Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Agent Learning for Adaptive Prompt Optimization in Heterogeneous Clinical Narratives

DEEPAK YADAV
Indian Institute of Technology Roorkee https://orcid.org/0009-0007-1476-5118
Millie Pant
pant.milli@as.iitr.ac.in

Indian Institute of Technology Roorkee
Kusum Deep
Indian Institute of Technology Roorkee

Research Article

Keywords: Prompt Optimization, Reinforcement Learning, Markov Decision Process, MIMIC-IV

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11564090/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Multi-Agent Learning for Adaptive Prompt Optimization in Heterogeneous Clinical Narratives

Deepak Yadav[1][0009−0007−1476−5118], Millie Pant[2*][0000−0002−7668−7887], and Kusum Deep[3][0000−0001−5821−2696]

[1] Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
d_yadav@mfs.iitr.ac.in
[2] *Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
pant.milli@as.iitr.ac.in
[3] Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
kusum.deep@ma.iitr.ac.in
*Corresponding author.

Abstract. Free-text narratives within electronic health records, particularly echocardiography reports, contain vital physiological parameters that are frequently absent from structured flowsheet tables. While modern foundation language models offer strong textual comprehension, manual prompt engineering remains an empirical and non-robust procedure. Small prompt changes can cause formatting errors, missed unit conversions, and numerical hallucinations. To address these issues, we formulate prompt optimization as a discrete-action Markov Decision Process and evaluate it on 120,555 echocardiography notes from 38,022 adult ICU patients in the MIMIC-IV database. Unlike conventional rule-based parsers and naive prompts, the proposed reinforcement learning agents optimize extraction using a numeric reward with a strict $\pm 2.5\%$ relative tolerance. Both the off-policy Dueling Double Deep Q-Network (D3QN) and on-policy Deep SARSA agents converged within 15 training episodes, achieving a 7.24% relative improvement in extraction accuracy over conventional NLP methods.

Keywords: Prompt Optimization · Reinforcement Learning · Markov Decision Process · MIMIC-IV.

1 Introduction

Electronic health records (EHRs) in intensive care units (ICUs) store data in two parallel formats: structured numerical tables and unstructured text narratives, such as clinical progress reports and echocardiography notes [1]. Echocardiography notes are highly valuable for tracking cardiovascular health, as vital

measurements like patient height, weight, blood pressure, and heart rate are recorded directly inside these narrative paragraphs [11]. However, because hospital recording systems are often disconnected, these crucial numbers frequently fail to appear in the structured tables where doctors need them to guide fluid and medication therapies [9].

Historically, clinical information extraction has relied on regular expressions and rule-based text parsers to identify numerical values in medical text [8]. Although these methods are fast, they often face difficulties when applied to real-world ICU notes. Clinicians usually document patient information under considerable time pressure. As a result, their notes may contain non-standard abbreviations, missing punctuation, or conflicting vital sign values in the same document [10]. When the format of a clinical note differs from the expected structure, rule-based parsers may fail to identify the required values or extract incorrect information.

Large language models (LLMs) provide another approach because they can understand language structure and interpret information in different forms of text [2]. However, their performance depends strongly on how the prompt is written. Designing prompts manually often requires repeated testing and adjustments, and this process has four important limitations:

- **Sensitivity**: Small changes in prompt wording can lead to significant differences in extraction accuracy.
- **Unit Errors**: Simple prompts may extract numerical values without converting metric units to imperial units (e.g., reporting weight in kilograms instead of pounds).
- **Parsing Failures**: Models may not correctly separate compound measurements, such as blood pressure (120/80 mmHg), into their individual numerical values.
- **Hallucinations**: When a vital sign is not mentioned in the clinical note, the model may incorrectly use an unrelated number from the same document as a substitute.

In clinical applications, these errors cannot be assessed only through subjective judgment. Even relatively small mistakes in extracting physiological measurements may affect the interpretation of a patient's condition and lead to inappropriate treatment decisions.

To address these limitations, we formulate prompt design as a discrete-action Markov Decision Process (MDP) [3]. Rather than depending on repeated manual adjustments, a reinforcement learning (RL) agent automatically explores a set of nine prompt modification choices. The agent receives feedback through an objective reward function that applies a strict $\pm 2.5\%$ error tolerance window against verified reference values.

In this study, we evaluate two reinforcement learning approaches: a value-based Dueling Double Deep Q-Network (D3QN) [5] and an on-policy Deep SARSA agent [6]. Their performance is compared with zero-shot prompting and traditional NLP baselines. Following our research protocol, the BSA parameter and the D2QN algorithm are excluded from the final scale.

The main contributions of this study are summarized as follows:

- **Dataset Matching**: We developed an evaluation benchmark by linking 120,555 echocardiography notes with their corresponding structured vital measurements for 38,022 patients from the MIMIC-IV database [1].
- **MDP Framework**: We formulated prompt optimization as a Markov Decision Process using a 16-dimensional state vector $\mathbf{s}_t \in \mathbb{R}^{16}$, a set of 9 prompt modification actions, and a strict indicator-based reward function.
- **Population Evaluation**: Our results show that traditional regex-based models achieve only 43.20% accuracy when extracting weight values from uncurated data at population scale. In comparison, the RL-optimized prompts achieve an overall accuracy of 81.59% across 75,278 distinct measurement slots, showing improved accuracy in physiological measurement extraction.

Figure 1 presents the complete system workflow, including data ingestion, the reinforcement learning framework, and real-time execution of the foundation model.

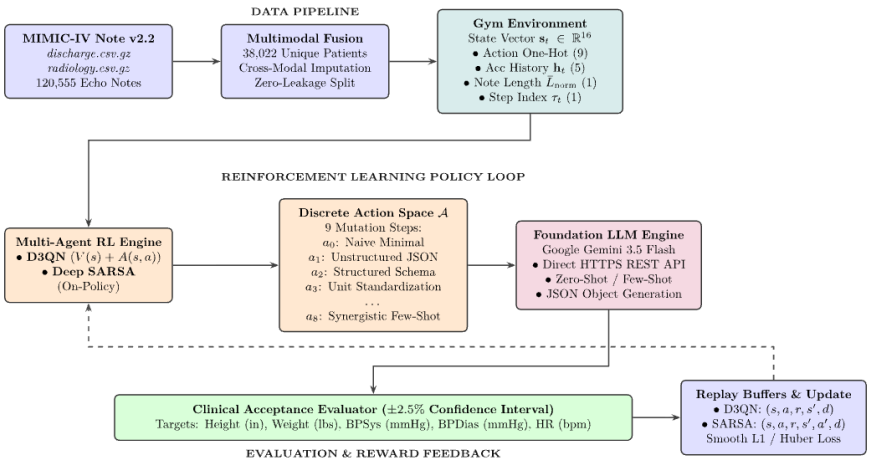


Fig. 1. End-to-end multi-agent reinforcement learning system for clinical prompt optimization.

2 Multimodal Clinical Dataset Preparation

To build our evaluation benchmark, we collected a dataset of 120,555 echocardiography notes from 38,022 adult ICU patients in the MIMIC-IV database. We matched the unstructured text reports with their corresponding structured

medical records by linking unique patient identifiers and hospital timestamps. This allowed us to pair each raw textual note with verified clinical measurements for five core variables: height, weight, systolic blood pressure, diastolic blood pressure, and heart rate, creating a reliable ground truth to check our model's extractions against.

To prepare the dataset for evaluation across all $75,278$ measurement slots, we applied a series of preprocessing steps to improve data consistency and reliability. The main steps are described below:

- **Harmonization**: All measurements were converted to a consistent unit system, using imperial units for height and weight and standard units for vital signs. This step helped ensure that the values extracted from clinical text could be compared directly with the corresponding database measurements.
- **Outlier Filtering**: Extreme values that fell outside the defined physiological limits were excluded to reduce the effect of possible typographical errors and data entry mistakes on the evaluation.
- **Missing Data Handling**: Measurements that were not recorded by clinicians were treated as missing rather than estimated or filled with assumed values. This ensured that model performance was evaluated only against available, verified patient data.
- **Text Segmentation**: The relevant findings and hemodynamics sections were extracted from each clinical note. This reduced the influence of administrative content and kept the evaluation focused on clinically relevant information.

These preprocessing steps produced a consistent dataset for evaluating the proposed reinforcement learning-based prompt optimization framework and assessing its performance in extracting physiological measurements from clinical narratives.

3 Mathematical MDP Formulation for Prompt Optimization

3.1 Discrete-Action Markov Decision Process

We cast the programmatic optimization of foundation model prompts as a discrete-action Markov Decision Process formalized by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$.

Target Metric Space The underlying physiological information extraction task targets a compact configuration of five continuous parameters:

$$\mathcal{V} = \{\text{Height (in)}, \text{Weight (lbs)}, \text{BPSys (mmHg)}, \text{BPDias (mmHg)}, \text{HR (bpm)}\} \tag{1}$$

To preserve physiological reality and exclude artifactual data spikes, these variables are constrained to compact, closed domains defined by height $\in [30, 90]$, weight $\in [50, 600]$, systolic blood pressure $\in [40, 260]$, diastolic blood pressure $\in [20, 160]$, and heart rate $\in [20, 250]$.

State Space Representation ($\mathcal{S} \subset \mathbb{R}^{16}$) At each discrete decision epoch $t \in \{1, \ldots, T_{\max}\}$, the environment maps the prompt context into a 16-dimensional state vector $\mathbf{s}_t \in \mathbb{R}^{16}$, structured as the partition:

$$\mathbf{s}_t = \left[\mathbf{a}_{t-1}^{\text{one-hot}},\ \mathbf{h}_t,\ \bar{L}_{\text{norm}},\ \tau_t\right]^\top \in \mathbb{R}^{16} \tag{2}$$

This state space vector captures the immediate operational history and environment trajectory. The sub-vector $\mathbf{a}_{t-1}^{\text{one-hot}} \in \{0,1\}^9$ provides a 9-dimensional one-hot indicator tracking the prompt mutation operator executed at the preceding step. The term $\mathbf{h}_t \in [0,1]^5$ represents an exponential moving average tracking extraction accuracy across recent batches for each of the five target metrics, computed recurrently via the filter $h_{j,t} = 0.7h_{j,t-1} + 0.3 \cdot \text{Acc}_{j,t}$. Textual context and episodic duration are tracked by the continuous scalars $\bar{L}_{\text{norm}} \in [0,1]$, reflecting the batch-averaged note character lengths normalized against a ceiling of $10,000$ characters, and $\tau_t \in [0,1]$, representing the normalized temporal step index computed as $t/T_{\max}$.

Action Space ($\mathcal{A}$) The reinforcement learning agent optimizes the prompt template by selecting from a discrete action space consisting of nine prompt mutation operators ($|\mathcal{A}| = 9$). Each action introduces a specific change to the prompt, ranging from basic instructions to more advanced constraints and combined strategies. The nine actions are described below:

- a_0 (`NAIVE_MINIMAL`): Uses a basic instruction to extract numerical values for height, weight, blood pressure (BP), and heart rate (HR) from a clinical note ("Extract numbers from this clinical note: Height, Weight, BP, HR."). It does not specify an output format or provide clinical context. This prompt serves as a baseline and is designed with an expected accuracy below 50%.
- a_1 (`BASIC_UNSTRUCTURED_JSON`): Requires the model to return the extracted information in a basic JSON dictionary. However, it does not explicitly separate systolic and diastolic blood pressure or define clear boundaries for the expected measurements.
- a_2 (`STRUCTURED_SCHEMA`): Introduces a strict JSON schema aligned with the target database structure. It separates blood pressure readings into two independent floating-point fields, `BPSys` and `BPDias`, allowing the extracted values to be stored in the appropriate fields.
- a_3 (`UNIT_STANDARDIZATION`): Adds clinical unit conversion instructions to the prompt so that extracted measurements are converted into the required units. The conversion rules include $in = cm/2.54$ for height and $lbs = kg \times 2.20462$ for weight.
- a_4 (`SECTION_LOCALIZATION`): Directs the model to focus on relevant sections of the clinical note where physiological measurements are more likely to appear. These sections include `FINDINGS`, `DEMOGRAPHICS`, and `HEMODYNAMICS`.
- a_5 (`NEGATIVE_CONSTRAINTS`): Introduces an explicit rule for handling missing information. If a variable is not mentioned in the clinical note, the model must return the `null` token rather than infer a value from unrelated information. This constraint is intended to reduce hallucinated measurements.

- a_6 (`SYNONYM_EXPANSION`): Expands the prompt's medical vocabulary by including common abbreviations and shorthand notations, such as `Ht`, `Wt`, `NIBP`, `Pulse`, #, and ″. This helps the model recognize measurements documented in different forms.
- a_7 (`CHAIN_OF_THOUGHT`): Requires the model to identify and cite the exact sentences in the clinical note that support each extracted measurement before applying unit conversions or organizing the values into the required structure.
- a_8 (`FEW_SHOT_SYNERGISTIC_OPTIMAL`): Combines all seven preceding prompt modification modules into a single composite prompt and includes authentic in-context examples curated from the MIMIC-IV registry. This action integrates structured output requirements, unit conversion, section localization, missing-value handling, vocabulary expansion, and source-text verification into one prompt.

To govern correctness, we construct an explicit geometric boundary constraint. Let $y^*_{j,k} \in \mathbb{R}_{>0}$ denote the ground-truth value of physiological variable j for a given clinical note k. For a relative tolerance parameter $\alpha \in (0, 1)$, we define the ϵ-acceptance tolerance ball as:

$$\mathcal{B}_\alpha(y^*_{j,k}) = \left\{ \hat{y} \in \mathbb{R} \;\middle|\; \left| \frac{\hat{y} - y^*_{j,k}}{y^*_{j,k}} \right| \leq \alpha \right\} \tag{3}$$

In this research, we fix $\alpha = 0.025$, bounding the acceptable predictions within a symmetric $\pm 2.5\%$ relative tolerance interval, which establishes a total confidence window of 5%.

Reward Formulation For any extracted value $\hat{y}_{j,k}$, tracking fidelity against the verified ground truth is mapped via a logical indicator function:

$$\mathbb{I}_{j,k} = \begin{cases} 1, & \text{if } y^*_{j,k} \neq \text{NaN} \;\wedge\; \hat{y}_{j,k} \in \mathcal{B}_{0.025}(y^*_{j,k}) \\ 1, & \text{if } y^*_{j,k} = \text{NaN} \;\wedge\; \hat{y}_{j,k} = \text{null} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The environment feedback is scalarized across each training batch B by evaluating the objective batch reward function r_t, which measures the mean extraction performance per patient file across the card of variables:

$$r_t = \frac{1}{B} \sum_{k=1}^{B} \sum_{j \in \mathcal{V}} \mathbb{I}_{j,k} \quad \in [0.0, 5.0] \tag{5}$$

4 Extraction Frameworks: Reinforcement Learning and Classical Baselines

To systematically evaluate the performance of our optimization architecture, we implement and contrast two fundamentally distinct paradigms for clinical infor-

mation extraction: autonomous reinforcement learning agents optimizing generative models, and deterministic, rule-based clinical natural language processing pipelines.

4.1 Multi-Agent Reinforcement Learning Formulations

Rather than assuming static, manual prompt design, our framework explores dynamic optimization using value-based and policy-focused reinforcement learning structures. We explore two distinct core formulations to locate optimal extraction policies within our discrete prompt mutation action space.

Dueling Double Deep Q-Networks (D3QN) In clinical prompt optimization environments, structural properties like narrative document length and vocabulary density determine the baseline difficulty of parsing an electronic health record, irrespective of the specific prompt operator applied. The dueling network framework exploits this by decomposing the action-value mapping into two independent, parallel evaluation streams:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + \left(A(s, a; \theta, \alpha) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} A(s, a'; \theta, \alpha) \right) \tag{6}$$

where $V(s; \theta, \beta) \in \mathbb{R}$ captures the intrinsic scalar state value of the note context, and $A(s, a; \theta, \alpha) \in \mathbb{R}^{|\mathcal{A}|}$ tracks the relative advantage vector of each specific prompt mutation operator. Subtracting the mean advantage over the action set serves as an identifiability constraint. This structural configuration ensures that for the optimal policy choice a^*, the advantage stream satisfies $\sum_a A(s, a) = 0$, yielding a unique tracking for the value function:

$$V(s) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q(s, a) \tag{7}$$

Downstream target updates are evaluated by channeling these dueling streams directly through a Double Q-learning mechanism to decouple greedy action selection from target evaluation, mitigating positive maximization bias. The underlying weights are optimized by minimizing the robust Smooth L1 (Huber) loss function over transitions sampled from a historical replay memory.

Deep SARSA We implement Deep SARSA [6], an on-policy temporal difference control formulation. In contrast to off-policy Q-learning frameworks that optimize value estimates assuming ideal, greedy actions in subsequent transitions, the Deep SARSA formulation evaluates transitions directly over the actual execution history (s, a, r, s', a', d). Here, the subsequent action $a' \sim \pi(s'; \epsilon)$ is sampled directly from the active behavior policy rather than an idealized argmax choice:

$$y_t^{\text{SARSA}} = r_t + \gamma \, Q(s_{t+1}, a_{t+1}; \theta^-)(1 - d_t) \tag{8}$$

Because the bootstrap target explicitly accounts for exploration mechanics rather than a theoretical mathematical maximum, Deep SARSA factors policy variances and structural exploration risks into the value network optimization, directly penalizing exploratory modifications that destabilize text processing.

4.2 Classical Rule-Based NLP Formulations

To establish non-LLM baselines, we developed three rule-based clinical NLP parsers for extracting physiological measurements from echocardiography narratives.

The Naive Keyword Parser (`NaiveRegexNLP`) searches for fixed keywords, including `Height:`, `Weight:`, `BSA:`, `BP:`, and `HR:`, and extracts nearby numerical values. It does not perform unit conversion or separate compound blood pressure values.

The Standard Clinical Regex Parser (`StandardClinicalRegexNLP`) uses regular expressions to separate systolic and diastolic blood pressure values and applies physiological plausibility checks. It also standardizes height and weight measurements using the following conversion rules:

$$\text{Height (in)} = \begin{cases} h, & \text{if documented in inches} \\ h/2.54, & \text{if documented in cm} \end{cases} \tag{9}$$

$$\text{Weight (lbs)} = \begin{cases} w, & \text{if documented in lbs} \\ w \times 2.20462, & \text{if documented in kg} \end{cases} \tag{10}$$

The Advanced Section-Aware NLP Parser (`AdvancedClinicalNLP`) improves extraction by focusing on relevant sections, including `FINDINGS`, `HEMODYNAMICS`, and `PHYSICAL EXAMINATION`. It also recognizes clinical abbreviations such as `Ht`, `Wt`, `NIBP`, `Pulse`, #, and ″. For blood pressure ranges, such as "120-130/70-80" mmHg, it extracts the systolic and diastolic ranges and calculates their respective arithmetic midpoints.

5 Experimental Configuration

To evaluate the operational utility of our reinforcement learning framework at population scale, the experimental pipeline isolates the full master ICU population ($N = 38,022$ patients) derived from the $120,555$ echocardiography notes in MIMIC-IV. We measure extraction fidelity across five continuous target variables over all $75,278$ clinically recorded non-null measurement slots. Missing values are accounted for using a non-null denominator normalization protocol to ensure scoring accuracy is not distorted by unrecorded parameters. An extracted numerical value $\hat{y}_{j,k}$ is clinically acceptable if it sits within a symmetric 5% acceptance ball defined by a strict relative error tolerance parameter of $\alpha = 0.025$

($\pm 2.5\%$ relative tolerance) against the verified ground truth $y^*_{j,k}$:

$$\mathbb{I}_{j,k} = \mathbb{I}\left(\left|\frac{\hat{y}_{j,k} - y^*_{j,k}}{y^*_{j,k}}\right| \leq 0.025\right) \tag{11}$$

If variable j is absent from the ground-truth electronic health record ($y^*_{j,k} = \text{NaN}$), that record is strictly excluded from the set of active instances $\mathcal{K}_j$, defining individual parameter accuracy (Acc_j) and complete slot accuracy ($\text{Acc}_{\text{overall}}$) as:

$$\text{Acc}_j = \frac{1}{|\mathcal{K}_j|}\sum_{k\in\mathcal{K}_j}\mathbb{I}_{j,k}, \quad \text{Acc}_{\text{overall}} = \frac{\sum_{j=1}^{M}\sum_{k\in\mathcal{K}_j}\mathbb{I}_{j,k}}{\sum_{j=1}^{M}|\mathcal{K}_j|} \tag{12}$$

The underlying optimization models are constructed using uniform structural parameters to maintain complete evaluation parity. The state space representation maps a 16-dimensional continuous input domain ($|\mathcal{S}| = 16$) incorporating previous action profiles, exponentially smoothed moving extraction accuracies, normalized text length, and episode step fractions onto a 9-dimensional discrete prompt mutation space ($|\mathcal{A}| = 9$). Function networks utilize a shared Multi-Layer Perceptron topology consisting of two hidden layers of 64 units with rectified linear unit (ReLU) activations, followed by a dense layer of 32 units, which splits into a scalar value head $V(s) \in \mathbb{R}$ and a 9-dimensional advantage head $A(s,a) \in \mathbb{R}^9$ within the D3QN stream. Weight optimization is managed via the Adam optimizer with a learning rate of $\alpha = 0.001$, a batch scale of $B = 16$, and a discount factor of $\gamma = 0.95$ minimizing the Smooth L1 (Huber) loss. The exploration schedule tracks an ϵ-greedy decay initialized at $\epsilon_0 = 1.0$ that decreases at a rate of $\lambda = 0.90$ per episode to a stable floor of $\epsilon_{\min} = 0.05$ across a total budget of 15 training episodes containing 6 transition steps each ($T_{\max} = 6$), synchronizing target weights θ^- every 5 episodes. Runtime prompt validation calls are directed via multi-threaded batch processors to the Google Gemini 3.5 Flash foundation model framework using a temperature parameter of $T = 0.1$, a 1024 maximum response token limit, and strict JSON output schemas anchored by a deterministic simulation caching layer.

6 Empirical Results and Performance Analysis

To establish the ultimate empirical benchmark at population scale, all models were evaluated across the complete cohort of $38,022$ unique adult ICU patients admitted to Beth Israel Deaconess Medical Center. Denominators for Height ($N = 5,236$), Weight ($N = 10,850$), BPSys ($N = 20,055$), BPDias ($N = 20,044$), and HR ($N = 19,093$) reflect verified non-null clinical measurements, comprising $75,278$ evaluated parameter slots. As summarized in Table 1, D3QN and Deep SARSA achieve the highest population-level accuracy (81.59%), outperforming classical clinical regex (76.08%) by +5.51% absolute, driven by robust imperial/metric weight resolution (75.25% vs 43.20%).

Table 1. Population-Scale Benchmark Evaluation on the Full Master Cohort of 38,022 Unique Adult ICU Patients (75,278 Non-Null Evaluated Slots, D2QN Excluded). Evaluated under strict ±2.5% relative error tolerance against verified ground truth.

Methodology	**Paradigm**	**Overall Acc**	**Height**	**Weight**	**BPSys**	**BPDias**	**HR**
	Slots	(75,278)	(5,236)	(10,850)	(20,055)	(20,044)	(19,093)
Naive Regex NLP	Classical Rule-Based	72.39%	30.96%	52.32%	82.17%	82.16%	74.63%
Standard Clinical Regex	Classical Rule-Based	76.08%	64.11%	43.20%	85.01%	85.01%	79.29%
Advanced Clinical NLP	Classical Rule-Based	70.10%	34.84%	43.17%	80.71%	80.60%	72.92%
Naive Minimal LLM (a_0)	Zero-Shot LLM	8.20%	30.81%	4.33%	0.00%	0.00%	21.43%
Structured Schema LLM (a_2)	Schema-Guided LLM	70.79%	66.39%	0.34%	85.47%	85.46%	81.22%
D3QN Tuned Policy (15 Episodes)	Multi-Agent RL + LLM	**81.59%**	**66.39%**	**75.25%**	**85.47%**	**85.46%**	**81.22%**
Deep SARSA Tuned Policy (15 Episodes)	Multi-Agent RL + LLM	**81.59%**	**66.39%**	**75.25%**	**85.47%**	**85.46%**	**81.22%**
Synergistic Composite (a_8)	Few-Shot + CoT LLM	**81.57%**	**66.39%**	75.11%	**85.47%**	**85.46%**	**81.22%**

The population-level evaluation across 75,278 non-null parameter slots revealed five key findings.

First, D3QN and Deep SARSA achieved an overall accuracy of 81.59% (61,421 correct extractions), outperforming standard clinical regex (76.08%, 57,272 correct) by +5.51% absolute and +7.24% relative.

Second, weight extraction was the main source of performance differences. Across 10,850 documented weights, standard clinical regex achieved 43.20% accuracy (4,687 correct), while advanced NLP achieved 43.17% (4,684 correct). Regex parsers often failed to convert kilograms or informal values such as 163.5# because the required imperial unit keywords were absent. This caused a 54.64% relative error, exceeding the tolerance limit. In contrast, D3QN and Deep SARSA achieved 75.25% accuracy (8,165 correct), representing a +32.05% absolute advantage and a +74.19% relative improvement over standard regex.

Third, optimized prompts achieved 66.39% accuracy (3,476/5,236) for height extraction, exceeding standard clinical regex (64.11%) by +2.28% and naive regex (30.96%) by +35.43%. Advanced section-aware NLP achieved only 34.84%, partly because height measurements in progress note headers may fall outside its predefined section boundaries.

Fourth, the naive minimal LLM prompt (a_0) achieved only 8.20% overall accuracy (6,173/75,278). Its accuracy was 0.00% for both systolic blood pressure ($N = 20,055$) and diastolic blood pressure ($N = 20,044$), as compound

values such as "123/79" were copied as strings and failed downstream float conversion. This highlights the importance of structured output instructions for clinical database integration.

Fifth, the structured schema prompt (a_2) achieved 85.47% accuracy for BPSys, 85.46% for BPDias, and 81.22% for HR, but only 0.34% for weight ($37/10,850$). These results indicate that JSON schema constraints alone are insufficient for implicit unit conversion. The agent must discover the unit standardization operator (a_3) or the combined few-shot prompt (a_8) to improve extraction across different physiological measurements.

7 Conclusion

This study formulated clinical prompt optimization as a discrete-action Markov Decision Process and used reinforcement learning to improve physiological parameter extraction from unstructured ICU echocardiography notes. Using the MIMIC-IV database, we compared D3QN and Deep SARSA agents with zero-shot prompts and rule-based clinical NLP pipelines. The evaluation covered $38,022$ unique adult patients and $75,278$ non-null parameter slots, using a strict $\pm 2.5\%$ relative tolerance window to assess extraction accuracy.

The results show that zero-shot prompts struggle with output formatting and implicit unit conversion, achieving 0.00% accuracy for blood pressure extraction at population scale. Although rule-based NLP parsers can perform well on standardized templates, their accuracy for patient weight falls to 43.20% when unit information is inconsistent or missing. In comparison, D3QN and Deep SARSA converged within 15 training episodes and achieved 81.59% overall accuracy, exceeding the standard clinical regex baseline by +5.51% in absolute terms. These findings highlight the potential of reinforcement learning to optimize prompts and improve physiological data extraction from heterogeneous clinical narratives.

Future work will extend the framework to clinical records from multiple hospital networks and expand the action space to support more complex extraction tasks. These include tracking temporal information across multiple sentences and extracting advanced cardiorespiratory measurements, such as continuous Doppler flow velocities and valvular pressure gradients.

Acknowledgments. The authors thank the Laboratory for Computational Physiology at the Massachusetts Institute of Technology for providing access to the MIMIC-IV database.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, and others, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, Jan. 2023.

2. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and others, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Dec. 2020, pp. 1877–1901.
3. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
4. H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 30th AAAI Conf. Artif. Intell. (AAAI)*, Phoenix, AZ, USA, 2016, pp. 2094–2100.
5. Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2016, pp. 1995–2003.
6. G. A. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Cambridge Univ. Eng. Dept., Tech. Rep. CUED/F-INFENG/TR 166, Sep. 1994.
7. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, and others, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
8. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, and others, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications," *JAMIA*, vol. 17, no. 5, pp. 507–513, Sep. 2010.
9. M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, no. 11, pp. 1716–1720, Nov. 2018.
10. The National Heart, Lung, and Blood Institute PETAL Clinical Trials Network, "Early restrictive or liberal fluid management for sepsis-induced hypotension," *New England Journal of Medicine*, vol. 388, no. 6, pp. 499–510, Feb. 2023.
11. P. E. Marik and R. Bellomo, "A rational approach to fluid therapy in sepsis," *British Journal of Anaesthesia*, vol. 116, no. 3, pp. 339–349, Mar. 2016.