RASPER: Reward-Aligned Summarization of Clinical Notes for EHR Outcome Prediction

Arya Hadizadeh Moghaddam, Mohsen Nayebi Kerdabadi
Chen Chen, Dongjie Wang, Zijun Yao*
Electrical Engineering and Computer Science, University of Kansas, USA
{a.hadizadehm, mohsen.nayebi, chenchen, wangdongjie, zyao}@ku.edu

Abstract

Unstructured discharge notes in Electronic Health Records (EHRs) often carry signal complementary to structured medical codes, holding patient-specific evidence that standardized cohort-level codes alone cannot capture. However, this evidence in notes is frequently buried in lengthy, noisy text that is not intentionally written with any specific clinical prediction in mind. Summarization is an obvious mitigation, but generic summaries, tuned for fluency rather than the outcome, routinely omit decisive evidence while retaining plausible but uninformative detail. To this end, we propose **RASPER**, a **R**eward-**A**ligned **S**ummarizer for **P**rediction in **E**H**R**, that optimizes note summarization directly against the downstream clinical task. RASPER employs a tunable LLM-based summarizer to extract task-relevant evidence from discharge notes and trains it via reinforcement learning from prediction feedback, using a reward derived from the downstream predictor's loss. To ground the summarizer, a longitudinal encoder converts structured codes into soft prompts that incorporate each patient's clinical context into note summarization. By rewarding the quality of the resulting multimodal prediction, RASPER encourages the summarizer to retain patient-specific evidence that complements, rather than duplicates, information captured by structured codes. RASPER consistently outperforms strong baselines on both readmission prediction and medication recommendation across MIMIC-III and MIMIC-IV.

1 Introduction

Electronic Health Records (EHRs) represent a patient's health trajectory with multimodal views: structured codes such as diagnoses capture standardized, cohort-level patterns across patients (Choi et al., 2016), while discharge notes record patient-specific symptom severity, treatment response, and clinician observations that codes cannot fully record (Tayefi et al., 2021).

Although discharge notes contain rich patient-specific information, they are primarily written for clinical documentation rather than predictive modeling. These raw notes are often lengthy, redundant, and sensitive to variation in documentation style (Liu et al., 2022). Directly feeding such notes into a model can therefore dilute the predictive signal and make the model sensitive to documentation artifacts rather than the underlying clinical state (H Moghaddam et al., 2026). A natural strategy is to summarize discharge notes with an LLM (Goyal et al., 2024; Gebreab et al., 2024), compressing noisy clinical narratives into a fluent, coherent summary before prediction. However, reducing note length and improving readability does not by itself ensure that the retained information is useful for a specific prediction task.

The predictive value of a note summary depends on both the downstream task and the patient-specific context. Although different clinical prediction tasks may rely on overlapping clinical information, the information that is most predictive of each task can differ (Yoon et al., 2025). Hence, a summary optimized only for generic coverage and denoising is optimized for the wrong objective (Yadav et al., 2021), since it might discard details that a specific prediction depends on while preserving plausible but uninformative clinical information. This mismatch suggests that discharge note summarization should not be treated as a separate preprocessing step, but should instead be aligned with the downstream prediction and patient-specific objective.

Existing EHR-LLM methods do not directly address this problem. Some approaches feed raw or hard-prompted notes into LLMs and inherit the noise of the original documentation (Jiang et al., 2023; Huang et al., 2019), and others rely on fixed

*Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 24519–24530
October 24-29, 2026 ©2026 Association for Computational Linguistics

summaries or extracted textual features that are never tuned to the target outcome (H Moghaddam et al., 2026). Some structured code models discard the note modality altogether (Jiang et al., 2024; Xu et al., 2024). As a result, prior methods generally do not align the summarization process with the downstream clinical objective.

To close this gap, we propose **RASPER**, a **R**eward-**A**ligned **S**ummarizer for **P**rediction in **EHR**. RASPER optimizes the summarization of notes directly against the downstream task: a tunable LLM summarizer is trained with reinforcement learning from prediction feedback (RLPF) (Wu et al., 2025), using the predictor's loss as the reward signal. A longitudinal EHR encoder maps structured codes into soft prompts that provide patient grounding. Because the reward reflects how well the note summary and the structured codes collaborate, the summarizer is encouraged to preserve note evidence that complements the structured code information and contributes to downstream prediction. The summarizer is first initialized through supervised fine-tuning on teacher-generated summaries, then optimized with this prediction-feedback objective to preserve the task-relevant evidence that improves predictive utility. Our approach makes the following contributions:

- We propose a reward-aligned note summarization framework that trains a tunable LLM summarizer directly against the downstream prediction objective, using reinforcement learning from prediction feedback rather than treating summaries as a fixed preprocessing step.
- We design an LLM-based predictor that jointly conditions on structured EHR soft prompts and generated note summaries, enabling prediction feedback to align the summarizer with both cohort-level code patterns and informative textual evidence.
- We conduct experiments on MIMIC-III and MIMIC-IV, showing consistent improvements over strong deep-learning and LLM-based baselines on two clinical prediction tasks.

2 Methodology

2.1 Problem Formulation

For each patient i, we use an ordered sequence of historical clinical visits $\{V_{i,t}\}_{t=1}^{T_i}$ as input, where $V_{i,t}$ denotes the t-th observed visit in chronological order and T_i denotes the number of historical visits available before the target visit. Each visit $V_{i,t}$ contains a set of medical codes, including diagnoses, medications, and procedures. Formally, the codes recorded at visit t are defined as $V_{i,t} = \{x_{j,t}^i\}_{j=1}^{|V_{i,t}|}$, where $|V_{i,t}|$ denotes the number of medical codes associated with that visit. In addition to structured medical codes, each visit is accompanied by discharge notes represented as text. We denote the token sequence of the discharge note for patient i at visit t as $S_{i,t} = \{s_{j,t}^i\}_{j=1}^{|S_{i,t}|}$, where each $s_{j,t}^i$ denotes a token and $|S_{i,t}|$ is the length of the note.

The target outcome is task-dependent and is defined as

$$Y_i \in \begin{cases} \{0,1\}, & \text{readmission prediction} \\ \{0,1\}^K, & \text{medication recommendation} \end{cases} \tag{1}$$

where K denotes the size of the medication vocabulary. For medication recommendation, $Y_{i,k} = 1$ indicates that medication k is prescribed at the target visit $T_i + 1$.

2.2 Model Overview

Figure 1 illustrates the overall architecture of **RASPER**. The model is designed around the idea that clinical-note summarization should not be optimized independently from the prediction task. The framework contains two interacting modules: an LLM-aware outcome predictor and a tunable LLM summarizer. The predictor combines structured EHR soft prompts with generated note summaries to model both cohort-level code patterns and patient-specific textual evidence. After supervised fine-tuning, the summarizer is further optimized through prediction feedback, so that its generated summaries preserve information that improves downstream clinical prediction.

2.3 LLM-Aware Outcome Predictor

Clinical prediction from EHR requires modeling both discharge notes and structured medical codes (H Moghaddam et al., 2026). Discharge notes describe patient condition, treatment, and clinical progression (Johnson et al., 2016). In contrast, diagnoses, procedures, and medications provide standardized cohort-level signals (Choi et al., 2016; H Moghaddam et al., 2024; Shang et al., 2019). Using only notes may expose the model to non-standardized textual evidence, and using only structured codes may miss important patient-specific clinical context. Therefore, we integrate

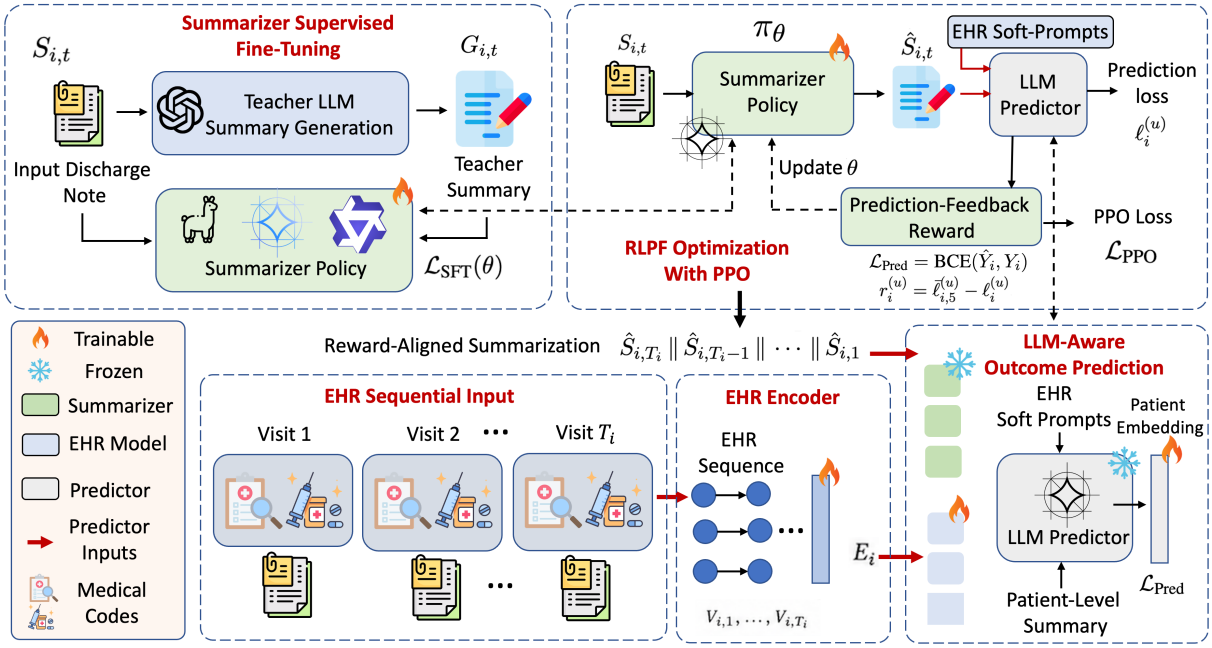

Figure 1: Overview of the proposed RASPER framework. A tunable LLM summarizer transforms discharge notes into standardized visit-level summaries, which are concatenated into a patient-level summary. An LLM-aware predictor combines this patient-level summary with soft prompt embeddings from structured EHR codes for outcome prediction. The summarizer is first initialized with supervised fine-tuning and then optimized with RLPF using prediction-loss-derived reward within the PPO objective.

both in a unified representation space.

Large language models can encode clinical text into contextualized representations (Yenduri et al., 2024), but many healthcare LLM approaches rely mainly on hard prompting over linearized patient records (Meskó, 2023; Zaghir et al., 2024). We therefore use prompt tuning (H Moghaddam et al., 2026), where the summarized discharge note embeddings are augmented with soft prompts from a structured EHR encoder. This allows the two modalities to complement each other in a lightweight and modular prediction framework.

Mathematically, the input to the LLM for patient i is defined as

$$\boldsymbol{I}_i = [\boldsymbol{E}_i; \text{Embed}(\hat{S}_i)] \in \mathbb{R}^{(P+C_i)\times D}. \quad (2)$$

The structured code sequence is first encoded by an encoder to obtain a patient-level representation $\boldsymbol{h}_i^{\text{ehr}}$. For the EHR encoder, we use RETAIN (Choi et al., 2016), a lightweight longitudinal model that captures visit-level and code-level patterns through attention (Further details on RETAIN are provided in Appendix A.1). We then project this representation into P continuous soft prompt embeddings:

$$\boldsymbol{E}_i = \text{Proj}_{\text{ehr}}(\boldsymbol{h}_i^{\text{ehr}}), \quad \boldsymbol{E}_i \in \mathbb{R}^{P\times D}. \quad (3)$$

where P is the number of soft prompt tokens and D is the hidden dimension of the LLM predictor.

$\text{Embed}(\hat{S}_i) \in \mathcal{R}^{C_i\times D}$ denotes the token embeddings corresponding to the patient summary in the *LLM-Aware Outcome Predictor*, where C_i is the number of summary tokens for patient i. As defined in Equation 4, the representation I_i is then provided to the predictor LLM, which outputs the final hidden representation $H_i \in \mathbb{R}^{(P+C_i)\times D}$:

$$\boldsymbol{H}_i = \text{LLM}_{\text{pred}}(\boldsymbol{I}_i). \quad (4)$$

As defined in Equation 5, we first apply average pooling over the final hidden representations H_i and then use a linear prediction head to obtain task-specific logits:

$$\mathbf{y}_i = \text{Linear}_{\text{pool}}(\boldsymbol{H}_i). \quad (5)$$

The prediction loss is binary cross-entropy:

$$\mathcal{L}_{\text{Pred}} = \text{BCE}\left(\hat{\boldsymbol{Y}}_i, Y_i\right), \quad (6)$$

where $\hat{Y}_i = \sigma(\mathbf{y}_i)$ denotes the predicted probability for patient i. The trainable components of the predictor include the EHR-to-soft-prompt layer, the prediction head, and the LoRA adapters of the LLM predictor. The backbone LLM parameters remain frozen.

You are a medical researcher. Summarize the following clinical note into a concise, high-signal summary for downstream prediction. Keep important diagnoses, medications, major findings, treatments, procedures, and time-relevant clinical facts. Avoid fluff, duplication, and formatting noise.

Discharge Note: {note}

Summarizer Prompt

Figure 2: Prompt used by the tunable LLM summarizer to extract relevant information from discharge notes.

2.4 Tunable LLM Summarizer

The predictive value of clinical information depends on the downstream task and the patient-level prediction context. Hence, a generic summarization objective may not preserve the evidence most useful for outcome modeling (Wu et al., 2025). This motivates using a summarizer that is not only a fixed preprocessing component, but can utilize prediction feedback from the downstream task.

The tunable LLM summarizer converts each raw discharge note into a shorter visit-level summary. It has two main goals: first, to reduce documentation noise and standardize the note format; second, to preserve patient-specific evidence that is useful for downstream prediction. After supervised initialization, the LLM-aware predictor provides a prediction-feedback reward by converting its per-sample prediction loss into a scalar reward, while the LLM summarizer is the policy model.

Given the discharge note $S_{i,t}$ for patient i at time step t, the summarizer generates a standardized clinical summary $\hat{S}_{i,t}$ using the prompt shown in Figure 2. We then construct the patient-level document as $\hat{S}_i = \hat{S}_{i,T_i} \| \hat{S}_{i,T_i-1} \| \cdots \| \hat{S}_{i,1}$, where $\|$ denotes concatenation. This ordering places the most recent visit first because recent clinical status is often more directly related to near-future outcomes, while earlier visits provide additional historical context.

2.4.1 Summarizer Supervised Fine-Tuning

Before applying reinforcement learning, we first perform supervised fine-tuning (SFT) to teach the local summarizer how to extract clinically relevant information from noisy discharge notes. Reinforcement-based alignment alone can be unstable when starting from an unadapted language model (Ouyang et al., 2022), especially in long-form clinical summarization settings where the output space is large, and rewards may be sparse or noisy (Rafailov et al., 2023). Therefore, SFT first trains the model to produce summaries that follow a structured clinical style before the summarizer is optimized for downstream predictive utility.

Formally, for each visit-level discharge note $\boldsymbol{S}_{i,t}$, we first use the GPT-5.4 mini model to generate a target summary $G_{i,t}$, due to the lack of annotated visit-level summaries. This creates a supervised fine-tuning dataset $\mathcal{D}_{\text{SFT}} = \{(S_{i,t}, G_{i,t})\}$, where $S_{i,t}$ is the input note and $G_{i,t}$ is the teacher-generated target summary used for supervised fine-tuning. We denote the summarizer policy as π_θ, where θ represents the trainable LoRA adapter parameters (Hu et al., 2021), while the backbone LLM weights remain frozen. The teacher-generated summary is written as

$$G_{i,t} = [g_{i,t}^1, g_{i,t}^2, \ldots, g_{i,t}^{L_{i,t}}], \tag{7}$$

where $L_{i,t}$ denotes the length of the target summary for patient i at visit t. At decoding step ℓ, the summarizer predicts the next token $g_{i,t}^\ell$ conditioned on the discharge note $S_{i,t}$ and the previously generated target tokens $g_{i,t}^{<\ell}$. Therefore, the supervised fine-tuning objective is defined as the negative log-likelihood of the GPT-generated target summary:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(S_{i,t},G_{i,t})\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{\ell=1}^{L_{i,t}} \log \pi_\theta\left(g_{i,t}^\ell \mid S_{i,t}, g_{i,t}^{<\ell}\right)\right]. \tag{8}$$

By minimizing $\mathcal{L}_{\text{SFT}}(\theta)$ for several epochs, denoted by $\mathcal{Q}_{\text{SFT}}$, the summarizer learns to generate clinically meaningful summaries before being further optimized with the reinforcement learning objective.

2.4.2 RLPF Optimization with PPO

After supervised fine-tuning, we further optimize the summarizer using Reinforcement Learning from Predictor Feedback (RLPF) (Wu et al., 2025). In this stage, the goal is to move beyond generic clinical summarization and encourage the model to retain information that improves downstream outcome prediction. The tunable summarizer is treated as a policy model π_θ initialized from the SFT-trained checkpoint. Given the discharge note $\boldsymbol{S}_{i,t}$ for patient i at visit t, the policy generates a summary token-by-token:

$$\hat{S}_{i,t} = [\hat{s}_{i,t}^1, \hat{s}_{i,t}^2, \ldots, \hat{s}_{i,t}^{M_{i,t}}] \sim \pi_\theta(\cdot \mid S_{i,t}), \tag{9}$$

where $M_{i,t}$ denotes the length of the generated summary. The generated visit-level summaries

are then concatenated to form the patient-level summary $\hat{S}_i$, following the same construction described in the previous subsection.

The LLM-aware predictor acts as the feedback model. Specifically, the generated patient-level summary $\hat{S}_i$ is passed to the LLM-aware outcome predictor defined in Equation 2, together with the structured EHR soft prompts. The predictor then produces the downstream prediction score following Equations 4 and 5. We use the resulting prediction loss to define the reward for optimizing the summarizer. This feedback helps the summarizer to produce clinical summaries that are more useful for prediction. We optimize the summarizer using Proximal Policy Optimization (PPO) (Schulman et al., 2017).

Let $\ell_i^{(u)}$ denote the per-sample binary cross-entropy loss for patient i at epoch u:

$$\ell_i^{(u)} = \text{BCE}(\boldsymbol{Y}_i^{(u)}, Y_i). \tag{10}$$

For each patient, we maintain a history of previous losses during warm-up in each epoch and define a strong reference loss as the average of the five lowest losses observed so far:

$$\bar{\ell}_{i,5}^{(u)} = \frac{1}{5}\sum_{m=1}^{5} \ell_{i,\text{lowest}}^{(u,m)}, \tag{11}$$

where $\ell_{i,\text{lowest}}^{(u,m)}$ denotes the m-th smallest historical loss for patient i observed during the LLM predictor warm-up stage, before the summarizer is optimized. Using the average of the lowest historical losses provides a reference of the strongest prediction performance observed for that patient during warm-up, and it makes the reward less sensitive to differences across patients. The raw reward is computed as the improvement over this reference:

$$r_i^{(u)} = \bar{\ell}_{i,5}^{(u)} - \ell_i^{(u)}. \tag{12}$$

This reward is used as the prediction-feedback signal for PPO optimization of the summarizer policy; a higher reward indicates that the generated summary improves downstream prediction relative to the patient's warm-up best historical baselines.

To stabilize PPO training, the patient-level rewards are normalized within each mini-batch $\mathcal{B}$. Let $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ denote the mean and standard deviation of rewards in the mini-batch. The normalized reward is

$$\tilde{r}_i^{(u)} = \frac{r_i^{(u)} - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}} + \epsilon}, \tag{13}$$

where ϵ is a small constant for numerical stability. The normalized reward is then clipped to avoid unstable updates:

$$\hat{r}_i^{(u)} = \text{clip}\left(\tilde{r}_i^{(u)}, -r_{\max}, r_{\max}\right). \tag{14}$$

Because the reward is computed at the patient level, the same reward is assigned to all generated visit summaries belonging to that patient. For each generated summary token, we compute the log-probability under the current policy and under the old policy that produced the sampled summary. Let $\pi_{\theta_{\text{old}}}$ denote the policy before the PPO update. The PPO probability ratio for token $\hat{s}_{i,t}^{\ell}$ is

$$\rho_{i,t}^{\ell}(\theta) = \frac{\pi_\theta\left(\hat{s}_{i,t}^{\ell} \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}\right)}{\pi_{\theta_{\text{old}}}\left(\hat{s}_{i,t}^{\ell} \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}\right)}. \tag{15}$$

In addition to the policy model, we use a value head V_ϕ on top of the summarizer hidden states to estimate the expected reward for each generated token. The token-level advantage is computed as

$$A_{i,t}^{\ell} = \hat{r}_i^{(u)} - V_\phi\left(\boldsymbol{h}_{i,t}^{\ell}\right), \tag{16}$$

where $\boldsymbol{h}_{i,t}^{\ell}$ is the hidden state corresponding to token $\hat{s}_{i,t}^{\ell}$. The PPO clipped policy loss is then defined as

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}\Big[\min\Big(\rho_{i,t}^{\ell}(\theta)A_{i,t}^{\ell}, \text{clip}\Big(\rho_{i,t}^{\ell}(\theta), 1-\epsilon_{\text{PPO}}, 1+\epsilon_{\text{PPO}}\Big)A_{i,t}^{\ell}\Big)\Big]. \tag{17}$$

where ϵ_{PPO} is the PPO clipping threshold.

The value head is trained to regress to the normalized and clipped reward:

$$\mathcal{L}_{\text{value}} = \mathbb{E}\left[\left(V_\phi\left(\boldsymbol{h}_{i,t}^{\ell}\right) - \hat{r}_i^{(u)}\right)^2\right]. \tag{18}$$

To prevent the policy from drifting too far from the SFT-initialized behavior, we use a frozen reference policy π_{ref}. We use a sampled KL penalty over the generated tokens to discourage the current policy from drifting too far from the frozen SFT reference policy (Ouyang et al., 2022):

$$\mathcal{L}_{\text{KL}} = \mathbb{E}\Big[\log \pi_\theta\big(\hat{s}_{i,t}^{\ell} \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}\big) - \log \pi_{\text{ref}}\big(\hat{s}_{i,t}^{\ell} \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}\big)\Big]. \tag{19}$$

We also include an entropy bonus to encourage sufficient exploration during summary generation (Zheng et al., 2023):

$$\mathcal{H}_{i,t}^{\ell} = -\sum_{v\in\mathcal{V}} \pi_\theta(v \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}) \log \pi_\theta(v \mid S_{i,t}, \hat{s}_{i,t}^{<\ell}), \tag{20}$$

Table 1: Dataset statistics.

Metric	MIMIC-III	MIMIC-IV
# of patients	7537	15874
# of medications per patient	79.30	113.55
# of diagnoses per patient	28.98	65.99
# of procedures per patient	7.37	9.08
# of history visits per patient	1.65	3.11
Positive rate for readmission	53.7%	53.5%

where $\mathcal{V}$ is the vocabulary.

The final PPO objective combines the clipped policy loss, value loss, entropy regularization, and KL regularization:

$$\mathcal{L}_{\text{PPO}} = \mathcal{L}_{\text{policy}} + \lambda_v \mathcal{L}_{\text{value}} - \lambda_H \mathbb{E}[\mathcal{H}_{i,t}^{\ell}] + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}, \tag{21}$$

where λ_v, λ_H, and λ_{KL} control the contributions of the value loss, entropy bonus, and KL penalty, respectively.

2.5 Staged Predictor–Summarizer Optimization

The RL optimization stage requires a reliable downstream predictor to provide meaningful feedback for the summarizer. When the predictor is under-trained, the reward signal can be noisy and may lead to unstable policy updates. To address this, we first train the predictor for Q_{warm} epochs using $\mathcal{L}_{\text{Pred}}$, and then optimize the summarizer using $\mathcal{L}_{\text{PPO}}$, while the backbone of the predictor LLM is frozen. The two modules use separate optimizers, $\mathcal{O}_{\text{Pred}}$ and $\mathcal{O}_{\text{PPO}}$:

$$\begin{cases} \mathcal{O}_{\text{Pred}}(\mathcal{L}_{\text{Pred}}), & e < Q_{\text{warm}} \\ \mathcal{O}_{\text{Pred}}(\mathcal{L}_{\text{Pred}}), \\ \mathcal{O}_{\text{PPO}}(\mathcal{L}_{\text{PPO}}) & e \geq Q_{\text{warm}} \end{cases} \tag{22}$$

After the warm-up stage, the predictor and summarizer are optimized with separate objectives and optimizers. The predictor continues minimizing $\mathcal{L}_{\text{Pred}}$, while the summarizer is updated using $\mathcal{L}_{\text{PPO}}$ based on the prediction-feedback reward. This staged training provides a more reliable reward signal before the summarizer policy is updated, and reduces the risk of optimizing the summarizer against under-trained predictor feedback.

3 Evaluation

3.1 Datasets and Setups

In this study, we use two real-world datasets to evaluate model performance on medication and readmission prediction tasks.

Table 2: Hyperparameters used in the proposed framework.

Component	Symbol	Value
Soft prompt encoder	P	8
LLM hidden size	D	2304
Max summary length	C_i	512
Reward normalization	ϵ	1×10^{-8}
Reward clipping	$r_{\max}$	2.0
PPO clipping	ϵ_{PPO}	0.2
Value loss weight	λ_v	0.5
Entropy weight	λ_H	0.01
KL weight	λ_{KL}	0.02
Max PPO generation	$M_{i,t}$	128
SFT epochs	Q_{SFT}	10
Predictor warm-up	Q_{warm}	20
Optimizers	$\mathcal{O}_{\text{Pred}}, \mathcal{O}_{\text{PPO}}$	AdamW

• **MIMIC-III** (Johnson et al., 2016) is an open-access database containing health records for over 40,000 patients admitted to critical care units between 2001 and 2012. In this study, we focus on patients with multiple visits and aim to predict binary outcomes.

• **MIMIC-IV** (Johnson et al., 2021) is a publicly available EHR dataset containing records from over 380,000 patients admitted to Beth Israel Deaconess Medical Center between 2008 and 2019. It extends MIMIC-III with a clearer modular structure, richer clinical detail, and improved data provenance.

Both MIMIC-III and MIMIC-IV are publicly available through PhysioNet under controlled-access terms for credentialed researchers. The dataset statistics are summarized in Table 1. In this study, we consider two key healthcare prediction tasks: hospital readmission prediction (binary classification) and medication recommendation (multi-label classification). Given patient information up to visit T_i, the model predicts the outcome at visit T_{i+1} for patient i based on T_i prior visits. Since these tasks require longitudinal data, we excluded patients with only a single recorded visit. We randomly split the data into 70% training and 30% testing sets and report the mean over 4 seeds. The source code is released.[1]

For both the LLM summarizer and the LLM-aware predictor, we utilize the Gemma-2-2B model. All experiments are implemented in Python, using PyTorch (Paszke et al., 2019) as the main deep learning framework. RASPER is also fully compatible with the PyHealth framework (Yang et al., 2023), which we use for imple-

[1]Code: https://github.com/aryahm1375/RASPER.

Table 3: Baseline performance comparison on MIMIC-III and MIMIC-IV. Best results are in bold, and underlined results denote the second-best results that are statistically lower at the 95% confidence level.

Model	MIMIC-III				MIMIC-IV			
	Readmission		Medication		Readmission		Medication	
	AUROC	PRAUC	F1	Jaccard	AUROC	PRAUC	F1	Jaccard
DL-based Methods								
Deepr (Nguyen et al., 2017)	0.671	0.702	0.255	0.160	0.621	0.651	0.228	0.189
RETAIN (Choi et al., 2016)	0.657	0.678	0.253	0.156	0.673	0.698	0.349	0.255
GRASP (Zhang et al., 2021)	0.618	0.635	0.153	0.096	0.535	0.578	0.267	0.195
AdaCare (Ma et al., 2020)	0.645	0.687	0.235	0.133	0.615	0.644	0.251	0.184
StageNet (Gao et al., 2020)	0.678	0.705	0.239	0.139	0.659	0.690	0.278	0.202
Adore (Cheong et al., 2023)	0.665	0.698	0.152	0.125	0.615	0.643	0.206	0.165
ARCI (H Moghaddam et al., 2024)	0.648	0.669	0.283	0.184	0.668	0.698	0.303	0.246
LLM-based Methods								
GraphCare (Jiang et al., 2024)	0.665	0.705	0.294	0.195	0.688	0.711	0.355	0.259
LINKO (Nayebi Kerdabadi et al., 2025)	0.674	0.711	0.304	0.202	0.695	0.715	0.367	0.264
RePrompT (H Moghaddam et al., 2026)	0.682	0.718	0.312	0.206	0.702	0.723	0.375	0.273
RASPER	**0.704**	**0.738**	**0.326**	**0.218**	**0.725**	**0.742**	**0.388**	**0.282**

Table 4: Ablation studies on MIMIC-III and MIMIC-IV. Best results are in bold, and underlined results denote the second-best results that are statistically lower at the 95% confidence level.

Model Variant	MIMIC-III				MIMIC-IV			
	Readmission		Medication		Readmission		Medication	
	AUROC	PRAUC	F1	Jaccard	AUROC	PRAUC	F1	Jaccard
RASPER w/o EHR encoder	0.661	0.698	0.288	0.187	0.685	0.704	0.353	0.258
RASPER w/o Summarization	0.674	0.712	0.298	0.191	0.691	0.713	0.363	0.262
RASPER w/o RL (GPT only)	0.689	0.725	0.309	0.201	0.708	0.723	0.371	0.272
RASPER w/o RL (SFT only)	0.684	0.721	0.305	0.198	0.701	0.718	0.367	0.269
RASPER	**0.704**	**0.738**	**0.326**	**0.218**	**0.725**	**0.742**	**0.388**	**0.282**

menting EHR baselines. For LLM tuning, we use the Hugging Face framework (Wolf et al., 2019). Experiments are conducted on a high-performance server with three NVIDIA A6000 GPUs.

The hyperparameters used in our experiments are shown in Table 2. These values were selected through a greedy search for the best performance.

3.2 Baselines

We compare RASPER with representative deep EHR models, including Deepr (Nguyen et al., 2017), RETAIN (Choi et al., 2016), GRASP (Zhang et al., 2021), AdaCare (Ma et al., 2020), StageNet (Gao et al., 2020), ADORE (Cheong et al., 2023), and ARCI (H Moghaddam et al., 2024). We also compare against recent LLM-based healthcare prediction methods, including GraphCare (Jiang et al., 2024), LINKO (Nayebi Kerdabadi et al., 2025), and RePrompT (H Moghaddam et al., 2026). These baselines cover sequential EHR modeling, ontology-enhanced prediction, patient-similarity learning, knowledge-graph augmentation, and LLM-based EHR representation learning. More information about baselines is available in Appendix A.2.

3.3 Discussion

3.3.1 Performance Comparison on Baselines

As shown in Table 3, the proposed method achieves the best performance across the baselines. RePrompT is the closest competing method, which is expected since it also uses a multi-modal LLM-based design. The improvement over RePrompT demonstrates the effectiveness of the overall RASPER framework for jointly leveraging structured EHR representations and prediction-aligned clinical summaries. The gap between RETAIN, RePrompT, and our method also shows the value of using clinical notes in EHR modeling. Compared with models that rely only on structured codes, adding note information clearly improves

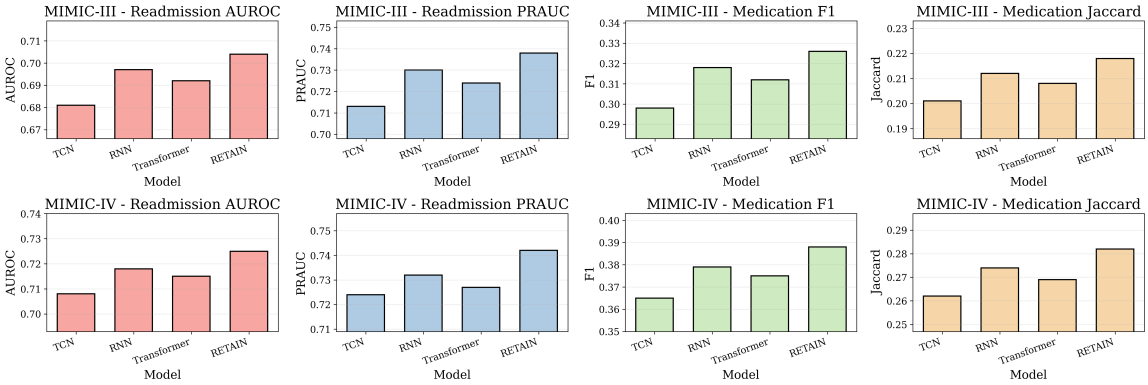

Figure 3: Performance comparison across different EHR models on MIMIC-III and MIMIC-IV.

Table 5: Comparison of LLM backbones on MIMIC-III and MIMIC-IV across two tasks.

Backbone	MIMIC-III				MIMIC-IV			
	Readmission		Medication		Readmission		Medication	
	AUROC	PRAUC	F1	Jaccard	AUROC	PRAUC	F1	Jaccard
Llama-3.2-1B	0.701	0.734	0.321	0.214	0.722	0.737	0.384	0.278
Qwen2.5-0.5B	0.697	0.730	0.315	0.209	0.716	0.731	0.379	0.276
Gemma-2-2B	**0.704**	**0.738**	**0.326**	**0.218**	**0.725**	**0.742**	**0.388**	**0.282**

the learned patient representations. In general, the LLM-based methods perform better than traditional deep learning baselines. Even when prior methods do not directly use clinical-note summarization, they suggest that LLM-based feature extraction (Jiang et al., 2024; Nayebi Kerdabadi et al., 2025) and multimodal learning (H Moghaddam et al., 2026) are useful for prediction.

3.3.2 Ablation Studies on Modules

As shown in Table 4, removing the EHR encoder leads to the largest performance drop, indicating that cohort-level structured information is essential for complementing patient-specific clinical notes. In addition, directly using noisy clinical notes as input leads to worse performance than the proposed method, supporting the main motivation of our paper: raw discharge notes contain redundant and noisy information that can distract the predictor. The GPT-based summarizer improves over raw-note input by reducing this noise, but it still relies on generic summarization objectives and is not explicitly optimized for the clinical prediction task. In contrast, our proposed method further improves performance by aligning the summaries with downstream prediction loss through RL-based optimization. This gap shows that prediction-aligned summarization adds value beyond generic note compression and helps preserve useful information for the final prediction.

3.3.3 Ablation Studies on Encoder Selection

The EHR encoder plays a central role in converting the structured medical codes in a patient's longitudinal history into cohort-level representations. To study this effect, we replace the structured encoder with different EHR modeling backbones and report the results in Figure 3. Among the compared encoders, RETAIN achieves the best overall performance. This is likely because RETAIN captures patient history at multiple levels. As a result, the generated soft tokens provide more informative structured signals to the LLM predictor. The gains are consistent across datasets and evaluation metrics, suggesting that the benefit is not limited to a single task setting. These results also indicate that the proposed framework can benefit from stronger structured EHR encoders.

3.3.4 LLM Model Selection

To examine the sensitivity of RASPER to the choice of LLM backbone, we conduct additional experiments using different LLMs, as shown in Table 5. Since RASPER relies on LLMs for both clinical note summarization and prediction, changing the backbone can affect the final performance. Overall, Gemma-2-2B achieves the best performance across all metrics, suggesting that a stronger backbone can provide better clinical rep-

resentations and improve downstream prediction. However, the performance gaps between Gemma-2-2B, Llama-3.2-1B, and Qwen2.5-0.5B are relatively small in most settings. This indicates that RASPER is not highly dependent on a specific LLM architecture, and the main performance gain comes from the proposed framework.

4 Related Work

A large body of work has studied how to learn predictive representations from Electronic Health Records for clinical decision support. Early neural models such as RETAIN (Choi et al., 2016) improve interpretability by applying reverse-time attention over visits and medical variables, allowing the model to emphasize clinically important parts of a patient's history. Other methods incorporate medical ontologies to reduce sparsity and improve the use of domain knowledge. For example, Hierarchical Attention Propagation (HAP) (Zhang et al., 2020) propagates attention through ontology structures to capture relationships between ancestor and descendant concepts. Temporal disease progression and patient heterogeneity have also been widely studied. AdaCare (Ma et al., 2020) models both short- and long-term variations in clinical signals using multi-scale convolutional modules, while StageNet (Gao et al., 2020) introduces stage-aware modeling to capture disease progression over time. These methods improve clinical prediction by better modeling structured EHR sequences, attention patterns, or disease stages. However, they mainly rely on structured medical codes or numerical features, and therefore do not fully exploit the rich information in unstructured clinical notes.

Recent advances in Large Language Models (LLMs) have motivated their use for EHR representation learning and clinical prediction. GraphCare (Jiang et al., 2024) constructs patient-specific knowledge graphs by combining structured knowledge bases with LLM-generated knowledge, and applies a bi-attention-augmented GNN for prediction. RAM-EHR (Xu et al., 2024) augments patient representations through LLM-powered dense retrieval from multiple knowledge sources, together with consistency regularization to improve robustness. Prompting-based studies (Zhu et al., 2024) further show that carefully designed prompts incorporating EHR-specific information can enable LLMs to perform clinical prediction in few-shot or zero-shot settings. LINKO (Nayebi Kerdabadi et al., 2025) uses LLM-initialized medical concept embeddings and performs dual-axis knowledge propagation, both within individual ontologies and across different ontology systems, and RePrompT (H Moghaddam et al., 2026) incorporates structured EHR encoders into LLMs through prompt tuning, using recurrent visit-level states and cohort-trained soft prompts.

5 Conclusion

We proposed **RASPER**, a reward-aligned framework that integrates structured EHR codes and discharge notes for clinical prediction. RASPER uses EHR soft prompts to capture cohort-level patterns and prediction-aligned summaries to retain useful patient-specific context. By optimizing the summarizer with downstream prediction feedback, RASPER produces summaries that are more directly useful for the target task. Experiments on MIMIC-III and MIMIC-IV show consistent improvements over strong baselines.

6 Limitations

Although RASPER shows strong performance, this study has two limitations. First, the experiments are conducted only on MIMIC-III and MIMIC-IV, so additional evaluation on datasets from other hospitals would better show its generalizability. Second, the current framework focuses on prediction performance and does not deeply analyze how different summary lengths or prompt formats affect the final results.

7 Ethical Considerations

This study uses publicly available, de-identified EHR datasets and does not involve direct interaction with patients. The proposed framework is intended to support clinical prediction research rather than replace clinical judgment. Because generated summaries may omit clinically relevant information or introduce inaccuracies, the proposed framework should not be used as an autonomous clinical decision-making system. We used AI-assisted tools to support language editing, grammar correction, and clarity improvement during manuscript preparation.

8 Acknowledgments

This work was supported by the National Science Foundation under Grant No. IIS-2544634.

References

Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. 2023. Adaptive integration of categorical and multi-relational ontologies with ehr data for medical concept embedding. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–20.

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

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

Senay A Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–7. IEEE.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1167–1168.

Arya H Moghaddam, Mohsen Nayebi Kerdabadi, Mei Liu, and Zijun Yao. 2024. Contrastive learning on medical intents for sequential prescription recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 748–757.

Arya H Moghaddam, Drew Ross, Mohsen Nayebi Kerdabadi, Dongjie Wang, and Zijun Yao. 2026. Reprompt: Recurrent prompt tuning for integrating structured ehr encoders with large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 26766–26778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

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

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. MIMIC-IV. *PhysioNet*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. "note bloat" impacts deep learning-based nlp models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

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

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638.

Mohsen Nayebi Kerdabadi, Arya H Moghaddam, Dongjie Wang, and Zijun Yao. 2025. Multi-ontology integration with dual-axis propagation for medical concept representation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2201–2211.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jiaxing Wu, Lin Ning, Luyang Liu, Harrison Lee, Neo Wu, Chao Wang, Sushant Prakash, Shawn O'Banion, Bradley Green, and Jun Xie. 2025. Rlpf: Reinforcement learning from prediction feedback for user summarization with llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25488–25496.

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

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. 2023. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5788–5789, New York, NY, USA. Association for Computing Machinery.

Gokul Yenduri, M. Ramalingam, G. Chemmalar Selvi, Y. Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G. Deepti Raj, Rutvij H. Jhaveri, B. Prabadevi, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2024. Gpt (generative pre-trained transformer)—a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE access*, 12:54608–54649.

WonJin Yoon, Boyu Ren, Spencer Thomas, Chanhwi Kim, Guergana K Savova, Mei-Hua Hall, and Timothy A Miller. 2025. Aspect-oriented summarization for psychiatric short-term readmission prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28037–28054.

Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. 2024. Prompt engineering paradigms for medical applications: scoping review. *Journal of Medical Internet Research*, 26:e60501.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. 2021. Grasp: generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 715–723.

Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 249–256.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

A Appendix

A.1 RETAIN Encoder

For the structured EHR encoder, we use RETAIN (Choi et al., 2016), which maps longitudinal medical code sequences into dense patient representations through a two-level attention mechanism: Visit-level and variable-level attention. The visit-level attention captures the relative importance of the visit embeddings $\{\boldsymbol{v}_{i,j}\}_{j=1}^{t}$:

$$\{\boldsymbol{q}_{i,j}\}_{j=1}^{T_i} = \mathrm{GRU}_{\alpha}(\{\boldsymbol{v}_{i,j}\}_{j=1}^{T_i}) \tag{23}$$

$$\{\boldsymbol{\alpha}_{i,j}\}_{j=1}^{t} = \mathrm{Softmax}(\boldsymbol{w}_{\alpha}\{\boldsymbol{q}_{i,j}\}_{j=1}^{t}) \tag{24}$$

The variable-level attention assigns feature-wise weights to the visit representation:

$$\{\boldsymbol{z}_{i,j}\}_{j=1}^{t} = \mathrm{GRU}_{\beta}(\{\boldsymbol{v}_{i,j}\}_{j=1}^{t}) \tag{25}$$

$$\{\boldsymbol{\beta}_{i,j}\}_{j=1}^{t} = \tanh(\boldsymbol{w}_{\beta}\{\boldsymbol{z}_{i,j}\}_{j=1}^{t}) \tag{26}$$

The final patient representation is computed by combining the two attention-weighted signals:

$$\boldsymbol{k}_{i,t} = \sum_{j=1}^{t} \boldsymbol{\alpha}_{i,j}\boldsymbol{\beta}_{i,j} \odot \boldsymbol{v}_{i,j} \tag{27}$$

$$\boldsymbol{h}_{i,t} = \boldsymbol{w}_{\mathrm{enc}}\boldsymbol{k}_{i,t} + b_{\mathrm{enc}} \tag{28}$$

The model outputs a sequence of visit-level embeddings $\boldsymbol{h}_{i,t}$, where the final embedding at visit T_i is used as the patient-level EHR representation, denoted by $\boldsymbol{h}_i^{\mathrm{ehr}} = \boldsymbol{h}_{i,T_i}$.

A.2 Baselines

In this research, we compare RASPER with both established healthcare deep learning methods and LLM-based approaches. The deep learning baselines include **Deepr** (Nguyen et al., 2017), which represents patient records as sequences of coded events with time gaps and hospital transfers and applies a CNN for prediction; **RETAIN** (Choi et al., 2016), which uses a dual-RNN attention mechanism to capture the interpretable influence of visits and medical features; **GRASP** (Zhang et al., 2021), which improves EMR-based prediction by identifying clinically similar patients; **AdaCare** (Ma et al., 2020), which captures both short- and long-term biomarker variations and adaptively emphasizes patient-specific risk factors; **StageNet** (Gao et al., 2020), a stage-aware neural network that models disease progression using LSTM and stage-adaptive convolution; **ADORE** (Cheong et al., 2023), which uses attention to adapt medical ontology category embeddings to EHR data for clinical prediction; and **ARCI** (H Moghaddam et al., 2024), which disentangles coexisting temporal medical intents across sequential visits. We also compare RASPER with three recent LLM-based healthcare prediction methods: **GraphCare** (Jiang et al., 2024), which extracts knowledge from LLMs and external biomedical knowledge graphs to construct patient-specific KGs and train a Bi-attention Augmented GNN for clinical prediction; **LINKO** (Nayebi Kerdabadi et al., 2025), which uses LLM-augmented initialization and dual-axis knowledge propagation to jointly learn medical concept representations across multiple ontology graphs; and **RePrompT** (H Moghaddam et al., 2026), which integrates structured EHR encoders with LLMs through time-aware prompt tuning, preserving longitudinal visit information and injecting cohort-level representations via trainable prompt tokens.

A.3 Inference Time Comparison

Table 6: Inference time comparison for a batch of 12 patients.

Model	Time (s)
Deepr	0.09
RETAIN	0.06
GRASP	0.17
AdaCare	0.05
StageNet	0.11
Adore	0.21
ARCI	0.92
GraphCare	0.34
LINKO	0.24
RePrompT	1.38
RASPER	1.87

In Table 6, we compare the proposed method with existing baselines to provide an inference time analysis. While our model is more complex than RePrompT, the observed performance gains suggest that this additional complexity is justified and contributes meaningful predictive benefit.