0
Protein 30.0 mg/dL None None
Yeast NONE None None None
WBC <1. None 0.0 5.0
Urobilinogen NEG. None 0.2 1.0
Urine Color Yellow. None None None
Specific Gravity >1.050*. None 1.001 1.035
RBC 3.0 #/hpf 0.0 2.0
pH 6.0 units 5.0 8.0
Urine Appearance Clear. None None None
Leukocytes NEG. None None None
Ketone NEG. None None None
Glucose NEG. None None None
Epithelial Cells 0.0 #/hpf None None
Blood TR. None None None
Bilirubin NEG. None None None
Bacteria NONE. None None None
Nitrite NEG. None None None
Calcium, Total 9.2 mg/dL 8.4 10.3
Magnesium 1.8 mg/dL 1.6 2.6
Phosphate 3.4 mg/dL 2.7 4.5

Microbiology Test Events

Item_name Valuestr
—— ——
Blood Culture, Routine NO GROWTH.
ANAEROBIC CULTURE NO ANAEROBES ISOLATED.
FLUID CULTURE PSEUDOMONAS AERUGINOSA, ENTEROCOCCUS SP.

Radiology Examinations

Exam_name Text
—— ——
LIVER OR GALLBLADDER US (SINGLE ORGAN) EXAMINATION: LIVER OR GALLBLADDER US (SINGLE ORGAN):

TECHNIQUE: Grey scale and color Doppler ultrasound images of the abdomen were obtained.

FINDINGS:

LIVER: The hepatic parenchyma appears within normal limits. The contour of the liver is smooth. Again seen is a hyperechoic lesion measuring approximately 6 mm in segment V of the liver, unchanged since prior study and likely represents a hemangioma. The additional hemangioma seen on previous study is not clearly identified on today's exam. Main portal vein is patent with hepatopetal flow. There is no ascites.

BILE DUCTS: There is no intrahepatic biliary dilation. The CBD measures 6 mm.

GALLBLADDER: The gallbladder is distended and filled with sludge. There is mild gallbladder wall edema measuring up to 6 mm. There is no pericholecystic fluid.

PANCREAS: The pancreas is not well seen secondary to overlying bowel gas.

KIDNEYS: The right kidney measures 8.1 cm. Survey views of the right kidney do not demonstrate any masses, hydronephrosis, or stones.

RETROPERITONEUM: Visualized portions of aorta and IVC are within normal limits. |

Case

Case 5. Free-text Input Example of EHR-Bench

Patient Demographics [None]

- Anchor_Age: 88
- Gender: F

Admissions [2127-04-18 16:53:00]

- Admission_Type: EW EMER.
- Admission_Location: PROCEDURE SITE
- Admission_Info: None

Provider Order Entry [2127-04-18 15:21:26]

Order_Type Order_Subtype
—— ——
Medications nan
ADT orders Admit
General Care Vitals/Monitoring
Nutrition Diet Order
General Care Other
Medications nan

Services [2127-04-18 16:54:06]

- Curr_Service: CMED

Transfers [2127-04-18 16:54:06]

- Eventtype: admit
- Careunit: Medicine/Cardiology

Pharmacy [2127-04-18 16:57:15]

Medication Proc_Type Status
—— —— ——
Potassium Chloride Unit Dose Discontinued via patient discharge
Potassium Chloride Unit Dose Discontinued via patient discharge
Oxybutynin Unit Dose Discontinued via patient discharge
Zolpidem Tartrate Unit Dose Discontinued via patient discharge
Nitroglycerin SL Unit Dose Discontinued
Sodium Chloride 0.9 Hydrochlorothiazide Unit Dose Discontinued via patient discharge
Potassium Chloride Unit Dose Discontinued via patient discharge
Lisinopril Unit Dose Discontinued via patient discharge
Multivitamins Unit Dose Discontinued via patient discharge
Acetaminophen Unit Dose Discontinued via patient discharge
Ferrous Sulfate Unit Dose Discontinued via patient discharge
Aluminum-Magnesium Hydrox.-Simethicone Unit Dose Discontinued via patient discharge
Pneumococcal Vac Polyvalent Unit Dose Discontinued
Simvastatin Unit Dose Discontinued via patient discharge
Aspirin Unit Dose Discontinued via patient discharge
Diltiazem Extended-Release Unit Dose Discontinued via patient discharge

Prescriptions [2127-04-18 17:00:00]

Drug Atc_Type Prod_Strength Dose_Val_Rx Dose_Unit_Rx
—— —— —— —— ——
Potassium Chloride potassium chloride 20mEq Packet 20 mEq

```
| Potassium Chloride | potassium chloride | 20mEq Packet | 40 | mEq | |
| Zolpidem Tartrate | zolpidem | 5mg Tablet | 5 | mg |
| Nitroglycerin SL | glyceryl trinitrate | 0.3mg SL Tablet Bottle | 0.3 | mg |
| Sodium Chloride 0.9| Potassium Chloride | potassium chloride | 20mEq Packet | 60 | mEq |
| Acetaminophen | paracetamol | 325mg Tablet | 650 | mg |
| Aluminum-Magnesium Hydrox.-Simethicone | aluminium hydroxide | 30 mL UDCup | 30 | mL |
| Pneumococcal Vac Polyvalent | nan | 25mcg/0.5mL Vial | 0.5 | mL |
```

Provider Order Entry [2127-04-18 17:13:19]

```
| Order_Type | Order_Subtype |
| —— | —— |
| Medications | nan |
| Medications | nan |
| Medications | nan |
```

Pharmacy [2127-04-18 17:26:01]

```
| Medication | Proc_Type | Status |
| —— | —— | —— |
| Fentanyl Citrate | Unit Dose | Expired |
| Ibuprofen | Unit Dose | Expired |
```

Provider Order Entry [2127-04-18 17:55:22]

```
| Order_Type | Order_Subtype |
| —— | —— |
| Radiology | Ultrasound |
| General Care | Other |
| Blood Bank | Blood tests |
| Lab | nan |
| Lab | nan |
| Radiology | General Xray |
| Cardiology | ECG |
| General Care | Other |
| General Care | Other |
```

Prescriptions [2127-04-18 18:00:00]

```
| Drug | Atc Type | Prod_Strength | Dose_Val_Rx | Dose_Unit_Rx |
| —— | —— | —— | —— | —— |
| Fentanyl Citrate | fentanyl | 100mcg/2mL Amp | 25 | mcg |
| Ibuprofen | ibuprofen | 600mg Tablet | 600 | mg |
```

Provider Order Entry [2127-04-18 18:12:48]

```
| Order_Type | Order_Subtype |
| —— | —— |
| General Care | Vitals/Monitoring |
| General Care | Other |
| General Care | Activity |
```

```

| General Care | Other |
| General Care | Other |
| IV therapy | IV fluids |
| Medications | nan |
| General Care | Other |

## Pharmacy [2127-04-18 18:44:04]
| Medication | Proc_Type | Status |
| — | — | — |
| nan | IV Large Volume | Expired |
| Atropine Sulfate | Unit Dose | Discontinued via patient discharge |

## Prescriptions [2127-04-18 19:00:00]
| Drug | Atc Type | Prod_Strength | Dose_Val_Rx | Dose_Unit_Rx | |
| — | — | — | — | — |
| 5| Sodium Bicarbonate | sodium bicarbonate | 50mEq Vial | 150 | mEq |
| Atropine Sulfate | atropine | 1mg/10mL Syinge | 0.5 | mg |

## Provider Order Entry [2127-04-18 19:11:50]
| Order_Type | Order_Subtype |
| — | — |
| Lab | nan |
| Medications | nan |

## Pharmacy [2127-04-18 19:14:10]
- Medication: Acetylcysteine 20- Proc_Type: Unit Dose
- Status: Discontinued via patient discharge

## Labotary Test Events [2127-04-18 19:19:00]
| Item_Name | Valuenum | Valueuom | Ref_Range_Lower | Ref_Range_Upper | Flag | Comments |
| — | — | — | — | — | — |
| INR(PT) | 1.2 | nan | 0.9 | 1.1 | abnormal | nan |
| PT | 13.7 | sec | 10.4 | 13.4 | abnormal | nan |
| PTT | 25.6 | sec | 22.0 | 35.0 | nan | nan |
| Alanine Aminotransferase (ALT) | 9.0 | IU/L | 0.0 | 40.0 | nan | nan |
| Albumin | 3.8 | g/dL | 3.4 | 4.8 | nan | nan |
| Alkaline Phosphatase | 61.0 | IU/L | 39.0 | 117.0 | nan | nan |
| Anion Gap | 13.0 | mEq/L | 8.0 | 20.0 | nan | nan |
| Aspartate Aminotransferase (AST) | 13.0 | IU/L | 0.0 | 40.0 | nan | nan |
| Bicarbonate | 27.0 | mEq/L | 22.0 | 32.0 | nan | nan |
| Bilirubin, Total | 0.2 | mg/dL | 0.0 | 1.5 | nan | nan |
| Chloride | 99.0 | mEq/L | 96.0 | 108.0 | nan | nan |
| Creatinine | 1.3 | mg/dL | 0.4 | 1.1 | abnormal | nan |
| Estimated GFR (MDRD equation) | nan | nan | nan | nan | Using this patient's age, gender, and serum creatinine value of 1.3.. Estimated GFR = 39 if non African-American (mL/min/1.73 m2). Estimated GFR = 47 if African-American (mL/min/1.73 m2). For comparison, mean GFR for age group 70+ is 75 (mL/min/1.73 m2). GFR<60 = Chronic Kidney Disease, GFR<15 = Kidney Failure.
|
| Glucose | 176.0 | mg/dL | 70.0 | 105.0 | abnormal | nan |
| Lactate Dehydrogenase (LD) | 134.0 | IU/L | 94.0 | 250.0 | nan | nan |
| Potassium | 3.8 | mEq/L | 3.3 | 5.1 | nan | nan |
| Sodium | 135.0 | mEq/L | 133.0 | 145.0 | nan | nan |

```

Urea Nitrogen 28.0 mg/dL 6.0 20.0 abnormal nan
Hematocrit 28.5 Hemoglobin 9.6 g/dL 12.0 16.0 abnormal nan
MCH 29.2 pg 27.0 32.0 nan nan
MCHC 33.7 MCV 87.0 fL 82.0 98.0 nan nan
Platelet Count 260.0 K/uL 150.0 440.0 nan nan
RDW 14.7 Red Blood Cells 3.29 m/uL 4.2 5.4 abnormal nan
White Blood Cells 4.6 K/uL 4.0 11.0 nan nan

Prescriptions [2127-04-18 20:00:00]

- Drug: Acetylcysteine 20- Atc Type: acetylcysteine
- Prod_Strength: 800mg/4ml Vial
- Dose_Val_Rx: 600
- Dose_Unit_Rx: mg

Labotary Test Events [2127-04-18 21:20:00]

Item_Name	Valueenum	Valueuom	Ref_Range_Lower	Ref_Range_Upper	Flag	Comments
— — — — — — —						
Bacteria nan nan nan nan nan NONE.						
Bilirubin nan mg/dL nan nan nan NEG.						
Blood nan nan nan nan nan MOD.						
Epithelial Cells 1.0 #/hpf nan nan nan nan						
Glucose nan mg/dL nan nan nan NEG.						
Ketone nan mg/dL nan nan nan NEG.						
Leukocytes nan nan nan nan nan NEG.						
Nitrite nan nan nan nan nan NEG.						
pH 7.5 units 5.0 8.0 nan nan						
Protein nan mg/dL nan nan nan NEG.						
RBC nan #/hpf 0.0 2.0 nan <1.						
Specific Gravity 1.015 1.001 1.035 nan nan						
Transitional Epithelial Cells nan #/hpf nan nan nan <1.						
Urine Appearance nan nan nan nan nan Hazy.						
Urine Color nan nan nan nan nan Straw.						
Urobilinogen nan mg/dL 0.2 1.0 nan NEG.						
WBC 3.0 #/hpf 0.0 5.0 nan nan						
Yeast nan nan nan nan nan nan						

Provider Order Entry [2127-04-19 02:55:46]

- Order_Type: Medications
- Order_Subtype: nan

Pharmacy [2127-04-19 02:58:06]

- Medication: Zolpidem Tartrate
- Proc_Type: Unit Dose
- Status: Discontinued via patient discharge

Prescriptions [2127-04-19 03:00:00]

- Drug: Zolpidem Tartrate
- Atc Type: zolpidem
- Prod_Strength: 5mg Tablet
- Dose_Val_Rx: 5
- Dose_Unit_Rx: mg

Provider Order Entry [2127-04-19 05:41:43]

- Order_Type: Lab
- Order_Subtype: nan

Labotary Test Events [2127-04-19 07:16:00]

Item_Name	Valuenum	Valueuom	Ref_Range_Lower	Ref_Range_Upper	Flag	Comments
Hematocrit	29.0		Creatinine	1.4 mg/dL	0.4 1.1	abnormal nan
Urea Nitrogen	28.0	mg/dL	6.0	20.0	abnormal	nan

Radiology Examinations [2127-04-19 09:49:00]

- Note_Type: RR
- Exam_Name: ['CHEST (PRE-OP PA & LAT)']
- Text: REASON FOR EXAMINATION: Preoperative evaluation of the patient with aortic stenosis before aortic valve replacement. PA and lateral chest radiograph was compared to _____. Heart size is normal. Mediastinal position, contour and width are stable except for dextroscoliosis, mild to moderate. The lungs are clear except for linear bibasilar opacities, unchanged since _____, consistent with scarring. There is no pleural effusion or pneumothorax. The lateral view demonstrates contrast material in expected location of the distal esophagus that might be related to recent study involving the administration of oral contrast, please correlate with clinical history. There is no pleural effusion or pneumothorax demonstrated.

Provider Order Entry [2127-04-19 12:11:28]

Order_Type	Order_Subtype
ADT orders	Discharge
ADT orders	Discharge

Transfers [2127-04-19 13:10:00]

- Eventtype: discharge
- Careunit: nan

Case Information	Subject ID: 14122030 Hadm ID: 25609889Task	Type: Decision Making Task: Diagnoses ICD
EHR Trajectory	<pre>## Patient Demographics [None] - Anchor_Age: 70 - Gender: F ## Admissions [2129-04-05 04:06:00] - Admission_Type: EU OBSERVATION - Admission_Location: EMERGENCY ROOM - Admission_Info: None ## Transfers [2129-04-04 20:41:00] - Eventtype: ED - Careunit: Emergency Department ## EDstay [2129-04-04 20:41:00] - Gender: F - Race: BLACK/CARIBBEAN ISLAND ## Triage [2129-04-04 20:41:01] - Temperature: 97.9 - Heartrate: 103.0 - Respirate: 16.0 - O2Sat: 100.0 - Sbp: 163.0 - Dbp: 91.0 - Pain: 10 - Acuity: 3.0 - Chiefcomplaint: s/p Fall, Facial injury <events>... ## Laboratory Test Events [2129-04-04 23:00:00] Item_Name Valuenum Value uom Ref_Range_Lower Ref_Range_Upper Flag Comments ----- ----- ----- ----- ----- ----- ----- ... Glucose 168.0 mg/dL 70.0 100.0 abnormal IF FASTING, 70-100 NORMAL, >125 Potassium 7.0 mEq/L 3.3 5.1 abnormal ____ <item info>... <events>...</pre>	<pre>## Radiology Examinations [2129-04-04 23:00:00] - Note_Type: RR - Exam_Name: [CT HEAD W/O CONTRAST] - Text: EXAMINATION: CT HEAD W/O CONTRAST Q111 CT HEAD INDICATION: History: ____ with fall, fractured dentures, facial pain, L-arm, L-hip/ankle, R-leg pain. Evaluate for traumatic injury. TECHNIQUE: Contiguous axial images of the brain were obtained without contrast. Coronal and sagittal reformations as well as bone algorithm reconstructions were provided and reviewed. <item info>... IMPRESSION: 1. No hemorrhage. 2. Aerosolized secretions in bilateral partially imaged sphenoid sinuses. 3. Please refer to the dedicated CT neck and CT sinus/maxillofacial exam performed concurrently for findings specific to these regions. <events>...</pre>
Task Label List	Laceration without foreign body of lip, initial encounter Type 2 diabetes mellitus without complications Nondisplaced fracture of lateral malleolus of left fibula, initial encounter for closed fracture Other fall on same level, initial encounter Hyperlipidemia, unspecified Essential (primary) hypertension	<pre>## Medrecon [2129-04-04 23:20:00] Name Atc Type ----- ----- glipizide glipizide Lipitor atorvastatin metformin nan *hctz None omeprazole omeprazole <events>...</pre>
Medical Relation	<pre>[Unspecified fracture of left talus, initial encounter for closed fracture], "Inverse is a", "Fractures", "is PCDC EWS authorized value for variable", "Yes (indicator)", "has seronet authorized value", "Hypertension", "May be treated by", "amlodipine hydrochloride", "Contraindicated class of", "Potassium"] [Unspecified fracture of left talus, initial encounter for closed fracture], "Inverse is a", "Fractures", "Is Related to", "Foot Injuries and Disorders", "Is Related to", "Difficulty walking", "Is Related to", "E Codes: Fall"] [[Laceration without foreign body of lip, initial encounter], "Is Related to", "Wounds and Injuries", "Is Related to", "Fall (on) (from) unspecified stairs and steps, init encntr"] [Laceration without foreign body of lip, initial encounter], "Is Related to", "Wounds and Injuries", "May treat", "zinc chloride", "Inverse is a", "Dietary Supplement", "Is Related to", "Dietary Potassium", "Mapped from", "Potassium"] [Laceration without foreign body of lip, initial encounter], "Is Related to", "Wounds and Injuries", "Is Related to", "Surgery", "is OORO PC value of", "Multi-Specialty Consultation", "has OORO PC value", "Radiology"] [Laceration without foreign body of lip, initial encounter], "Is Related to", "Wounds and Injuries", "Is Related to", "E Codes: Fall"] </pre>	<pre>## ED Diagnoses on International Classification of Diseases [2129-04-05 15:51:00] Icd_Title Ccs_Type ----- ----- Oth fracture of upper and lower end of left fibula, init Fracture of lower limb Fall (on) (from) unspecified stairs and steps, init encntr E Codes: Fall </pre>
Reasoning Chain	<pre>## Extraction 1. **Triage [2129-04-04 20:41:01]**: - Chief complaint: s/p Fall, Facial injury - Pain: 10+ - This indicates a fall leading to a facial injury, which may relate to fractures or lacerations. 2. **Radiology Examinations [2129-04-04 23:00:00]**: - CT HEAD W/O CONTRAST: No fracture or hemorrhage, but aerosolized secretions in sinuses. - CT SINUS/MANDIBLE/AXILOFACIAL W/O CONTRAST: Maxillary teeth fractures, soft tissue hematoma, no mandibular fracture. - KNEE (AP, LAT & OBLIQUE) RIGHT: No fracture, minimal degenerative changes. - CT C-SPINE W/O CONTRAST: Possible chronic fracture, moderate degenerative changes. - ANKLE (AP, MORTISE & LAT) LEFT: Nondisplaced fracture of the distal tibia and talus avulsion fracture. - GLENO-HUMERAL SHOULDER (W/Y VIEW) LEFT: No fracture, moderate degenerative changes. 3. **Laboratory Test Events [2129-04-04 23:00:00]**: - Potassium: 7.0 mEq/L (abnormal, hyperkalemia). - Glucose: 168.0 mg/dL (abnormal, potential diabetes). - These lab results point towards metabolic issues such as diabetes and hyperkalemia. 4. **ED Diagnoses on International Classification of Diseases [2129-04-05 15:51:00]**: - Oth fracture of upper and lower end of left fibula, init - Fall (on) (from) unspecified stairs and steps, init encntr - This confirms a fracture due to a fall. 5. **Medrecon [2129-04-04 23:20:00]**: - Medications: glipizide, Lipitor, metformin, hctz, omeprazole. - Glipizide and metformin indicate management of diabetes.</pre> <p>## Reasoning</p> <ol style="list-style-type: none"> The patient's fall and subsequent injuries, including facial and dental trauma, are consistent with the diagnosis of "Laceration without foreign body of lip, initial encounter" due to the presence of facial pain and fractured teeth. The elevated glucose level of 168 mg/dL, along with the use of diabetes medications (glipizide and metformin), strongly supports the diagnosis of "Type 2 diabetes mellitus without complications." The radiology findings of a nondisplaced fracture of the distal tibia and talus avulsion fracture confirm the presence of a "Nondisplaced fracture of lateral malleolus of left fibula, initial encounter for closed fracture." The fall from unspecified stairs leading to multiple injuries aligns with the diagnosis of "Other fall on same level, initial encounter." The patient's medication list includes atorvastatin (Lipitor), commonly used for managing hyperlipidemia, which supports the diagnosis of "Hyperlipidemia, unspecified." The use of hydrochlorothiazide (hctz), a diuretic often used for hypertension, supports the diagnosis of "Essential (primary) hypertension." <p>## Final Results</p> Laceration without foreign body of lip, initial encounter Type 2 diabetes mellitus without complications Nondisplaced fracture of lateral malleolus of left fibula, initial encounter for closed fracture Other fall on same level, initial encounter Hyperlipidemia, unspecified Essential (primary) hypertension	

Supplementary Figure 1 | A case study of EHR Trajectory, Medical Relation, and Reasoning Chain. (a) EHR Trajectory for a patient, where `<events>...` and `<item info>...` represent the omission of a large amount of information for display purposes. (b) Medical Relation, showing the connections between the context medical entities and target items. (c) Reasoning Chain, detailing the process of inferring a diagnosis from the patient's EHR data. The parts highlighted in bold are the content commonly found in the EHR Trajectory, Medical Relation, and Reasoning Chain. This indicates that the medical graph is effective in identifying valid medical entities from the trajectory and using them to enhance reasoning.