MORE-CLEAR: Multimodal Offline Reinforcement learning for Clinical notes Leveraged Enhanced State Representation

Yooseok Lim[1] * **ByoungJun Jeon**[1] * **Seong-A Park**[1] **Jisoo Lee**[3] **Sae Won Choi**[2]
Chang Wook Jeong[1,3] **Ho-Geol Ryu**[1] **Hongyeol Lee**[1] † **Hyun-Lim Yang**[1,3] †
[1]Seoul National University Hospital [2]Bucheon Sejong Hospital
[3]Seoul National University

Abstract

Sepsis, a life-threatening inflammatory response to infection, causes organ dysfunction, making early detection and optimal management critical. Previous reinforcement learning (RL) approaches to sepsis management rely primarily on structured data, such as lab results or vital signs, and on a dearth of a comprehensive understanding of the patient's condition. In this work, we propose a Multimodal Offline REinforcement learning for Clinical notes Leveraged Enhanced stAte Representation (MORE-CLEAR) framework for sepsis control in intensive care units. MORE-CLEAR employs pre-trained large-scale language models (LLMs) to facilitate the extraction of rich semantic representations from clinical notes, preserving clinical context and improving patient state representation. Gated fusion and cross-modal attention allow dynamic weight adjustment in the context of time and the effective integration of multimodal data. Extensive cross-validation using two public (MIMIC-III and MIMIC-IV) and one private dataset demonstrates that MORE-CLEAR significantly improves estimated survival rate and policy performance compared to single-modal RL approaches. To our knowledge, this is the first to leverage LLM capabilities within a multimodal offline RL for better state representation in medical applications. This approach can potentially expedite the treatment and management of sepsis by enabling reinforcement learning models to propose enhanced actions based on a more comprehensive understanding of patient conditions.

1 Introduction

Sepsis, characterized by a dysregulated host response to infection, causes organ dysfunction with a high mortality risk. It remains the leading cause of death in critically ill patients [50, 60]. Early identification of sepsis and prompt intervention in the intensive care unit (ICU) are pivotal for improving patient outcomes, highlighting the importance of developing effective therapeutic strategies. However, timely recognition of sepsis is challenging due to the multitude of potential etiologies and the rapid progression of the disease [48, 38]. Furthermore, inotropes and fluid therapy can be administered to resuscitate patients from dangerously low hypotensive condition, however, the appropriate timing and amount of these treatments remain controversial [34, 61, 66, 9, 37].

In recent years, reinforcement learning (RL) has emerged as a promising approach for identifying optimal treatment strategies in complex and uncertain clinical environments [30, 31, 14]. By for-

*These authors contributed equally.
†Corresponding author.
takumama@naver.com, hlyang@snu.ac.kr

Preprint.

mulating the patient care as a sequential decision-making problem—comprising states, actions, and rewards—RL facilitates the development of data-driven, personalized treatment policies. RL research in sepsis management relies predominantly on structured data, such as laboratory testing (lab) results and vital signs [25, 4, 62, 35]. While these data are valuable from a clinical perspective, they are often characterized by substantial missingness, noise, and irregular sampling, which collectively undermine the reliability of state representations and hinder effective policy learning [42]. Although recent approaches have sought to alleviate sparsity and bias in electronic medical records (EMRs) [68, 2], a fundamental limitation remains in that structured data alone is insufficient for capturing the complexity of patient states [49, 47].

In clinical practice, decisions regarding treatment planning and medication administration are often informed by clinical notes, which provide nuanced contextual records, including medical history, symptom evolution, and response to interventions [46, 39]. Incorporating clinical notes can compensate for structured data limitations and enhance patient state representations. Although prior research in the domain of supervised learning has evidenced the efficacy of such multimodal integration [56], the potential of this approach within RL remains under-explored. In particular, given the irregular documentation of clinical notes during events and the imbalanced distribution of information over time, effective integration of multimodal clinical records that accounts for these characteristics poses a significant challenge, one that must be surmounted to facilitate robust policy learning.

To address these challenges, we propose MORE-CLEAR (Multimodal Offline REinforcement learning for Clinical notes Leveraged Enhanced stAte Representation), a novel multimodal offline RL framework tailored for sepsis management in the ICU. MORE-CLEAR employs a pre-trained large language model (LLM) to extract enriched embedding vectors containing semantic representations from clinical notes. Motivated by clinical reasoning, in which early patient information guides understanding prior treatment trajectories, we model the initial-time note as a context vector injected at each decision to mitigate information sparsity and preserve temporal coherence. A gated fusion mechanism is introduced to facilitate effective integration of the context vectors and observation vectors. Embeddings for structured data, i.e., lab results and vital signs, are obtained using a refined multi-layer perception (MLP)-based encoder to handle missing values and complex feature interactions. A bidirectional cross-modal attention mechanism is then employed to integrate salient information from structured and unstructured modalities, forming the final state representation for RL. The effectiveness of MORE-CLEAR was evaluated on two publicly available datasets, MIMIC-III and MIMIC-IV, and a private ICU dataset. Results demonstrate that incorporating clinical notes through MORE-CLEAR leads to substantial improvements in survival rate estimation and policy performance compared to unimodal RL approaches.

Extracting comprehensive patient state representations for robust RL remains a fundamental challenge in the medical field. MORE-CLEAR addresses this issue through the following key contributions: (i) We show that integrating clinical notes with lab results using LLMs significantly enhances policy performance compared to unimodal baselines. (ii) We propose a context-aware gated fusion that encodes the initial patient information as a persistent context vector and fuses it with temporal observations to enhance RL performance. (iii) A bidirectional cross-modal attention mechanism integrating clinical notes and lab tests is shown to yield a statistically significant improvement in policy performance. (iv) Extensive cross-evaluations on MIMIC-III, MIMIC-IV, and a private dataset demonstrate that MORE-CLEAR consistently outperforms unimodal frameworks in off-policy evaluation (OPE) metrics and predicted survival rates based on enhanced state representation.

2 Related Work

2.1 Sepsis Treatment Optimization via RL

Prior work on RL for sepsis management has primarily focused on optimizing vasopressor and intravenous fluid dosages [25, 4, 35, 59, 21, 10], or devising strategies for antibiotic administration [62] using conventional off-policy algorithms, such as Deep Q-Networks (DQN) and Double DQN (DDQN). Although many studies have incorporated rule-based constraints in their training to ensure reliability in real-world clinical settings [62, 10], the fundamental limitation of relying solely on structured data hinders capturing contextual factors, such as patient comorbidities and treatment history. Additionally, the robustness and efficacy of the model in various clinical settings and cohorts have not been thoroughly evaluated, as only one-way evaluations have been considered.

2.2 Offline RL

Offline RL provides a practical framework for policy learning in data-constrained domains such as healthcare, autonomous driving, and robotics [23, 51]. Recently, it has been explored for treating sepsis to mitigate distributional shifts between the learned and behavioral policies [59, 21, 10]. A key challenge in Offline RL is reducing the overestimation of Q-value for out-of-distribution (OOD) actions compared to the behavior policy. Although various methods, such as Batch-Constrained Q-Learning (BCQ) [12], Bootstrapping error accumulation reduction [27], Conservative Q-Learning (CQL) [28], or Implicit Q-Learning [26], have been proposed to address this challenge, a one-size-fits-all solution for all scenarios has yet to be found. In the medical domain, several ideas, such as variations of BCQ [22], critical patient state [11], and sampling strategies [40], have been proposed to address the problem; however, investigating the influence of multimodal state representations on policy performance remains under-explored.

2.3 Multimodal fusion in Medicine

Multimodal learning has garnered increasing attention in the medical domain, aiming to integrate diverse data types such as EMRs, clinical notes, imaging, and genomics [52, 33, 67, 1, 24, 3, 45]. Notable contributions include image fusion studies employing multiple modalities such as MRI, PET, and CT scans [52, 33, 67, 1]; analyses leveraging biosignals like EEG and ECG [24, 3]; and integrative approaches involving genomic and other omics data [45]. Additionally, researchers have explored models that combine lab results with clinical notes to enhance predictive performance [56]. Despite encouraging outcomes in predictive tasks, these methods remain confined to classifying specific data, hindering optimal policy identification in uninterrupted clinical settings.

3 Background

3.1 Multimodal Markov Decision Process

RL provides a principled framework for addressing sequential decision-making problems, typically formalized as a Markov decision process (MDP) [53]. In this study, we model the progression of patient states as an MDP with multiple modalities of observation. The multimodal MDP is specified by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, with $\mathcal{S} = \prod_{i=1}^{d} \mathcal{O}^{\mathrm{M_i}}$, which is the joint observation space formed by d modalities, where $\mathcal{O}^{M_i}$ denotes the observation space of the i-th modality. $\mathcal{A}$ denotes the set of all possible actions. The transition probability function $\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is defined such that, for any joint observation $o_t = \left(o_t^{M_1}, o_t^{M_2}, \ldots, o_t^{M_d}\right)$ and action a_t, the probability $P\big(o_{t+1} \mid o_t, a_t\big)$ represents the likelihood of transitioning to the next joint observation $o_{t+1} = \left(o_{t+1}^{M_1}, o_{t+1}^{M_2}, \ldots, o_{t+1}^{M_d}\right)$. The function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes a reward function providing the expected reward. $\gamma \in [0, 1)$ is the discount factor. The goal of an RL is to find a policy $\pi : S \times A \to [0, 1]$ that maximizes the expected cumulative discounted return $\sum_{t=0}^{\infty} \gamma^t r_t$.

We conduct multimodal RL in an offline setting [32], where the policy π is learned from a fixed dataset $\mathcal{D}$ collected under a behavior policy π_β, without any additional online interaction.

3.2 Q-Learning and Deep Q-Networks (DQN)

Q-Learning [64] is an off-policy, model-free reinforcement learning algorithm designed to approximate the optimal action-value function $Q^*(s, a)$ without requiring knowledge of the environment's transition dynamics. Q-Learning is fundamentally grounded in the Bellman optimality equation,

$$Q^*(s, a) = \mathbb{E}_{s'}\left[r(s, a) + \gamma \max_{a'} Q^*(s', a')\right], \tag{1}$$

which it approximates through sample-based updates. Specifically, for each observed transition (s_t, a_t, r_t, s_{t+1}), the Q-value is updated as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right], \tag{2}$$

where $\eta \in (0, 1)$ is the learning rate. Through repeated interaction with the environment, this update rule allows the Q-values to progressively converge, ultimately enabling the agent to derive the optimal policy $\pi^*(s) = \arg\max_a Q^*(s, a)$.

However, Q-learning becomes impractical in high-dimensional or continuous state spaces. To address this issue, DQN utilizes a deep neural network to approximate the Q-function $Q(s, a; \theta)$, where θ denotes the parameters of the Q-network. In DQN, the target y used for training is defined as:

$$y = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-), \tag{3}$$

where θ^- represents the parameters of a fixed target network from the previous iteration, and the $\max$ operator selects the action that yields the highest Q-value in the next state. Based on this target, the current network parameters θ are updated to minimize the following loss function:

$$L(\theta) = \mathbb{E}_{(s_t, a_t) \sim \pi_b} \Big[\big(y - Q(s_t, a_t; \theta) \big)^2 \Big]. \tag{4}$$

4 Method

4.1 Problem Formulation

In this section, we formalize the problem setting of multimodal RL for optimizing sepsis treatment. The patient state $\mathcal{S} = O^{M_l} \times O^{M_n}$ comprises observations from two distinct modalities: structured data, i.e., lab results and vital signs, O^{M_l} and unstructured clinical notes O^{M_n}. The action space $\mathcal{A}$ is defined as a discrete set of treatment options, specifically representing dosage levels of vasopressors and intravenous fluids. Each action $a_t \in \mathcal{A}$ denotes the treatment administered at time step t. The reward function $r(s_t, a_t)$ quantifies the clinical outcome of applying a_t in state s_t, which follows the primary outcome of this study. The transition function $\mathcal{P}$ specifies the evolution of the patient's state over time, and an episode τ represents a sequence of clinician–patient interactions. The treatment-optimization objective is to identify a policy π that maximizes the expected discounted cumulative clinical benefit: $\arg\max_\pi \ \mathbb{E}_{\tau \sim \pi} \Big[\sum_{(s,a) \in \tau} \gamma \, r(s, a) \Big]$.

4.2 Conservative Q-Learning (CQL)

Our framework is motivated by CQL [28], an offline RL algorithm that has demonstrated strong performance in the sepsis treatment task [59, 21, 40]. CQL modifies the standard Bellman objective by adding a conservative regularization term to mitigate the tendency of conventional DQN to overestimate Q-values.

$$\begin{aligned} L_{\mathrm{CQL}}(\theta) = \tfrac{1}{2} \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \Big[\Big(Q_\theta(s, a) - \big(r + \gamma \, \mathbb{E}_{a' \sim \pi} [Q_{\bar{\theta}}(s', a')] \big) \Big)^2 \Big] \\ + \ \alpha \Big(\mathbb{E}_{s \sim \mathcal{D},\, a \sim \mu} [Q_\theta(s, a)] \ - \ \mathbb{E}_{(s,a) \sim \mathcal{D}} [Q_\theta(s, a)] \Big). \end{aligned} \tag{5}$$

μ denotes a distribution over actions intended to represent OOD behavior at state s. By suppressing the Q-values of actions sampled from μ, CQL prevents the policy from overestimating the value of actions that are not well represented in the offline dataset. Conversely, by encouraging higher Q-values for state–action pairs observed in the dataset $\mathcal{D}$, the learned Q-function is guided to assign sufficiently high values to actions actually taken in the data.

The hyperparameter $\alpha > 0$ controls the strength of the conservative regularization. Larger α intensifies the OOD penalty, promoting more rigorous suppression of unobserved action values and enhancing robustness. In contrast, excessively small α diminishes the conservative effect, potentially reviving the overoptimistic bias commonly observed in standard DQN formulations. Consequently, α plays a crucial role in balancing OOD-action suppression and learning stability.

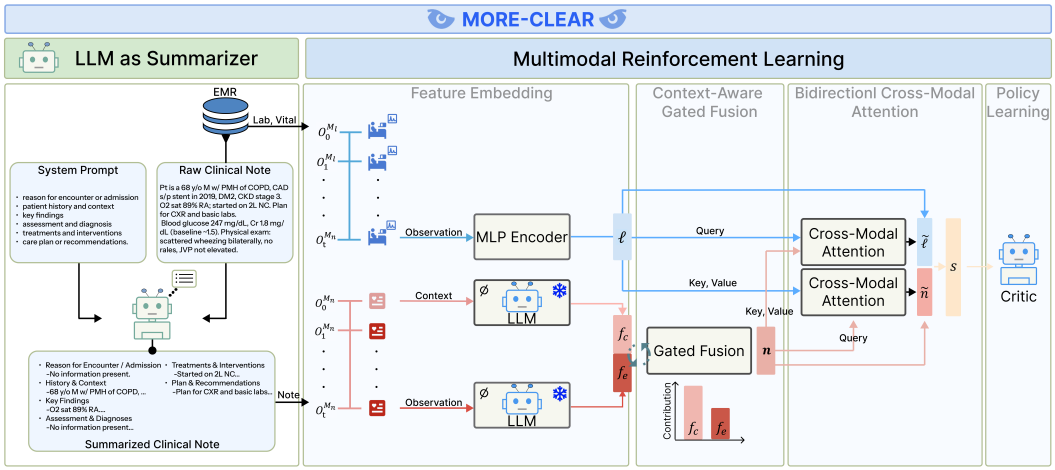


Figure 1: MORE-CLEAR framework

4.3 MORE-CLEAR framework

MORE-CLEAR (Figure 1) integrates structured data, i.e, lab results and vital signs, and clinical notes to derive a more enriched and comprehensive state representation, enhancing the quality of offline RL-based policy optimization. In our framework, raw clinical notes are initially summarized using an LLM, and the generated summaries are encoded into dense vectors utilizing LLM-based encoders. The clinical note is further categorized into context and observation components. Each component is independently encoded and subsequently fused using a gated fusion mechanism. Concurrently, structured data is processed using a data-specific MLP-based encoder motivated by MLP-Mixer [57]. Ultimately, the bidirectional cross-modal attention module integrates embeddings from the textual and structured modalities to construct a unified state representation that facilitates robust policy learning.

4.4 Structured Summarization with LLM for Multimodal Embedding

Our method constructs sequential trajectories by concatenating all notes recorded within fixed time intervals. This process often produces long sequences containing a mixture of diverse information, which complicates the identification of key content and occasionally results in inputs that exceed the token limits. To address this, we employ an open-source LLM to summarize raw clinical notes and generate meaningful note embeddings. A structured prompting guides LLM to summarize with organized content into clinically meaningful sections: reason for encounter or admission, patient history and context, key findings, assessment and diagnosis, treatments and interventions, and care plan or recommendations. Gemma-3-27B-it [54] is employed for the note summarization LLM according to our experiments in Appendix A. Subsequent to the summarization, the structured text is fed into the LLMs, which employ average pooling on the hidden state of the final layer to extract the latent representation

4.5 Context-Aware Gated Fusion

We introduce a context-aware gated fusion mechanism to extract enriched patient state representations from clinical notes o^{M_n}. Henceforth, the initial note $o_0^{M_n}$ refer the context note and $o_t^{M_n}$ refer as the event observation note at time t. The context note often offers implicit contextual cues that critically inform subsequent clinical decision-making, while the observation note captures local, time-specific information, including the patient's evolving condition and treatment progress. Therefore, context-aware gated fusion is designed to maintain contextual notes to retain information about the patient's underlying medical conditions, history, presenting complaints, and the cause of hospitalization, while dynamically integrating them with event observation notes from each time frame to yield more expressive patient state representations.

In our context-aware gated fusion mechanism, $o_0^{M_n}$ and $o_t^{M_n}$ are projected into a high-dimensional vector space by a pre-trained LLM encoder ϕ.

$$f_c = \phi\big(o_0^{M_n}\big) \in \mathbb{R}^{d_n}, \quad f_e = \phi\big(o_t^{M_n}\big) \in \mathbb{R}^{d_n}. \tag{6}$$

d_n denotes the dimensionality of the note embedding. The final note representation $\mathbf{n}_t \in \mathbb{R}^d$ is computed as an adaptive, weighted combination of the context vector f_c and the event observation vector f_e. A gating mechanism modulates this combination. $\boldsymbol{\psi}_t$ is parameterized by a learnable weight $W \in \mathbb{R}^{d\times 2d}$ and bias $\mathbf{b} \in \mathbb{R}^d$.

$$\begin{aligned} \boldsymbol{\psi}_t &= \sigma\big(W\,[f_c; f_e] + \mathbf{b}\big) \in (0,1)^d, \\ \mathbf{n}_t &= \boldsymbol{\psi}_t \odot f_c \;+\; (1 - \boldsymbol{\psi}_t) \odot f_e. \end{aligned} \tag{7}$$

Here, $\sigma(\cdot)$ is the sigmoid function, $[\cdot\,;\,\cdot]$ denotes concatenation, and $\odot$ is the element-wise product. The context-aware representation $\mathbf{n}_t$ selectively integrates episode-level context with time-specific information, enabling enriched state representations for RL optimization.

4.6 Bidirectional Cross-Modal Attention

Clinical notes offer therapeutic context and insights, while also encompassing diverse content such as auxiliary procedures, clinician–patient communications, and detailed patient behaviors. In contrast, structured data, i.e., lab results and vital signs, provide quantitative measurements of a patient's physiological state at specific time points, though often limited by missing values and irregular sampling. In clinical practice, accurate patient assessment depends on the integration of these two complementary modalities [41, 7]. Motivated by this insight, we introduce a bidirectional cross-modal attention module that captures the interaction between the two modalities to selectively extract clinically meaningful information.

Let $\mathbf{n}, \ell \in \mathbb{R}^d$ denote the clinical note from gated fusion and structured data embedding vectors, respectively. We employ a bidirectional cross-modal attention mechanism to capture both structured data (l, v)→note and note→(l,v) interactions. Each modality is first projected into the query, key, and value spaces, as defined below:

$$Q^{(\mathrm{l,v})\to\mathrm{note}} = W_\ell^Q\,\ell, \qquad K^{(\mathrm{l,v})\to\mathrm{note}} = W_{\mathbf{n}}^K\,\mathbf{n}, \qquad V^{(\mathrm{l,v})\to\mathrm{note}} = W_{\mathbf{n}}^V\,\mathbf{n}, \tag{8}$$

$$Q^{\mathrm{note}\to(\mathrm{l,v})} = W_{\mathbf{n}}^Q\,\mathbf{n}, \qquad K^{\mathrm{note}\to(\mathrm{l,v})} = W_\ell^K\,\ell, \qquad V^{\mathrm{note}\to(\mathrm{l,v})} = W_\ell^V\,\ell, \tag{9}$$

where each $W^{\cdot} \in \mathbb{R}^{d_k\times d}$ is learnable and d_k is the attention dimensionality. For each direction $u \in \{(\mathrm{l,v})\to\mathrm{note},\ \mathrm{note}\to(\mathrm{l,v})\}$, attention weights α and attention-weighted features A are computed as follows:

$$\alpha^{(u)} = \mathrm{softmax}\!\left(\frac{Q^{(u)}\,(K^{(u)})^\top}{\sqrt{d_k}}\right), \quad A^{(u)} = \alpha^{(u)}\,V^{(u)} \in \mathbb{R}^{1\times d}. \tag{10}$$

Next, each original feature $(\mathbf{n}, \ell)$ is concatenated with its corresponding attention vector and then passed through a linear transformation.

$$\tilde{\ell} = W_\ell\,\big[\ell;\, A^{\mathrm{note}\to(\mathrm{l,v})}\big] \in \mathbb{R}^d, \tag{11}$$

$$\tilde{n} = W_{\mathbf{n}}\,\big[\mathbf{n};\, A^{(\mathrm{l,v})\to\mathrm{note}}\big] \in \mathbb{R}^d, \tag{12}$$

where $W_\ell, W_{\mathbf{n}} \in \mathbb{R}^{d\times 2d}$. Finally, the fused state representation is calculated as:

$$s = \big[\tilde{\ell};\, \tilde{\mathbf{n}}\big] \in \mathbb{R}^{2d}. \tag{13}$$

Variable	MIMIC-III (n=11,114)	MIMIC-IV (n=10,203)	Private Dataset (n=599)
Age (Mean ± SD)	64.14 ± 16.94	64.86 ± 16.26	66.59 ± 16.15
Weight (kg, Mean ± SD)	83.29 ± 24.63	83.13 ± 24.95	57.16 ± 13.40
SOFA (Mean ± SD)	5.39 ± 3.30	5.16 ± 2.84	6.41 ± 3.35
GCS (Mean ± SD)	12.48 ± 3.31	12.95 ± 3.29	9.87 ± 3.35
Female (%)	56	58	37
90-day mortality (%)	22	26	24

Table 1: Dataset Statistics

5 Experimental setups

5.1 Sepsis Treatment Optimization

5.1.1 Dataset and ethical approval

We employed two publicly available datasets, MIMIC-III [18] and MIMIC-IV [20], along with one private dataset for experiments. A private dataset (PD) was retrieved from the prospective registry of ICU patients at Seoul National University Hospital in the Republic of Korea from April 2022 to March 2025 via the clinical data warehouse. The collection of data and subsequent analysis were approved by the institutional review board of Seoul National University Hospital (IRB-000-000-000), with a waiver for written informed consent due to the study's retrospective design and data anonymity. The inclusion and exclusion criteria were identical to those in previous research [25, 4]. Table 1 summarizes key feature statistics related to sepsis.

5.1.2 Task

The patient state s_t is defined at four-hour intervals and consists of two primary modalities: structured data O^{M_l}, which comprises 42 clinical variables, and clinical notes O^{M_n}, which encompass nursing records, physician documentation, discharge summaries, and other information. Each episode spans from 24 hours prior to 48 hours after the suspected onset of sepsis, capturing the critical period for therapeutic intervention. The action a_t denotes a joint decision over intravenous (IV) fluid and vasopressor administration, each discretized into five levels, yielding 25 unique treatment combinations. The reward r_t is defined such that all intermediate time steps are assigned a reward of 0, whereas the terminal step (T) yields a reward of $+1$ if the patient survives for 90 days and -1 otherwise. The reward function is defined as

$$r_t = \begin{cases} 0, & t < T, \\ +1, & \text{if the patient survives for 90 days after discharge,} \\ -1, & \text{if the patient dies,} \end{cases} \tag{14}$$

5.1.3 Preprocessing

All datasets were preprocessed using a unified pipeline adapted from public implementations [3] [4], with minor modifications to ensure consistency. Specifically, four variables from the original set of 47 clinical features [25]—readmission, Elixhauser score, mechanical ventilation, and cumulative fluid balance—were excluded. The vasopressor drugs included vasopressin, norepinephrine, and dopamine. IV fluids included the same or similar categories as normal saline, plasma solution, albumin, and Hartmann's solution. Apart from these adjustments, all preprocessing steps were consistent with the original pipeline.

5.2 Evaluation Metrics

A rigorous evaluation of the MORE-CLEAR's policy was conducted by employing four off-policy evaluation (OPE) metrics, including weighted importance sampling (WIS) [36], doubly robust (DR) estimator [15], fitted Q-evaluation (FQE) [13], and offline policy evaluation with re-weighted

[3] `https://github.com/matthieukomorowski/AI_Clinician`

[4] `https://github.com/cmudig/AI-Clinician-MIMICIV`

aggregates (OPERA) [43]. Beyond these quantitative measures, we further examined the policy's clinical relevance by introducing the Behavioral Discrepancy Estimated Survival Rate (BDESR), a metric designed to capture alignment between RL policy actions and clinician actions.

5.2.1 Behavioral Discrepancy Estimated Survival Rate

We define the BDESR as a means to estimate the clinical effectiveness of an RL policy. This metric quantifies the degree of discrepancy between actions administered by the RL policy and clinician actions recorded in the batch dataset, with the corresponding survival rate estimated based on behavioral inconsistency.

For each patient episode i, divergences between the RL policy and clinician actions over time are computed separately for the two actions: IV fluid and vasopressor administration.

$$m_{\text{iv}}^{(i)} = \frac{1}{T_i}\sum_{t=1}^{T_i}\big|\hat{A}_{\text{iv},t}^{(i)} - A_{\text{iv},t}^{(i)}\big|, \quad m_{\text{vaso}}^{(i)} = \frac{1}{T_i}\sum_{t=1}^{T_i}\big|\hat{A}_{\text{vaso},t}^{(i)} - A_{\text{vaso},t}^{(i)}\big|, \tag{15}$$

where $\hat{A}$ denotes the actions recommended by the RL policy and A denotes the actions taken by the clinician. A weighted average of $m_{\text{iv}}^{(i)}$ and $m_{\text{vaso}}^{(i)}$ is used to compute the overall discrepancy:

$$m^{(i)} = \alpha\, m_{\text{iv}}^{(i)} + \beta\, m_{\text{vaso}}^{(i)}, \quad \alpha+\beta=1,\ \alpha,\beta\geq 0. \tag{16}$$

To evaluate the clinical impact of adherence to the RL policy, we compare survival rates between episodes with low and high policy discrepancies. We identify the most extreme $N\%$ of episodes by sorting the discrepancy scores $m^{(i)}$ and computing the $p = N/2$ and $(100-p)$ percentiles, denoted Q_p and Q_{100-p}, where p is a hyperparameter. The low and high-discrepancy cohort is defined as:

$$L = \big\{\, i \mid m^{(i)} \leq Q_p \big\}, \quad H = \big\{\, i \mid m^{(i)} \geq Q_{100-p} \big\}. \tag{17}$$

These cohorts represent the lower and upper extremes of the discrepancy distribution. Survival status for each episode is encoded as a binary indicator $S^{(i)} \in \{0,1\}$, where 0 indicates mortality. Survival rate is defined as follows:

$$\begin{aligned} \text{High-BDESR} &= \frac{1}{|H|}\sum_{i\in H}\mathbf{1}\{S^{(i)}=1\}, \\ \text{Low-BDESR} &= \frac{1}{|L|}\sum_{i\in L}\mathbf{1}\{S^{(i)}=1\}. \end{aligned} \tag{18}$$

BDESR facilitates a more precise evaluation of the clinical effectiveness of an RL policy by enabling a comparison of survival between patients with well-behaved (Low-BDESR) and significantly deviant (High-BDESR) behaviors.

5.3 Implementation details

All experiments were conducted with fixed hyperparameter settings for consistency. We trained all models using a batch size of 256, a learning rate of 0.0001, and the Adam optimizer. The BCQ threshold was set to 0.3, and the CQL regularization coefficient to 2.0. Both BCQ and CQL employed a Dueling DQN architecture [63]. The structured data-only RL model employed a neural network consisting of three fully connected layers with 512 units each. The text-based model employed an LLM for embedding, integrated with a Q-network. Textual embeddings were obtained from the LLM with all weights frozen during training.

6 Results and Discussion

6.1 Performance Comparison Across Modalities

The OPE and BDESR performances for each modality are reported in Tables 2 and 3, respectively. Table 2 shows that the MORE-CLEAR generally outperforms unimodal baselines across a broad

Dataset	Metric	Structured data		Text			Multimodal (MORE-CLEAR)		
		BCQ	CQL	Bert	CB	Llama	Bert+CQL	CB+CQL	Llama+CQL
MIMIC-III	↑ OPERA	0.889 ± 0.02	2.925 ± 0.03	0.512 ± 0.09	0.607 ± 0.09	0.450 ± 0.09	$\underline{3.335 \pm 0.14}$	3.329 ± 0.10	$\underline{3.382 \pm 0.14}$
	↑ DR	0.917 ± 0.04	2.419 ± 0.05	0.574 ± 0.08	0.587 ± 0.07	0.546 ± 0.10	2.968 ± 0.10	$\underline{2.975 \pm 0.06}$	3.007 ± 0.09
	↑ FQE	0.472 ± 0.12	0.622 ± 0.07	0.320 ± 0.28	0.485 ± 0.32	0.181 ± 0.08	$\underline{1.382 \pm 0.74}$	$\underline{1.346 \pm 0.38}$	0.899 ± 0.22
	↑ WIS	0.563 ± 0.25	0.682 ± 0.03	0.721 ± 0.08	$\underline{0.738 \pm 0.02}$	$\underline{0.728 \pm 0.08}$	0.693 ± 0.02	0.702 ± 0.04	0.683 ± 0.03
MIMIC-IV	↑ OPERA	0.885 ± 0.09	$\underline{3.862 \pm 0.04}$	0.436 ± 0.07	0.461 ± 0.07	0.714 ± 0.36	3.861 ± 0.16	$\underline{3.877 \pm 0.16}$	3.810 ± 0.14
	↑ DR	0.963 ± 0.05	$\underline{3.791 \pm 0.03}$	0.521 ± 0.24	0.866 ± 0.18	0.750 ± 0.29	3.648 ± 0.10	$\underline{3.677 \pm 0.09}$	3.596 ± 0.12
	↑ FQE	0.466 ± 0.10	1.243 ± 0.13	0.406 ± 0.22	0.210 ± 0.05	0.535 ± 0.24	$\underline{2.597 \pm 2.07}$	$\underline{7.522 \pm 3.33}$	1.322 ± 0.35
	↑ WIS	0.731 ± 0.08	0.753 ± 0.03	0.566 ± 0.31	0.736 ± 0.12	0.739 ± 0.12	0.734 ± 0.04	$\underline{0.758 \pm 0.04}$	$\underline{0.766 \pm 0.04}$
Private Dataset	↑ OPERA	0.782 ± 0.05	2.390 ± 0.22	0.602 ± 0.10	0.599 ± 0.12	0.717 ± 0.22	$\underline{2.781 \pm 0.17}$	$\underline{2.930 \pm 0.14}$	1.969 ± 0.28
	↑ DR	0.952 ± 0.07	2.487 ± 0.20	0.752 ± 0.09	0.601 ± 0.09	0.689 ± 0.06	$\underline{2.790 \pm 0.23}$	$\underline{2.739 \pm 0.09}$	1.893 ± 0.27
	↑ FQE	0.591 ± 0.13	1.290 ± 0.28	0.692 ± 0.15	0.510 ± 0.33	0.585 ± 0.07	1.517 ± 0.71	$\underline{1.681 \pm 0.96}$	$\underline{1.811 \pm 0.42}$
	↑ WIS	0.675 ± 0.26	0.685 ± 0.02	0.422 ± 0.22	0.413 ± 0.28	0.464 ± 0.40	$\underline{0.742 \pm 0.01}$	$\underline{0.729 \pm 0.04}$	0.627 ± 0.25

Table 2: OPE metric performance across all modalities. The means and standard deviations of the results across five seeds are reported. The best two results are underlined for each metric. CB: Clinical Bert, Llama: Llama3.1-8B

Dataset	Metric	Structured data		Text			Multimodal (MORE-CLEAR)		
		BCQ [4]	CQL	Bert	CB	Llama	Bert+CQL	CB+CQL	Llama+CQL
MIMIC-III	Low-BDESR	0.848 ± 0.04	0.847 ± 0.03	0.834 ± 0.03	0.815 ± 0.03	0.836 ± 0.01	0.832 ± 0.02	$\underline{0.861 \pm 0.03}$	$\underline{0.862 \pm 0.02}$
	High-BDESR	0.744 ± 0.06	0.633 ± 0.02	0.682 ± 0.06	0.660 ± 0.06	0.722 ± 0.08	0.631 ± 0.01	$\underline{0.619 \pm 0.02}$	$\underline{0.614 \pm 0.01}$
MIMIC-IV	Low-BDESR	0.843 ± 0.05	$\underline{0.851 \pm 0.01}$	0.827 ± 0.02	0.839 ± 0.02	0.790 ± 0.03	0.840 ± 0.01	$\underline{0.856 \pm 0.02}$	0.837 ± 0.02
	High-BDESR	0.733 ± 0.04	0.635 ± 0.02	0.688 ± 0.02	0.700 ± 0.02	0.560 ± 0.02	0.631 ± 0.03	$\underline{0.618 \pm 0.03}$	$\underline{0.569 \pm 0.01}$
Private Dataset	Low-BDESR	0.889 ± 0.08	0.893 ± 0.06	0.856 ± 0.06	0.901 ± 0.00	$\underline{0.930 \pm 0.05}$	0.810 ± 0.09	$\underline{0.921 \pm 0.12}$	0.891 ± 0.04
	High-BDESR	0.722 ± 0.13	$\underline{0.606 \pm 0.05}$	0.636 ± 0.12	$\underline{0.578 \pm 0.14}$	0.644 ± 0.09	0.644 ± 0.09	0.614 ± 0.06	0.711 ± 0.06

Table 3: Performance of the BDESR metric across all modalities. The means and standard deviations of the results across five seeds are reported. The top two results are underlined for each metric.

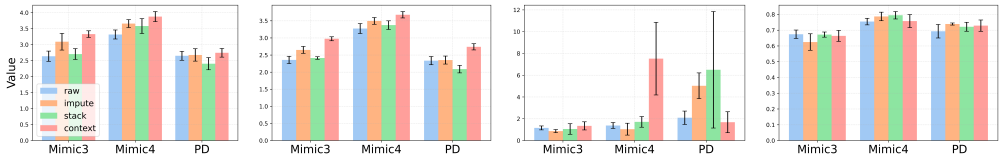
evaluation metrics. The multimodal models (Bert+CQL, CB+CQL, and Llama+CQL) outperform both structured data-only and text-only baselines, particularly concerning the OPERA (0.450 $\rightarrow$ 3.382 in Llama+CQL), DR (0.546 $\rightarrow$ 3.007 in Llama+CQL), and FQE (0.320 $\rightarrow$ 1.382). Dataset-specific trends further reinforce the advantage of the MORE-CLEAR framework. On MIMIC-III, Llama+CQL and Bert+CQL show the highest scores on OPERA (3.382), DR (3.007), and FQE (1.382), respectively. Furthermore, Llama+CQL achieves the best outcomes for WIS (0.766), whereas CB+CQL dominates the outstanding performance on the FQE score of 7.522 in MIMIC-IV. On the private dataset, the superiority of multimodal models remains evident. Across all settings, CB+CQL emerges as the most robust combination, with Bert+CQL also exhibiting strong and consistent results. Combining text representations with conventional structured data-based RL algorithms provides clear performance benefits. These results also provide substantial evidence supporting the efficacy of the MORE-CLEAR framework, which offers a meaningful synergy of multimodal data.

Table 3 presents estimated survival rates based on the BDESR, which stratifies patient episodes into two cohorts: those closely aligned with the RL policy's recommended actions (Low-BDESR) and those that deviate significantly (High-BDESR). We set p to 20. Across all datasets, survival rates for the Low-BDESR group were consistently higher than those for High-BDESR, indicating that the RL policy has been trained in a clinically beneficial direction. Within the Low-BDESR group, multimodal configurations demonstrated superior survival rates in every dataset, with CB+CQL consistently achieving the highest rates of 0.861, 0.856, and 0.921 for MIMIC-III, MIMIC-IV, and private dataset, respectively. This finding suggests that the integration of clinical notes and structured data contributes meaningfully to the identification of effective treatment strategies. Furthermore, structured data-based models outperformed their text-only counterparts, suggesting that structured data may provide more informative patient state representations for policy learning than the clinical note itself. Conversely, survival rates in the High-BDESR group exhibited a decline across all modalities, likely attributable to including a greater number of trajectories that culminated in mortality. The evidence suggests that the RL policy is inclined to suggest courses of action that diverge from adverse outcomes, thereby indicating its capacity to circumvent detrimental treatment patterns.

6.2 Contextual Representation Performance

We evaluate four strategies for integrating clinical notes: (1) raw note embedding (raw), (2) simple token imputation (impute), (3) concatenation within a fixed time window (stack), and (4) encoding

the initial note as a static context vector (context). In raw note embedding, the note is utilized as input if it exists within the time frame (4 hours); otherwise, no textual input is provided. Simple token imputation involves forward-filling the note if it does not exist in the current time frame. Concatenation within a fixed time window stacks notes of the previous 12 hours (window size(W)=3; see Appendix B for more details). Figure 2 shows experimental results across several OPE metrics from CB+CQL policy learning. OPERA scores consistently increase from raw to context in MIMIC-III (≈2.6→≈3.3) and MIMIC-IV(≈3.3→≈3.8). On the PD dataset, context achieves the highest OPERA score (≈2.7). Context also yields the best performance on the DR metric, particularly in PD. FQE differences are negligible in MIMIC-III but favor context in MIMIC-IV, albeit with high variance. In PD, the stack method attains the highest FQE score. WIS scores remain comparable across all approaches, with context performing slightly better. These results indicate that leveraging the initial clinical note as a context vector using our context-aware gated fusion improves policy performance, though some metrics remain sensitive to high variance.



(a) OPERA (b) DR (c) FQE (d) WIS

Figure 2: Performance of policies under different clinical note integration strategies

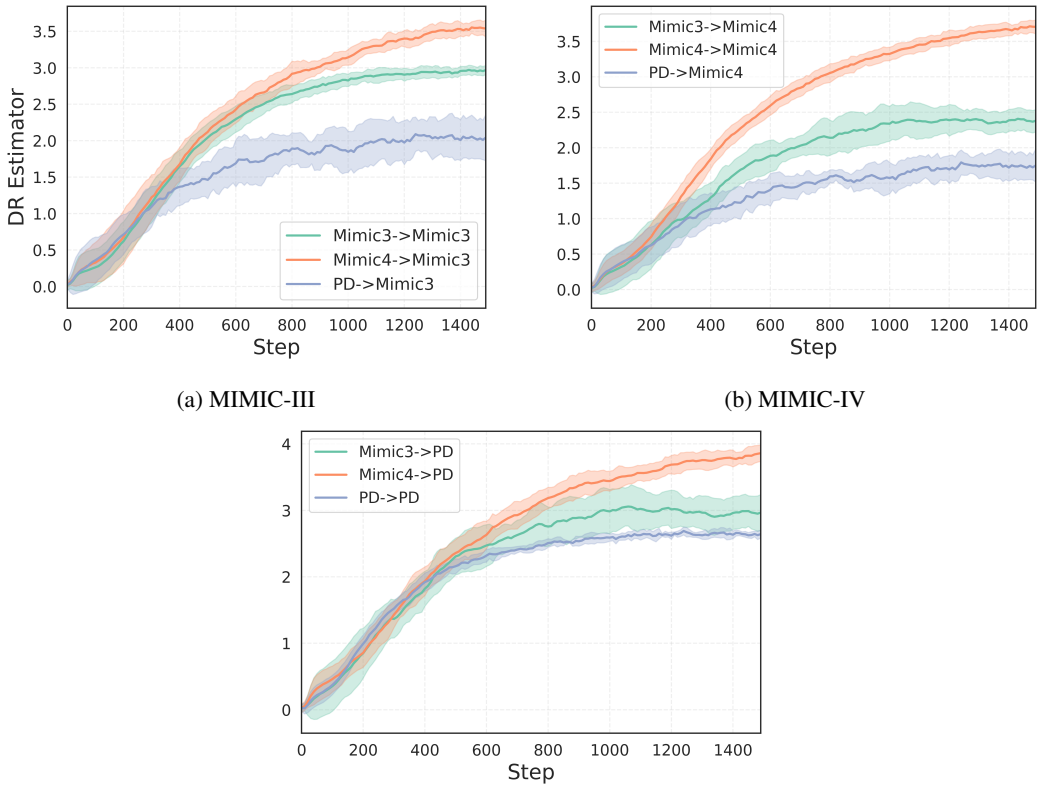
6.3 External Validation

Table 4 reports cross-dataset validation results of the MORE-CLEAR framework in conjunction with CB+CQL. When evaluated on MIMIC-IV, the policy trained on MIMIC-III yields OPERA 3.004, FQE 2.116, and WIS 0.689, while the policy trained using PD achieves a slightly higher OPERA, a substantially lower FQE, and a comparable WIS. In the reverse setting, the MIMIC-III-based policy demonstrates suboptimal performance on PD, while the MIMIC-IV-based policy exhibits marginally superior generalization. On MIMIC-III, the policy trained with PD yields OPERA 2.830, FQE 1.150, and WIS 0.601, while the policy trained with MIMIC-IV attains the highest OPERA score of 3.974 across all settings. Overall, the policy trained using MIMIC-IV demonstrates consistent performance in terms of high expected returns and low-variance estimates across datasets, thereby indicating superior generalization. Conversely, the MIMIC-III-based policy demonstrates a high degree of sensitivity to distributional shifts, resulting in performance degradation in the external dataset.

In order to assess whether policies improve consistently across different datasets, the DR estimates are further visualized across training iterations in Figure 3. Our observations indicate that policy learning exhibits a positive reinforcement trend across all datasets. It is noteworthy that the model trained on MIMIC-IV consistently attains the highest performance when evaluated on other datasets, followed by models trained on MIMIC-III and PD, respectively. This finding indicates that the data distribution of MIMIC-IV effectively captures the underlying characteristics of both MIMIC-III and PD.

The results of these trends can be interpreted in light of the characteristics of each dataset. For instance, the data collection process for MIMIC-IV has been more refined in comparison to that of MIMIC-III, resulting in higher-quality data [20]. The clinical notes themselves exhibit greater consistency, as MIMIC-IV is collected on a radiology report basis. Furthermore, the MIMIC-IV, which was collected from 2008 to 2022, may contain a more homogeneous sepsis treatment trajectory than the MIMIC-III, which was collected from 2001 to 2012, because it has more data since the treatment guidelines were revised to emphasize the importance of sepsis [9].

As such, offline RL is susceptible to distributional shifts, influenced by the nature of the data. Rigorous validation in diverse cohorts is imperative prior to its implementation in a clinical setting. We have demonstrated that our proposed MORE-CLEAR framework performs robustly in cross-validation based on a wide variety of data sets, including those from disparate collection periods and countries, thereby substantiating the model's viability for implementation in real-world clinical settings.



(a) MIMIC-III

(b) MIMIC-IV

(c) PD

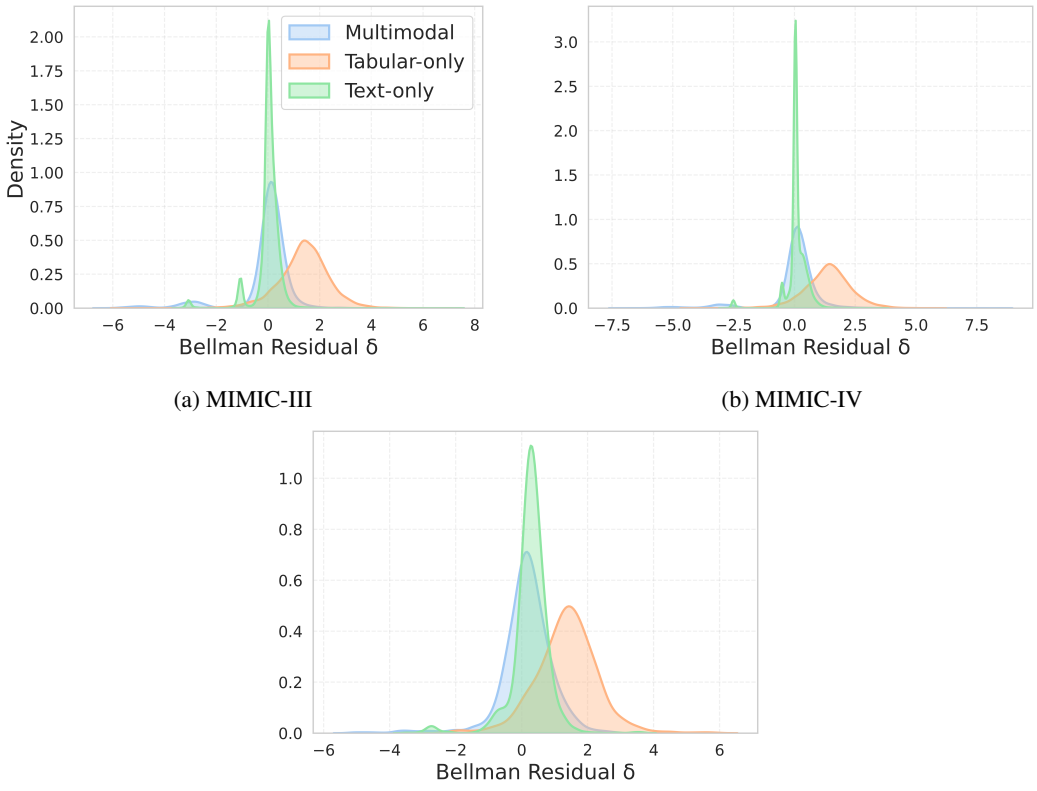
Figure 3: DR estimator performance across training iterations under cross-dataset validation

Train→Test	OPERA	FQE	WIS
MIMIC-III → MIMIC-IV	3.004 ± 0.148	2.116 ± 1.293	0.689 ± 0.033
PD → MIMIC-IV	3.107 ± 0.303	1.188 ± 0.110	0.677 ± 0.017
MIMIC-III → PD	2.011 ± 0.162	0.835 ± 0.363	0.731 ± 0.065
MIMIC-IV → PD	2.909 ± 0.148	0.897 ± 0.395	0.744 ± 0.049
PD → MIMIC-III	2.830 ± 0.088	1.150 ± 0.223	0.601 ± 0.038
MIMIC-IV → MIMIC-III	3.974 ± 0.115	1.488 ± 0.576	0.649 ± 0.018

Table 4: Cross-dataset validation results (CB+CQL)

6.4 Overestimation Analysis

Figure 4 presents the estimated probability density functions of the Bellman residuals for three policies: Multimodal (blue), Structured data-only (orange), and Text-only (green). It directly compares overestimation bias and estimation stability across datasets. In Figure 4a, the multimodal policy exhibits a distribution tightly concentrated around zero with minimal variance, indicating low bias and high reliability in value estimation. By contrast, the density of structured data-only policies is shifted to the right and substantially wider, reflecting a pronounced positive bias (overestimation) and unstable predictions. The text-only policy shows a sharp peak at zero but with a narrow spread, suggesting overconfident estimates that may under-represent actual estimation error. Figure 4b and Figure 4c further demonstrate that the structured data-only policy exhibits overestimation, while the text-only policy shows excessively high confidence in its estimates. In contrast, the multimodal approach achieves a more balanced value estimation between exploration and exploitation. These results demonstrate that integrating structured data with clinical notes via the MORE-CLEAR framework substantially reduces both bias and variance in Bellman residuals, yielding more stable and trustworthy value-function estimates than either unimodal strategy.



(a) MIMIC-III

(b) MIMIC-IV

(c) PD

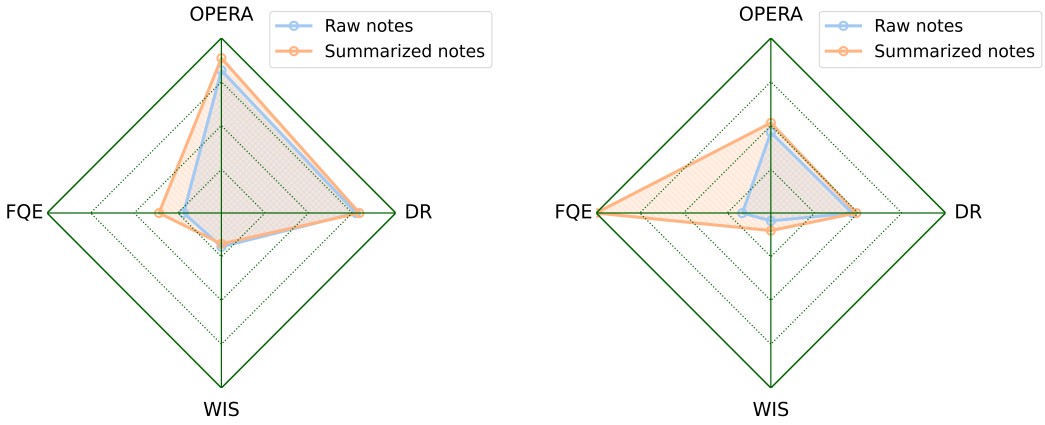
Figure 4: Bellman residual distributions across modalities

6.5 Summarization Performance

The MORE-CLEAR framework employs LLM-based summarization to mitigate the irregular distribution of unstructured information during sampling. Figure 5 presents results of OPE metrics for policies trained on either raw clinical notes or structured summarization notes. The policies leveraging summarized notes exhibit consistently larger envelopes in the radar plots. The OPERA scores extend further toward the outer grid, indicating a higher expected return under the learned policy. The DR also improves, reflecting more accurate value estimation. Improvement in FQE and WIS further suggests that utilizing summarized notes leads to stable estimation of both bias and variance. These results indicate that structured summarization of clinical notes enables more consistent learning than raw clinical notes.

6.6 Ablation Study

Table 5 presents the results of an ablation study that progressively integrates key components of the MORE-CLEAR framework, evaluating their impact on OPERA, DR, and WIS metrics. The primary results already established the efficacy of CQL. Therefore, we adopt BCQ as the base algorithm to better isolate and quantify the contribution of each individual module. The baseline configuration, CB+BCQ, achieves a modest OPERA score of approximately 0.66 in the PD cohort, reflecting limited expected return. Introducing the bidirectional cross-modal attention (BCMA) module yields a substantial improvement, elevating OPERA. The subsequent inclusion of the context-aware gated fusion (GF) mechanism further boosts OPERA by approximately 1.6 in the MIMIC-III cohort, although the improvement in PD is marginal. The addition of GF slightly decreases the WIS score by about 0.2. These findings suggest that BCMA is the primary contributor to performance gains. At the same time, GF offers auxiliary benefits by enhancing the quality of modality fusion, particularly within the MIMIC-III setting, albeit with dataset-dependent variability.



(a) MIMIC-III (b) MIMIC-IV

Figure 5: Radar plots comparing OPE metrics between raw and summarized clinical notes

	PD			MIMIC-III		
Model	**OPERA**	**DR**	**WIS**	**OPERA**	**DR**	**WIS**
CB+BCQ	0.663 ± 0.04	0.714 ± 0.11	0.528 ± 0.06	0.712 ± 0.13	0.657 ± 0.15	0.791 ± 0.05
CB+BCQ+BCMA	2.217 ± 0.15	2.262 ± 0.15	0.539 ± 0.25	2.190 ± 0.20	2.090 ± 0.19	0.688 ± 0.09
CB+BCQ+BCMA+GF	2.245 ± 0.11	2.395 ± 0.15	0.367 ± 0.08	3.730 ± 0.20	3.436 ± 0.14	0.692 ± 0.05

Table 5: Ablation study in the MORE-CLEAR framework

7 Conclusion

In this work, we proposed MORE-CLEAR, a multimodal offline reinforcement learning framework for clinical decision-making by integrating structured tabular data with unstructured clinical notes. The results demonstrate that the effective fusion of heterogeneous modalities enhances policy learning. The incorporation of textual modalities through bidirectional cross-modal attention and context-aware gated fusion has been demonstrated to enhance the expressiveness of patient state representations, thereby leading to the development of more robust and generalizable treatment policies. In particular, the use of context vectors from structured summarization of clinical notes contributes to policy performance. While the present evaluation concentrated on sepsis cohorts, the proposed framework is inherently generalizable and can readily be applied to other medical tasks.

References

[1] Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos one*, 17(4):e0266516, 2022.

[2] Mohammad Al Olaimat, Serdar Bozdag, and Alzheimer's Disease Neuroimaging Initiative. Ta-rnn: An attention-based time-aware recurrent neural network architecture for electronic health records. *Bioinformatics*, 40(Supplement_1):i169–i179, 2024.

[3] Siwar Chaabene, Brahim Haroun Hassan, Amal Boudaya, Lotfi Chaari, and Bassem Bouaziz. New mci detection method based on transformer and eeg data. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1200–1204. IEEE, 2023.

[4] Yunho Choi, Songmi Oh, Jin Won Huh, Ho-Taek Joo, Hosu Lee, Wonsang You, Cheng-mok Bae, Jae-Hun Choi, and Kyung-Joong Kim. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. *Communications Medicine*, 4(1):245, 2024.

[5] ChuGyouk. Kormedconceptsqa. `https://huggingface.co/datasets/ChuGyouk/KorMedConceptsQA`, 2024.

[6] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097.

[7] Tiago K Colicchio and James J Cimino. Clinicians' reasoning as reflected in electronic clinical note-entry and reading/retrieval: a systematic review and qualitative synthesis. *Journal of the American Medical Informatics Association*, 26(2):172–184, 2019.

[8] ContactDoctor. Bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

[9] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49(11):e1063–e1143, 2021.

[10] Nan Fang, Guiliang Liu, and Wei Gong. Offline inverse constrained reinforcement learning for safe-critical decision making in healthcare. *arXiv preprint arXiv:2410.07525*, 2024.

[11] Mehdi Fatemi, Mary Wu, Jeremy Petch, Walter Nelson, Stuart J Connolly, Alexander Benz, Anthony Carnicelli, and Marzyeh Ghassemi. Semi-markov offline reinforcement learning for healthcare. In *Conference on Health, Inference, and Learning*, pages 119–137. PMLR, 2022.

[12] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[13] Geoffrey J Gordon. *Approximate solutions to Markov decision processes*. Carnegie Mellon University, 1999.

[14] Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7(1):337, 2024.

[15] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

[16] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

[17] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

[18] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

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

[21] Pramod Kaushik, Sneha Kummetha, Perusha Moodley, and Raju S Bapi. A conservative q-learning approach for handling distribution shift in sepsis treatment strategies. *arXiv preprint arXiv:2203.13884*, 2022.

[22] Taylor W Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. *arXiv preprint arXiv:2011.11235*, 2020.

[23] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926, 2021.

[24] Onur Kocak, Tuncay Bayrak, Aykut Erdamar, Levent Ozparlak, Ziya Telatar, and Osman Erogul. Automated detection and classification of sleep apnea types using electrocardiogram (ecg) and electroencephalogram (eeg) features. *Advances in Electrocardiograms-Clinical Applications*, pages 211–230, 2012.

[25] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[26] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[27] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019.

[28] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33: 1179–1191, 2020.

[29] Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations, 2024. URL `https://arxiv.org/abs/2403.01469`.

[30] Hong Yeul Lee, Soomin Chung, Dongwoo Hyeon, Hyun-Lim Yang, Hyung-Chul Lee, Ho Geol Ryu, and Hyeonhoon Lee. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. *npj Digital Medicine*, 7(1):325, 2024.

[31] Hyeonhoon Lee, Hyun-Kyu Yoon, Jaewon Kim, Ji Soo Park, Chang-Hoon Koo, Dongwook Won, and Hyung-Chul Lee. Development and validation of a reinforcement learning model for ventilation control during emergence from general anesthesia. *npj Digital Medicine*, 6(1):145, 2023.

[32] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[33] Yi Li, Junli Zhao, Zhihan Lv, and Zhenkuan Pan. Multimodal medical supervised image fusion method by cnn. *Frontiers in neuroscience*, 15:638976, 2021.

[34] Yuting Li, Hongxiang Li, and Dong Zhang. Timing of norepinephrine initiation in patients with septic shock: a systematic review and meta-analysis. *Critical Care*, 24:1–9, 2020.

[35] Weijie Liang and Jinzhu Jia. Reinforcement learning using neural networks in estimating an optimal dynamic treatment regime in patients with sepsis. *Computer Methods and Programs in Biomedicine*, page 108754, 2025.

[36] A Rupam Mahmood, Hado P Van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 27, 2014.

[37] Paul E Marik, Walter T Linde-Zwirble, Edward A Bittner, Jennifer Sahatjian, and Douglas Hansell. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive care medicine*, 43:625–632, 2017.

[38] Greg S Martin, David M Mannino, Stephanie Eaton, and Marc Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 348(16): 1546–1554, 2003.

[39] Alexander Mathioudakis, Ilona Rousalova, Ane Aamli Gagnat, Neil Saad, and Georgia Hardavella. How to keep good clinical records. *Breathe*, 12(4):369–373, 2016.

[40] Mila Nambiar, Supriyo Ghosh, Priscilla Ong, Yu En Chan, Yong Mong Bee, and Pavitra Krishnaswamy. Deep offline reinforcement learning for real-world treatment optimization applications. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4673–4684, 2023.

[41] Karthik Natarajan, Daniel Stein, Samat Jain, and Noémie Elhadad. An analysis of clinical queries in an electronic health record search utility. *International journal of medical informatics*, 79(7):515–522, 2010.

[42] Peter C Nauka, Jason N Kennedy, Emily B Brant, Matthieu Komorowski, Romain Pirracchio, Derek C Angus, and Christopher W Seymour. Challenges with reinforcement learning model transportability for sepsis treatment in emergency care. *npj Digital Medicine*, 8(1):1–5, 2025.

[43] Allen Nie, Yash Chandak, Christina Yuan, Anirudhan Badrinath, Yannis Flet-Berliac, and Emma Brunskill. Opera: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators. *Advances in Neural Information Processing Systems*, 37:103652–103680, 2024.

[44] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL `https://arxiv.org/abs/2203.14371`.

[45] Gaurav Paraskar, Sankha Bhattacharya, and Anitha Kuttiappan. The role of proteomics and genomics in the development of colorectal cancer diagnostic tools and potential new treatments. *ACS Pharmacology & Translational Science*, 2025.

[46] Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Summits on Translational Science Proceedings*, 2014:218, 2014.

[47] Yucheng Ruan, Daniel J Tan, See Kiong Ng, Ling Huang, and Mengling Feng. Evidence-based multimodal fusion on structured ehrs and free-text notes for icu outcome prediction. *arXiv preprint arXiv:2501.04389*, 2025.

[48] James A Russell. Management of sepsis. *New England Journal of Medicine*, 355(16):1699–1713, 2006.

[49] Satya Narayan Shukla and Benjamin M Marlin. Integrating physiological time series and clinical notes with deep learning for improved icu mortality prediction. *arXiv preprint arXiv:2003.11059*, 2020.

[50] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[51] Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990, 2022.

[52] Juan Song, Jian Zheng, Ping Li, Xiaoyuan Lu, Guangming Zhu, and Peiyi Shen. An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis. *Frontiers in digital health*, 3:637386, 2021.

[53] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[54] Gemma Team. Gemma 3. 2025. URL `https://goo.gle/Gemma3Report`.

[55] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[56] Ariel Soares Teles, Ivan Rodrigues de Moura, Francisco Silva, Angus Roberts, and Daniel Stahl. Ehr-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Information Fusion*, page 102981, 2025.

[57] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

[59] Rui Tu, Zhipeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5(1):63–76, 2025.

[60] Jean-Louis Vincent, John C Marshall, Silvio A Ñamendys-Silva, Bruno François, Ignacio Martin-Loeches, Jeffrey Lipman, Konrad Reinhart, Massimo Antonelli, Peter Pickkers, Hassane Njimi, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations (icon) audit. *The lancet Respiratory medicine*, 2(5):380–386, 2014.

[61] Jason Waechter, Anand Kumar, Stephen E Lapinsky, John Marshall, Peter Dodek, Yaseen Arabi, Joseph E Parrillo, R Phillip Dellinger, Allan Garland, Cooperative Antimicrobial Therapy of Septic Shock Database Research Group, et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Critical care medicine*, 42(10):2158–2168, 2014.

[62] Yuan Wang, Anqi Liu, Jucheng Yang, Lin Wang, Ning Xiong, Yisong Cheng, and Qin Wu. Clinical knowledge-guided deep reinforcement learning for sepsis antibiotic dosing recommendations. *Artificial Intelligence in Medicine*, 150:102811, 2024.

[63] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[64] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[65] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Huan He, Lucila Ohno-Machido, Yonghui Wu, Hua Xu, and Jiang Bian. Me llama: Foundation large language models for medical applications, 2024.

[66] Hye Ju Yeo, Young Seok Lee, Tae Hwa Kim, Jin Ho Jang, Heung Bum Lee, Dong Kyu Oh, Mi Hyeon Park, Chae-Man Lim, Woo Hyun Cho, et al. Vasopressor initiation within 1 hour of fluid loading is associated with increased mortality in septic shock patients: analysis of national registry data. *Critical Care Medicine*, 50(4):e351–e360, 2022.

[67] Menghao Zhang, Minghao Xue, Shuying Li, Yun Zou, and Quing Zhu. Fusion deep learning approach combining diffuse optical tomography and ultrasound for improving breast cancer classification. *Biomedical Optics Express*, 14(4):1636–1646, 2023.

[68] Yinghao Zhu, Zixiang Wang, Long He, Shiyun Xie, Xiaochen Zheng, Liantao Ma, and Chengwei Pan. Prism: Mitigating ehr data sparsity via learning from missing feature calibrated prototype patient representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3560–3569, 2024.

A Data availability

The MIMIC datasets are publicly available through Physionet (https://physionet.org/about/database). Due to the ethical restrictions imposed by the IRB, the private dataset is only available upon reasonable request. The source code of the MORE-CLEAR framework can be found at : https://anonymous.4open.science/r/MORE-CLEAR-FD32

B LLM Performance in Note Summarization

The clinical notes contain a mixture of medical jargon in English and other languages. In order to assess the efficacy of LLM models in a bilingual setting, evaluations on question answering and summarization were conducted (Table 6). The Gemma3-27B-it performed well overall, thus it was employed for our note summarization.

LLMs	**Question Answering**										**Text Summarization**			
	PubMedQA[17]		MedMCQA[44]		MedQA[16]		KorMedMCQA[29]		KorMedConceptsQA[5]		Pubmed[6]		MIMIC-CXR[19]	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	R-1	BertS	R-1	BertS
Llama3.2-3B-it [58]	**0.734**	**0.514**	0.522	0.514	0.542	0.539	0.359	0.354	0.299	0.271	**0.381**	**0.837**	0.068	0.816
Llama3.1-8B-it [58]	**0.758**	**0.531**	**0.570**	**0.569**	**0.611**	**0.608**	**0.484**	**0.479**	**0.561**	**0.555**	**0.384**	0.836	0.078	0.825
Med-Llama3-8B [65]	0.660	0.332	0.480	0.479	0.563	0.281	0.198	0.064	0.538	0.531	0.335	**0.825**	**0.213**	**0.854**
Bio-Medical Llama3-8B [8]	0.478	0.346	0.333	0.320	0.426	0.421	0.200	0.084	0.267	0.245	0.325	0.826	0.080	0.828
Gemma3-27B-it [54]	0.474	0.423	**0.560**	**0.557**	**0.685**	**0.679**	**0.660**	**0.662**	0.342	0.303	0.363	**0.840**	**0.179**	**0.855**
Qwen3-32B [55]	0.454	0.309	0.297	0.296	0.637	0.454	0.318	0.271	**0.572**	0.469	0.355	0.830	0.091	0.834

Table 6: LLM Evaluation on Medical QA and Summarization Benchmarks. R-1: Rouge-1, BertS: BertScore

C Window size for note stack

Table 7 shows an ablation of context-window size on OPE in three cohorts. When the window size W is 3, the highest OPERA scores and DR estimates are achieved, indicating the highest expected return and minimal bias. As W increases, OPERA and DR decline monotonically, suggesting that compact contextual windows more effectively guide policy optimization. This tendency appears consistently in both FQE and WIS. However, WIS remains invariant comparably, implying that importance-weight variance is largely unaffected by window length.

Dataset	**Metric**	**Window Size**		
		W=3	**W=5**	**W=7**
MIMIC-III	↑ OPERA	2.705 ± 0.169	2.551 ± 0.154	2.316 ± 0.064
	↑ DR	2.412 ± 0.038	2.161 ± 0.124	1.906 ± 0.088
	↑ FQE	1.058 ± 0.489	1.666 ± 0.544	0.795 ± 0.208
	↑ WIS	0.672 ± 0.016	0.653 ± 0.024	0.630 ± 0.031
MIMIC-IV	↑ OPERA	3.581 ± 0.233	3.478 ± 0.173	3.002 ± 0.220
	↑ DR	3.375 ± 0.126	3.302 ± 0.113	2.757 ± 0.182
	↑ FQE	1.711 ± 0.491	1.389 ± 0.261	1.363 ± 0.393
	↑ WIS	0.794 ± 0.023	0.793 ± 0.035	0.772 ± 0.033
Private Dataset	↑ OPERA	2.403 ± 0.190	1.986 ± 0.079	1.465 ± 0.137
	↑ DR	2.088 ± 0.110	1.573 ± 0.061	1.112 ± 0.066
	↑ FQE	6.495 ± 5,347	3.910 ± 1.349	3.141 ± 1.468
	↑ WIS	0.721 ± 0.028	0.697 ± 0.039	0.719 ± 0.040

Table 7: Effect of stacking window size on performance