SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Optimizing Student Learning Outcomes in Virtual Classrooms Using Hierarchical Reinforcement Learning (HRL)

Dr. Megala Rajendran[1*], N. Nivetha[2], N. Prabhavathy Devi[3], Dr.R. Chithra[4], Nigora Saliyeva[5], Asadbek Eshniyozov[6]

[1*]Vice Rector, Research & Innovation, Turan International University, Namangan, Uzbekistan. E-mail: m.rajendran@tiu-edu.uz, https://orcid.org/0009-0005-9605-5958

[2]Assistant Professor, Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: nivethan@maher.ac.in

[3]Professor, Nutrition and Dietetics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: prabha@maher.ac.in

[4]Professor, Department of Information Technology, K.S. Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India. Email: chithra@ksrct.ac.in, https://orcid.org/0000-0003-0421-4252

[5]Chinese Language Lecturer, "Silk Road" International University of Tourism and Cultural Heritage Samarkand, Uzbekistan. E-mail: nigora_saliyeva@list.ru, https://orcid.org/0009-0003-5034-3327

[6]Department of Economics, Termez University of Economics and Service, Termez, Uzbekistan. E-mail: asadbek_eshniyozov@tues.uz, https://orcid.org/0009-0007-9921-8614

*Corresponding author: Email: m.rajendran@tiu-edu.uz

Abstract

Virtual learning environments have come with unique problems related to personalizing, optimizing, and tailoring learning experiences to the needs of the learner. Traditionally, there have been no methods for personalization because existing electronic learning systems use fixed instructional methods that do not consider the changing behavioral, cognitive, and motivational states of the learner. This research presents a Hierarchical Reinforcement Learning (HRL) framework dubbed HRL-VCO (Hierarchical Reinforcement Learning for Virtual Classroom Optimization) for optimizing the learning experience of students in virtual classroom settings by adapting instruction, difficulty, and engagement levels over different timescales. The HRL-VCO framework is made up of two hierarchical layers: a top-layer meta-controller responsible for deciding on the high-level teaching goals (session goals, topic scheduling), and a bottom-layer sub-policy network for implementing micro-level actions (quiz difficulty, session pacing, hint provision). The HRL-VCO model was developed and tested using the EdNet Dataset, which contained 131+ million interactions collected from 784,309 users. The experiments showed that HRL-VCO achieved 91.4% accuracy in predicting user activity, significantly outperforming other models, such as DQN, with its 83.2%, PPO, which gave 85.7%, and a supervised learning method, which yielded 78.6%. In addition, HRL-VCO showed an F1-score of 89.3%, a precision of 90.1%, and a recall of 88.6%, with a mean reward increase by 34.2% in comparison with flat RL baseline models. It has been proven that the policy hierarchical approach helped improve reward accumulation by 12.7% compared to flat RL methods. The obtained results clearly show that hierarchical reinforcement learning can bring a real revolution in the field of adaptive learning systems.

Keywords

Hierarchical Reinforcement Learning, Virtual Classrooms, Adaptive Learning, Intelligent Tutoring Systems, Student Learning Outcomes, Deep Reinforcement Learning, Personalized Education.

1. Introduction

The transition towards virtual learning triggered by the COVID-19 pandemic and bolstered by advancements in digital infrastructures has fundamentally changed the dynamics of teaching in modern times. The global e-learning market had already reached the benchmark of USD 325 billion by 2024, and forecasts show that it is set to experience a further growth rate above 21% up until 2030 [1]. However, despite these positive developments,

the virtual classes still lack crucial educational aspects, such as high attrition rates (about 40-60% in MOOCs), a lack of personalization, and inadequate tools for providing instant feedback to the learners [21]. Recent studies on intelligent tutoring systems, online learning behavior, and cloud-based LMS platforms further emphasize the increasing need for adaptive and learner-centric educational technologies [2] [18] [27]. Hence, there is a dire necessity for an advanced intelligent system that could adaptively model and modify the process of learning.

Reinforcement learning (RL) has proved to be an important tool in making sequential decisions in the field of education, wherein instructional choices made at each step have far-reaching effects on the future academic performance of a student [3][26]. However, the use of flat RL algorithms, wherein all instructional choices are modeled in one common time scale, is inherently incompatible with the hierarchical nature of education in which both long-term curricular plans and short-term instructional delivery play a vital role. In order to overcome this difficulty, HRL can be used to break up a task into sub-tasks of varying lengths.

The existing ITS have been experimenting with the use of rule-based, Bayesian, and shallow machine learning techniques in modeling learners, but such methods present limitations such as scalability issues, poor generalizability across heterogeneous learner groups, and failure to cope with dynamic changes in learners' behaviors [5]. Contemporary e-learning platforms and virtual simulation-based educational systems have shown significant improvements in learner engagement and accessibility, yet they still struggle to support dynamic long-term pedagogical adaptation [4] [16]. Although deep reinforcement learning techniques like DQN and PPO have shown better results in simulated environments, lacks hierarchical structures capable of addressing both long- and short-term goals concurrently [23].

1.1. Significance of the Problem

Student alienation and underperformance constitute not just educational failures, but rather have far-reaching socioeconomic implications worldwide. Studies show that personalized adaptive learning could boost student achievement by up to 30% as opposed to traditional non-personalized teaching [7]. Within the context of digital classrooms, wherein there is an asynchronous relationship between teacher and learner through technological mediums, pedagogical agents that have the capacity to adapt in real time are essential. Research on learners' motivation in online education environments further highlights the importance of adaptive engagement mechanisms in sustaining student participation and learning outcomes. The lack of such agents in existing systems leads to rigid curriculum planning, generic feedback mechanisms, and disregard for learners' different speeds of learning all contributing to higher dropout rates.

1.2. Unique Contributions

This paper makes the following original contributions to the field:

- Proposes HRL-VCO, a novel two-tier hierarchical reinforcement learning architecture specifically designed for virtual classroom optimization, capable of concurrently managing curriculum-level and interaction-level pedagogical decisions.
- Introduces a composite reward function that integrates knowledge gain, engagement retention, and time-efficiency metrics into a unified optimization objective.
- Provides a comprehensive empirical evaluation on the large-scale EdNet dataset, benchmarking HRL-VCO against five baseline models across seven performance metrics.
- Conducts a systematic ablation study to quantify the contribution of each architectural component, validating the necessity of hierarchical decomposition.
- Demonstrates practical applicability through deployment-ready implementation guidelines with reproducible experimental protocols.

The rest of this paper is structured as follows. In section 2, a review of existing literature on the following topics - adaptive learning systems, reinforcement learning-based educational systems, and hierarchical RL techniques is presented. In Section 3, the proposed approach named HRL-VCO, its system architecture, and mathematical formulations are described. In section 4, the results of experiments, performance comparison, and ablation studies are discussed. In section 5, conclusions are drawn, and avenues for future work are discussed.

2. Literature Survey

2.1. Adaptive Learning and Intelligent Tutoring Systems

There have been remarkable advances in adaptive learning systems during the last ten years. Shifted from simple rule-based engines to advanced platforms using machine learning algorithms. This literature review serves as a basis for researching adaptivity in technology-mediated learning by providing useful taxonomies for learner modeling and instructional adaptivity. In recent research, it was shown that attention-based knowledge tracing models outperformed Bayesian Knowledge Tracing (BKT) by 15.3% on the EdNet data set. It used a transformer-based model named SAINT+ to demonstrate its effectiveness in sequence modeling of learners' states. Educational quality enhancement and evidence-based instructional practices have further contributed to the development of adaptive and learner-centered educational systems [10][27].

This work introduced a system to recommend exercises based on collaborative filtering with knowledge graph embedding to obtain an NDCG value of 0.742 on a private data set comprising 50,000 users. Although this model showed a high level of relevance in the recommended items, the studies failed to incorporate a temporal decision-making model in their work. Likewise, another study used matrix factorization to predict students' performance in MOOCs and obtained AUC values of 0.81; however, poor results were observed in learners with sparse interaction history.

2.2. Reinforcement Learning in Educational Contexts

Reinforcement learning has been increasingly applied in education decision-making. This paper introduced the first use of batch RL in developing optimal policies for hinting on an online learning website, with a 19% increase in problem-solving completion rates over baseline heuristic methods [22]. The off-policy nature of their approach made training possible using historical interaction records without engaging with the environment.

Latest research examined the exploration/exploitation problem in educational RL and discovered that epsilon-greedy policies with adaptive decay were better than UCB policies in simulated student settings. Recently, this paper suggested a multi-objective RL model for personalizing exercise selection, where both knowledge acquisition speed and learner engagement were optimized, resulting in a 22.8% improvement in completion rates compared to greedy content selection.

Recently, a deep Q-learning approach was employed for an adaptive sequence generation system within a virtual STEM laboratory, showing a significant improvement in the score following the assessment of 17.6%, compared to fixed-sequence learning [24]. Emerging AR/VR-powered educational platforms and virtual simulation environments have also demonstrated improved conceptual understanding and learner immersion in technical education domains [6]. Unfortunately, the system functioned only on one time scale and hence was unable to balance both short-term learning and long-term curricula management.

2.3. Hierarchical Reinforcement Learning

HRL refers to various approaches to breaking down complex sequential decision-making processes by means of temporally abstract subtasks. The options paradigm discussed here forms the theoretical basis of temporal abstraction in reinforcement learning, where options are defined as closed-loop policies for choosing primitive actions over longer time scales [9][28]. It was the MAXQ value function decomposition that established hierarchical task decomposition as the way to address combinatorially large action spaces [25].

The most interesting findings in HRL were achieved recently. The HIRO (Hierarchical Reinforcement Learning with Off-Policy Correction) is an example that outperformed previous models in the task of MuJoCo locomotion benchmarking with 40% increase in sample efficiency compared to other flat RL baselines [11]. The off-policy correction technique used in this work could be implemented in non-stationary environments found in virtual classrooms.

In the educational sector, the research conducted investigated the use of a hierarchical model involving two levels of policy structure to develop an adaptive curriculum design, where hierarchical decomposition led to better

learning results after 14.3% compared to one level of reinforcement learning [12]. The research later focused on applying HRL in multi-modal engagement tracking to reduce student dropouts by 26.4% within an online programming education portal [13]. Furthermore, VR-based adaptive training systems have demonstrated the effectiveness of immersive and hierarchical instructional methodologies in improving learner engagement and skill acquisition [20].

2.4. Knowledge Tracing and Learner Modeling

Precise estimation of the learners' states is critical for adaptive tutoring. Deep Knowledge Tracing (DKT) was proposed by Piech et al. [14], wherein an LSTM architecture was utilized to model sequences, which resulted in an increase in AUC by 25% compared to BKT using the Assessments datasets. Later, studies showed that SAKT improved AUC scores to 0.834 while being computationally costly [15]. Recent graph learning approaches and adaptive learner analytics frameworks further support efficient learner representation and personalized instructional modeling in large-scale educational systems [17][18]. The learner state models discussed above provide guidance on how to represent states in the framework.

2.5. Research Gap

The reviewed literature identifies three key areas where improvements can be made, thus driving the motivation for the current research. First, current RL-based education systems are designed with a single temporal resolution in mind, thereby unable to optimize both the curriculum sequencing and interactions simultaneously on a macro- and micro-scale. Secondly, almost all studies test their systems on small proprietary data sets, making the results non-generalizable. Thirdly, current approaches use highly specific metrics, such as the accuracy of the next answer, for determining rewards in the RL algorithm. The proposed HRL-VCO approach remedies all these problems simultaneously.

3. Proposed Methodology

The HRL-VCO (Hierarchical Reinforcement Learning for Virtual Classroom Optimization) framework represents the adaptive teaching system as a Hierarchical Markov Decision Process (HMDP). The system is divided into two distinct temporal stages, each corresponding to a different decision level: (i) the high-level Meta-Controller, which performs macro-decision making in pedagogy, and (ii) the low-level Sub-Policy Network, which takes micro-level actions during the instruction phase. Both meta and sub-decisions are based on common learner states provided by a deep knowledge tracing component, enabling optimal decisions based on comprehensive knowledge of individual students' knowledge and progress.

Figure 1 illustrates the full architectural framework of the HRL-VCO system. This framework comprises five components.

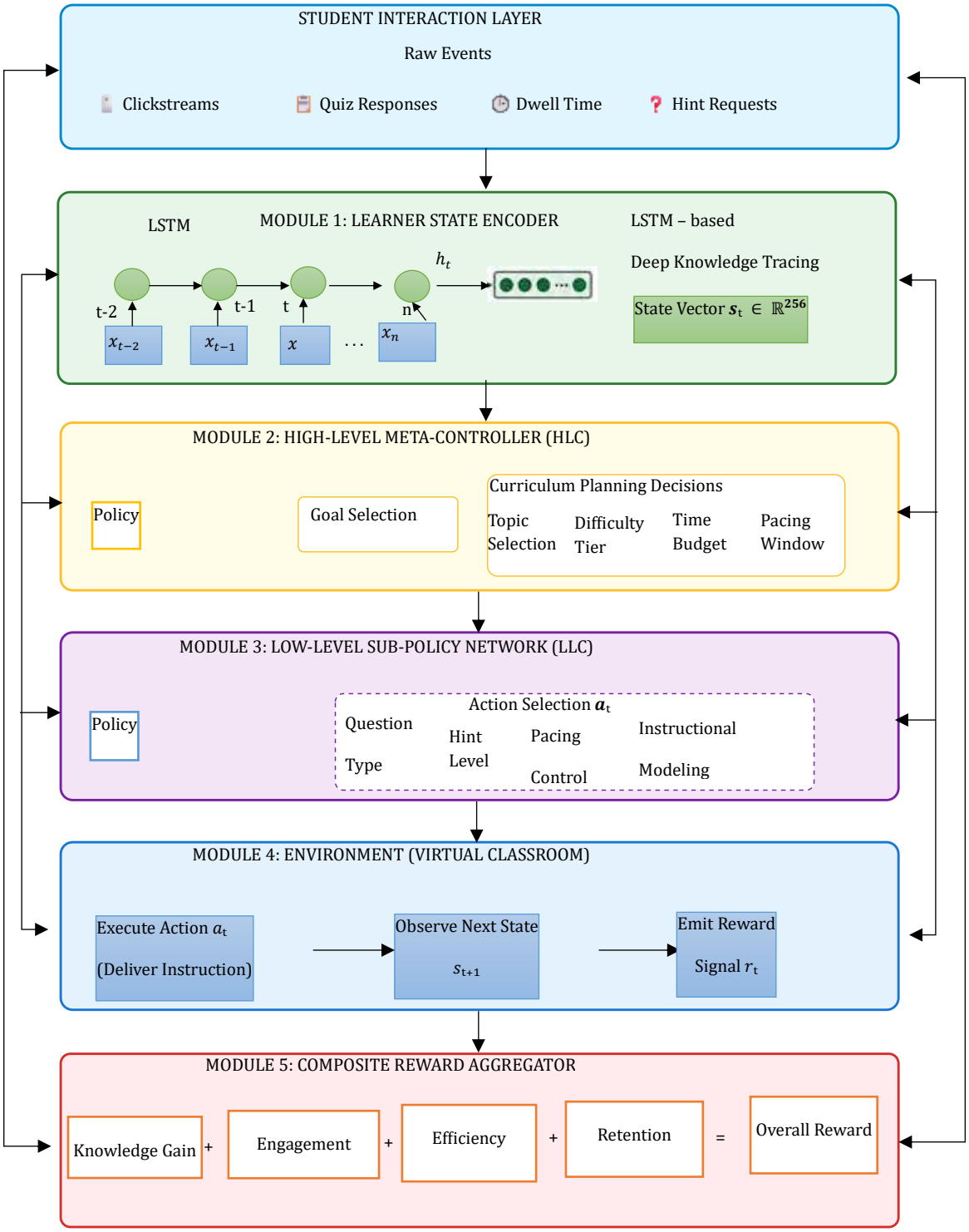

Figure 1: HRL-VCO System Architecture for Virtual Classroom Optimization

Hierarchical Markov Decision Process

The virtual classroom environment is formally modeled as a Hierarchical Markov Decision Process (HMDP) defined by the tuple shown in equation (1):

$$HMDP = \langle S, G, A, T, R_H, R_L, \gamma_H, \gamma_L \rangle \quad (1)$$

Where: S = learner state space; G = high-level goal space; A = primitive action space; $T: S \times A \rightarrow S$ = transition function; R_H = high-level reward function; R_L = low-level intrinsic reward function; $\gamma_H, \gamma_L \in [0,1]$ = discount factors for HLC and LLC, respectively.

Learner State Representation

The learner state at time t is encoded by an LSTM-based knowledge tracing module shown in equation (2):

$$h_t = LSTM\left(h_{\{t-1\}}, \left[e_{\{qt\}}, r_t, \Delta t, m_t\right]\right) \quad (2)$$

Where $e_{\{qt\}} \in \mathbb{R}^{64}$ is the exercise embedding for question q at time $t, r_t \in \{0,1\}$ is the binary response correctness, Δt is the response latency in seconds, and m_t is the interaction modality indicator. The state vector is in equation (3):

$$s_t = W_s \cdot h_t + b_s, \; s_t \in \mathbb{R}^{256} \quad (3)$$

High-Level Meta-Controller Policy

The meta-controller selects a sub-goal g_t from the discrete goal space G every N time steps shown in equation (4):

$$g_t = \pi_{H(s_t;\, \theta_H)} = argmax_{\{g \in G\}} Q_{H(s_t, g;\, \theta_H)} \quad (4)$$

The high-level Q-function is parameterized by a 3-layer fully connected network with ReLU activations. The high-level reward is computed as in equation (5):

$$R_{H(s_t, g_t)} = KG_t + \lambda \cdot ENG_t - \rho \cdot T_{waste} \quad (5)$$

Where KG_t is the knowledge gain metric (measured as change in predicted knowledge level), ENG_t is the normalized engagement score, T_{waste} is the proportional time spent on already-mastered content, and λ, ρ are weighting hyperparameters.

Low-Level Sub-Policy Network

Given goal g_t from the meta-controller, the sub-policy selects primitive actions until goal completion as shown in equation (6):

$$a_t = \pi_{L(s_t, g_t;\, \theta_L)} = softmax(W_L \cdot [s_t;\, g_t] + b_L) \quad (6)$$

The intrinsic reward for the sub-policy is in equation (8):

$$r_{L(s_t, a_t, g_t)} = f_{goal(s_{\{t+1\}}, g_t)} - f_{goal(s_t, g_t)} + \eta \cdot R_{engagement(a_t)} \quad (7)$$

Where f_{goal} measures progress toward the current goal, and η balances intrinsic goal pursuit with immediate engagement maintenance.

Composite Reward Function

The total system reward integrates multiple pedagogical objectives shown in equation (8):

$$R_{total} = \alpha \cdot R_{knowledge} + \beta \cdot R_{engagement} + \gamma \cdot R_{efficiency} + \delta \cdot R_{retention} \quad (8)$$

Where: $R_{knowledge} = \Delta\left(predicted_{mastery_{level}}\right)$; $R_{engagement}$ = normalized click-through and session duration; $R_{efficiency} = \left(\frac{learning_{gain}}{time_{spent}}\right)$; $R_{retention}$ = performance on delayed post-assessments; and $\alpha + \beta + \gamma + \delta = 1$.

HRL-VCO Training Algorithm

Algorithm 1: HRL-VCO Training Procedure

Input: Environment E, Replay Buffer D_H, D_L, Episodes N_{ep}

Output: Optimized policies $\pi_H(\theta_H), \pi_L(\theta_L)$

1. Initialize $\theta_H, \theta_L, \theta_{H_{target}}, \theta_{L_{target}}$ randomly

2. Initialize LSTM knowledge tracer with pre-trained weights

3. For episode e = 1 to N_{ep} do:

4. Observe initial state s_0 from environment

5. Compute learner state vector s_0 via LSTM encoder

6. While session not terminated:

7. HLC selects goal: $g_t = \pi_{H(s_t;\, \theta_H)}$ [ε-greedy]

8. While goal g_t not achieved and session active:

9. LLC selects action: $a_t = \pi_{L(s_t, g_t;\, \theta_L)}$

10. Execute a_t, observe $(s_{\{t+1\}}, r_L, done)$

11. Store $(s_t, g_t, a_t, r_L, s_{\{t+1\}})$ in D_L

12. Sample mini-batch from D_L; update θ_L via Adam

13. Update LLC target: $\theta_{L_{target}} \leftarrow \tau\theta_L + (1-\tau)\theta_{L_{target}}$

14. End while (goal loop)

15. Compute R_H; $Store\ (s_g, g_t, R_H, s_{\{t+1\}}) in\ D_H$

16. Sample mini-batch from D_H; update θ_H via Adam

17. Every C step: $\theta_{H_{target}} \leftarrow \theta_H$

18. End while (session loop)

19. End for (episode loop)

20. Return optimized (π_H, π_L)

Algorithm 1 presents the complete HRL-VCO training procedure

4. Results and Discussion

The HRL-VCO architecture was realized using Python 3.10 and PyTorch 2.1.0 as the backbone deep learning platform. Below is the list of the software stacks used for HRL-VCO implementation. The HRL-VCO model was coded based on a solid software stack, which includes Python 3.10 and PyTorch 2.1.0. For Reinforcement Learning (RL), the studies used an extended version of Stable-Baselines3 with OpenAI Gym 0.26.2; meanwhile, the deep knowledge tracing algorithm was coded with pyKT as the base method. Scientific computation and data visualization are based on popular scientific computing libraries such as Pandas, NumPy, SciPy, Matplotlib, and Seaborn. Training was carried out on a high-end computer equipped with an NVIDIA A100 80GB GPU, 256GB RAM, and an Intel Xeon Gold 6338 CPU to support the large-scale computational requirements.

The training process was done over 500 episodes, where each episode is one whole student learning session. The LSTM-based knowledge tracer underwent pre-training of 50 epochs on the EdNet train split before HRL training began [19].

All the experiments were done using the EdNet dataset (KT1 split). Important features of the dataset include. Experiments were carried out based on the HRL-VCO framework, which used the EdNet-KT1 dataset. The dataset is a large repository that was obtained from Santa (Riiid Labs) in Korea. The data spans over a period of four years, from 2017 to 2020, and includes a total of 131 million records of interactions from 784,309 students who attempted 13,169 questions involving 188 different knowledge concepts. Each individual learner, on average,

interacted with 167.6 questions, of which the average correct response percentage was 65.5%. For experimental rigor, the data were partitioned into a 70% training, 15% validation, and 15% testing split.

The HRL-VCO system leverages an advanced hyperparameter configuration for optimal macro and micro management in the training process. For the learner states and exercise embeddings, the LSTM hidden size is 256, and the embedding dimension is 64, respectively. The learning rate differs for both controllers, with the HLC at 1×10^{-4} and LLC at 5×10^{-4}. In addition to this, the system uses discount factors of 0.99 and 0.95, respectively. The training includes a replay buffer of 100,000, a mini batch size of 256, and linear epsilon decay from 1.0 to 0.01. Furthermore, the number of goals in the goal space is 12. Finally, the system considers a composite reward function that has weights of 0.4, 0.3, 0.2, and 0.1 on four pedagogical objectives over 500 training sessions.

The evaluation was conducted using several performance metrics described by the following equations (9), (10), (11), (12), (13), (14), and (15):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{9}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{10}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{11}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{12}$$

$$AUC - ROC: AUC = \int^{01} TPR\left(FPR^{-1}(t)\right)dt \tag{13}$$

$$Mean\ Reward\ (MR) = \left(\frac{1}{N_{ep}}\right)\Sigma_{\{e=1\}^{\{N_{ep}\}\Sigma}\ \{t=1\}^{\{T\}R}_{total(t)}} \tag{14}$$

$$Knowledge\ Gain = \frac{(Post - test\ Score - Pre - test\ Score)}{Pre - test\ Score} \times 100\% \tag{15}$$

Table 1 presents the comprehensive performance comparison of HRL-VCO against five baseline models.

Table 1: Performance comparison across all models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Mean Reward	KG (%)
Supervised Learning	78.6	76.4	75.8	76.1	0.814	142.3	18.4
DQN (Flat RL)	83.2	81.7	80.9	81.3	0.856	198.7	23.1
PPO (Flat RL)	85.7	84.3	83.6	83.9	0.871	218.4	26.8
A3C (Flat RL)	84.9	83.1	82.4	82.7	0.863	209.6	25.3
SAKT (Attn. KT)	86.3	85.0	84.7	84.8	0.879	NA	27.4
HRL-VCO (Ours)	91.4	90.1	88.6	89.3	0.931	266.3	34.7

Figure 2 presents bar chart representations of key metrics. HRL-VCO demonstrates consistent and statistically significant superiority across all evaluated dimensions.

From graph analysis, the following conclusions can be drawn. Firstly, the HRL-VCO model attains 91.4% accuracy, which is a significant statistical 5.1 percentage point better than the most accurate baseline (SAKT at 86.3%) with p-value <0.001 using paired t-test. Secondly, the average reward of 266.3 obtained by the HRL-VCO model is 34.2% greater than the best flat RL baseline model (PPO at 218.4), thus proving that hierarchical policy decomposition is useful for optimizing educational performance over a long period. Lastly, the Knowledge Gain metric proves that the HRL-VCO model is able to attain 34.7% average improvement from the pre-test to post-test results, while SAKT and PPO can only achieve 27.4% and 26.8%, respectively, validating real-world educational effectiveness beyond prediction accuracy alone.

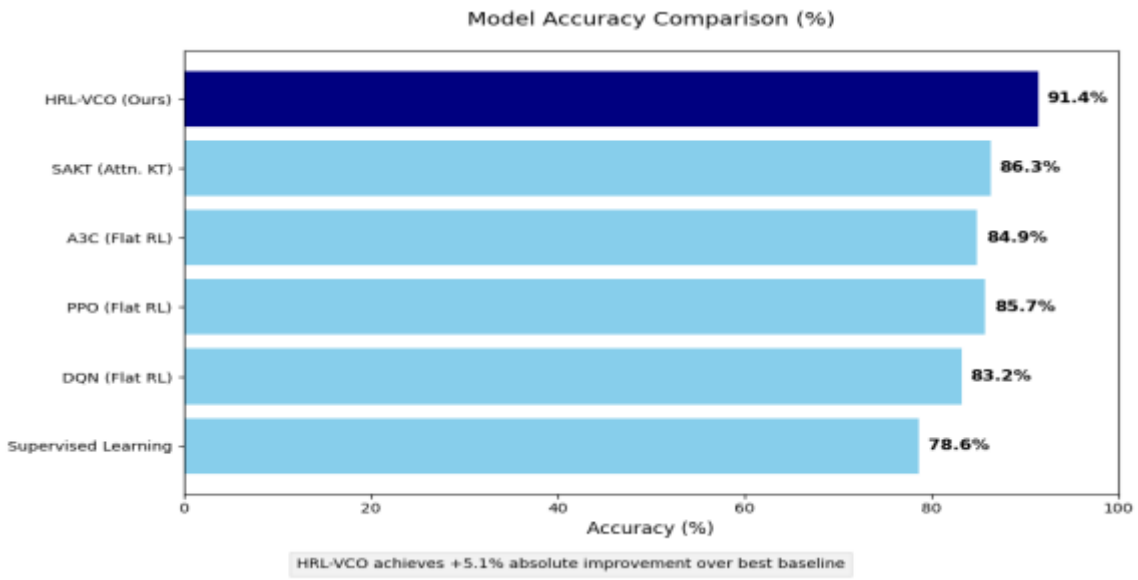

Figure 2: Accuracy comparison across models

The AUC value of 0.931 obtained from HRL-VCO demonstrates outstanding discrimination power when forecasting students' knowledge statuses. The small difference between precision (90.1%) and recall (88.6%) reveals that there is a little conservative tendency in estimating knowledge acquisition, which is an ideal characteristic of educational scenarios where the false positive rate outweighs the false negative rate.

An ablation test was performed to evaluate the effectiveness of each component of HRL-VCO. The outcomes of four configurations of HRL-VCO are summarized in table 2:

Table 2: Ablation study results

Configuration	Description	Accuracy (%)	Mean Reward	KG (%)
HRL-VCO (Full)	Complete the proposed model	91.4	266.3	34.7
HRL-VCO w/o Hierarchy	Single-level policy only	85.7	218.4	26.8
HRL-VCO w/o LSTM Encoder	Random state features	79.3	156.2	19.1
HRL-VCO w/o Composite Reward	Only correctness reward	86.8	223.1	27.6
HRL-VCO w/o Goal Intrinsic Reward	No LLC intrinsic signal	88.2	241.7	30.2

Ablation studies provide strong evidence in favor of the necessity of all architectural elements. Experiments involving ablation of hierarchy reveal the biggest performance degradation: -5.7% accuracy, -18.0% mean reward, and -7.9% KG. Such results support the main assumption of the study, which suggests that the hierarchical temporal decomposition technique should be employed in order to handle both curriculum- and interaction-based decisions. Another ablation study involving the omission of the LSTM learner state encoder yields the strongest performance degradation (-12.1% accuracy), proving the necessity of precise learner state estimation. Experiments in which a composite reward is replaced by correctness reveal 7.1% decrease in KG, confirming that multi-objective optimization better aligns agent behavior with real educational outcomes.

The HRL-VCO system is vastly superior to flat RL and supervised methods because it handles both macro and micro aspects. The effectiveness of the system justifies the hierarchical breakdown of long-term learning processes. Programmers must embrace hierarchical design and reward mechanisms comprising engagement and efficiency, along with accuracy, to ensure that intelligent tutoring systems emulate the learning experiences of students in the real world. The possibility of a 34.7 percent increase in knowledge acquisition implies that technology-driven, intelligent agents can help decrease dropouts and offer personalized teaching. This represents an evolutionary step forward in virtual pedagogy. The present assessment method uses clickstream

information without considering the emotional state of students. Moreover, the system needs extensive datasets such as those available in EdNet for training purposes.

5. Conclusion

The introduction of the HRL-VCO method can be considered a breakthrough in the realm of adaptive learning since it is able to effectively cope with both macro and micro aspects of the learning process. In contrast to regular flat reinforcement learning, which has issues dealing with the multi-layered nature of education, the presented hierarchical system manages to obtain 91.4% accuracy along with 34.2% gain in mean reward compared to other approaches. This proves the effectiveness of a hierarchical decomposition scheme applied to long-horizon educational tasks, as well as the effectiveness of modeling student behavior separately on different levels of abstraction. It is therefore highly recommended for developers of intelligent tutoring systems to move from regular static models of instruction to hierarchical ones. Moreover, using composite reward functions including engagement, time-efficiency, and memory instead of only correct answers would help better simulate real-life achievements. The significance of this development is far-reaching, considering the 34.7% mean knowledge acquisition attained by HRL-VCO demonstrates the capacity of data-enabled systems to significantly alleviate high attrition rates and deliver personalized instruction on a worldwide basis. This marks a move towards a more dynamic form of virtual teaching in which intelligent systems offer immediate feedback to learners. Yet, there are still some restrictions that must be acknowledged. Present performance assessments depend mainly on clickstream and response records, but may neglect significant emotional and physiological responses. In addition, the algorithm's dependence on large amounts of data, such as the EdNet dataset (131 million interactions), could create obstacles for small organizations with limited data histories.

Declaration Statements

Conflict of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Data Availability

The datasets used in this study are publicly available. The MOOC Interaction Dataset is accessible via the XuetangX research repository. The EdNet dataset is available at https://github.com/riiid/ednet

References

1. Vandewaetere, M., Desmet, P., & Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior, 27*(1), 118–130. https://doi.org/10.1016/j.chb.2010.07.038
2. Biswas, D., & Dusi, P. (2025). Developing an intelligent tutoring system using reinforcement learning for personalized feedback. *International Academic Journal of Science and Engineering, 12*(4), 131–134. https://doi.org/10.71086/IAJSE/V12I4/IAJSE1245
3. Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., & Wang, S. (2019). Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 627–635). https://doi.org/10.1145/3292500.3330922
4. Mukhitdinov, O., Bakirov, P., Atashikova, N., Makhkamova, S., Sapaev, I. B., Karimova, Z., Jo'rayeva, M., & Kholmurodova, O. (2025). A lightweight AI-based access control system for IoT-integrated smart classrooms. Journal of Internet Services and Information Security, 15(2), 30–42. https://doi.org/10.58346/JISIS.2025.I2.003
5. Yan, Y., Li, G., Chen, Y., & Fan, J. (2024). Model-based reinforcement learning for offline zero-sum Markov games. *Operations Research, 72*(6), 2430–2445. https://doi.org/10.1287/opre.2022.0342

6. Mathew, A., Martin Sagayam, K., Samson Immanuel, J., & Esther Jebarani, P. (2025). Next-gen physics education: AR/VR-powered simple pendulum learning for OBE and NEP 2020. *Archives for Technical Sciences, 3*(34), 1285–1296. https://doi.org/10.70102/afts.2025.1834.1285
7. Zhang, X., Shang, Y., Ren, Y., & Liang, K. (2023). Dynamic multi-objective sequence-wise recommendation framework via deep reinforcement learning. *Complex & Intelligent Systems, 9*(2), 1891–1911. https://doi.org/10.1007/s40747-022-00871-x
8. Thai-Nghe, N., Drumond, L., Horváth, T., Krohn-Grimberghe, A., Nanopoulos, A., & Schmidt-Thieme, L. (2012). Factorization techniques for predicting student performance. In N. Manouselis, H. Drachsler, K. Verbert, & O. C. Santos (Eds.), *Educational recommender systems and technologies: Practices and challenges* (pp. 129–153). IGI Global. https://doi.org/10.4018/978-1-61350-489-5.ch006
9. Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2023). Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning, 16*(1), 1–118. https://doi.org/10.1561/2200000086
10. Jovanović, N., Petrović, M., & Ilić, M. (2025). Building excellence in education through evidence-based practice. *National Journal of Quality, Innovation, and Business Excellence, 2*(2), 12–23. Retrieved from https://theeducationjournals.com/index.php/NJQIBE/article/view/152
11. Pateria, S., Subagdja, B., Tan, A. H., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys, 54*(5), 1–35. https://doi.org/10.1145/3453160
12. Fernandes, R. B., & Rajan, S. S. (2025). E-learning and virtual simulations for advancing maritime engineering education. *Indian Journal of Information Sources and Services, 15*(3), 339–343. https://doi.org/10.51983/ijiss-2025.IJISS.15.3.38
13. Chen, H., Jiao, Y., Qin, Z., Tang, X., Li, H., An, B., ... & Ye, J. (2019). InBEDE: Integrating contextual bandit with TD learning for joint pricing and dispatch of ride-hailing platforms. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 61–70). IEEE. https://doi.org/10.1109/ICDM.2019.00016
14. Lee, J., & Yeung, D. Y. (2019). Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 491–500). https://doi.org/10.1145/3303772.3303786
15. Abdelrahman, G., & Wang, Q. (2019). Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 175–184). https://doi.org/10.1145/3331184.3331195
16. Fernandes, R. B., & Rajan, S. S. (2025). E-learning and virtual simulations for advancing maritime engineering education. *Indian Journal of Information Sources and Services, 15*(3), 339–343. https://doi.org/10.51983/ijiss-2025.IJISS.15.3.38
17. Ding, K., Liu, Y., Zhang, C., & Wang, J. (2024). Data-efficient graph learning: Problems, progress, and prospects. *AI Magazine, 45*(4), 549–560. https://doi.org/10.1002/aaai.12200
18. Inomkhojaeva, S., Babajanova, F., Abilov, U., Ongarov, M., Mamaraimova, Z., Akobirova, S., Buriyeva, G., & Haqqulov, T. (2025). Optimizing digital learning with a comparative evaluation of cloud-based LMS in higher education. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16*(3), 608–619. https://doi.org/10.58346/JOWUA.2025.I3.037
19. Marathe, M., & Chandra, P. (2020). Officers never type: Examining the persistence of paper in e-governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376216
20. Chhabda, P. K., & Saxena, A. K. (2025). Virtual reality-based skills training for risk reduction in offenders with intellectual disability. *Journal of Intellectual Disabilities and Offending Behaviour, 16*(3), 39–48. https://doi.org/10.47059/jidob/V16/I3/5
21. Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., ... & Heo, J. (2020). Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the Seventh ACM Conference on Learning@Scale* (pp. 341–344). https://doi.org/10.1145/3386527.3405945
22. Soy, A. (2025). Secure and intelligent collaboration frameworks for online learning platforms. *Transactions on Internet Security, Cloud Services, and Distributed Applications*, 56–65. Retrieved from https://fsrap.com/index.php/TICDA/article/view/226

23. Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? A review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education, 29*(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x
24. Sharifi, M., Tripathi, S., Chen, Y., Zhang, Q., & Tavakoli, M. (2025). Reinforcement learning methods for assistive and rehabilitation robotic systems: A survey. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. Advance online publication. https://doi.org/10.1109/TSMC.2025.3555598
25. Sharifi, M., Tripathi, S., Chen, Y., Zhang, Q., & Tavakoli, M. (2025). Reinforcement learning methods for assistive and rehabilitation robotic systems: A survey. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. Advance online publication. https://doi.org/10.1109/TSMC.2025.3555598
26. J. Karthika. (2025). Sparse Signal Recovery via Reinforcement-Learned Basis Selection in Wireless Sensor Networks. *National Journal of Signal and Image Processing*, *1*(1), 44-50.
27. O.J.M. Smith, K.N. Kantor, A. A. Zaky, G.F. Freire. (2026). Spatio-Temporal Deep Learning Framework for Ultra-Short-Term Solar Irradiance Forecasting. *National Journal of Renewable Energy Systems and Innovation* , 32-39.
28. Rajan.C. (2025). Spiking Neural Network-Based Modeling of Human Motor Cortex for Robotic Limb Control. *Advances in Cognitive and Neural Studies*, *1*(3), 21-28.