FLAG: Feature Learning-Enhanced Generation for Large Language Models in Early Warning of ICUs

Anonymous ACL submission

Abstract

Early risk prediction in intensive care units (ICUs) requires properly modeling native electronic health records (EHRs), which are often fraught with pervasive noise. While traditional machine learning (ML) models struggle with cross-ICU transferability, Large Language Models (LLMs) offer a promising alternative but underperform when processing high-dimensional, sparse tabular data. To bridge this gap, we propose **FLAG**, a novel early risk prediction framework tailored for native EHRs. FLAG synergizes the strengths of both paradigms by leveraging SHAP-based feature attributions from ML models to enhance LLM reasoning. Specifically, it transforms native EHRs into an LLM-friendly textual format, annotating noisy features to enrich context, and constructs personalized risk and protective factors to guide the LLM's predictions. Experiments on two real-world datasets demonstrate that FLAG improves the average performance of LLMs by 8.98% on MIMIC-III and 7.76% on PICU, exhibiting exceptional data efficiency and cross-domain generalizability. Our code is available at `https://anonymous.4open.science/r/ehr_prediction-195E/`.

1 Introduction

Intensive care units (ICUs) are dedicated to managing critically ill patients with complex and rapidly evolving conditions, who face an high risk of mortality. Early warning of high-risk patients within the critical 24-hour window after ICU admission is paramount for improving prognosis (Lim et al., 2024; Shi et al., 2025). During this period, patients generate high-dimensional electronic health records (EHRs) encompassing demographics, vital signs and laboratory tests. Distinct from past predictions, EHRs collected in the early stages suffer from pervasive noise, such as missing values and data redundancy (Tang et al., 2024), posing formidable challenges to predictive modeling.

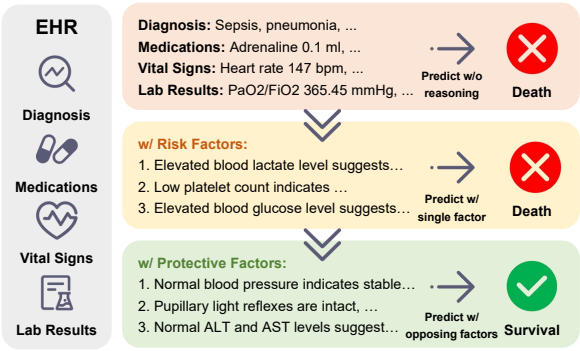

Figure 1: Reasoning of LLMs guided by two factors

Although traditional machine learning (ML) models are widely deployed in clinical prediction due to their proficiency in capturing complex patterns from tabular data (Brown et al., 2024; Chen et al., 2024a), they still exhibit significant limitations in ICU scenarios (Nguyen et al., 2024; Zhang et al., 2022). First, these models heavily rely on specific feature engineering, where applying statistical imputation strategies to handle large-scale missingness is prone to introducing selection bias, thereby masking the clinical information inherently carried by non-random missing values themselves. Second, they are constrained by fixed input schemas and exhibit limited generalization capabilities in small-sample scenarios, hindering their cross-ICU transferability across different EHR systems. This naturally leads to a core question: **How can we leverage noisy native EHRs to achieve robust and transferable risk prediction?**

In recent years, Large Language Models (LLMs) (Achiam et al., 2023) have demonstrated powerful generalization and semantic reasoning capabilities across various medical tasks, such as clinical text classification and summarization (Chen et al., 2025; Zhou et al., 2023; Cabral et al., 2024). This provides an opportunity to utilize LLMs as a unified reasoning engine to interpret native EHRs. However, most existing approaches adopt a hybrid

paradigm (Jiang et al., 2023; Xu et al., 2024; Zhu et al., 2024; Nguyen et al., 2024), where LLMs serve as retrievers of prior knowledge, while the final prediction still depends on traditional models. Such a design fails to fully unleash the reasoning potential of LLMs and instead inherits the inherent limitations of conventional methods.

Furthermore, recent benchmarks indicate that the direct application of LLMs to EHR-based clinical prediction tasks yields unsatisfactory performance (Chen et al., 2024a; Brown et al., 2024). This performance gap highlights the need to enhance the ability of LLMs to model native EHRs. However, achieving this objective still faces the following challenges:

- **C1**: Although LLMs excel at textual inference, they perform poorly when processing high-dimensional and sparse tabular data. The noise within these records severely restricts their reasoning performance (Lovón-Melgarejo et al., 2025; Wang et al., 2024).
- **C2**: The high heterogeneity of ICU patients' conditions demands personalized attention to their clinical features (Zhang et al., 2023), whereas LLMs trained on general corpora struggle to effectively focus on key factors from a comprehensive set of features (Shi et al., 2024).

To address these challenges, we propose **FLAG**, an early risk prediction framework meticulously designed for native EHRs. Specifically, we leverage SHAP techniques to extract the feature contributions learned by ML models and convert them into two types of semantic information, thereby assisting LLMs in understanding EHR data. The core insight is that the complementary strengths of ML models and LLMs can be synergized when processing EHRs. To tackle **C1**, we introduce a method that transforms native EHRs into text format, utilizing specific semantic labels to annotate different types of noise within the records to enrich input semantics. To address **C2**, inspired by clinical risk scoring systems (Quintairos et al., 2023), we further exploit feature contributions to construct two categories of predictive factors along with their corresponding scores, serving as critical cues to guide LLM reasoning (Figure 1).

Our contributions can be summarized as follows:

- We emphasize the necessity of properly modeling native EHRs and internal noisy data for prediction, as well as the potential of LLMs in understanding EHR. To the best of our knowledge, this is the first work to propose and attempt to address this problem.
- We design **FLAG**, an early risk prediction framework for native EHRs, which incorporates a strategy that leverages ML-derived feature information to overcome the reasoning limitations of LLMs on clinical tabular data.
- We conduct extensive experiments on two real-world ICU datasets, demonstrating that FLAG improves the average performance of LLMs by 8.98% on MIMIC-III and 7.76% on PICU datasets. Further investigations validate the effectiveness and generalizability of FLAG.

2 Overview

2.1 Problem Formulation

Definition 1 (EHR Dataset). To simulate early warning (Fullerton et al., 2012), we utilize the EHR data collected within the first 24-hour window after ICU admission. To maximize data completeness, all measurements within this window are aggregated into a single record, represented as $v_{t=24} = \{f_1, f_2, \cdots, f_n\}$, where f_i denotes clinical features, and its value corresponds to the most recent measurement observed within this window.

Definition 2 (Early Risk Prediction). Given the aggregated EHR record $v_{t=24}$, the task is to predict the patient's outcome during the current ICU stay. The label $y \in 0, 1$, where $y = 1$ denotes death and $y = 0$ denotes survival.

2.2 Framework

As shown in Figure 2, **FLAG** includes two stages:

- **# Stage-1: Feature learning** converts native EHR data into textual representations, while modeling potential noisy features within the records into LLM-friendly formats to further enrich the semantic information of EHRs.
- **# Stage-2: Clinical reasoning** leverages key factors and their scores to guide LLM reasoning on EHRs, preventing attention dispersion over high-dimensional data.

2.3 Stage-1: Feature Learning

This stage aims to model the native EHR into an LLM-friendly format to enhance its interpretability for LLMs. To achieve this, we first identify

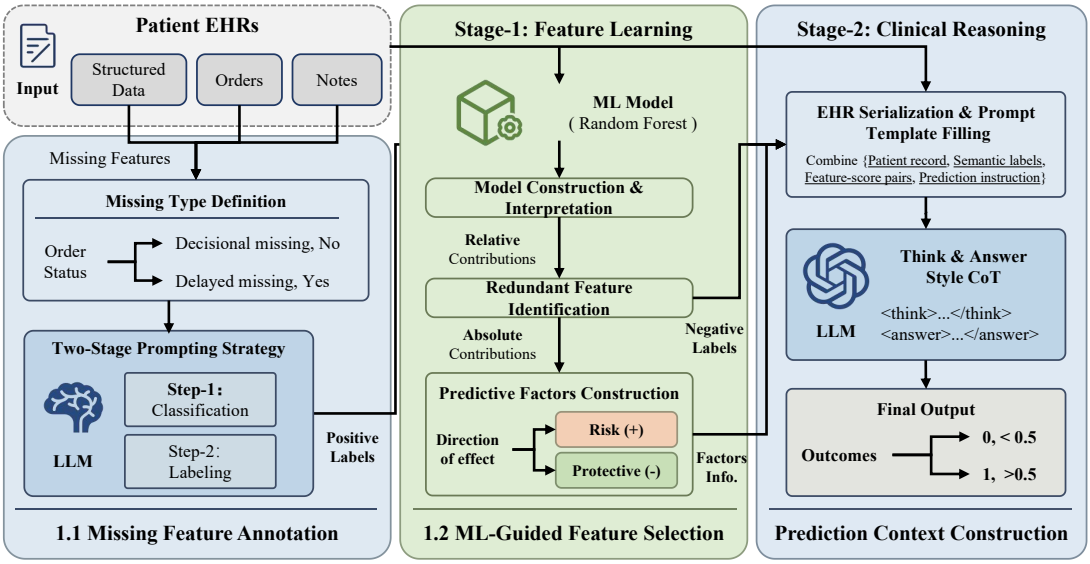

Figure 2: Overall of FLAG.

potential redundant features in high-dimensional numerical data under the guidance of an ML model, and then respectively model redundant and missing features during the transformation of structured EHRs into textual representations.

Missing Feature Annotation. To convert native EHRs into textual format, we concatenate all features in the form of <*Feature name: Feature value*>. Specifically, for demographic variables, we concatenate the variable descriptors (e.g., age, gender) with their corresponding values to form textual inputs such as "Gender: Male." For ICD codes, we map each code to its corresponding disease name. For laboratory values, we generate textual inputs by joining the test name (e.g., "Glucose"), the measured value (e.g., "116"), and the unit (e.g., "mg/dL"). For redundant or missing features, we retain them but annotate with explicit labels, such as <*Irrelevant feature*> and <*Missing feature*>. This design preserves the original semantics of the EHR and avoids the bias introduced by feature removal or imputation, thus enabling interpretable predictions.

ML-Guided Feature Selection. To appropriately model the numerical component of EHRs rather than merely converting them, we employ a machine learning (ML) model combined with SHapley Additive exPlanations (SHAP) [1] to identify potentially redundant features.

- **Model Construction & Interpretation.** We first instantiate an ML model on the training set and tune it based on the internal validation set. Then, for each patient in the test set, we perform online prediction using the trained model and extract feature contributions via SHAP. This approach enables dynamic, patient-specific interpretability and thus facilitates personalized prediction.

- **Redundant Feature Identification.** The strategy for identifying redundant features is as follows: First, each feature importance score w_i from the SHAP-based contributions is normalized via L1 normalization. Next, the features are ranked in ascending order based on their normalized contribution $\hat{w}_i$. Finally, an iterative procedure is applied to progressively select features with the lowest contributions until the cumulative contribution of the removed features surpasses a predefined threshold τ. Features whose cumulative contribution is below this threshold τ are identified as irrelevant features, denoted by $\mathcal{F}_{\text{irrelevant}}$. The complete procedure is detailed in Algorithm 1.

2.4 Stage-2: Clinical Reasoning

This stage aims to guide the reasoning process of LLMs on EHR with significant features identified by ML models. To achieve this, we compute an influence score for each feature and extract those with strong impacts as key factors. These factors help LLMs focus on meaningful information, while two types of factors with opposite effects on mor-

[1] https://github.com/shap/shap

Algorithm 1 Redundant Feature Selection

Input: Feature set $\mathcal{F}$, feature importance score w_i, threshold τ
Output: Irrelevant feature set $\mathcal{F}_{\text{irrelevant}}$
Step 1: Normalization
$\hat{w}_i \leftarrow |w_i| / \sum_{j=1}^{n} |w_j|$ for all $f_i \in \mathcal{F}$
Step 2: Rank
Sort $\mathcal{F}$ in ascending order of $\hat{w}_i$
Step 3: Iterative selection
$\mathcal{F}_{\text{irrelevant}} \leftarrow \emptyset$, $C \leftarrow 0$
for each $f_i \in \mathcal{F}$ in sorted order **do**
    $C \leftarrow C + \hat{w}_i$
    if $C > \tau$ **then break**
    end if
    $\mathcal{F}_{\text{irrelevant}} \leftarrow \mathcal{F}_{\text{irrelevant}} \cup \{f_i\}$
end for
Return $\mathcal{F}_{\text{irrelevant}}$

tality risk are introduced to encourage reasoning from complementary perspectives.

- **Predictive Factors Construction.** Inspired by clinical risk scoring systems, we transform SHAP-based feature contributions into a set of discrete features with integer scores. This design avoids introducing additional numerical complexity, enabling LLMs to better understand and leverage the information.

- **Predictive Reasoning Generation.** Unlike prior approaches that only consider risk factors, we provide LLMs with two types of clinical factors: *risk factors* that increase mortality risk and *protective factors* that decrease it. This dual-factor design allows LLMs to reason about both adverse and favorable conditions, thereby enhancing interpretability.

Finally, we construct a chain-of-thought prompt based on the EHR of patient and the aforementioned information. Detailed examples of our prompt design are provided in the Appendix.

3 Experiments

3.1 Experiment Setups

Datasets. We conduct experiments on the public MIMIC-III dataset (Johnson et al., 2016) and a privately-owned PICU dataset. Specifically, we perform adult mortality prediction on the former and pediatric mortality prediction on the latter.

Table 1: Statistics of the datasets.

Stats	MIMIC-III	PICU
# of health records	2440	1809
# of clinical features	114	98
# of mortality rate	17.0%	9.7%
# of age range	18-88	0-17

The PICU (Pediatric Intensive Care Unit) dataset was collected from 54 pediatric intensive care infection centers in China, comprising de-identified EHRs from 2022 to 2023.

In this study, we focus on patients with infections in the ICU, as the mortality prediction in severe infections remains a primary focus of intensive care (Weng et al., 2021). Table 1 provide detailed statistics for two datasets. We then employ stratified strategy to randomly split the datasets into training, validation, and test sets (7:1:2), ensuring consistency with the real-world distribution. Further dataset details can be found in Appendix B.

Evaluation Metrics. Following Choi et al. (2017), we employ the Macro-F1 score, AUROC (Area Under the Receiver Operating Characteristic curve), and AUPRC (Area Under the Precision-Recall Curve) as our evaluation metrics.

Baselines. We compare FLAG with a wide range of baselines from three perspectives.

- **ML-based methods.** We include widely used ML algorithms as baseline, such as Logistic Regression, Random Forest, XGBoost, and Neural Networks.

- **Prompt-based methods.** We include a Direct baseline, where the model directly outputs the final answer without any intermediate reasoning. We also include Think then Answer-style CoT(Guo et al., 2025), which prompts the model to reason inside <think> tags before outputting a final answer in <answer>.

- **Open-source LLMs and backbone variants.** To assess the generality of FLAG, we instantiate it on multiple backbone models, including GPT-4o series (Achiam et al., 2023), Qwen2.5 series (Yang et al., 2024), Llama3 series (Dubey et al., 2024), Gemma-2 (Team et al., 2024), Mistral-v0.3 (Chaplot, 2023). We also compare against prominent open-source medical LLMs such as Apollo series (Zheng et al., 2024) and MMed-Llama-3.

Implementation Details. In FLAG, we set the top-k features to 0.33 and select Random Forest with SHAP method. The choice of top-k, ML models, and model explainability method are thoroughly analyzed in section 4.

We leveraged vLLM [2] to accelerate the inference of general LLMs and directly loaded medical LLMs. The following hyperparameters were configured: $max_new_tokens = 4096$; $repetition_penalty = 1.05$, and $temperature = 0.01$. Each experiment was executed three times, and the results were averaged to ensure reliability. More details are listed in Appendix C.

3.2 Main Experimental Results

Table 2 presents our evaluation results of various models on two datasets.

(1) ML models generally outperform prompt-based methods. On the MIMIC-III dataset, ML models take the lead, while on the PICU dataset, CoT-based approaches using Qwen2.5-14B, Meta-Llama-3.3-70B, and GPT-4o obtain comparable results to the strongest ML models. This highlights the necessity of properly modeling native EHR data for LLMs. Notably, all AUPRC scores below 50%, indicating that achieving accurate early risk prediction in ICUs remains a challenging task.

(2) Leading general LLMs outperform medical LLMs. While general LLMs, such as Gemma-2-9B, Qwen2.5-14B, Meta-Llama-3-70B, and GPT-4o, demonstrate competitive performance, existing medical LLMs perform poorly. Even the best-performing medical LLMs, Apollo-7b, achieves average scores of only 55.26% and 49.45% on the two datasets, respectively. Additionally, general LLMs benefit more from our approach than medical LLMs due to their stronger contextual understanding capabilities.

(3) Our methods combines data and medical reasoning. In zero-shot setting, FLAG improves the average scores of LLMs on two datasets by 8.98% and 7.76%, outperforming both direct prediction and CoT methods. Additionally, we observe that on the MIMIC-III dataset, FLAG performs comparably to ML models, while on the PICU dataset, FLAG achieves superior average performance, particularly in terms of the AUPRC score. We attribute this to the greater imbalance in the PICU dataset, where fewer high-risk patient samples limit the performance of data-driven ML models. In contrast, our method leverages the medical knowledge and reasoning capabilities of LLMs, demonstrating better generalization ability. Additionally, we observed that our method offers limited improvement for Llama series LLMs and even shows a slight decline on MIMIC-III dataset. We speculate that Llama series LLMs tend to rely more on their internal knowledge for prediction.

(4) The CoT method may impair performance, which is contrary to our expectations. We observed that the CoT method resulted in varying degrees of performance degradation for most LLMs except Qwen2.5-7B, Qwen2.5-14B, MMed-Llama-3-8B, and Meta-Llama-3.3-70B. We hypothesize that this may be due to the increased complexity of ICU mortality prediction, which requires stronger medical reasoning capabilities, while the CoT training data of Qwen and Llama series LLMs may contain similar downstream tasks. Notably, Llama series LLMs autonomously utilize clinical risk scores during reasoning to assist in prediction.

4 Detailed Study

4.1 Ablation Study

To evaluate the effectiveness of each design in our framework, we conducted comprehensive ablation studies on both datasets. We introduced three variants to investigate the impact of each component on model performance as follows:

- **w/ Risk Factors Only**: This variant only provides the risk factors for LLMs to validate the necessity of protective factors.
- **w/ LIME (Ribeiro et al., 2016)**: This variant substitutes SHAP with LIME (Local Interpretable Model-agnostic Explanations) to compare the explanation quality of different model explanation methods.
- **w/ Other ML Models**: We replace Random Forest used in the main experiments sequentially with Logistic Regression, XGBoost, and Neural Networks to compare the feature capture capabilities of different ML algorithms.

As shown in Table 3, we observe that all components contribute to the improvement. (i) Directly removing factors leads to a significant decline in overall performance, particularly in terms of F1 Score and AUPRC. (ii) Providing only risk factors results in an average performance degradation, specifically manifested as an increase in AUROC but decreases

[2] https://github.com/vllm-project/vllm

Table 2: Aggregated performance across all metrics on MIMIC-III and PICU. "SL" represents Supervised Learning. **Bold** denotes the best performance. Underline denotes the second performance. Asterisk (*) denotes the best method.

Model	Settings	MIMIC-III				PICU			
		Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.	Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.
Traditional Machine Learning Models									
Logistic Regression	SL	64.65	72.73	40.68	59.35	58.93	69.89	23.57	50.80
Random Forest	SL	65.94	73.40	42.07	60.47†	62.05	76.54	33.69	57.43†
XGBoost	SL	65.85	73.15	42.08	60.36	62.05	75.62	33.40	57.02
Neural Network	SL	65.01	71.08	41.04	59.04	58.33	68.87	25.78	50.99
Open-source General Large Language Models									
Qwen2.5-7B (Yang et al., 2024)	Direct	47.55	65.89	23.90	45.78	38.48	67.44	14.17	40.03
	CoT	49.59	69.27	27.65	48.84	42.15	74.22	19.71	45.36
	Ours	64.09	72.06	30.66	55.60†	54.82	77.45	30.53	54.27†
Mistral-7B-v0.3 (Chaplot, 2023)	Direct	26.41	68.22	27.69	40.77	22.94	65.96	24.56	37.82
	CoT	47.55	49.72	18.86	38.71	40.83	50.00	20.62	37.15
	Ours	48.79	72.59	39.36	53.58†	47.46	73.63	29.67	50.25†
Meta-Llama-3.1-8B (Dubey et al., 2024)	Direct	52.90	68.32	26.01	49.08	20.70	59.81	14.46	31.66
	CoT	47.82	66.36	29.16	47.78	20.52	50.97	10.58	27.36
	Ours	48.55	72.29	30.27	50.37†	30.27	78.50	28.47	45.75†
Gemma-2-9B (Team et al., 2024)	Direct	55.55	64.04	23.76	47.78	31.62	72.85	28.02	44.16
	CoT	51.17	64.43	23.9	46.50	23.61	71.89	16.73	37.41
	Ours	63.75	74.10	40.82	59.56†	59.71	74.91	31.47	55.36†
Qwen2.5-14B (Yang et al., 2024)	Direct	64.63	70.98	34.64	56.75	58.82	77.75	32.32	56.30
	CoT	64.16	76.30	40.09	60.18	58.11	82.34	34.47	58.31
	Ours	66.73	**77.55**	41.34	61.87†	59.18	**83.74**	35.52	59.48†
Meta-Llama-3.3-70B (Dubey et al., 2024)	Direct	65.38	73.12	32.24	56.91	56.12	80.39	25.00	53.84
	CoT	63.99	74.59	36.67	58.42†	57.64	80.96	27.43	55.34
	Ours	62.24	75.30	34.80	57.45	58.87	82.38	33.70	58.32†
Closed-source Large Language Models									
GPT-4o-mini (Achiam et al., 2023)	Direct	51.96	72.60	31.82	52.13	32.84	74.54	21.43	42.94
	CoT	29.65	71.27	29.97	43.63	19.61	74.08	19.58	37.76
	Ours	62.44	75.43	42.66	60.18†	45.58	75.55	30.11	50.41†
GPT-4o (Achiam et al., 2023)	Direct	63.11	75.45	36.54	58.37	59.20	77.05	33.00	56.57
	CoT	59.41	74.37	38.99	57.59	56.64	76.13	39.45	57.41
	Ours	**71.26**	76.42	**43.14**	**63.61**†	**62.53**	78.92	**41.43**	**60.96**†
Open-source Medical Large Language Models									
Apollo2-7B (Zheng et al., 2024)	Direct	62.29	66.12	28.50	52.30	40.16	69.20	19.19	42.85
	CoT	59.34	66.25	27.69	51.09	50.77	46.66	11.98	36.47
	Ours	60.86	70.05	34.86	55.26†	57.24	66.18	24.93	49.45†
MMed-Llama-3-8B	Direct	48.79	45.61	19.04	37.81	52.57	52.26	11.97	38.93
	CoT	49.27	51.00	19.23	39.83†	52.17	53.29	13.16	39.54
	Ours	46.59	50.81	19.06	38.82	57.46	53.01	19.67	43.38†
Apollo2-9B (Zheng et al., 2024)	Direct	45.35	45.67	16.01	35.68	49.58	55.99	12.09	39.22
	CoT	49.84	50.20	20.68	40.24	35.77	57.84	11.54	35.05
	Ours	52.74	50.60	21.18	41.51†	53.14	59.71	21.46	44.77†

in F1-score and AUPRC. This indicates that considering only risk factors improves the LLMs' ability to identify majority classes, yet reduces its accuracy in distinguishing minority classes (at-risk patients), which appears counter-intuitive. (iii) The variant using the LIME explanation technique performed worse than using SHAP, possibly due to the higher quality of feature explanation results provided by the SHAP algorithm. (iv) We observe that the variant using XGBoost demonstrates competitive performance, as both XGBoost and Random Forest are decision tree-based algorithms. In contrast, the variants employing Neural Networks exhibit significantly inferior performance due to the inherent interpretability challenges. Notably, these variants all outperformed the baseline, further demonstrating the effectiveness of our proposed framework.

4.2 Analysis on the Number of Shots

To systematically investigate the impact of the number of shots on mortality prediction, we conducted experiments with different shot settings across three methods. A subset of patient data (N=20) was randomly selected from the training set to build an example library, ensuring a balanced ratio of deceased to survived cases (1:1). In contrast to the conclusions drawn on other tasks, we observed that adding shots reduced the average performance of LLMs in this task. In particular, the macro F1 scores of almost all methods significantly decreased. We attribute this phenomenon to two potential reasons: (1) The selection of shots inevitably introduces bias into LLMs' predictions, and ICU mortality prediction is a task highly sensitive to data distribution, where such bias can be amplified and even become harmful; (2) The increase in the number of shots leads to rapid growth of context

Table 3: Ablation study of FLAG on both datasets. **Bold** denotes the best performance.

Method	MIMIC-III				PICU			
	Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.	Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.
FLAG	**71.26**	76.42	**43.14**	**63.61**	62.53	78.92	**41.43**	**60.96**
w/o Feature Selection Strategy	67.22	77.05	41.83	62.03	**64.63**	76.36	38.02	59.67
w/o Facors Scoring Mechanism	61.61	75.78	38.52	58.64	60.19	78.76	40.59	59.85
w/o All Designs	63.11	75.45	36.54	58.37	59.20	77.50	33.00	56.57
Variant								
w/ Risk Factors Only	65.05	78.15	40.31	61.17	59.45	81.35	40.29	60.36
w/ LIME	66.52	76.55	37.54	60.20	59.68	78.80	35.27	57.92
w/ Logistic Regression	65.60	77.77	39.84	61.07	60.99	**83.24**	34.79	59.67
w/ XGBoost	69.28	**78.64**	42.84	63.59	60.16	80.56	36.02	58.91
w/ Neural Network	69.04	76.52	40.37	61.98	62.68	70.42	30.98	54.69

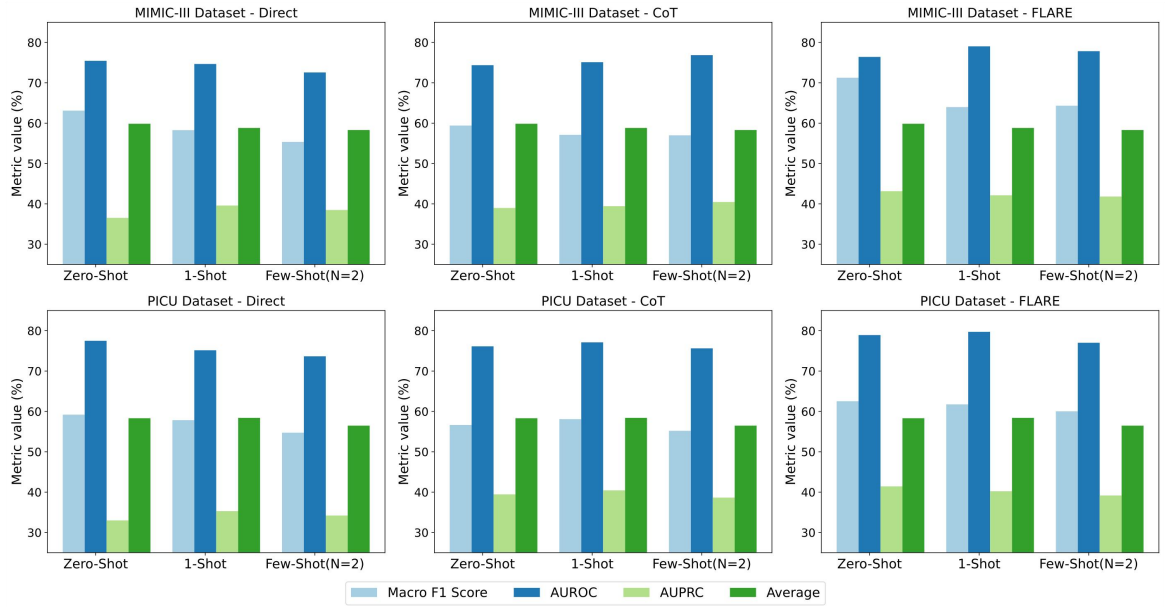

Figure 3: Performance comparison under different shot settings on MIMIC-III and PICU datasets.

length while introducing more noisy data. Notably, compared with other methods, FLAG demonstrates relatively stable performance and achieves optimal results under various settings, indicating better robustness of our approach.

4.3 Analysis of Top-K Features as Factors

We explored the impact of the number of factors by varying the top-k value on two datasets. The MIMIC-III dataset showed higher sensitivity, while the PICU dataset remained more stable, likely due to the more concentrated distribution of pediatric data compared to adult data. In Figure 5, both datasets revealed an inflection point around 33%, indicating optimal performance. As the value of top-k increased, AUROC showed minimal fluctuation, while AUPRC decreased more noticeably, suggesting that an increased number of low-contribution features negatively impacts the ability to identify high-risk patients. This aligns with our expectations: a low value of top-k leads to the loss of key feature information, while a high value of top-k significantly increases the number of low-contribution features, interfering with the final prediction. Notably, when all features are included, the average performance of LLMs declines rapidly, even falling below the baselines.

4.4 Data Efficiency

In the medical field, due to strict ethical regulations and the high cost of annotation, labeled EHR data is often scarce. We here investigate the data efficiency of our framework under low data availability conditions. Specifically, we utilized only a portion of the training set, with data sizes of 20%, 40%, 60%, 80%, and 100%, to train the machine learning models. As shown in Figure 6, FLAG outperforms the baselines across all training data

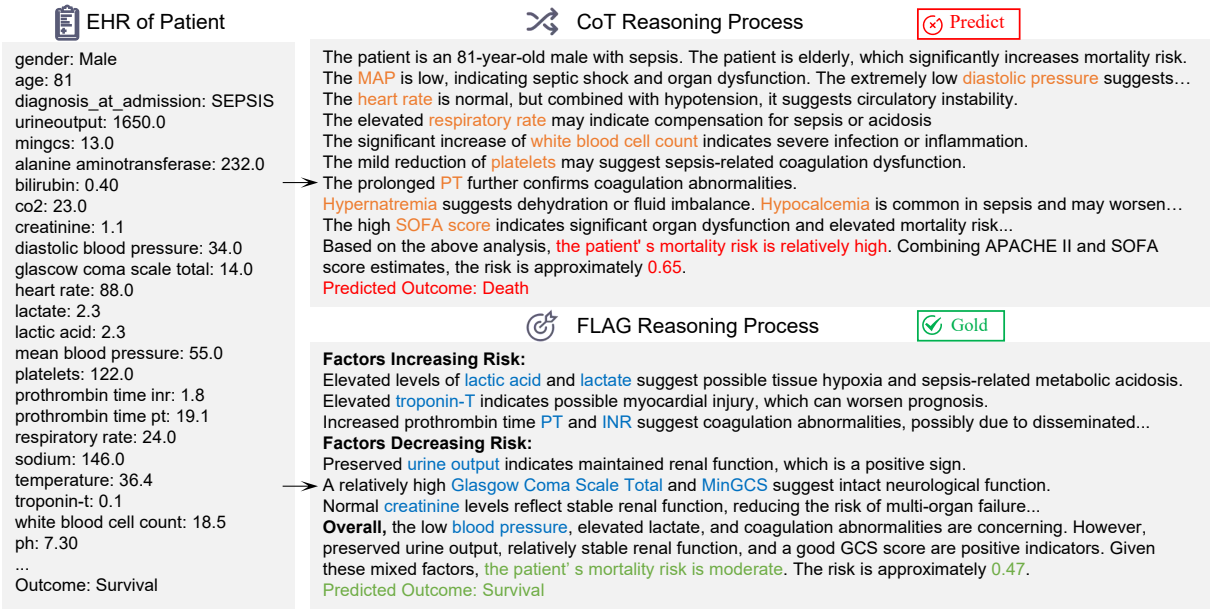

Figure 4: Comparison of examples for two reasoning methods. Features involved in the reasoning process are marked in different colors. Green indicates correct results, red indicates incorrect results.

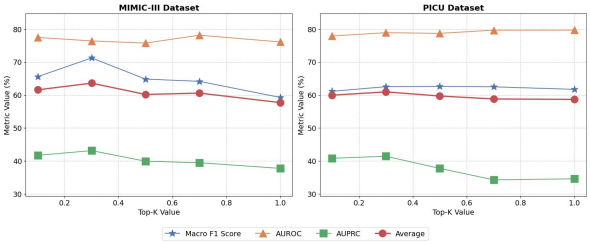

Figure 5: Performance comparison of FLAG across different numbers of factor on both datasets.

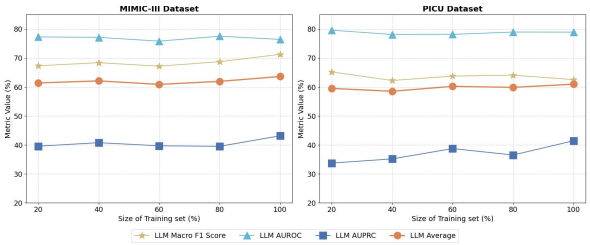

Figure 6: Performance comparison of FLAG across different size of training set on both datasets.

scales. Additionally, we observe that the decline in the framework's average performance remains relatively modest as the amount of training data decreases. Even with only 20% of the training data, the framework still achieves F1 scores and AUROC close to the optimal setting. This highlights the generalization capability of our framework, which incorporates both influence factors and the reasoning ability of LLMs, instead of relying solely on data-driven machine learning methods.

4.5 Case Study

We present examples of the CoT and FLAG reasoning process, as illustrated in Figure 4. We observe that the reasoning of CoT typically analyzes all features sequentially according to clinical categories, ignoring varying priorities among features. Moreover, it exhibits confirmation bias, where LLMs selectively seek evidence that supports their predictions. In contrast, FLAG provides feature information to guide LLMs in reasoning. By ignoring noisy features and prioritizing the analysis of predictive factors, it provides a more transparent and reliable basis.

5 Conclusion

In this paper, we proposed **FLAG**, a novel early risk prediction framework for native EHRs in ICUs. To overcome the inherent challenges LLMs face with noisy, high-dimensional tabular data, FLAG leverages SHAP-based feature attributions from ML models to guide LLM reasoning. By transforming EHRs into an annotated textual format and constructing personalized risk and protective factors, FLAG successfully bridges the gap between ML interpretability and LLM semantic reasoning. Extensive experiments on the MIMIC-III and PICU datasets confirm that FLAG significantly improves prediction performance, data efficiency, and cross-domain generalizability, offering a robust and transferable solution for clinical prediction.

Limitations

Although FLAG has achieved encouraging results, we acknowledge several limitations in this study. First, we conducted mortality prediction tasks on two ICU datasets, but the method has the potential to be applied to a broader range of clinical prediction tasks, including readmission prediction, disease prediction, and health event prediction. Second, our experimental results indicate a correlation between transfer efficiency and explanation quality, suggesting that the performance of FLAG may be influenced by the choice of different model explanation methods. Finally, our study focuses solely on predictions based on textual EHR data, and its applicability to multimodal data remains to be explored in future work.

Ethical Considerations

To ensure the ethical use of credential data (MIMIC-III and PICU) with GPT-based online services, we have signed and strictly adhered to the PhysioNet Credentialed Data Use Agreement[3]. To prevent sensitive information from being shared with third parties, we follow the guidelines[4] for responsible use of EHR data in online services. This includes opting out of human review of the data through the Azure OpenAI Additional Use Case Form[5]. Additionally, the examples we used in the paper were all synthesized using LLMs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. 2024. Not the models you are looking for: Traditional ml outperforms llms in clinical prediction tasks. *medRxiv*, pages 2024–12.

Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. 2024. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine*, 184(5):581–583.

[3] https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150

[4] https://physionet.org/news/post/gpt-responsible-use

[5] https://aka.ms/oai/additionalusecase

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*.

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

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

James N Fullerton, Charlotte L Price, Natalie E Silvey, Samantha J Brace, and Gavin D Perkins. 2012. Is the modified early warning score (mews) superior to clinician judgement in detecting critical illness in the pre-hospital environment? *Resuscitation*, 83(5):557–562.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. 2024. When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? *arXiv preprint arXiv:2408.11854*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma,

Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.

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

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):1–9.

HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, et al. 2024. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint arXiv:2404.05144*.

Leerang Lim, Ukdong Gim, Kyungjae Cho, Dongjoon Yoo, Ho Geol Ryu, and Hyung-Chul Lee. 2024. Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation. *Critical Care*, 28(1):76.

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G Moreno, and Lynda Tamine. 2024. Revisiting the mimic-iv benchmark: Experiments using language models for electronic health records. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 189–196.

Jesús Lovón-Melgarejo, Martin Mouysset, Jo Oleiwan, Jose G Moreno, Christine Damase-Michel, and Lynda Tamine. 2025. Evaluating llm abilities to understand tabular electronic health records: A comprehensive study of patient data extraction and retrieval. In *European Conference on Information Retrieval*, pages 153–168. Springer.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752.

Tuan Nguyen, Thanh Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2024. Carer-clinical reasoning-enhanced representation for temporal health risk prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10392–10407.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Amanda Quintairos, David Pilcher, and Jorge IF Salluh. 2023. Icu scoring systems. *Intensive Care Medicine*, 49(2):223–225.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Songchang Shi, Lihui Zhang, Shujuan Zhang, Jinyang Shi, Donghuang Hong, Siqi Wu, Xiaobin Pan, and Wei Lin. 2025. Developing a rapid screening tool for high-risk icu patients of sepsis: integrating electronic medical records with machine learning methods for mortality prediction in hospitalized patients—model establishment, internal and external validation, and visualization. *Journal of Translational Medicine*, 23(1):97.

Tongyue Shi, Jun Ma, Zihan Yu, Haowei Xu, Minqi Xiong, Meirong Xiao, Yilin Li, Huiying Zhao, and Guilan Kong. 2024. Stochastic parrots or icu experts? large language models in critical care medicine: A scoping review. *arXiv preprint arXiv:2407.19256*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Alice S Tang, Sarah R Woldemariam, Silvia Miramontes, Beau Norgeot, Tomiko T Oskotsky, and Marina Sirota. 2024. Harnessing ehr data for health research. *Nature Medicine*, 30(7):1847–1855.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Jiankun Wang, Sumyeong Ahn, Taykhoom Dalal, Xiaodan Zhang, Weishen Pan, Qiannan Zhang, Bin Chen, Hiroko H Dodge, Fei Wang, and Jiayu Zhou. 2024. Augmented risk prediction for the onset of alzheimer's disease from electronic health records with large language models. *arXiv preprint arXiv:2405.16413*.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235.

Jie Weng, Ruonan Hou, Xiaoming Zhou, Zhe Xu, Zhiliang Zhou, Peng Wang, Liang Wang, Chan Chen, Jinyu Wu, and Zhiyi Wang. 2021. Development and validation of a score to predict mortality in icu patients with sepsis: a multicenter retrospective study. *Journal of translational medicine*, 19:1–12.

Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2023. Hypergraph transformers for ehr-based clinical predictions. *AMIA Summits on Translational Science Proceedings*, 2023:582.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2414–2423.

Zhihao Yu, Yujie Jin, Yongxin Xu, Xu Chu, Yasha Wang, and Junfeng Zhao. 2024. Intellicare: Improving healthcare analysis with variance-controlled patient-level knowledge from large language models. *arXiv preprint arXiv:2408.13073*.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. 2022. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature biomedical engineering*, 6(12):1330–1345.

Jin-jin Zhang, Xiang Dong, Guang-hui Liu, and Yadong Gao. 2023. Risk and protective factors for covid-19 morbidity, severity, and mortality. *Clinical reviews in allergy & immunology*, 64:90–107.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

A Related work

External Knowledge Integration in Healthcare Prediction

In the healthcare domain, many studies have attempted to leverage external knowledge, such as medical ontologies, knowledge graphs, or unstructured text, to enhance representation learning and thereby improve predictive performance. Initial efforts focused on mining structured data. For instance, Choi et al. (2017); Ma et al. (2018) supplemented EHRs with medical ontology information through attention mechanisms. Yang et al. (2023); Xu et al. (2023); Jiang et al. (2023) leveraged graphs to understand complex relationships among medical concepts to support predictive tasks. Later approaches employed unstructured text as auxiliary information. For example, Ye et al. (2021); Zhu et al. (2024); Xu et al. (2024) used LLMs to summarize existing external medical databases, including knowledge graphs and medical literature, to provide more comprehensive knowledge for each medical entity. Alternatively, Nguyen et al. (2024); Yu et al. (2024) employed LLMs to analyze patient health conditions, incorporating generated clinical reasoning into EHR modeling. However, knowledge at the medical code level does not directly guide predictive tasks, and knowledge directly generated by LLMs faces challenges of hallucination. In contrast, our proposed method uses feature information directly related to predictions from expert models to improve clinical risk prediction.

Large Language Models in Healthcare Domain

LLMs have demonstrated impressive performance across various general tasks. Recently, an increasing number of studies have applied LLMs to the healthcare domain (Peng et al., 2023; He et al., 2025), particularly in medical question answering (**?**Han et al., 2023), summarization generation (Ghosh et al., 2024; Van Veen et al., 2024), and clinical text embedding tasks (Gao et al., 2024; Singhal et al., 2023). For instance, Jung et al. (2024) utilized LLMs to generate discharge summaries for cardiac patients, improving medical documentation workflows to enhance clinical efficiency. Chen et al. (2024b) introduced HuatuoGPT-o1, enabling LLMs to perform complex reasoning to improve the resolution of medical problems. They explored the application of LLMs in embedding and generation tasks, highlighting their potential to support clinical decision-making. However, the application of LLMs to EHR-based clinical prediction tasks remains underexplored. Although some studies have evaluated the performance of LLMs in EHR-based clinical prediction tasks (Chen et al., 2024a; Lovon-Melgarejo et al., 2024), research on how to improve LLMs' performance in clinical prediction tasks is still scarce. Our study addresses this gap by exploring the application of LLMs in clinical risk prediction tasks and proposing novel methods.

B Dataset Details

We selected patients with infections based on their ICU admission diagnoses, with the scope of diseases shown in Table 4. Then, we preprocess the MIMIC-III dataset following the approach described by (Wang et al., 2020). In addition, we provided supplementary statistical information related to these groups.

Table 4: Details of two datasets.

Dataset	MIMIC-III	PICU
Average Age	65.28	4.68
Gender (Male:Female)	1352:1088	1090:719
Infectious Disease Scope	Sepsis, Meningitis, Osteomyelitis, Peritonitis, Endocarditis, Myocarditis, Pneumonia, Pyelonephritis, Cholecystitis, ...	

C Implementation Details

We conducted all experiments on a Linux server equipped with NVIDIA A100 GPUs. For the usage of the GPT series models, we used Azure OpenAI Service, ensuring sensitive patient data was processed in a protected environment.

Model Training. The ML models were trained via twenty epochs on the training set. We optimize models based on their performance on the validation set. The macro-F1 score is designated as the primary metric due to the class imbalance in the dataset.

Table 5: Hyperparameters for ML Model Training.

Model	Hyperparameters
LogisticRegression	`max_iter=1000`
RandomForest	`n_estimators=100, max_depth=6`
XGBoost	`eval_metric='logloss', max_depth=6`
NeuralNetwork	`hidden_layer_sizes=(100,), max_iter=1000`

D Supplementary Experimental Results

Table 6: Performance comparison under different shot settings on two datasets. **Bold** denotes the best performance.

Method	Settings	MIMIC-III				PICU			
		Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.	Macro F1 Score (%)	AUROC (%)	AUPRC (%)	Avg.
Direct	Zero-Shot	63.11	75.45	36.54	58.37	59.20	77.50	33.00	56.57
	1-Shot	58.28	74.67	39.59	57.51	57.84	75.15	35.30	56.10
	Few-Shot(N=2)	55.36	72.57	38.48	55.47	54.73	73.66	34.20	54.20
CoT	Zero-Shot	59.41	74.37	38.99	57.59	56.64	76.13	39.45	57.41
	1-Shot	57.11	75.11	39.44	57.22	58.09	77.12	40.44	58.55
	Few-Shot(N=2)	56.99	76.86	40.46	58.10	55.20	75.63	38.65	56.49
Ours	Zero-Shot	**71.26**	76.42	**43.14**	**63.61**	**62.53**	78.92	**41.43**	**60.96**
	1-Shot	63.99	**79.04**	42.13	61.72	61.75	**79.72**	40.23	60.57
	Few-Shot(N=2)	64.33	77.84	41.82	61.33	60.03	77.01	39.17	58.74

Table 7: Performance across different parameter settings on MIMIC-III and PICU datasets. **Bold** denotes the best performance.

Top-K Value	MIMIC-III			**Avg.**	PICU			**Avg.**
	Macro F1 Score (%)	AUROC (%)	AUPRC (%)		Macro F1 Score (%)	AUROC (%)	AUPRC (%)	
0.1	65.58	77.48	41.71	61.59	61.14	77.91	40.82	59.96
0.3	**71.26**	76.42	**43.14**	**63.61**	62.53	78.92	**41.43**	**60.96**
0.5	64.84	75.76	39.93	60.18	**62.64**	78.72	37.75	59.70
0.7	64.16	**78.16**	39.47	60.60	62.49	79.68	34.28	58.82
1.0	59.27	76.11	37.75	57.71	61.73	**79.70**	34.59	58.67

Table 8: Performance comparison of FLAG across different size of training set on both datasets. **Bold** denotes the best performance.

Size of Training Set	MIMIC-III			**Avg.**	PICU			**Avg.**
	Macro F1 Score (%)	AUROC (%)	AUPRC (%)		Macro F1 Score (%)	AUROC (%)	AUPRC (%)	
20%	67.32	77.28	39.57	61.39	**65.19**	**79.59**	33.74	59.51
40%	68.38	77.12	40.78	62.09	62.24	78.13	35.17	58.51
60%	67.18	75.82	39.69	60.90	63.75	78.16	38.74	60.22
80%	68.73	**77.52**	39.53	61.93	64.09	78.96	36.52	59.86
100%	**71.26**	76.42	**43.14**	**63.61**	62.53	78.92	**41.43**	**60.96**

E Comprehensive Prompt Templates for Mortality Prediction Across Both Datasets

Prompt Template of Direct in MIMIC-III dataset

Instruction:
You are an ICU doctor. There is a simulated patient. I will provide you with a record of the patient's first 24 hours in the ICU, including demographics, vital signs, laboratory results, interventions, and other clinical information. Please complete the following task based on the information I provide.

First 24-hour ICU record:
gender: female
age: 74
height_first: 175.3
vent: yes
rrt: no
urineoutput: 62.0
mingcs: 12.0
...

Task Description:
Please assess the likelihood of the patient's death during this ICU stay based on the above information in PICU dataset.
Answer with only a floating-point number between 0 and 1, where a larger number indicates a higher risk of death.

Prompt Template of CoT in PICU dataset

Instruction:
You are an ICU doctor. There is a simulated patient. I will provide you with a record of the patient's first 24 hours in the ICU, including demographics, vital signs, laboratory results, interventions, and other clinical information. Please complete the following task based on the information I provide.

First 24-hour ICU record:
Gender: Male
Age (in months): 12
Underlying condition: Post-Kasai procedure
Infection-related diagnosis at inclusion: Sepsis
Organ dysfunction: Brain, kidneys, lungs, and circulatory system
...

Task Description:
Please assess the likelihood of the patient's death during this ICU stay based on the above information.
Answer with only a floating-point number between 0 and 1, where a larger number indicates a higher risk of death.
Let's think step by step to come up with the answer, and return the reply in the following format:
Thought: Provide your thought process here.
Answer: Provide your answer here.

Prompt Template of FLAG in MIMIC-III dataset

Instruction:
You are an ICU doctor. There is a simulated patient. I will provide you with a record of the patient's first 24 hours in the ICU, including demographics, vital signs, laboratory results, interventions, and other clinical information. Please complete the following task based on the information I provide.

First 24-hour ICU record:
gender: female
age: 54
height_first: 172.0 < *Irrelevant feature*>
urineoutput: 61.0
glascow coma scale total: 10.0
prothrombin time inr: ? < *Missing feature*>
...

Feature explanation:
Features increasing risk:
glascow coma scale total: 9.30%
lactic acid: 8.46%
peak inspiratory pressure: 7.64%
urineoutput: 5.82%
lactate: 4.68%
...

Features decreasing risk:
prothrombin time pt: -2.76%
ph: -2.42%
prothrombin time inr: -2.23%
partial thromboplastin time: -2.05%
alanine aminotransferase: -1.96%
...

Task Description:
Please assess the likelihood of the patient's death during this ICU stay based on the above information.
Answer with only a floating-point number between 0 and 1, where a larger number indicates a higher risk of death.
Let's think step by step to come up with the answer, and return the reply in the following format:
Thought: Provide your thought process here.
Answer: Provide your thought answer here.
Explanation: Provide your explanation here.