Toward Better EHR Reasoning in LLMs: Reinforcement Learning with Expert Attention Guidance

Yue Fang[1,2 *], **Yuxin Guo**[1,2 *], **Jiaran Gao**[1,2 *], **Hongxin Ding**[1,2 *], **Xinke Jiang**[1,2], **Weibin Liao**[1,2], **Yongxin Xu**[1,2], **Yinhao Zhu**[1,2], **Zhibang Yang**[1,2], **Liantao Ma**[1,2†], **Junfeng Zhao**[1,2,3†], **Yasha Wang**[2,3,4†],

[1] School of Computer Science, Peking University, Beijing, China
[2] National Engineering Research Center For Software Engineering, Peking University, Beijing, China
[3] Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[4] Peking University Information Technology Institute (Tianjin Binhai)
yuefang25@stu.pku.edu.cn; malt@pku.edu.cn

Abstract

Improving large language models (LLMs) for electronic health record (EHR) reasoning is essential for enabling accurate and generalizable clinical predictions. While LLMs excel at medical text understanding, they underperform on EHR-based prediction tasks due to challenges in modeling temporally structured, high-dimensional data. Existing approaches often rely on hybrid paradigms, where LLMs serve merely as frozen prior retrievers while downstream deep learning (DL) models handle prediction, failing to improve the LLM's intrinsic reasoning capacity and inheriting the generalization limitations of DL models. To this end, we propose **EAG-RL**, a novel two-stage training framework designed to intrinsically enhance LLMs' EHR reasoning ability through expert attention guidance, where expert EHR models refer to task-specific DL models trained on EHR data. Concretely, EAG-RL first constructs high-quality, stepwise reasoning trajectories using expert-guided Monte Carlo Tree Search to effectively initialize the LLM's policy. Then, EAG-RL further optimizes the policy via reinforcement learning by aligning the LLM's attention with clinically salient features identified by expert EHR models. Extensive experiments on two real-world EHR datasets show that EAG-RL improves the intrinsic EHR reasoning ability of LLMs by an average of 14.62%, while also enhancing robustness to feature perturbations and generalization to unseen clinical domains. These results demonstrate the practical potential of EAG-RL for real-world deployment in clinical prediction tasks.

Code — https://github.com/devilran6/EAG-RL

1 Introduction

Large Language Models (LLMs) have shown strong capabilities across a wide spectrum of unstructured medical text processing tasks such as clinical note classification and report summarization (Jahan et al. 2024; Chen et al. 2023; Zhou et al. 2023), showing potential to assist physicians in accurate diagnosis (Zhang et al. 2025; Liu et al. 2024; Xu et al. 2025c) and treatment planning (Jiang et al. 2023b).

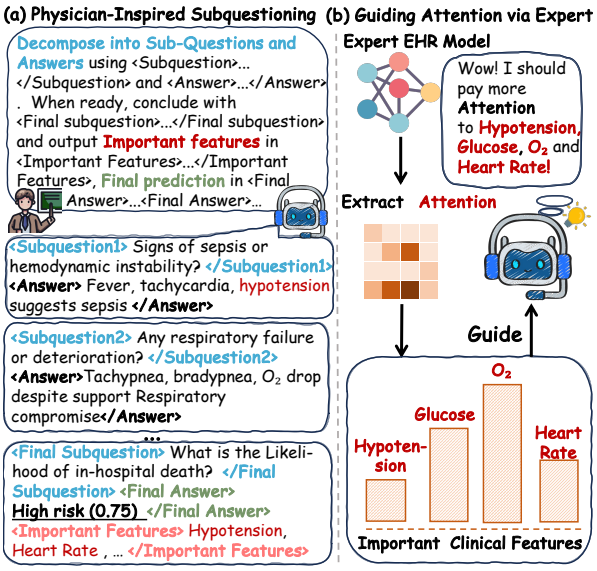

Figure 1: (a) Physician-inspired subquestioning, (b) Guiding attention via expert.

While effective in above tasks, LLMs still underperform on clinical prediction tasks based on **E**lectronic **H**ealth **R**ecords (EHR) (Brown et al. 2024; Chen et al. 2024a; Liao et al. 2025), which contain time-series values prevalent in healthcare systems and convey crucial physiological signals of patient health. Recent studies (Brown et al. 2024; Chen et al. 2024a; Wang et al. 2023; Fang et al. 2023) show that conventional deep learning models, referred to as *expert EHR models* due to their task-optimized architectures, remain the dominant choice for clinical prediction. Although significantly outperforming LLMs on EHR tasks (Zhu et al. 2024a), their reliance on fixed input schemas renders them

*These authors contributed equally.
†Corresponding Author.

brittle to variations in feature order, availability, and encoding, which are common across institutions with heterogeneous EHR systems and ultimately limit their generalizability. LLMs, by contrast, exhibit stronger generalization and hold promise as unified reasoning engines capable of learning transferable paradigms for robustly interpreting heterogeneous EHR data. Yet most existing approaches (Jiang et al. 2023a; Xu et al. 2025b, 2024; Nguyen et al. 2024; Zhu et al. 2024b) adopt hybrid, tool-using paradigms, where LLMs serve primarily as static prior retrievers, while downstream expert models handle final prediction, failing to improve the LLM's intrinsic reasoning capacity and inheriting traditional models' generalization limits. This gap highlights the urgent need to **intrinsically strengthen LLMs' EHR reasoning capacity** for clinical decision-making.

Recently, **R**einforcement **L**earning (RL) (Sutton, Barto et al. 1999; Kaelbling, Littman, and Moore 1996) has emerged as a powerful paradigm to enhance reasoning capacity, with models like OpenAI-O1 (OpenAI 2025) and DeepSeek-R1 (Guo et al. 2025a) achieving strong performance through learned stepwise reasoning policies. Inspired by these advances, we ask: *Can RL similarly benefit LLMs in EHR-based clinical prediction tasks?* Notably, real-world clinical prediction reflects a *hypothetico-deductive reasoning* process (Norman and Eva 2010), where physicians iteratively pose diagnostic subquestions and integrate evidence through stepwise reasoning to reach a well-supported conclusion. This leads to our first motivation(Figure 1(a)): **M#1.** *To enhance LLMs' EHR reasoning capacity via RL by imitating physicians' hypothesis refinement through stepwise subquestioning and evidence integration.*

However, the high-dimensional and temporally evolving nature of EHR demands reasoning policies that can dynamically attend to clinically salient features, similar to how physicians iteratively analyze important features by assessing their trends and clinical implications. Yet, sparse outcome rewards provide limited guidance for learning such fine-grained attention behaviors. Interestingly, prior works (Ma et al. 2023; Xu et al. 2023a,b) show that transformer-based expert EHR models can capture clinically salient features through attention mechanisms. This insight motivates us to distill the attention patterns of expert EHR models as auxiliary reward signals, guiding LLMs to assign greater focus to clinically salient features during training. This leads to our second motivation(Figure 1(b)): **M#2.** *To distill attention from expert EHR models to guide LLMs in attending to clinically salient features.*

These dual motivations give rise to our central goal: **to enhance LLMs' EHR reasoning capabilities via expert-attention guided RL**. However, realizing this goal remains unexplored and faces substantial challenges:

❶ **Challenge#1: How to construct high-quality stepwise trajectories for effective policy initialization?** RL on reasoning tasks often suffers from unstable convergence and poor sample efficiency when initialized from a weak policy (Wang et al. 2025; Xu et al. 2025a). Prior work demonstrates that strong initialization via **S**upervised **F**ine-**T**uning(SFT) can substantially improve RL efficiency (Chu et al. 2025; Chen, Gao, and Wu 2025; Wang et al. 2025). However, the scarcity of real-world multi-step SFT data showing how clinicians reason over patient records poses a major obstacle. The challenge lies in constructing high-quality, stepwise trajectories that capture realistic clinical reasoning patterns for effective policy initialization.

❷ **Challenge#2. How to extract reliable supervision from expert attention to guide EHR reasoning?** Although expert EHR models can highlight clinically salient features via attention (Yang et al. 2023; Ma et al. 2020; Xu et al. 2023a,b), directly aligning LLM attention with expert models is nontrivial due to semantic and architectural mismatches (Hao et al. 2023b), which risks introducing spurious supervision and leading to suboptimal policy updates. Additionally, extracting attention from LLMs imposes prohibitive computational overhead (Dao et al. 2022), making it impractical for frequent RL sampling. This calls for a scalable and semantically coherent alignment strategy that bridges model discrepancies while preserving training efficiency.

❸ **Challenge#3. How to encourage exploration of informative clinical patterns?** Recent studies indicate that RL algorithms like GRPO (Shao et al. 2024) are prone to entropy collapse, which leads to converging prematurely on low-entropy reasoning trajectories while failing to explore diverse, high-entropy alternatives, trapping in local optima (Yu et al. 2025). In EHR reasoning, LLMs risk converging on attending to a narrow set of high-confidence clinical features while overlooking truly critical ones. The challenge lies in designing reward mechanisms that adaptively amplify the influence of under-attended yet clinically meaningful features, thereby encouraging the model to explore high-entropy but clinical meaningful reasoning patterns.

To address these challenges, we propose **E**xpert-**A**ttention **G**uided **RL** (**EAG-RL**), a two-stage framework designed to intrinsically strengthen LLMs' EHR reasoning capacity via expert-guided policy optimization. For **Challenge#1**, we introduce *Expert-Guided Trajectory Distillation*, which constructs high-quality, stepwise reasoning trajectories via *Expert-guided Monte Carlo Tree Search*, enabling effective policy initialization. For **Challenge#2**, we introduce *Attention-Aligned Policy Optimization*, an RL-based stage that leverages expert attention as auxiliary reward through a lightweight alignment strategy, which quantifies the overlap between clinically salient features identified by the LLM and expert model via Jaccard similarity. For **Challenge#3**, we introduce *Entropy-Aware Adaptive Up Clipping*, which adaptively adjusts the clipping bound based on the entropy of salient features within reasoning trajectories, encouraging exploration of high-entropy yet meaningful reasoning patterns. In summary, our contributions are as follows:

- **Insightfully**, we demonstrate that expert EHR models can serve as effective policy supervisors for LLM, and propose EAG-RL as the first framework leveraging expert attention for enhancing LLMs' EHR reasoning capacity.
- **Technically**, we design a novel two-stage optimization framework: (1) a *warm-up stage* utilizing expert-guided MCTS to distill high-quality reasoning trajectories for effective policy initialization, and (2) a *reinforcement stage* integrating expert attention alignment and entropy-aware adaptive up clipping to encourage exploration of clinically

meaningful reasoning patterns.
- **Experimentally**, we conduct extensive experiments to validate EAG-RL on two real-world EHR datasets, demonstrating superior performance over state-of-the-art baselines. Further ablation and analysis substantiate the reasonableness and generalizability of EAG-RL.

2 Task Definition

Definition 1 (EHR Dataset). A patient's EHR is represented as a sequence of T time-ordered visits $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, where each visit $\mathbf{x}_t = \{l_{t,1}, l_{t,2}, \cdots, l_{t,n_t}\}$ contains n_t lab test features.

Definition 2 (Mortality Prediction). Given $\mathbf{X}$, predict whether the patient will survive the hospital stay. The label $y \in \{0, 1\}$ denotes death ($y = 1$) or survival ($y = 0$).

Definition 3 (Readmission Prediction). Given $\mathbf{X}$, predict whether the patient will be readmitted within 30 days after discharge. The label $y \in \{0, 1\}$ indicates readmission ($y = 1$) or not ($y = 0$).

3 Methodology

3.1 Overview

As illustrated in Figure 2, EAG-RL includes two stages:

- **# Stage 1: Expert-Guided Trajectory Distillation** constructs high quality, step-wise expert-guided reasoning trajectories to effectively initialize the LLM's policy.
- **# Stage 2: Attention-Aligned Policy Optimization** further optimizes LLM's policy via RL, guided by attention alignment with the *expert EHR model*.

3.2 Expert-Guided Trajectory Distillation

This stage aims to initialize the LLM's EHR reasoning ability through expert-guided, high-quality reasoning trajectories. To achieve this, we first generate multi-step trajectories using *Expert-Guided Monte Carlo Tree Search*, followed by *Trajectory-Level Supervised Fine-Tuning* to instill clinical reasoning patterns into the model.

Prompt-based Question Decomposition. Clinical reasoning often follows a hypothetico-deductive process (Norman and Eva 2010), where clinicians iteratively pose subquestions to refine diagnostic hypotheses. Inspired by this, we design prompt $\mathcal{P}_{\text{QD}}(\mathbf{X})$ to guide LLMs in decomposing complex EHR tasks into a sequence of subquestions and intermediate answers, simulating step-by-step clinical reasoning. Each step is represented using standardized tags: `<Subquestion>` and `<Answer>`, forming a structured reasoning unit. Once the model determines sufficient evidence has been gathered, it generates a final prediction using `<Final subquestion>`, `<Important Features>`, and `<Final answer>`. A reasoning trajectory is defined as an ordered sequence of subquestion-answer pairs, $\tau = \{(q_1, a_1), (q_2, a_2), \ldots, (q_T, a_T)\}$, where each a_i is either an intermediate reasoning output $\mathcal{S}(q_i)$ for $i < T$, or a final prediction $(y, \mathcal{C})$ at step $i = T$. Here, y denotes the predicted outcome, and $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ is a set of salient clinical features explicitly identified by the LLM as salient evidence in support of the prediction.

Expert-Guided Monte Carlo Tree Search. To construct high-quality reasoning trajectories aligned with clinical logic, we adopt Monte Carlo Tree Search (MCTS) (Kocsis and Szepesvári 2006; Coulom 2006), which explores the space of subquestion-answer paths defined by $\mathcal{P}_{\text{QD}}$ and efficiently balances exploration and exploitation to identify high-reward inference traces. To guide exploration toward salient clinical features, we further incorporate attention signals from a pretrained *expert EHR model* $\mathcal{M}_{\text{exp}}$, instantiated as Concare (Ma et al. 2020), a lightweight Transformer-based model specifically designed for EHR prediction and capable of capturing clinically salient features. The expert model's attention serves as an interpretable proxy for clinical relevance, steering MCTS toward salient clinical features and diagnostically meaningful subquestions. Specifically, *Expert-Guided MCTS* iteratively constructs a reasoning tree, where each node represents a reasoning state $s \in \mathcal{S}$ defined by a partial trajectory $\tau_{1:t} = \{(q_1, a_1), \ldots, (q_t, a_t)\}$. An edge corresponds to adding a new subquestion-answer pair, extending the trajectory to $\tau_{1:t+1}$. At each state s, the action space $\mathcal{A}(s)$ consists of candidate subquestions q_{t+1} generated by the LLM, conditioned on the current trajectory. The goal is to discover a complete trajectory $\tau_{1:T}$ that yields an accurate and clinically meaningful prediction.

- **Selection.** At each iteration, we select a leaf node for expansion by traversing the tree using the Upper Confidence Bound (UCT) criterion (Kocsis and Szepesvári 2006), which balances exploitation and exploration. For a node s and action $a \in \mathcal{A}(s)$, the next subquestion is chosen by maximizing:

$$\text{UCT}(s, a) = Q(s, a) + \lambda \cdot \sqrt{\frac{\log N(s)}{N(s, a)}}, \tag{1}$$

  where $Q(s, a)$ is the estimated action value, $N(s)$ is the visit count of node s, $N(s, a)$ is the count for action a, and λ controls exploration strength.
- **Expansion.** If the selected node is non-terminal, we generate d candidate subquestions $\{q_{t+1}^{(j)}\}_{j=1}^{d}$, each conditioned on the current trajectory $\tau_{1:t}$. For each $q_{t+1}^{(j)}$, the LLM produces an answer $a_{t+1}^{(j)}$, which is either an intermediate reasoning output or a final prediction, based on a termination indicator $\delta^{(j)} \in \{0, 1\}$:

$$a_{t+1}^{(j)} = \begin{cases} \mathcal{S}(q_{t+1}^{(j)}), & \text{if } \delta^{(j)} = 1, \\ (y^{(j)}, \mathcal{C}^{(j)}), & \text{if } \delta^{(j)} = 0, \end{cases} \tag{2}$$

  Each resulting pair $(q_{t+1}^{(j)}, a_{t+1}^{(j)})$ extends the trajectory and is added as a new child node. If the selected node is terminal, expansion is skipped.
- **Simulation.** To estimate the expected future reward of an expanded node efficiency, we perform a single-step rollout, recursively extending the trajectory until reaching a terminal state. At each step, the LLM generates d candidate subquestions and selects the most helpful one based on a local reward signal, following (Hao et al. 2023a) to reduce noise and cost. The local reward evaluates the usefulness of a candidate q_{t+1} under the current prefix $\tau_{1:t}$, using a dedicated evaluation prompt $\mathcal{P}_{\text{H}}$:

$$r(s, a) = \mathcal{M}_{\mathcal{P}_{\text{H}}}\left(\tau_{1:t} \cup \{(q_{t+1}, \varnothing)\}\right), \tag{3}$$

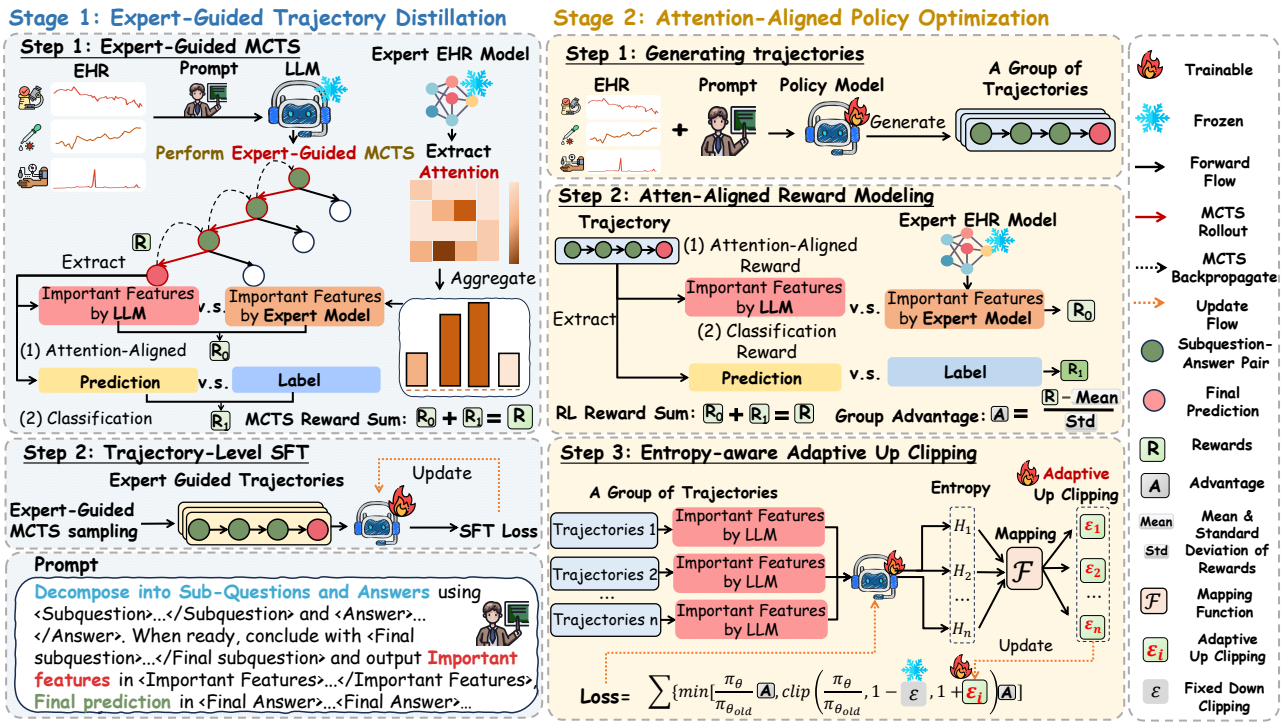

Figure 2: Illustration of EAG-RL.

where $\mathcal{M}$ denotes the LLM and $\varnothing$ indicates the subquestion is scored without an answer. This encourages exploration of informative, clinically relevant reasoning paths.

- **Backpropagation:** After reaching a terminal node, Expert-Guided MCTS backpropagates the cumulative reward to update all visited nodes. We integrate two complementary reward signals:

(1) *Classification reward* $\mathcal{R}_{\text{cls}} \in \mathbb{R}$ captures the model's prediction confidence and its directional margin. Let $\hat{y} \in (0, 1)$ denote the predicted probability, $y^\star \in \{0, 1\}$ the ground-truth label, and $\theta = 0.5$ the decision threshold:

$$\mathcal{R}_{\text{cls}} = \log\left(y^\star \cdot \hat{y} + (1 - y^\star)(1 - \hat{y})\right) + \Delta, \tag{4}$$

where Δ is a margin-aware bonus:

$$\Delta = \begin{cases} \beta(\hat{y} - \theta), & \text{if } y^\star = 1 \wedge \hat{y} > \theta, \\ \beta(\theta - \hat{y}), & \text{if } y^\star = 0 \wedge \hat{y} < \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

(2) *Attention alignment reward* $\mathcal{R}_{\text{att}} \in [0, 1]$ provides a semantically coherent and computationally lightweight strategy to guide the model's attention toward clinically meaningful information. Formally, $\mathcal{R}_{\text{att}}$ is computed as the Jaccard similarity between the model-extracted features $\mathcal{C}$ and the expert-highlighted features $\mathcal{C}_{\text{exp}}$:

$$\mathcal{R}_{\text{att}} = \frac{|\mathcal{C} \cap \mathcal{C}_{\text{exp}}|}{|\mathcal{C} \cup \mathcal{C}_{\text{exp}}|}. \tag{6}$$

This reward encourages alignment with expert insight while maintaining training efficiency.

The final reward is computed as a convex combination:

$$\mathcal{R} = \lambda \cdot \mathcal{R}_{\text{cls}} + (1 - \lambda) \cdot \mathcal{R}_{\text{att}}, \tag{7}$$

where $\lambda \in [0, 1]$ balances predictive accuracy and clinical alignment. The reward $\mathcal{R}$ is backpropagated along the search path to promote trajectories that are both outcome-correct and clinically meaningful.

Trajectory-Level Supervised Fine-Tuning. To distill high-quality reasoning behavior, we construct a SFT dataset $\mathcal{D}_{\text{SFT}}$ by selecting the top-k reward trajectories from *Expert-Guided MCTS*. Each trajectory τ represents a complete step-by-step reasoning path. Given the original clinical question $\mathcal{P}_{\text{QD}}(\mathbf{X})$, the model $\mathcal{M}$ is trained to generate the entire trajectory:

$$\mathcal{L}_{\text{SFT}} = -\sum_{\tau \in \mathcal{D}_{\text{SFT}}} \log \mathcal{M}(\tau \mid \mathcal{P}_{\text{QD}}(\mathbf{X})), \tag{8}$$

3.3 Attention-Aligned Policy Optimization

This stage aims to further optimize the LLM's reasoning policy through RL, guided by expert model's attention. To achieve this, an *Attention-Aligned Reward* is introduced to encourage the LLM to focus on meaningful features identified by the expert, and an *Entropy-Aware Adaptive Up Clipping* mechanism is applied to amplify learning signals for high-entropy reasoning paths, promoting exploration of uncertain but potentially informative clinical features.

Attention-Aligned Reward Modeling. To supervise policy optimization, we employ a composite reward that encourages both accurate predictions and clinically grounded reasoning. Specifically, we combine the outcome-based reward $\mathcal{R}_{\text{cls}}$, reflecting prediction confidence and correctness (Eq. 4), and the attention alignment reward $\mathcal{R}_{\text{att}}$, quantifying consistency with expert-highlighted features from $\mathcal{M}_{\text{exp}}$

(Eq. 6). The total reward $\mathcal{R}$ is computed as a convex combination (Eq. 7) and provides fine-grained feedback to guide clinically meaningful trajectory refinement.

Entropy-Aware Adaptive Up Clipping. While DAPO partially mitigates entropy collapse by decoupling the clipping mechanism to encourage exploration (Yu et al. 2025) , it still applies static upper and lower bounds across all trajectories, lacking adaptation to individual uncertainty. Consequently, it fails to strengthen learning signals for trajectories that are rare yet clinically informative. To address this, we introduce *Entropy-Aware Adaptive Up Clipping* to adaptively adjust the clipping bound based on token-level entropy over model-attended clinical features. This adaptive strategy enables trajectory-specific modulation of update strength, amplifying learning signals for uncertain, promising paths while constraining updates for overconfident ones.

For each trajectory τ, let $\mathcal{C}$ denote the set of key clinical tokens. We compute the average predictive entropy:

$$\bar{H}(\tau) = \frac{1}{|\mathcal{C}|} \sum_{t \in \mathcal{C}} H_t, \quad H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v), \tag{9}$$

where $\mathcal{V}$ is the vocabulary and $p_t(v)$ the predictive distribution at token t. Then, we map $\bar{H}(\tau)$ to a trajectory-specific clipping bound $\epsilon(\tau) \in [\epsilon_{\min}, \epsilon_{\max}]$ using min-max normalization across all g sampled trajectories:

$$\epsilon(\tau) = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot \frac{\bar{H}(\tau) - H_{\min}}{H_{\max} - H_{\min}}, \tag{10}$$

where $H_{\min}$ and $H_{\max}$ denote the minimum and maximum entropy values in the group. We use $\epsilon_{\min} = 0.2$, $\epsilon_{\max} = 0.4$ in our implementation.

RL Training. Building on the reward design and entropy-adaptive clipping strategy, we optimize the model policy using a GRPO-style objective (Shao et al. 2024), further enhanced with trajectory-level adaptive clipping. Following DAPO (Yu et al. 2025), we adopt an asymmetric clipping scheme. However, instead of using static bounds, we introduce a trajectory-specific adaptive upper bound $\epsilon(\tau)$ (defined in Eq. 10). The clipping function is:

$$\phi(r_t; \varepsilon, \epsilon(\tau)) = \max\left(1 - \varepsilon,\ \min(r_t,\ 1 + \epsilon(\tau))\right), \tag{11}$$

where ε is a fixed lower bound and r_t is the token-level importance ratio. Based on this, the optimization objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\sum_{t=1}^{|\tau|} \min\left(r_t \cdot \hat{A},\ \phi(r_t; \varepsilon, \epsilon(\tau)) \cdot \hat{A} \right) \right], \tag{12}$$

where $\hat{A}$ denotes the group-normalized advantage.

4 Experiments

In this section, we provide detailed information on the experimental setup, further analysis to validate the performance and rationality of EAG-RL.

4.1 Experimental Setup

Datasets. We conduct experiments on two real-world public EHR datasets: *MIMIC-IV* (Johnson et al. 2023), containing de-identified ICU records (2008–2019), and *TJH* (Yan et al. 2020), comprising structured inpatient data with clinical annotations. Following prior work (Gao et al. 2024; Zhu et al. 2024c), we apply temporal aggregation, LOCF imputation (Wells et al. 2013), and sequence visits at the patient level. We select patients with at least two visits and use the last visit for prediction. Stratified train/validation/test splits are used.

Baselines. We compare EAG-RL with a wide range of baselines from three perspectives.

- **Prompt-based methods.** We include a *Vanilla* baseline, where the model directly outputs the final answer without any intermediate reasoning. We also include *Think then Answer* (Guo et al. 2025b), which prompts the model to reason inside `<think>` tags before outputting a final answer in `<answer>`, and our proposed *Question Decomposition* prompting, which encourages sub-question generation to support progressive reasoning.
- **Training-based methods.** We include standard SFT (Ding et al. 2024; Zhang et al. 2023; Zelikman et al. 2022) as a strong baseline for aligning models with task-specific data. To evaluate the benefit of *Stage-1*, we compare against SFT directly. Additionally, to ensure a fair comparison in the RL stage, we compare EAG-RL with GRPO(Shao et al. 2024) and DAPO (Yu et al. 2025) under the *same Stage-1 initialization*, allowing us to isolate the effect of our reinforcement phase.
- **Open-source LLMs and backbone variants.** To assess the generality of our method, we instantiate EAG-RL on multiple backbone models, including Qwen2.5-7B-Instruct (Yang et al. 2024), LLaMA3.1-8B-Instruct (Dubey et al. 2024), and Qwen2.5-3B-Instruct (Yang et al. 2025). We also compare against several powerful open-source medical and reasoning LLMs such as HuatuoGPT-o1-7B (Chen et al. 2024b), OpenBioLLM-8B (Pal and Sankarasubbu 2024), and DeepSeek-R1-7B (Guo et al. 2025b).

Evaluation Metrics and Strategy. We employ two widely used evaluation metrics to measure the performance, namely, Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). Higher scores in these metrics indicate better predictive performance. Model selection is performed on the validation set. To assess variability, we apply bootstrapping with 100 resamples on the test set and report the mean and standard deviation in Table 1.

4.2 Experimental Results

Performance Comparison. Table 1 presents the performance of EAG-RL and baselines across two datasets. Overall, across all evaluation metrics on the two datasets, especially AUPRC, which is the most informative primary evaluation metric when dealing with highly imbalanced datasets, **EAG-RL consistently outperforms the current state-of-the-art methods**. Compared to prompt-based approaches, EAG-RL achieves an average improvement of 14.62% across models and tasks, validating the intrinsic EHR reasoning enhancement from our two-stage framework. Notably, our *Question Decomposition* also surpasses

Method		TJH Mortality		MIMIC-IV Mortality		MIMIC-IV Readmission	
Category	Variant	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)
Qwen2.5-7B-Instruct							
Prompt-based	Vanilla	71.99±2.78	64.27±4.44	53.90±2.74	10.68±2.58	57.76±4.29	26.53±3.91
	Think & Answer	79.83±2.68	70.87±4.61	61.57±7.12	13.58±3.17	55.86±3.98	25.32±4.02
	Question Decomposition	81.66±2.85	73.26±4.53	63.49±6.14	15.08±4.48	59.29±3.85	26.26±3.68
SFT	SFT	81.20±3.06	75.14±4.39	64.70±6.63	16.99±5.96	51.93±4.81	23.25±3.84
	EAG-RL(Stage-1)	83.64±2.75	77.60±4.16	69.33±5.74	17.90±5.20	60.00±4.67	28.00±4.47
RL	EAG-RL(Stage-1)+GRPO	85.82±2.72	78.52±4.40	65.78±6.62	19.53±6.87	53.67±4.62	26.03±4.39
	EAG-RL(Stage-1)+DAPO	85.67±2.62	77.88±4.38	73.93±5.18	19.71±5.71	57.80±4.25	26.74±3.94
	EAG-RL(Stage-1+Stage-2)	**87.70±2.49**	**82.95±3.64**	**77.21±5.04**	**23.99±6.97**	**61.09±4.48**	**29.92±0.46**
LLaMA3.1-8B-Instruct							
Prompt-based	Vanilla	71.98±3.58	67.75±4.59	58.73±6.33	11.77±3.26	52.88±5.06	25.75±4.36
	Think & Answer	57.34±3.75	50.26±4.18	60.79±6.29	10.33±2.99	55.41±6.51	10.85±2.91
	Question Decomposition	74.15±3.41	64.88±4.83	58.88±6.84	12.71±3.70	54.90±5.18	27.06±4.53
SFT	SFT	76.45±3.36	67.84±4.91	47.24±7.69	9.79±3.19	52.49±4.85	24.11±3.92
	EAG-RL(Stage-1)	80.67±3.03	74.39±4.37	57.31±7.39	13.16±3.83	55.45±3.87	27.04±4.44
RL	EAG-RL(Stage-1)+GRPO	80.49±3.08	72.17±4.75	56.82±4.37	**15.64±5.80**	52.98±4.24	25.52±3.94
	EAG-RL(Stage-1)+DAPO	78.89±3.20	72.19±4.40	59.26±6.43	13.82±4.08	55.12±3.99	26.83±4.30
	EAG-RL(Stage-1+Stage-2)	**84.34±2.71**	**76.36±4.64**	**62.17±5.26**	12.51±3.00	**55.65±4.28**	**27.99±4.51**
Qwen2.5-3B-Instruct							
Prompt-based	Vanilla	66.27±2.74	57.63±4.11	51.10±6.33	10.06±3.41	54.76±3.79	26.27±4.19
	Think & Answer	59.35±4.04	53.42±4.88	54.88±5.88	10.88±2.74	57.92±4.85	27.09±4.58
	Question Decomposition	70.97±3.34	59.07±4.52	57.88±7.78	12.69±5.60	56.82±4.96	28.35±5.29
SFT	SFT	69.01±3.24	59.55±4.60	55.60±6.38	13.47±5.27	51.84±4.72	24.83±4.39
	EAG-RL(Stage-1)	77.49±6.88	67.91±10.08	67.87±4.95	14.62±4.04	61.15±4.46	29.62±4.75
RL	EAG-RL(Stage-1)+GRPO	75.78±5.76	68.70±8.11	67.95±5.87	16.18±4.45	56.23±4.53	26.17±3.83
	EAG-RL(Stage-1)+DAPO	79.33±4.90	72.32±7.30	65.64±6.43	15.42±4.57	57.68±4.83	28.14±4.32
	EAG-RL(Stage-1+Stage-2)	**80.31±4.89**	**73.46±7.48**	**70.40±5.46**	**17.95±0.51**	**61.23±4.91**	**31.76±5.21**

Table 1: Performance comparison on TJH and MIMIC-IV datasets. We report AUROC and AUPRC (%) for each task. **Bold** indicates the best-performing method, and underline denotes the second-best across all methods.

Method	TJH Mortality		MIMIC-IV Mortality	
	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)
EAG-RL	**87.70±2.49**	**82.95±3.64**	**77.21±5.04**	**23.99±6.97**
w/o Stage-1	85.34±2.58	79.03±4.06	76.23±4.26	20.83±5.85
w/o Stage-2	83.64±2.75	77.60±4.16	69.33±5.74	17.90±5.20
w/o $\mathcal{R}_{\text{att}}$	85.20±2.62	76.64±4.40	75.58±4.28	20.32±5.61
w/o $\epsilon(\tau)$	80.84±2.96	72.71±4.53	66.24±5.12	14.82±4.03

Table 2: Ablation study results of our proposed EAG-RL.

Think then Answer baseline, suggesting that guiding reasoning via subquestion decomposition is a more effective prompting strategy in clinical contexts. Secondly, compared to vanilla SFT, the warm-up stage of EAG-RL using expert-guided MCTS trajectories leads to consistent improvements. This underscores the importance of high-quality, stepwise initialization for aligning the LLM's policy with clinical reasoning patterns. Finally, given the *same Stage-1 initialization*, EAG-RL consistently outperforms state-of-the-art RL methods. This fair comparison underscores the effectiveness of our reinforcement stage, which integrates expert attention and *Entropy-Aware Adaptive Up Clipping* to guide exploration toward uncertain but informative clinical features.

Ablation Study. To investigate the effectiveness of each component in EAG-RL, we construct several ablated variants. In *w/o Stage-1*, we removes *Expert-Guided Trajectory Distillation* and trains the model directly via RL. In *w/o Stage-2*, we remove *Attention-Aligned Policy Optimization*. In *w/o* $\mathcal{R}_{\text{att}}$, we removes the *Attention-Alignment Reward*. In *w/o* $\epsilon(\tau)$,we disable the *Entropy-Aware Adaptive Up Clipping* during policy optimization. As shown in Table 2, removing either *Stage-1* or *Stage-2* leads to substantial performance degradation, confirming that both the expert-guided warm-up and the reinforcement refinement are essential for achieving strong EHR reasoning. In particular, the removal of *Stage-2* causes a larger drop, indicating that reward-based policy refinement plays a critical role beyond trajectory initialization. Furthermore, removing $\mathcal{R}_{\text{att}}$ results in a noticeable decline in performance, validating the importance of leveraging expert attention as auxiliary supervision to guide the model's focus toward clinically meaningful features. Similarly, without $\epsilon(\tau)$, the model fails to sufficiently explore uncertain yet informative patterns, leading to lower robustness and reduced predictive accuracy. Overall, the ablation results highlight the necessity of each component in EAG-RL and demonstrate the synergistic benefits of expert-guided initialization and reward-aware policy optimization.

Method	TJH Mortality		MIMIC-IV Mortality		MIMIC-IV Readmission	
	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)
Qwen2.5-7B-Instruct	79.83	70.87	61.57	13.58	55.86	25.32
HuatuoGPT-o1-7B	85.34	77.31	70.39	20.33	50.54	24.30
Deepseek-R1-7B	52.70	47.89	40.94	9.43	53.19	24.53
LLaMA3.1-8B-Instruct	57.34	50.26	60.79	10.33	55.41	10.85
OpenBioLLM-8B	56.75	49.76	58.69	12.85	50.21	24.23
Ours + LLaMA3.1-8B-Instruct	84.34	76.36	62.17	12.51	55.65	27.99
Ours + Qwen2.5-7B-Instruct	**87.70**	**82.95**	**77.21**	**23.99**	**61.09**	**29.92**

Table 3: Performance comparison across open-source models. Best results for each metric are highlighted in **bold**.

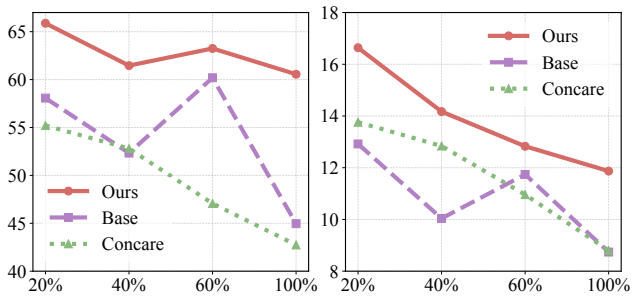

Figure 3: Robustness to feature order perturbation measured by AUROC (left) and AUPRC (right).

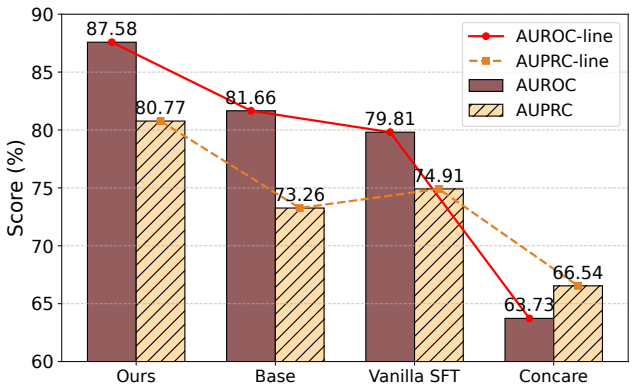

Figure 4: Cross-Dataset generalization from MIMIC-IV to TJH on mortality.

4.3 Analysis

EXP#1: Intrinsic EHR Reasoning Capability. To quantify how much our method enhances LLMs' intrinsic EHR reasoning, we compare EAG-RL with several competitive open-source models on mortality and readmission prediction tasks across TJH and MIMIC-IV datasets. As shown in Table 3, **EAG-RL consistently outperforms all competitive models across tasks and metrics.** Notably, EAG-RL significantly surpasses HuatuoGPT-o1-7B and OpenBioLLM-8B, both trained on medical corpora or enhanced via domain-specific RL, demonstrating the effectiveness of EAG-RL in intrinsically strengthening EHR reasoning.

EXP#2: Robustness Against Feature Order Perturbation. To evaluate whether EAG-RL captures clinically meaningful reasoning beyond superficial feature ordering, we conduct a robustness test on the *mortality prediction* task by perturbing feature order at inference. This *simulates real-world deployment scenarios* where EHR features may arrive in inconsistent orders due to variability in data collection, preprocessing, or integration pipelines. We compare against Concare, the expert EHR model, and Qwen2.5-7B, the untrained LLM backbone. For each MIMIC-IV test case, we randomly shuffle a proportion ($p\%$) of features while keeping the rest fixed. As shown in Figure 3, Concare's performance drops sharply under moderate perturbations, indicating its reliance on fixed input order. In contrast, EAG-RL maintains robust performance across all disruption levels and continues to outperform the base model even under full permutation. These results suggest that **EAG-RL learns semantically grounded, order-invariant reasoning strategies, enabling stronger generalization.**

EXP#3: Cross-Dataset Generalization. We assess EAG-RL's generalization via an out-of-distribution (OOD) test, training on MIMIC-IV and evaluating on TJH for *mortality prediction*. TJH differs from MIMIC-IV in patient demographics and coding schemas, simulating deployment across heterogeneous medical environments. We compare against Concare, the untrained base model (Qwen2.5-7B), and Vanilla SFT. As shown in Figure 4, EAG-RL achieves the highest performance, significantly outperforming all baselines in both AUROC and AUPRC. These results suggest that **EAG-RL captures transferable clinical patterns instead of relying on dataset-specific artifacts**.

5 Conclusions and Future Works

We present **EAG-RL**, a novel two-stage training framework that enhances the intrinsic EHR reasoning ability of LLMs through expert-guided attention. Experiments show that EAG-RL achieves an average improvement of 14.62% across multiple EHR prediction tasks, while also improving robustness to input perturbations and generalization to unseen clinical domains. These results highlight the potential of EAG-RL for real-world clinical deployment. In future work, we plan to explore richer forms of supervision beyond attention from expert EHR models, and extend our framework to incorporate multi-expert distillation to better capture diverse clinical reasoning patterns.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20468), Peking University Clinical Medicine Plus X (Young Scholars Project-the Fundamental Research Funds for the Central Universities PKU2025PKULCXQ024; Pilot Program-Key Technologies Project 2024YXXLHGG007), and Peking University "TengYun" Clinical Research Program (TY2025015). Liantao Ma was supported by Beijing Natural Science Foundation (L244063, L244025), Beijing Municipal Health Commission Research Ward Excellence Clinical Research Program (BRWEP2024W032150205), and Beijing Traditional Chinese Medicine Science and Technology Development Fund (BJZYZD-2025-13).

References

Brown, K. E.; Yan, C.; Li, Z.; Zhang, X.; Collins, B. X.; Chen, Y.; Clayton, E. W.; Kantarcioglu, M.; Vorobeychik, Y.; and Malin, B. A. 2024. Not the models you are looking for: Traditional ML outperforms LLMs in clinical prediction tasks. *medRxiv*.

Chen, C.; Yu, J.; Chen, S.; Liu, C.; Wan, Z.; Bitterman, D.; Wang, F.; and Shu, K. 2024a. ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction? *arXiv preprint arXiv:2411.06469*.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Chen, Q.; Du, J.; Hu, Y.; Kuttichi Keloth, V.; Peng, X.; Raja, K.; Zhang, R.; Lu, Z.; and Xu, H. 2023. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv e-prints*, arXiv–2305.

Chen, Y.; Gao, J.; and Wu, J. 2025. Towards Revealing the Effectiveness of Small-Scale Fine-tuning in R1-style Reinforcement Learning. *arXiv preprint arXiv:2505.17988*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Coulom, R. 2006. Efficient selectivity and backup operators in Monte-Carlo tree search. In *International conference on computers and games*, 72–83. Springer.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359.

Ding, H.; Fang, Y.; Zhu, R.; Jiang, X.; Zhang, J.; Xu, Y.; Chu, X.; Zhao, J.; and Wang, Y. 2024. 3DS: Decomposed Difficulty Data Selection's Case Study on LLM Medical Domain Adaptation. *arXiv preprint arXiv:2410.10901*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fang, Y.; Xu, Y.; Ding, H.; Zhao, J.; Wang, Y.; and Jin, H. 2023. A method and practice for menopausal disease prediction based on knowledge graph. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 10–18. IEEE.

Gao, J.; Zhu, Y.; Wang, W.; Wang, Z.; Dong, G.; Tang, W.; Wang, H.; Wang, Y.; Harrison, E. M.; and Ma, L. 2024. A comprehensive benchmark for COVID-19 predictive modeling using electronic health records in intensive care. *Patterns*, 5(4).

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hao, S.; Gu, Y.; Ma, H.; Hong, J. J.; Wang, Z.; Wang, D. Z.; and Hu, Z. 2023a. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023b. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36: 79570–79582.

Jahan, I.; Laskar, M. T. R.; Peng, C.; and Huang, J. X. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171: 108189.

Jiang, P.; Xiao, C.; Cross, A.; and Sun, J. 2023a. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Jiang, X.; Zhang, R.; Xu, Y.; Qiu, R.; Fang, Y.; Wang, Z.; Tang, J.; Ding, H.; Chu, X.; Zhao, J.; and Wang, Y. 2023b. Think and Retrieval: A Hypothesis Knowledge Graph Enhanced Medical Large Language Models.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Kocsis, L.; and Szepesvári, C. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*, 282–293. Springer.

Liao, W.; Zhu, Y.; Zhang, Z.; Wang, Y.; Wang, Z.; Chu, X.; Wang, Y.; and Ma, L. 2025. Learnable prompt as pseudo-imputation: Rethinking the necessity of traditional ehr data imputation in downstream clinical prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 765–776.

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Zhang, Z.; Tian, F.; and Zheng, Y. 2024. Large Language Model Distilling Medication Recommendation Model. arXiv:2402.02803.

Ma, L.; Zhang, C.; Gao, J.; Jiao, X.; Yu, Z.; Zhu, Y.; Wang, T.; Ma, X.; Wang, Y.; Tang, W.; et al. 2023. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients. *Patterns*, 4(12).

Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 833–840.

Nguyen, T. D.; Huynh, T. T.; Phan, M. H.; Nguyen, Q. V. H.; and Le Nguyen, P. 2024. CARER-ClinicAl Reasoning-Enhanced Representation for Temporal Health Risk Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10392–10407.

Norman, G. R.; and Eva, K. W. 2010. Diagnostic error and clinical reasoning. *Medical education*, 44(1): 94–100.

OpenAI. 2025. OpenAI-o1. Accessed: 2025-05-16.

Pal, M. S. A.; and Sankarasubbu, M. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sutton, R. S.; Barto, A. G.; et al. 1999. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1): 126–134.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.; Yang, J.; Zhang, Z.; et al. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wang, Z.; Li, R.; Dong, B.; Wang, J.; Li, X.; Liu, N.; Mao, C.; Zhang, W.; Dong, L.; Gao, J.; et al. 2023. Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. *arXiv preprint arXiv:2306.01499*.

Wells, B. J.; Chagin, K. M.; Nowacki, A. S.; and Kattan, M. W. 2013. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3): 1035.

Xu, H.; Zhu, Q.; Deng, H.; Li, J.; Hou, L.; Wang, Y.; Shang, L.; Xu, R.; and Mi, F. 2025a. KDRL: Post-Training Reasoning LLMs via Unified Knowledge Distillation and Reinforcement Learning. *arXiv preprint arXiv:2506.02208*.

Xu, R.; Shi, W.; Yu, Y.; Zhuang, Y.; Jin, B.; Wang, M. D.; Ho, J. C.; and Yang, C. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

Xu, Y.; Chu, X.; Yang, K.; Wang, Z.; Zou, P.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023a. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, 2819–2830.

Xu, Y.; Jiang, X.; Chu, X.; Qiu, R.; Feng, Y.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2025b. DearLLM: Enhancing Personalized Healthcare via Large Language Models-Deduced Feature Correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 941–949.

Xu, Y.; Yang, K.; Zhang, C.; Zou, P.; Wang, Z.; Ding, H.; Zhao, J.; Wang, Y.; and Xie, B. 2023b. VecoCare: Visit Sequences-Clinical Notes Joint Learning for Diagnosis Prediction in Healthcare Data. In *IJCAI*, volume 23, 4921–4929.

Xu, Y.; Zhang, R.; Jiang, X.; Feng, Y.; Xiao, Y.; Ma, X.; Zhu, R.; Chu, X.; Zhao, J.; and Wang, Y. 2025c. Parenting: Optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11643–11662.

Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; et al. 2020. An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence*, 2(5): 283–288.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, G.; Yu, S.; Sheng, Y.; and Yang, H. 2023. Attention and feature transfer based knowledge distillation. *Scientific Reports*, 13(1): 18369.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35: 15476–15488.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Li, J.; Chen, G.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Zhang, J.; Fang, Y.; Ding, H.; Liao, W.; Ye, M.; Chu, X.; Zhao, J.; and Wang, Y. 2025. ADEPT: Continual Pretraining via Adaptive Expansion and Dynamic Decoupled Tuning. *arXiv preprint arXiv:2510.10071*.

Zhou, H.; Liu, F.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S. S.; Zhou, P.; Liu, J.; et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Zhu, Y.; Gao, J.; Wang, Z.; Liao, W.; Zheng, X.; Liang, L.; Bernabeu, M. O.; Wang, Y.; Yu, L.; Pan, C.; et al. 2024a. ClinicRealm: Re-evaluating Large Language Models with Conventional Machine Learning for Non-Generative Clinical Prediction Tasks. *arXiv preprint arXiv:2407.18525*.

Zhu, Y.; Ren, C.; Wang, Z.; Zheng, X.; Xie, S.; Feng, J.; Zhu, X.; Li, Z.; Ma, L.; and Pan, C. 2024b. Emerge: Integrating rag for improved multimodal ehr predictive modeling. *arXiv preprint arXiv:2406.00036*.

Zhu, Y.; Wang, W.; Gao, J.; and Ma, L. 2024c. Pyehr: A predictive modeling toolkit for electronic health records.