Adaptive STEM Learning Pathway Optimisation via Reinforcement Learning

1st Nhu-Tai Do
Dept. of Information Technology
Saigon University, Vietnam
Ho Chi Minh City, Vietnam
dntai@sgu.edu.vn

2st Nguyen Huu Loc
Dept. of Information Technology
Saigon University, Vietnam
Ho Chi Minh City, Vietnam
lockbkbang@gmail.com

3st Van Tuan Kiet
Dept. of Information Technology
Saigon University, Vietnam
Ho Chi Minh City, Vietnam
vankiet27012004@gmail.com

4st Nguyen Thi Ngoc Thanh
Dept. of Information Technology
Ho Chi Minh City Open University, Vietnam
Ho Chi Minh City, Vietnam
thanh.ntn@ou.edu.vn

***Abstract*—In the context of education 4.0, conventional learning management systems often lack effective personalization mechanisms, typically enforcing a uniform learning path for all learners. To address this limitation in STEM education, this paper proposes an adaptive learning framework based on the Q-learning algorithm, integrated into the Moodle platform via the LTI 1.3 standard. The learning process is modeled as a Markov decision process, combined with behavioral clustering to construct a multidimensional learner state space. To mitigate data scarcity and the cold-start problem, we introduce a data-driven simulation strategy that preserves the distributional characteristics of real learner behavior. Experimental results from 500 simulation iterations demonstrate that the proposed approach significantly outperforms traditional methods, achieving a 22.5% improvement in average scores and a 51.0% reduction in weak skills, thereby highlighting its potential for scalable, personalized STEM education.**

***Index Terms*—Reinforcement learning, Q-learning, personalized learning, STEM education, data-driven simulation, Markov decision process**

I. Introduction

The rapid advancement of artificial intelligence is profoundly reshaping multiple domains, including education. According to the seminal study by Frey and Osborne, approximately 47% of traditional occupations are at risk of automation [1], highlighting the urgent need to equip the workforce with new competencies, particularly in STEM. STEM education emphasizes the development of critical thinking and problem-solving skills; however, its effective implementation remains challenging due to the substantial heterogeneity in learners' abilities, prior knowledge, and learning paces.

One of the most pressing challenges in contemporary education is the realization of large-scale personalized adaptive learning. Conventional learning management systems (LMS), such as Moodle and Blackboard, primarily function as repositories for learning materials and assessment records, offering limited capability for behavioral analysis and timely pedagogical intervention [2]. In the Vietnamese context, existing studies on AI applications in education have predominantly focused on predictive tasks, such as forecasting dropout risks or final examination outcomes [3], [4], while comparatively little attention has been paid to prescriptive approaches that actively recommend pedagogical actions to improve learning outcomes.

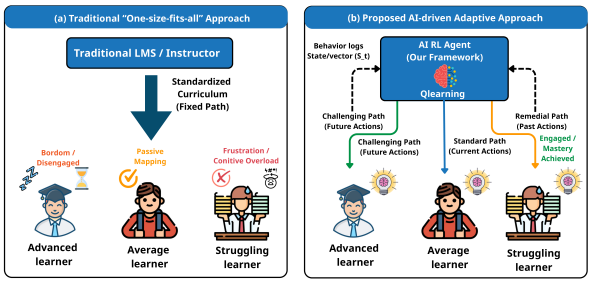

Figure 1. Comparison of learning approaches. (a) Traditional methods apply a uniform learning path to all learners, often resulting in suboptimal engagement due to boredom or cognitive overload. (b) The proposed framework employs an AI agent to infer learner states from interaction logs and generate personalized action recommendations (review, standard, or enrichment) to maximize engagement and knowledge mastery.

To bridge this gap, there is a growing demand for personalized STEM teaching and learning support systems based on reinforcement learning (RL), as shown in Fig.1. Unlike traditional rule-based systems, RL enables an intelligent agent to autonomously explore and optimize instructional strategies through a trial-and-error learning paradigm, continuously adapting its policy based on learner feedback and observed outcomes [5].

This study contributes to the field of personalized adaptive learning through the following three main aspects:

Corresponding author: thanh.ntn@ou.edu.vn

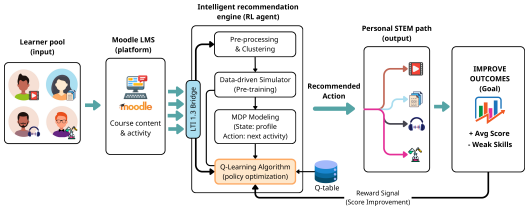

Figure 2. Proposed method overview. Moodle behavioral logs are transformed into learner states. Based on the current state, the agent selects an action and receives feedback in the form of rewards.

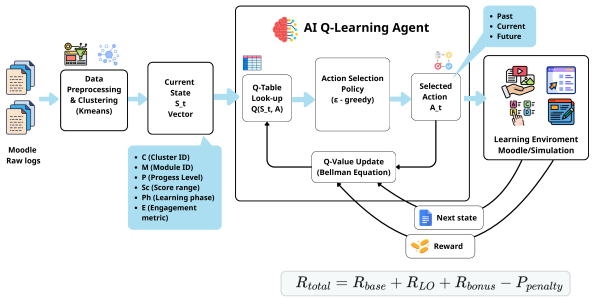

Figure 3. Detailed flowchart of the Q-learning procedure for generating pedagogical recommendations.

(1) For the adaptive learning framework, we propose a reinforcement learning framework that models the learning process as a Markov Decision Process (MDP) with a multi-objective reward function. The framework integrates learner clustering results with behavioral learning theory, specifically the ICAP framework, to capture both cognitive engagement and performance-oriented objectives [6]; (2) With a data-driven simulation process, to address the challenges of data scarcity and the cold-start problem commonly encountered in educational settings, we design a data-driven simulation environment parameterized by statistical characteristics extracted from real course data, enabling a realistic approximation of learner behavior dynamics [7]; (3) About experimental validation, the effectiveness of the proposed approach is verified through A/B testing on a simulated dataset, where the Q-learning-based policy consistently outperforms traditional strategies in terms of academic performance and learner engagement, demonstrating its potential for scalable personalized STEM education.

II. Proposed Method

A. Problem Overview

To overcome the limited personalization capabilities of conventional learning management systems, this study proposes an adaptive learning framework based on the Q-learning algorithm. The proposed pipeline starts from raw behavioral logs recorded by Moodle, applies preprocessing and feature engineering to construct a compact learner state space, and then trains an AI agent to select pedagogical actions that maximize long-term learning benefits. Fig. 2 summarizes the overall workflow.

Personalized learning pathway recommendation is formulated as a MDP defined by the triplet $\langle S, A, R \rangle$, where S denotes the state space, A the action space, and R the reward function.

1) State Space S: At time step t, the system observes the learner state S_t. To ensure generality and extensibility, we represent S_t as a d-dimensional feature vector:

$$S_t = \{f_1, f_2, \ldots, f_d\}. \tag{1}$$

In this study, the learner state is characterized by a compact yet informative feature set that captures both behavioral patterns and learning progress. Specifically, learners are first assigned to a behavioral cluster C obtained through unsupervised clustering techniques. The current learning context is represented by the module index M and the corresponding completion level P. Academic performance is summarized by a discretized cumulative score level Sc. In addition, the learning process is contextualized by the learning phase Ph, aligned with the ICAP framework, and by the engagement level E, which reflects the intensity and quality of learner interactions derived from activity logs.

2) Action Space A: Given S_t, the agent chooses an action $a_t \in A$ from a finite set of m pedagogical actions $A = \{a_0, a_1, \ldots, a_{m-1}\}$. Actions are organized along a temporal axis (past, present, and future) to support different instructional intents, including remedial review for struggling learners and enrichment for advanced learners.

3) Reward Function R: The objective is to maximize the accumulated reward over time. To reflect multiple pedagogical goals, we define a multi-objective reward:

$$R_{\text{total}} = R_{\text{base}} + R_{\text{LO}} + R_{\text{bonus}} - P_{\text{penalty}}. \tag{2}$$

Here, R_{base} provides a baseline incentive for meaningful engagement, R_{LO} captures learning-outcome attainment, R_{bonus} rewards beneficial behavior sequences, and P_{penalty} penalizes inefficient or unproductive actions.

B. Data Processing and Clustering

Raw LMS logs are typically noisy and unstructured. Before being used by the RL agent, the logs are preprocessed (filtering, normalization, and feature extraction) to obtain a compact behavioral representation. We then apply K-means to partition learners into K clusters with similar behavioral characteristics.

The optimal number of clusters is selected using three complementary validation criteria: the Elbow method (inertia), the Silhouette coefficient (inter-cluster separation), and the Davies-Davies–Bouldin (cluster compactness). A majority voting scheme aggregates the three criteria to determine the final K. The resulting cluster ID is incorporated as a key component of the state vector S_t.

C. Q-learning Algorithm

We adopt Q-learning to learn the optimal state and state-action function (Q-table) using Bellman's update rule shown in Fig.3:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right], \tag{3}$$

where α is the learning rate and γ is the discount factor. An ϵ-greedy strategy is employed to balance exploration and exploitation during training.

D. Explainability Framework

To mitigate the black-box nature of the learned Q-table, we integrate SHAP [8] to quantify how each state feature contributes to the agent's decision.

1) Mathematical Basis: The SHAP value $\phi_i(s)$ measures the marginal contribution of feature i to the predicted utility:

$$\phi_i(s) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]. \tag{4}$$

The additive property ensures:

$$f(s) = \phi_0 + \sum_{i=1}^{6} \phi_i(s). \tag{5}$$

To assess global importance, we compute the mean absolute SHAP value over N randomly sampled states:

$$I_i = \frac{1}{N} \sum_{j=1}^{N} |\phi_i(s_j)|. \tag{6}$$

2) KernelExplainer Implementation: Shapley value approximation is performed in three steps:

1) **Prediction function:** Define a wrapper function $f(s) = \max_a Q(s,a)$ that maps a state s to its maximum expected utility.
2) **Background sampling:** Use a background set D_{bg} ($N_{\text{bg}} = 100$) to approximate the baseline expectation $E[f(x)]$.
3) **Computation:** For each test state, KernelExplainer evaluates all $2^6 = 64$ feature coalitions, yielding accurate attributions with a computational cost of $O(N \cdot 2^D)$.

E. Data-driven Simulation Framework

To address the *cold-start* issue and promote stable convergence before real deployment, we propose a **parameter mining** procedure that transforms historical logs into probabilistic parameters to operate a classroom *digital twin*.

1) Action Mapping and Contextualization: Low-level Moodle events are normalized into a unified action space A comprising 15 canonical actions. To refine decision context, each action is labeled with a temporal context derived from learner progress P_t:

$$Context(a_t) = \begin{cases} \text{Past}, & \text{if } P_t < 25\%, \\ \text{Current}, & \text{if } 25\% \le P_t < 85\%, \\ \text{Future}, & \text{if } P_t \ge 85\%. \end{cases} \tag{7}$$

2) Transition Dynamics Modeling: Using historical data $\mathcal{D}_{history}$, we estimate a transition probability matrix (TPM) $\mathcal{P}$. For each state–action pair (s,a), the probability of transitioning to the next state s' is estimated by:

$$P(s' \mid s,a) = \frac{count(s,a,s')}{\sum_{s^*} count(s,a,s^*)}. \tag{8}$$

The matrix $\mathcal{P}$ governs the simulated environment, ensuring that its responses are consistent with observed learner dynamics.

3) Param Policy Baseline: For benchmarking, we construct a **Param Policy** baseline π_{param} from historical action frequency distributions, stratified by learning phase and learner cluster:

$$\begin{aligned} \pi_{\text{param}}(a \mid s) &= P(a \mid \text{Phase}(s), \text{Cluster}(s)) \\ &= \frac{\text{count}\big(\text{Phase}(s), \text{Cluster}(s), a\big)}{\sum_{a'} \text{count}\big(\text{Phase}(s), \text{Cluster}(s), a'\big)}. \end{aligned} \tag{9}$$

This baseline mimics learners' natural historical tendencies and serves as a reference point against the optimized Q-learning policy π^*.

4) Training and Evaluation Procedure: The simulation is executed in a closed-loop manner:

1) **Initialization:** Instantiate student agents whose behavioral profiles are parameterized from cluster statistics.
2) **Training:** The RL agent interacts with student agents in the environment $\mathcal{P}$ to optimize the Q-table via trial-and-error.
3) **Comparison:** Evaluate and compare performance (Reward, Score, LO Mastery) between the learned policy π^* and the baseline π_{param}.

III. Experiments and Results

A. Simulation Setup

1) Learning Style Modeling: Virtual student agents in the simulation are not assumed to behave randomly; instead, they are endowed with heterogeneous and non-linear learning styles in order to better approximate authentic classroom behavior. Based on empirical observations from the dataset, the population is configured with three dominant learning styles: the majority are linear learners (70%), who follow a conventional sequential pathway by progressing through learning materials and activities in the prescribed order; a smaller group of practice-first learners (10%), who prioritize attempting quizzes or exercises before engaging

with theoretical content; and a group of video/reading-first learners (20%), who prefer passive content consumption such as videos or readings prior to active interaction.

2) Simulation settings and stochasticity: To ensure objectivity and reproducibility, the simulation environment is configured with explicit control over scale, randomness, and termination conditions.

Scale and reproducibility: Training is conducted over 500 episodes. In each episode, a population of 100 virtual student agents is initialized, resulting in a total of 50,000 simulated interaction trajectories. This scale provides sufficient coverage of the discretized state space. All experiments use a fixed random seed of 42 to ensure reproducibility across runs.

Noise modeling: Stochastic components are incorporated to reflect variability in real-world learner behavior. Score variation σ obtained after each action is perturbed by uniform noise to simulate performance fluctuations:

$$S_{\text{real}} = \text{clip}\left(S_{\text{base}} + \mathcal{U}(-\sigma_c, \sigma_c), 0, 1\right), \tag{10}$$

where σ_c denotes the variability characteristic of each learner group. The Weak group exhibits the highest variability ($\sigma = 0.18$), followed by the Medium group ($\sigma = 0.10$), while the Strong group shows the lowest variability ($\sigma = 0.05$).

Time variation required to complete an action is randomly sampled from a uniform distribution to capture differences in information processing speed:

$$T \sim \mathcal{U}(5, 30) \text{ minutes}. \tag{11}$$

Termination conditions: Each episode terminates when the learner agent completes all $N = 6$ modules of the course or when a maximum step limit of 100 is reached. This constraint prevents infinite loops during the early exploration phase of training.

Table I
SUMMARY OF SIMULATION CONFIGURATION PARAMETERS

Parameter	Description	Value
N_{episodes}	Number of training episodes	500
Seed	Random seed	42
N_{modules}	Number of course modules	6
Cluster-dependent parameters:		
P_{success} (Weak)	Baseline success probability	0.72
P_{success} (Medium)	Baseline success probability	0.78
P_{success} (Strong)	Baseline success probability	0.90
α_{learn} (Weak)	Learning rate	0.22
α_{learn} (Medium)	Learning rate	0.32
α_{learn} (Strong)	Learning rate	0.30

B. Experimental Setup

1) Ground-Truth Dataset: This study uses the publicly available *Moodle Log & Grades* dataset [9]. Among the available courses, we selected **Course ID 670** as the empirical basis for constructing the data-driven simulation environment.

This choice is motivated by its balanced data characteristics. The course contains 13,995 interaction events from 23 students, with an approximately normal score distribution ($\mu = 7.64$, $\sigma = 2.95$). In contrast, several other courses exhibit highly skewed score distributions (e.g., Course ID 42 with a mean score of 1.07), which can obscure behavioral differences and hinder the learning agent's ability to distinguish strategies across competency groups (e.g., Excellent, Average, and Weak). Therefore, Course ID 670 provides a reliable foundation for both clustering analysis and policy learning.

2) Data Preprocessing and Clustering: We apply K-means clustering due to its efficiency and suitability for numerical feature spaces shown in Fig.4. To improve robustness and avoid subjective parameter tuning, the clustering pipeline incorporates (i) a two-stage feature filtering procedure and (ii) a majority voting mechanism for selecting the optimal number of clusters.

1) Feature selection and filtering: All numerical features are first standardized using Z-score normalization. We then refine the feature space using two complementary filters:

- **Variance filtering:** Remove features with variance below 0.01, as such low-variability features provide limited discriminative power.
- **Correlation filtering:** Remove highly correlated features (Pearson correlation > 0.95) to reduce redundancy and mitigate multicollinearity. This step eliminates 78 duplicated features and stabilizes the clustering structure.

2) Cluster optimization K: The number of clusters K is determined by aggregating three validation criteria shown in Fig. 4: the elbow method (inertia), the silhouette coefficient (separation), and the Davies-Bouldin index (compactness). The three criteria are combined via a majority voting rule.

- **Elbow method:** The inertia curve exhibits a clear knee point at $K = 6$.
- **Silhouette coefficient:** Peaks at $K = 2$ (0.42) but remains acceptable at $K = 6$ (≈ 0.35).
- **Davies–Bouldin index:** Shows a favorable local minimum at $K = 6$ (lower is better).

3) Rationale for choosing $K = 6$: Although $K = 2$ yields the highest Silhouette score, partitioning learners into only two groups (e.g., Excellent vs. Weak) is overly coarse and does not provide sufficient granularity for the agent to optimize differentiated pedagogical actions. Conversely, selecting a large number of clusters (e.g., $K \geq 10$) substantially increases complexity and leads to state-space explosion. Therefore, it $K = 6$ offers a practical trade-off, capturing six representative learner competency profiles while remaining computationally tractable.

During the cluster inspection, one cluster was identified as corresponding to instructor behavior, characterized by a high frequency of grading-related and forum moderation events. This cluster was excluded from the training data. The remaining five student clusters are subsequently mapped

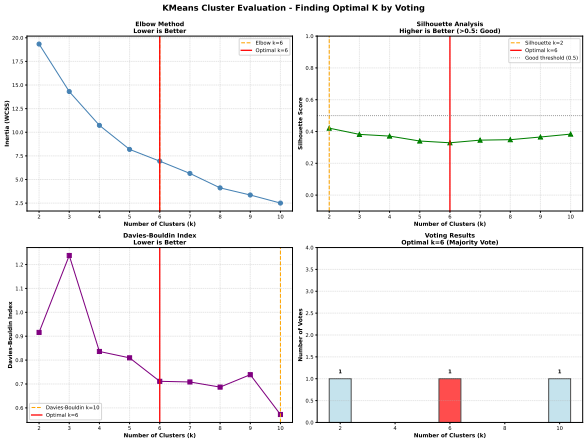

Figure 4. Clustering validation metrics. The Elbow and Davies-Bouldin criteria indicate $K = 6$, while the Silhouette coefficient remains acceptable at this value.

Table II
PROBABILITY OF ACTION SELECTION BY LEARNING PHASE

Phase	**Action** (a_t)	**Probability** (P)
Phase 0 (Pre-learning)	view_assignment	56.4%
	view_content	43.6%
Phase 1 (Active)	**attempt_quiz**	**74.4%**
	submit_assignment	25.6%
Phase 2 (Reflective)	post_forum	100.0%

into three pedagogical groups (weak, medium, and strong) to support adaptive reward shaping in later stages.

3) Modeling transition dynamics and baseline: The fidelity of the proposed simulation environment is governed by the transition probability matrix (TPM), which captures how learners evolve from one state to another in response to pedagogical actions. Transition probabilities are estimated from historical interaction data through detailed frequency analysis. The resulting probability distributions are organized into structured tables and subsequently used as input parameters for constructing the parametric baseline policy (π_{param}), which serves as a reference model of natural learner behavior.

1) Action distribution across learning phases: Table II summarizes the empirical distribution of learner actions across different learning phases. A clear progression in behavioral patterns can be observed: during the pre-learning phase, learners predominantly engage in passive activities (viewing assignments and content); in the active phase, behavior shifts toward constructive actions such as quiz attempts and assignment submissions; and in the reflective phase, interaction becomes exclusively discussion-oriented through forum posting.

2) Action distribution by engagement level: Table III reports action selection patterns conditioned on learner engagement levels. Highly engaged learners exhibit a

Table III
ACTION PROBABILITY DISTRIBUTION BY ENGAGEMENT LEVEL

Engagement Level	**View Assign**	**View Content**	**Quiz**	**Submit**	**Forum**
High	52.2%	40.5%	5.8%	1.5%	0.02%
Medium	51.7%	41.9%	4.3%	2.2%	–
Low	**57.6%**	34.0%	4.8%	3.6%	–

balanced strategy, alternating between content consumption and task completion. In contrast, learners with low engagement predominantly focus on viewing assignment requirements (*view_assignment*), indicating a more passive and preparatory interaction style.

The extracted probability distributions are exported as structured JSON configurations and used to drive the stochastic yet directed behavior of virtual student agents during simulation. This design ensures that the baseline policy and transition dynamics closely mirror empirically observed learner behaviors, providing a realistic benchmark against which the reinforcement learning policy can be evaluated.

4) Model Parameters: To ensure stable learning dynamics and convergence of the Q-learning algorithm, the model parameters were carefully configured as follows.

State Space (S): The learner state at time t, denoted by S_t, is represented as a 6-dimensional contextual vector:

$$S_t = (C, M, P, Sc, Ph, E) \tag{12}$$

where C denotes the learner behavior cluster obtained from the K-means model with $K = 6$. After excluding one cluster corresponding to instructor activity, the remaining five student clusters are encoded as $C \in \{0, 1, 2, 3, 4\}$. M represents the current module index; P and Sc denote the learner's progress and score level, respectively, both discretized into four ordinal levels. Ph indicates the learning phase (0: pre-learning, 1: active learning, 2: reflective learning), while E captures the engagement level (low, medium, high). Under this formulation, the total state space comprises $5 \times 6 \times 4 \times 4 \times 3 \times 3 \approx 4{,}320$ distinct states, balancing representational richness and computational tractability.

Action Space (A): The action space consists of $m = 15$ pedagogical actions, selected based on the ICAP framework and filtered using a Pareto threshold ($> 1\%$ occurrence frequency). For interpretability and policy design, actions are organized according to their temporal context (past, current, future), as summarized in Table V.

Q-learning Hyperparameters: The learning rate is set to $\alpha = 0.1$, and the discount factor to $\gamma = 0.95$. An ϵ-greedy exploration strategy with a monotonically decreasing ϵ is employed to balance early-stage exploration and late-stage exploitation, thereby facilitating convergence toward an optimal policy.

5) State Space Modeling: The adaptive learning problem is formulated as a MDP, in which the learner state at time

Table IV
DETAILED DEFINITION AND VALUE DOMAIN OF THE STATE SPACE

Symbol	Component	Definition / Discretization (Bins)	Size
C	Cluster	Student behavior group after removing the instructor cluster from the original K-means model ($ID \in \{0, 1, 2, 3, 4\}$)	5
M	Module	Current lesson index ($ID \in \{0, \ldots, 5\}$)	6
P	Progress	Beginner ($< 25\%$) Learning (25%–50%) Nearly finished (50%–99%) Completed (100%)	4
Sc	Score	Weak (< 2.5) Average ($2.5 \leq s < 5.0$) Fair ($5.0 \leq s < 7.5$) Excellent (≥ 7.5)	4
Ph	Phase	0: Pre-learning (passive) 1: Active learning (interactive/exercises) 2: Reflective learning (review/discussion)	3
E	Engagement	Low ($S_{total} < 8$) Medium ($8 \leq S_{total} < 16$) High ($S_{total} \geq 16$)	3

t is represented by a compact yet expressive 6-dimensional vector that integrates both static learner characteristics and dynamic behavioral signals:

$$S_t = (C, M, P, Sc, Ph, E) \tag{13}$$

Here, it C denotes the learner behavior cluster, M the current module index, P the learning progress, Sc the score level, Ph the learning phase, and E the engagement level. This joint representation enables the agent to capture not only where the learner is in the course structure but also how they are learning.

To guarantee tractable learning and ensure convergence of the Q-table under limited computational resources, all continuous variables are discretized using pedagogically motivated thresholds. The detailed definitions and value domains of each state component are summarized in Table IV. This discretization strategy strikes a balance between preserving meaningful behavioral distinctions and avoiding state-space explosion.

6) Engagement Scoring Mechanism: Unlike prior approaches that rely solely on action counts, this study introduces a composite engagement score grounded in the ICAP theoretical framework. The total engagement score S_{total} integrates three complementary dimensions: action quality, time investment, and behavioral consistency.

$$S_{total} = S_{weighted} + S_{time} + S_{consistency}. \tag{14}$$

Action quality $S_{weighted}$: represents the cumulative ICAP-based weights of actions performed within an observation window. Constructive actions (e.g., *submit_assignment*) are assigned higher weights ($w = 5$), whereas passive actions (e.g., *view*) receive lower weights ($w = 1$):

$$S_{weighted} = \sum_{i=1}^{n} w(\text{action}_i).$$

Time efficiency S_{time}: evaluates the ratio between actual interaction time T_{real} and expected time T_{exp}:

$$S_{time} = \begin{cases} 2, & \text{if } T_{real} \geq 0.5\,T_{exp}, \\ 1, & \text{if } 0.3\,T_{exp} \leq T_{real} < 0.5\,T_{exp}, \\ 0, & \text{otherwise.} \end{cases}$$

Table V
COMPLETE ACTION SPACE (15 PEDAGOGICAL ACTIONS)

Group	ID	Action Code	Pedagogical Meaning
PAST (Review)	0	view_assign_past	Review previous assignment requirements
	1	view_content_past	Review previous lecture content
	2	attempt_quiz_past	Retry a previous quiz
	3	review_quiz_past	Analyze previous quiz errors
	4	post_forum_past	Discuss prior topics
CURRENT (In-progress)	5	view_assign_curr	View current assignment requirements
	6	view_content_curr	Study current learning materials
	7	submit_assign_curr	Submit the main assignment
	8	attempt_quiz_curr	Attempt the current quiz
	9	submit_quiz_curr	Submit quiz answers for grading
	10	review_quiz_curr	Review submitted quiz results
	11	post_forum_curr	Discuss current topics
FUTURE (Preparation)	12	view_content_fut	Preview upcoming content
	13	attempt_quiz_fut	Attempt a future quiz
	14	post_forum_fut	Explore upcoming discussion topics

Regularity $S_{consistency}$: measures temporal regularity based on the average inter-action interval Δt:

$$S_{consistency} = \begin{cases} 2, & \text{if } 1 \text{ min} \leq \Delta t \leq 60 \text{ min}, \\ 1, & \text{if } 30 \text{ s} \leq \Delta t \leq 2 \text{ h}, \\ 0, & \text{otherwise.} \end{cases}$$

Through this formulation, engagement is treated as a multidimensional construct rather than a simple activity count, enabling the agent to differentiate between superficial participation and meaningful learning behavior. Together with the discretization scheme described above, the resulting state space remains computationally manageable while retaining sufficient expressiveness for the Q-learning algorithm to converge reliably.

C. Evaluation Metrics

The effectiveness of the proposed model is assessed using a set of complementary quantitative metrics that reflect both learning performance and policy quality.

Total reward: measures the cumulative reward obtained by the agent over an episode, serving as an indicator of policy convergence and overall optimization effectiveness.

Average Score: the final course score achieved by learners, reported on a 10-point grading scale, which reflects academic performance.

Weak Skills Count: The number of learning outcomes (LOs) with mastery levels below 0.5, indicating residual knowledge gaps after completing the course.

D. Data generation and convergence analysis

1) Data generation process with simulation: The simulation is conducted over 500 training iterations (episodes). In each episode, the environment initializes a cohort of 100 virtual learner agents whose performance distribution mirrors a realistic classroom setting: 20% weak, 60% medium, and 20% strong learners. Overall, the learning agent is trained on interactions generated by 50,000 virtual learners, providing sufficient coverage of thestate-actionn space for policy learning.

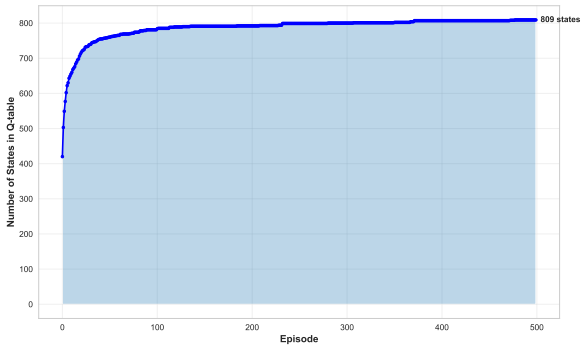

Figure 5. Cumulative reward convergence over 500 training episodes. Stable convergence behavior becomes evident after approximately episode 350.

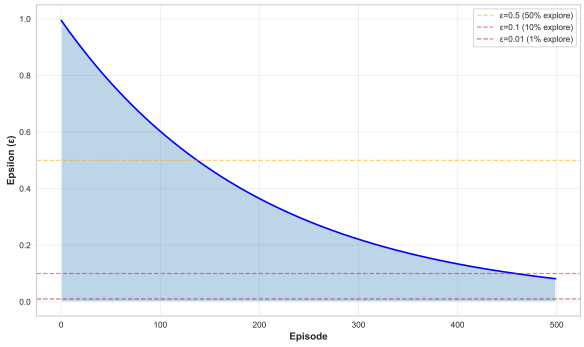

Figure 6. Progressive decay of the exploration parameter ϵ across three training stages: exploration, transition, and exploitation.

2) Convergence and Coverage Analysis: Experimental results indicate that the learned Q-table occupies approximately 219 KB of memory. Within the theoretical state space of 4,320 states, the agent actively explored and optimized 802 core states, which correspond to the most frequently encountered and pedagogically relevant learning scenarios. Rare or infrequent states were effectively handled through the generalization capability induced by the ϵ-greedy exploration strategy during the early training phase.

Fig. 5 and 6 illustrate a clear inverse relationship between the exploration rate ϵ and the cumulative reward. As ϵ gradually decays to 0.01 during the exploitation phase, the average reward increases steadily, indicating that the agent has successfully learned a stable and near-optimal policy for maximizing long-term learning outcomes.

E. A/B Testing Comparison Results

The comparative evaluation was conducted over 500 training episodes using fixed hyperparameters $\alpha = 0.1$ and $\gamma = 0.95$. The aggregated quantitative outcomes reported in Table VI, together with the cluster-wise performance visualization in Fig. 7, consistently demonstrate the superiority of the proposed Q-learning policy over the parametric baseline (π_{param}).

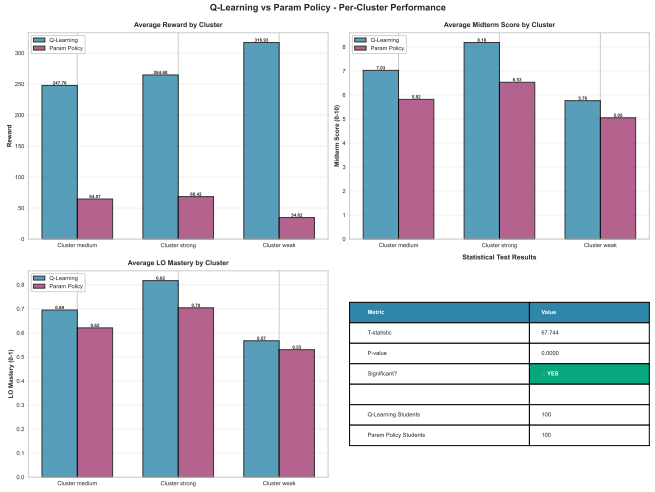

Figure 7. Performance comparison between Q-learning and the Param Policy baseline in terms of cumulative reward, average score, and learning outcome (LO) mastery.

Table VI
AVERAGE PERFORMANCE COMPARISON RESULTS

Metric	Param Policy	Q-learning	Improvement
Average Reward	59.95 ± 12.38	264.26 ± 27.33	+340.8%
Average Score (10-point scale)	5.82 ± 0.48	7.14 ± 0.82	+22.5%
LO Proficiency	0.621 ± 0.056	0.707 ± 0.085	+13.9%
Number of Weak Skills	3.02	1.48	−51.0%

Overall, the Q-learning agent achieves substantially higher cumulative rewards, significantly improves average academic performance, and reduces the number of weak learning outcomes by more than half. These results confirm the effectiveness of reinforcement learning in optimizing long-term pedagogical strategies compared to behavior-driven heuristic policies.

F. Feature Importance Analysis

To interpret the learned policy, SHAP values were computed over 802 representative states extracted from the optimized Q-table. For each state feature, two complementary importance indicators were calculated:

- **Mean Absolute SHAP**:

$$I_i = \frac{1}{N} \sum_{j=1}^{N} |\phi_i(s_j)|,$$

  which measures the average magnitude of feature i's contribution to the estimated Q-value.
- **SHAP Variance**:

$$V_i = \text{Var}(\phi_i),$$

  which captures the variability of a feature's influence across contexts, with higher variance indicating stronger context dependence.

Table VII
FEATURE IMPORTANCE RANKING BASED ON SHAP ANALYSIS OVER 802 CORE STATES

Feature	Mean |SHAP|	Variance	Rank
Module ID	28.32	1171.49	1
Engagement	26.53	995.79	2
Student Cluster	17.42	461.44	3
Score Level	11.39	431.01	4
Learning Phase	9.12	149.83	5
Progress Level	7.42	95.96	6

The numerical rankings are summarized in Table VII and visually illustrated by the beeswarm plot in Figure 8. The analysis reveals that *Module ID* (mean $|\text{SHAP}| = 28.32$) and *Engagement* (26.53) are the two most influential factors guiding the agent's decision-making process.

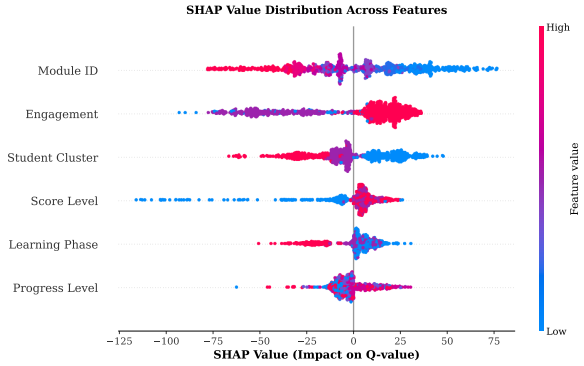

Figure 8. Distribution of SHAP values across 802 core states. Each point represents a state; color denotes feature magnitude (red = high, blue = low), while the horizontal axis indicates the direction and strength of influence on the Q-value.

The beeswarm visualization further highlights strong non-linear effects. In particular, engagement exhibits the highest variance ($\sigma^2 = 995.79$), with SHAP values ranging approximately from -50 to $+50$. This indicates pronounced context sensitivity: high engagement substantially increases the Q-value when aligned with the appropriate module and learner cluster but may decrease it when engagement is misdirected (e.g., excessive passive content consumption prior to assessment).

In contrast, *the progress level* shows the lowest mean absolute SHAP value (7.42) and limited variance (95.96), with values tightly centered around zero. This confirms that mere completion progress has a relatively minor impact on optimal action selection. These findings empirically support the pedagogical hypothesis that *the quality of engagement outweighs the quantity of completed activities* in determining effective learning outcomes.

G. Pedagogical Impact Analysis

As illustrated in Fig. 7, the proposed adaptive learning system demonstrates clear pedagogical benefits across different learner groups, indicating its ability to tailor instructional strategies in a meaningful and effective manner.

Weak Group: Learners in this group predominantly received remedial recommendations focused on review and consolidation. This strategy resulted in the highest growth in cumulative reward and a substantial reduction in knowledge gaps, with the number of weak learning outcomes decreasing by 51% [**?**]. These results suggest that the system effectively supports struggling learners by reinforcing foundational concepts and preventing early disengagement.

Strong Group: High-performing learners were consistently guided toward more challenging and exploratory activities (advanced actions). As a result, this group achieved the highest final performance, with an average score of $8.18/10$, reflecting successful promotion of mastery-oriented learning.

An independent two-sample T-test confirms that the observed differences between learner groups are statistically significant ($p < 0.001$), with a very large effect size (Cohen's $d = 6.78$). This provides strong evidence that the adaptive strategies learned by the Q-learning agent yield not only statistically reliable improvements but also pedagogically meaningful impacts.

IV. CONCLUSION

This paper proposes a reinforcement learning–based framework for personalized STEM education, featuring an open microservices architecture integrated via LTI 1.3, a behavior- and engagement-aware learner modeling pipeline, and a Q-learning algorithm with adaptive reward shaping. Experimental results demonstrate that the proposed approach improves average academic performance by 22.5% while substantially reducing learning gaps among weak students by 51%, thereby supporting the pedagogical principle of inclusive and differentiated learning.

Despite these promising results, the current evaluation is limited to a simulated environment and relies on discretized state representations inherent to tabular Q-learning. Future work will focus on extending the framework with deep reinforcement learning methods such as DQN and PPO to handle high-dimensional and continuous state spaces more effectively [10], validating the system through real-world deployment in university-level STEM courses, and exploring federated learning paradigms to enable privacy-preserving collaborative training across multiple institutions [7].

REFERENCES

[1] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" *Technological forecasting and social change*, vol. 114, pp. 254–280, 2017.
[2] N. Capuano and S. Caballé, "Adaptive learning technologies," *Ai Magazine*, vol. 41, no. 2, pp. 96–98, 2020.
[3] T. B. Thuấn, "Ứng dụng machine learning dự báo sinh viên diện cảnh báo học tập tại trường Đại học kinh tế huế," *Tạp chí Khoa học Quản lý và Kinh tế*, no. 21, 2022.
[4] L. H. Sang, N. T. Hải, T. T. Điện, and N. T. Nghe, "Dự báo kết quả học tập bằng kỹ thuật học sâu với mạng nơ-ron đa tầng," *Tạp chí Khoa học Đại học Cần Thơ*, vol. 56, no. 3, pp. 20–28, 2020.
[5] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[6] M. T. Chi and R. Wylie, "The icap framework: Linking cognitive engagement to active learning outcomes," *Educational psychologist*, vol. 49, no. 4, pp. 219–243, 2014.

[7] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: A literature review," *Education Sciences*, vol. 13, no. 12, p. 1216, 2023.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[9] M. Sneiders, "Moodle log & grades dataset," 2021. [Online]. Available: https://www.kaggle.com/datasets/martinssneiders/moodle-grades-and-action-logs

[10] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.