ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on dynamic decision-making model of school-enterprise collaborative education driven by reinforcement learning

Wensheng Liu[1,*]

[1] School of Marxism of Yantai Nanshan University, Longkou 265713, Shandong, China

SUMMARY: *Aiming at the problems of insufficient perception of students 'ability evolution, lagging response of enterprise job demand and dependence on experience for training path adjustment in traditional mode, this paper constructs a dynamic decision-making model of school-enterprise collaborative education driven by reinforcement learning. The model encodes students 'learning behavior, course status, practical tasks, enterprise job requirements and collaborative feedback into 112-dimensional state features, and combines GRU time series modeling, job demand graph, cross-domain attention mechanism and action mask strategy to form an executable action space for course recommendation, practical task allocation, tutor matching and path adjustment. In the strategy optimization stage, the PPO policy network and multi-objective reward function were introduced to complete the continuous decision-making by comprehensive ability improvement, job adaptation, learning participation, enterprise satisfaction and resource consumption. The experiment was carried out based on 1260 students, 86 courses, 42 types of posts and 26780 effective interaction samples. The results show that the matching degree of the training program of the model in this paper reaches 91.7%, the job suitability increases to 89.1%, the average reward is stable at 0.86, and the stability score is 0.88. The research results provide technical support for the intelligent decision-making of school-enterprise collaborative education.*

KEYWORDS: *Reinforcement learning; Cooperative education between schools and enterprises; Dynamic decision model; PPO policy network*

1 Introduction

1.1 The realistic demand for dynamic decision-making of school-enterprise collaborative education

With the digital transformation of the industry and the rapid change of the post ability structure, the traditional school-enterprise collaborative education mode gradually shows the problem of lagging response in terms of training program adjustment, curriculum resource allocation, practical task arrangement and student ability tracking [1]. The school side has the data of students' course performance, learning behavior and practical performance, while the enterprise side has the data of post standards, project tasks, skill needs and employee feedback. However, the two types of data are often scattered in different systems, and it is difficult to form a continuous linkage decision-making basis [2, 3]. School-enterprise collaborative education needs to be dynamically adjusted according to the growth of students' ability, the change of enterprise demand and the deviation of training objectives, and it is difficult to

*maidoudeyanlei@163.com
https://doi.org/10.65102/is2026907 Published: 30 April, 2026

balance multiple factors in time by simply relying on manual experience [4]. Therefore, constructing a dynamic decision-making model that can perceive multi-source data, identify student status, match job requirements and continuously optimize training paths has become an important requirement to improve the accuracy and collaborative efficiency of talent training [5].

1.2 The technological evolution of artificial intelligence Driving educational decision-making

Artificial intelligence technology has promoted educational decision-making from static evaluation to data-driven and intelligent optimization. Early education management mainly relies on statistical analysis and rule judgment, which can complete score summary, index evaluation and simple early warning, but its ability to express the evolution process of students' ability, the fluctuation of enterprise job demand and the long-term effect of training programs is limited [6]. With the development of machine learning, knowledge graph, deep learning and recommendation algorithms, education systems begin to be able to extract features from learning behaviors, course grades, practice records and post labels to support personalized resource recommendation and learning path planning [7, 8]. Reinforcement learning further expands the dynamic optimization ability of educational decision-making. It can incorporate students' ability state, training intervention action and stage feedback into the continuous decision-making process, and constantly revise the strategy through the reward signal, so that the model can gradually form a better training program adjustment mechanism in the process of multiple rounds of school-enterprise collaboration [9-11].

1.3 Main Contributions and research problems of this paper

Focusing on the problems of slow decision response, extensive adjustment of training path, and insufficient utilization of enterprise demand feedback in the process of school-enterprise collaborative education, this paper constructs a dynamic decision-making model driven by reinforcement learning. In this study, students 'learning performance, ability portrait, practical task completion, enterprise job demand and collaborative feedback were uniformly encoded into state representation, and the action space for course recommendation, practical task allocation, post ability reinforcement and training program adjustment was designed. Moreover, a multi-objective reward function was introduced to comprehensively measure ability improvement, job fitness, learning engagement and enterprise satisfaction. The model used PPO policy network to complete continuous decision optimization, and realized policy calibration and dynamic update through online feedback loop. This paper focuses on answering three questions: how to construct a computable state space in the school-enterprise collaboration scenario, how to use reinforcement learning to realize the dynamic choreography of training paths, and how to verify the effectiveness of the model in terms of matching effect, convergence stability and education effectiveness.

2 Related work

Artificial intelligence education research has gradually shifted from technical feasibility discussion to actual decision support in teaching scenarios. Holmes and Tuomi pointed out that the application of AI in education has covered learning analysis, intelligent recommendation, automatic feedback and learning support, but its core value is not only to improve the level of automation, but also to enhance the adaptability and explanation ability in the education process [12]. Celik et al. reviewed the research on AI education from the perspective of teachers, and concluded that artificial intelligence can assist teachers to complete learning diagnosis, resource organization and feedback generation, but it also faces challenges such as insufficient data literacy, low algorithm transparency and unclear boundaries of teaching responsibility [13]. Chounta et al. further found that teachers' acceptance of AI tools was closely related to whether they could be embedded in real teaching processes, and simply providing technical tools could not be naturally translated into effective teaching decisions [14].

In the aspect of human-computer collaborative learning, Molenaar proposed the concept of hybrid human-computer regulation, emphasizing that the AI system should be combined with the self-regulation process of learners, and support the adjustment of learning objectives and strategy selection through dynamic data feedback [15]. The subsequent research further pointed out that future learning technologies need to form a collaborative mechanism between human judgment and algorithmic decision making, so that AI can not only deal with complex learning data, but also retain the value judgment in the educational context [16]. However, after a critical analysis of AI discourse in higher education, Bearman et al. pointed out that existing studies tend to strengthen the technology-centric tendency and pay insufficient attention to educational equity, learning subjectivity and institutional adaptation [17]. Crompton and Burke's summary of the current research status of AI in higher education also shows that although intelligent teaching, learning analysis and automatic evaluation develop rapidly, the research on dynamic adjustment of training programs and collaborative decision-making of organization is still relatively insufficient [18]. Gašević et al. emphasized that in the era of artificial intelligence, learner empowerment should be based on data interpretation, feedback understanding and decision-making participation, and educational AI systems need to help learners form the ability of sustainable development [19]. Focusing on the educational application of large language models, Kasneci et al. pointed out that generative AI can support personalized tutoring, content generation and learning feedback, but its output reliability, ethical risk and teaching adaptation still need to be constrained by systematic mechanisms [20].

Personalized learning and adaptive system provide an important technical basis for dynamic education decision-making. Major et al. meta-analysis shows that personalized learning supported by technology can improve learning outcomes, but the effects are affected by data quality, instructional design and implementation conditions [21]. Park et al. pointed out from the perspective of adaptive learning analysis dashboard that learner behavior sequence and reflection data can reveal the self-regulation process and provide basis for subsequent strategy recommendation [22]. Hemmler et al. classified the adaptive learning system and proposed the dimensions of system adaptation object, adaptation basis, adaptation time and adaptation mode, which provided reference for the structural design of educational decision-making model [23].

In terms of enterprise cooperation and vocational ability cultivation in schools, Hiim showed through action research that deep cooperation between schools and workplaces can enhance the real context of vocational education, but the collaborative process requires continuous communication and co-design of curriculum tasks [24]. Jackson and Dean pointed out that different types of work-integrated learning have differentiated effects on graduates' employability, and project practice, industry internship and curriculum embedded tasks can improve career readiness from different dimensions [25]. Jackson and Rowe further proved that the integration of learning and extracurricular activities would affect the labor market outcomes of graduates, indicating that there was a continuous correlation between practical experience in the education process and employment performance [26]. Tushar and Sooraksa reviewed the global employability in the 21st century and pointed out that communication and collaboration, digital ability, problem solving and adaptability have become important dimensions of job evaluation [27]. On the whole, the existing research provides a theoretical basis for AI education decision-making, personalized learning and school-enterprise collaborative education. However, most of the methods still stay at the static evaluation, rule recommendation or stage analysis level, and lack a dynamic decision-making mechanism that can continuously optimize according to the evolution of students 'ability, the change of enterprise needs and collaborative feedback. Therefore, it is of further research value to introduce reinforcement learning into the school-enterprise collaborative education scene and construct the closed loop of state perception, strategy selection, reward feedback and online update.

3 Dynamic decision-making model design of school-enterprise collaborative education driven by reinforcement learning

3.1 Reinforcement Learning state-aware Modeling for School-enterprise collaboration scenarios

The school-enterprise collaborative education decision is a dynamic process of continuous evolution with the growth of students 'ability, the advancement of courses, the change of enterprise post requirements and practical feedback. In order to enable reinforcement learning agents to form effective decisions in complex education scenarios, this paper abstracts the school-enterprise collaborative education process as a partially observable Markov decision process, and constructs a state perception model around four core objects of "student-school-enterprise-feedback". The goal of state modeling is to transform the discrete and scattered educational management data, learning behavior data and enterprise employment data into a unified, continuous and computable state vector, which can provide reliable input for subsequent action selection and strategy optimization.

In the data layer, the student side mainly collected course scores, knowledge point mastery, online learning activity, project completion quality and practice participation frequency. The school side mainly collects curriculum resource supply, training module configuration, teacher guidance intensity and training progress. The enterprise side mainly collects job skill requirements, task difficulty, tutor evaluation and job adaptation standards. The feedback side includes student satisfaction, enterprise recognition, task completion timeliness and ability gain. Considering the heterogeneity and time series of these data, this paper uses a combination of normalization, one-hot encoding and embedded representation to complete the preprocessing, and concatenates the multi-source observation information into the original observation vector at time t:

$$o_t = [e_t^{stu} \parallel e_t^{sch} \parallel e_t^{ent} \parallel e_t^{fb}] \tag{1}$$

where, o_t represents the original observation vector at time t, e_t^{stu} represents the student feature embedding, e_t^{sch} represents the school training environment feature embedding, e_t^{ent} represents the enterprise job demand feature embedding, e_t^{fb} represents the historical feedback feature embedding, and "$\parallel$" represents the vector splicing operation. This formula realizes the unified expression of multi-source heterogeneous information and lays the foundation for subsequent state extraction.

Because the importance of data from different sources to education decision is not the same, if the concatenation vector is directly used as the state input, it is easy to cause the key features to be submerged by noise. To this end, this paper introduces an attention mechanism to weight multi-source observations and automatically learn the relative contributions of student ability, course environment, job requirements and collaborative feedback in the current decision stage. The corresponding attention weights are calculated as follows.

$$\alpha_i^t = \frac{\exp\left((W_q o_t)^\top (W_k x_i^t)/\sqrt{d}\right)}{\sum_{j=1}^{n} \exp\left((W_q o_t)^\top (W_k x_j^t)/\sqrt{d}\right)} \tag{2}$$

where, α_i^t represents the attention weight of the ith feature at time t, x_i^t represents the corresponding feature block, W_q and W_k are the learnable parameter matrices, d is the scaling factor, and n is the number of feature blocks. This formula can highlight the observation information that is more relevant to the current decision, and make the state representation more suitable for the real collaborative situation.

Considering the continuity of student ability change and enterprise demand adjustment, only relying on the current time information is still not enough to describe the state evolution process. In this paper, the gated recurrent unit is further introduced to recursively encode the weighted feature sequence, and the reinforcement learning state vector with both current observation and historical memory is constructed：

$$s_t = \mathrm{GRU}\left(\sum_{i=1}^{n} \alpha_i^t W_v x_i^t, s_{t-1}\right) \tag{3}$$

where s_t represents the state vector at time t, W_v is the feature mapping matrix, s_{t-1} represents the hidden state at the previous time, and GRU(·) represents the gated cyclic update function. This formula can preserve the growth trajectory of international students, the change trend of job demand and the history of school-enterprise interaction in the status coding, so as to improve the perception ability of reinforcement learning agent for long-term decision value. The overall architecture of the school-enterprise collaborative education dynamic decision-making model driven by reinforcement learning is shown in Figure 1.

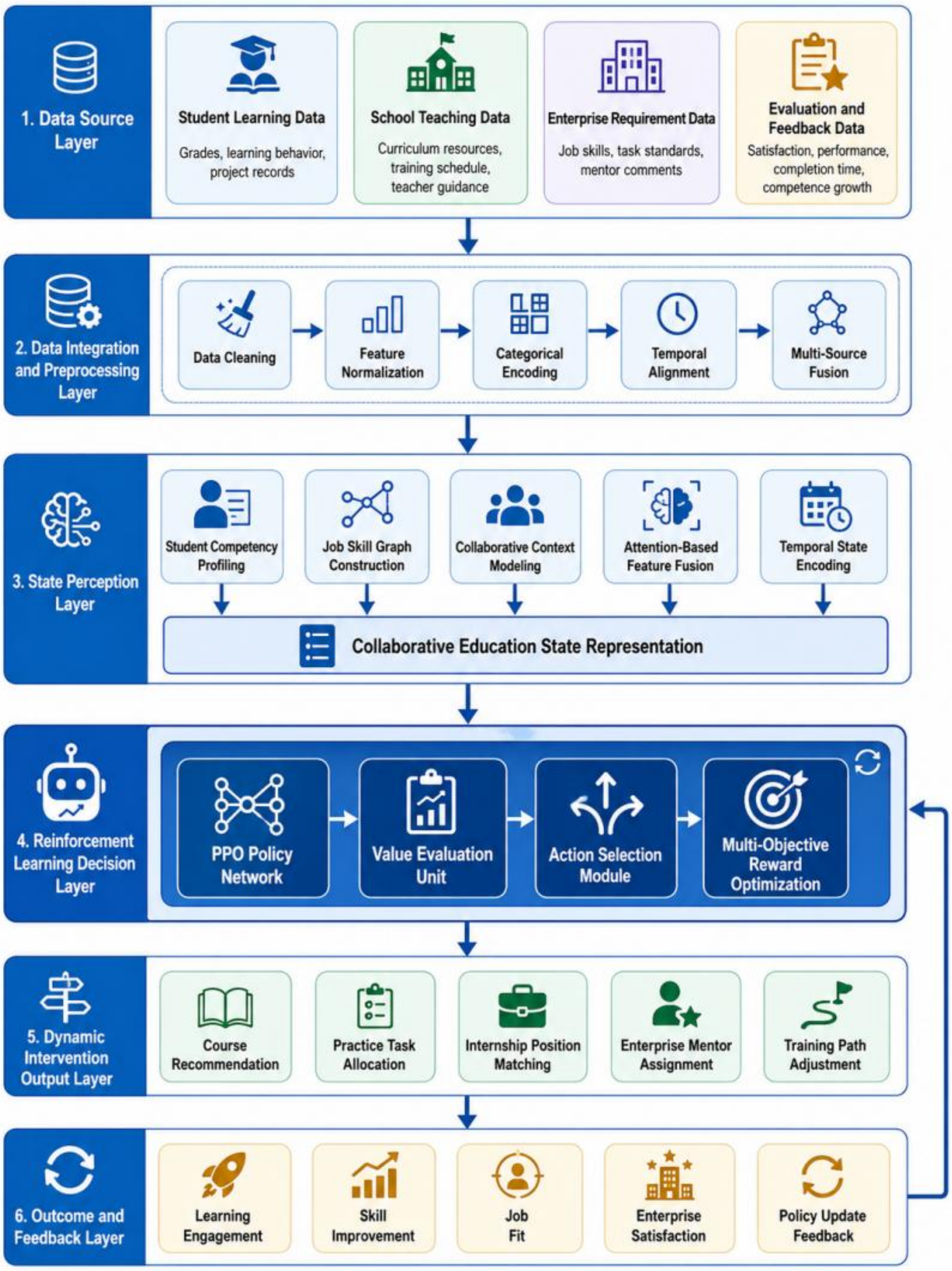

Figure 1: The overall architecture diagram of the dynamic decision-making model of school-enterprise collaborative education driven by reinforcement learning

Figure 1 shows that the state-aware modeling converts the collaborative information of students, schools and enterprises into high-quality state representations that can be directly invoked by the reinforcement learning strategy network through multi-source feature fusion, attention weighting and time series coding, which provides key support for the dynamic choreography and decision optimization of subsequent training paths. On the whole, the state-aware model constructed in this section realizes the unified expression and dynamic coding of multi-source heterogeneous data in the school-enterprise collaborative education scene, so that the reinforcement learning method can carry out strategy learning in a more real and continuous education environment. Subsequent action space design will be developed based on this state representation.

3.2 State representation coding integrating students' ability evolution and job requirements

After the multi-source state perception is completed, the model also needs to further characterize the dynamic relationship between "how students' ability changes" and "how enterprise job requirements change". This paper does not regard student portraits as fixed labels, but models them as temporal ability vectors that are continuously updated with course learning, project practice and enterprise feedback. At the same time, the enterprise job demand is represented as a dynamic graph consisting of positions, skills, tasks and ability levels, so that the reinforcement learning agent can identify the gap between the current ability of students and the enterprise demand before making decisions.

The evolution vector of students' ability is jointly updated by course performance, practical performance, project completion quality and learning behavior, and its temporal coding process is expressed as follows:

$$c_t = \mathrm{GRU}(W_c[x_t^{learn}||x_t^{prac}||x_t^{proj}||x_t^{beh}], c_{t-1}) \quad (4)$$

where, c_t represents the student ability state vector at time t, $x_t^{learn}, x_t^{prac}, x_t^{proj}$ and x_t^{beh} represent the characteristics of course learning, practical training, project task and learning behavior respectively, W_c is the feature mapping matrix, and ct−1 is the ability state at the previous stage. The coding method can retain the growth trajectory of students 'ability and avoid judging the training state only according to a single score.

The demand side of enterprise jobs is structurally expressed by graph neural network. The job requirement graph is formed by job skill nodes, task requirement nodes and ability level nodes, and the job requirement embedding is obtained by neighborhood aggregation:

$$h_i^{job} = \sigma\left(\sum\nolimits_{j\in\mathcal{N}(i)} \beta_{ij} W_g h_j\right) \quad (5)$$

where, h_i^{job} represents the embedding representation of the ith job demand node, $\mathcal{N}(i)$ represents the set of neighborhood nodes connected to the node, β_{ij} represents the correlation weights between nodes, W_g is the graph convolution parameter matrix, and $\sigma(\cdot)$ is the nonlinear activation function. In this way, the model can transform the skill combination, task association and ability level in enterprise positions into computable graph structure features.

In order to establish the fine-grained connection between students' ability and job requirements, this paper introduces the cross-domain attention mechanism to calculate the matching relationship between the ability dimension and the job skill dimension:

$$\lambda_i^t = \frac{\exp\left((W_q c_t)^\top \left(W_k h_i^{job}\right)/\sqrt{d}\right)}{\sum_{r=1}^{m} \exp\left((W_q c_t)^\top \left(W_k h_r^{job}\right)/\sqrt{d}\right)} \quad (6)$$

where, λ_i^t represents the attention weight of students' current ability status to the type i job demand, W_q and W_k are query and key mapping matrices, m is the number of job demand nodes, and d is the scaling dimension. The higher the weight is, the more critical the job skill is to the current training decision, and the model can identify the ability direction that needs to be strengthened preferentially.

On this basis, students' ability evolution information, job demand graph information and their difference relationship are fused into the state representation of reinforcement learning agent:

$$z_t = \text{LayerNorm}\left(W_s[c_t \| h_t^{job} \| c_t \odot h_t^{job} \| |c_t - h_t^{job}|]\right) \tag{7}$$

where, z_t represents the final state representation vector,h_t^{job} represents the job demand aggregation vector after attention weighting, $\odot$ represents element-by-element multiplication, $|c_t - h_t^{job}|$ represents the difference characteristics of ability demand, and W_s is the fusion mapping matrix. The state representation also contains the information of students 'ability, post goal demand, ability matching degree and ability gap, which can be used as the direct input for subsequent reinforcement learning action selection.

Through the above coding process, the model transformed the "learning process data" and "job demand data" in school-enterprise collaborative education into a unified state space, so that the subsequent strategy network could carry out continuous decisions around curriculum recommendation, practical task allocation and job adaptation optimization.

3.3 Action Space Design of Reinforcement Learning for dynamic choreography of training paths

In the scene of school-enterprise collaborative education, the action of reinforcement learning agent is not a simple "recommend courses" or "assign tasks", but a phased arrangement of student training paths. The action space needs to simultaneously cover school curriculum supply, enterprise practice resources, post ability requirements and student growth rhythm, so that the model can output executable training intervention strategies in different states. In this paper, the action space is designed as a hierarchical structure of "action type-resource object-execution intension-feedback cycle", which not only ensures that the action has educational business meaning, but also facilitates the discrete selection and parameterized control of the computer model.

The action space mainly includes the categories of course recommendation, practical task allocation, enterprise project participation, tutor matching, post ability reinforcement and training path adjustment. Each type of action is not directly used as an isolated label, but is further bound to the course number, task difficulty, enterprise position direction, training cycle and target ability dimensions. In order to avoid the reinforcement learning agent selecting actions that do not meet the training conditions, this paper introduces a feasible action screening mechanism to dynamically generate a candidate action set according to the current ability state of students, the relationship between course prerequisites, enterprise resource capacity and job skill gap:

$$\mathcal{A}_t^{\text{valid}} = \{a_k \mid \phi(a_k, z_t) = 1,\ \psi(a_k, r_t) \le \tau\} \tag{8}$$

where, $\mathcal{A}_t^{\text{valid}}$ represents the set of actions that can be executed at time t, a_k represents the candidate training intervention action, $\phi(a_k, z_t)$ represents the adaptation judgment function between the action and the current student state representation, $\psi(a_k, r_t)$ represents the occupation evaluation of the action on enterprise resources, course capacity and tutor load, and τ represents the resource constraint threshold. This mechanism can reduce invalid decisions and make the action selection conform to the real school-enterprise collaboration conditions.

In order to enhance the ability of action expression, each training intervention action is encoded as a vector representation integrating action categories, object resources and execution parameters:

$$u_k = \mathrm{Emb}(\mathrm{type}_k) + \mathrm{Emb}(\mathrm{obj}_k) + W_p p_k \tag{9}$$

where, u_k represents the vector representation of the KTH action, $\mathrm{Emb}(\mathrm{type}_k)$ represents the embedding of the action category, $\mathrm{Emb}(\mathrm{obj}_k)$ represents the embedding of resource objects such as courses, tasks, positions or mentors, pk represents the continuous parameters such as training intensity, execution cycle and difficulty level, and W_p is the parameter mapping matrix. With this encoding, the reinforcement learning agent can understand the structural differences between different training interventions, rather than just making a mechanical choice of action numbers. Table 1 shows the correspondence between the action space of reinforcement learning and the specific training intervention strategy, which is used to illustrate how the model output is transformed into executable operations in school-enterprise collaborative education.

Table 1: Correspondence table between reinforcement learning action space and training intervention strategy

Action type	Intervention strategy	Trigger basis	Execution object	Decision function
Course recommendation	Recommend core courses, extended courses, or online learning resources	Insufficient knowledge mastery and deviation from course objectives	Students and course resources	Fill theoretical knowledge gaps
Practice task assignment	Assign enterprise cases, project training, or staged tasks	Insufficient practical ability and lack of project experience	Students and training tasks	Strengthen engineering practice ability
Job ability enhancement	Push special training modules and skill improvement tasks	Obvious job skill gaps	Students and job skill points	Improve job adaptability
Enterprise mentor matching	Match enterprise mentors, technical consultants, or project leaders	Student development direction close to enterprise positions	Students and enterprise mentors	Enhance guidance in real scenarios
Internship position matching	Recommend internship positions or enterprise project groups	Ability profile meeting basic job requirements	Students and enterprise positions	Promote transformation of learning outcomes
Training path adjustment	Adjust course sequence, practice cycle, and training intensity	Abnormal learning progress or feedback fluctuation	Training plans and teaching arrangements	Achieve dynamic path optimization

Table 1 shows that the action space not only includes the recommendation of learning resources, but also covers the key links of collaborative education such as enterprise participation, job adaptation and path adjustment, so that the decision results of reinforcement learning can directly serve the dynamic execution of the training program.

In the action selection phase, we use a mask strategy to filter infeasible actions, and calculate the selection probability in the set of valid actions:

$$P(a_k|z_t) = \frac{M_k \exp(g_\theta(z_t, u_k))}{\sum_{a_j \in \mathcal{A}_t^{valid}} M_j \exp(g_\theta(z_t, u_j))} \quad (10)$$

where $P(a_k|z_t)$ represents the probability of selecting action a_k under state z_t, M_k represents the action feasibility mask, and $g_\theta(zt,uk)$ represents the scoring function of the policy network for the matching degree between state and action. The design enables the model to select training interventions that are more conducive to ability improvement and post adaptation on the basis of ensuring the executability of actions.

Through hierarchical action modeling, resource constraint screening and masked strategy selection, this paper constructs a reinforcement learning action space oriented to the dynamic choreography of training path, which makes the school-enterprise collaborative education decision-making change from static resource matching to an intelligent choreography process that can be feedback, adjustable and sustainable optimization.

3.4 Multi-objective reward decision optimization mechanism based on PPO policy network

After the state representation encoding and action space construction, the model needs to further solve the strategy optimization problem of "which kind of training intervention is more conducive to the long-term education effect". The decision-making of school-enterprise collaborative education has obvious multi-objective characteristics. The single pursuit of course performance improvement may weaken the effect of post practice, and the single emphasis on enterprise post adaptation may also cause unbalanced development of students 'basic ability. Therefore, this paper uses PPO strategy network to construct a stable reinforcement learning optimization mechanism, and integrates ability improvement, job adaptation, learning participation, enterprise satisfaction and resource consumption into the reward calculation, so that the model can find a more balanced training path in multiple rounds of interaction.

In this paper, the comprehensive reward at time t is designed as a multi-objective weighted form:

$$R_t = \omega_1 G_t^{cap} + \omega_2 G_t^{job} + \omega_3 G_t^{eng} + \omega_4 G_t^{ent} - \omega_5 C_t^{res} \quad (11)$$

where, R_t represents the comprehensive reward obtained by the current decision action, G_t^{cap} represents the benefit of improving students' ability, G_t^{job} represents the benefit of job adaptation, G_t^{eng} represents the benefit of learning participation, G_t^{ent} represents the benefit of enterprise evaluation, C_t^{res} represents the cost of resource consumption, and ω_1 to ω_5 are the weight coefficients of different objectives. The reward function incorporated the education quality objective and enterprise collaboration cost into the strategy learning process at the same time to avoid the model only favoring short-term score improvement.

In order to reduce the interference of single-step reward fluctuation on policy update, this paper introduces generalized dominance estimation to smooth the calculation of long-term payoff:

$$A_t = \sum_{l=0}^{L-1} (\gamma\lambda)^l (R_{t+l} + \gamma V_\theta(z_{t+l+1}) - V_\theta(z_{t+l})) \quad (12)$$

where, A_t represents the estimated advantage value At time t, γ represents the discount factor of future benefits, λ represents the smoothing coefficient of advantages, L represents the length of the trajectory, and $V_\theta(\cdot)$ represents the estimate of the long-term benefits of the state by the value network. The formula can integrate immediate feedback and subsequent training results, so that the model pays more attention to the cumulative education effect brought by continuous decisions.

PPO policy network improves training stability by limiting the update range of new and old policies, and its pruning objective function is expressed as follows:

$$L^{clip}(\theta) = \mathbb{E}_t\left[\min(\rho_t(\theta)A_t, \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)A_t)\right] \tag{13}$$

where, $L^{clip}(\theta)$ represents the PPO clipping objective function, $\rho_t(\theta)$ represents the probability ratio of the new and old strategies selecting the same action, ϵ represents the clipping threshold, and $E_t[\cdot]$ represents the expectation by time step. This mechanism can prevent the strategy from changing drastically under a small amount of feedback, which is suitable for the scenario of school-enterprise collaborative education with long feedback cycle and high decision-making risk.

In order to optimize policy selection, value estimation and exploration ability simultaneously, this paper combines policy loss, value loss and entropy regularization term as an overall training objective:

$$L(\theta) = L^{clip}(\theta) - \eta_1(R_t - V_\theta(z_t))^2 + \eta_2 H(\pi_\theta(\cdot|z_t)) \tag{14}$$

where $L(\theta)$ represents the final optimization objective, η_1 is the value loss weight, η_2 is the entropy regularization weight, and $H(\pi_\theta(\cdot|z_t))$ represents the entropy of the policy distribution. The value loss is used to improve the accuracy of state benefit estimation, and the entropy regularization is used to maintain a certain exploration ability to avoid the model being prematurely fixed on a few training intervention actions. Figure 2 shows the optimization process of multi-objective reward decision based on PPO policy network.

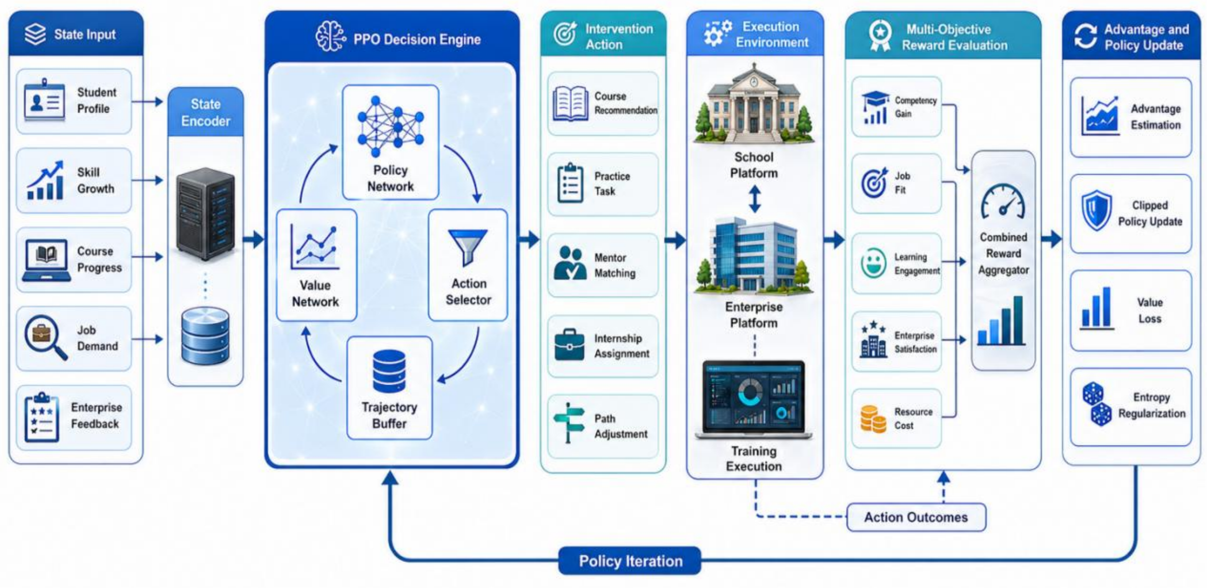

Figure 2: Flow chart of multi-objective reward decision optimization based on PPO policy network

Figure 2 shows that the PPO decision optimization process takes the state representation as input, generates training intervention actions through the policy network, calculates multi-objective rewards according to student performance, job adaptation and enterprise feedback, and uses the tailoring update mechanism to continuously revise the strategy. This process enables the model to realize the dynamic optimization of the training path while ensuring the training stability.

In the actual training, the system collects student learning records, enterprise practice feedback and stage evaluation results in batches to form continuous trajectory samples, and re-evaluates the action effect after each round of strategy update. When a certain type of action brings high ability improvement but excessive resource consumption, the reward function will automatically reduce its long-term selection probability. When a certain type of practical task can improve student participation and job suitability at the same time, the strategy network will gradually improve the output probability of the action in similar states. Through this multi-objective reward constraint and PPO stable update mechanism, the school-enterprise collaborative education decision-making can shift from artificial experience driven to data feedback driven, which provides a reliable basis for subsequent online strategy calibration and model continuous update.

3.5 Online Policy Calibration and Model Update in Reinforcement Learning feedback Loops

In the real operation, the school-enterprise collaborative education decision will continue to receive feedback such as students 'learning behavior, the evaluation of enterprise tutors, the results of training tasks and the changes of job requirements. In order to avoid policy fixation after model training, this paper constructs a reinforcement learning feedback loop, which connects online log collection, feedback credibility evaluation, policy drift detection and security update control, so that the decision model can be continuously calibrated during the actual training process. The system adopts the calculation process of "front-end behavior embedding point - message queue - feature service - policy evaluation - gray update", writes student click, task completion, course score, enterprise score and job demand change into the feedback cache, and generates interaction samples for policy update according to the time window.

In order to reduce the interference of abnormal feedback on model update, this paper first calculates the feedback credibility weight:

$$\kappa_t = \mathrm{Sigmoid}(W_f[q_t, l_t, v_t, h_t] + b_f) \tag{15}$$

where κ_t represents the feedback credibility weight at time t, q_t represents the data integrity score, lt represents the feedback delay feature, v_t represents the multi-source evaluation consistency, h_t represents the historical behavior stability, W_f and b_f are the learnable parameters. This weight is used to control the influence strength of feedback samples when entering the strategy update, so that the model focuses more on stable, complete and credible school-enterprise collaborative feedback.

Considering the possible distribution deviation of enterprise post requirements and students 'learning status over time, this paper sets the state drift detection index:

$$D_t = \mathrm{KL}(P_t^{new} \| P^{ref}) + \xi \left\| \bar{z}_t - \bar{z}^{ref} \right\|_2 \tag{16}$$

where, D_t represents the state drift intensity of the current window, Ptnew represents the state distribution in the new feedback window, Pref represents the historical reference distribution,

$\bar{z}_t$ represents the current state mean vector, $\bar{z}^{ref}$ represents the reference state mean vector, ξ represents the drift amplitude adjustment coefficient. The index can identify the change of post ability demand, the fluctuation of student group ability and the deviation of course implementation effect, and provide a trigger basis for online update.

In the policy update phase, this paper does not directly replace the original policy, but uses the security fusion method to generate the deployment policy:

$$\pi_t^{deploy} = (1 - \chi_t)\pi_t^{old} + \chi_t\pi_t^{new}, \quad \chi_t = \min(\chi_{max}, \kappa_t e^{-D_t}) \tag{17}$$

where, π_t^{deploy} denotes the actual deployment policy, π_t^{old} denotes the current stable policy, π_t^{new} denotes the new training policy, χ_t denotes the policy fusion coefficient, and χ_{max} denotes the maximum update ratio. This mechanism increases the weight of the new policy when the feedback is reliable and the state drift is small, and reduces the update amplitude when the data is abnormal or the scene changes dramatically, so as to enhance the security of model deployment.

Through the above online calibration mechanism, the reinforcement learning model can continuously revise the training path decision according to the real collaborative feedback. When the enterprise feedback shows that a certain type of practical task can improve the job suitability, the system will gradually increase the selection probability of relevant actions. When student engagement decreases or resource load is too high, the model will automatically reduce the corresponding intervention intensity. Therefore, the school-enterprise collaborative education decision forms a closed-loop update mechanism that can be perceived, calibrated, and rolled back.

4 Results

4.1 Construction of experimental data set and explanation of sample characteristics

In order to verify the effectiveness of the dynamic decision-making model of school-enterprise collaborative education driven by reinforcement learning, this paper constructs an experimental data set covering students 'learning process, curriculum resource allocation, enterprise job demand, practical task execution and collaborative feedback evaluation. The data sources include school teaching management platform, online learning system, enterprise training management system and school-enterprise joint evaluation records. The experimental objects are students of information technology related majors in an application-oriented university, and the data period is 16 weeks. The original data includes students' basic information, course scores, learning behavior logs, project training records, enterprise tutor ratings, job skill requirements and stage feedback results. In order to ensure data quality, duplicate logs, missing fields, abnormal scores and invalid interaction records are cleaned, and normalization, category coding and time window segmentation methods are used to complete feature processing. Finally, 1260 students, 86 courses, 42 types of enterprise positions, 7840 practical task records and 9360 collaborative feedback records were retained, forming 26780 effective interaction samples. The samples were divided into training set, validation set and test set according to the student number, with the proportion of 70%, 15% and 15% to avoid the same student data appearing across sets. Each processed sample contained 112 dimensional features such as students 'ability, learning behavior, course status, job requirements, practical performance and feedback evaluation, which could provide continuous state input for PPO strategy network, and also provide experimental basis for subsequent training program matching, strategy convergence and model stability verification.

4.2 Comparison of matching effects of training programs under different decision-making models

In order to verify the matching advantages of the proposed model in the dynamic decision-making of school-enterprise collaborative education, this paper selects five models of rule recommendation, collaborative filtering, deep recommendation, DQN and PPO for comparative experiments, and takes the matching degree of training program as the core evaluation index. The matching degree was calculated by integrating students 'ability status, course goal achievement degree, job skill coverage rate and enterprise feedback consistency. The experimental results show that the traditional rule recommendation is limited by static condition matching, which is difficult to adapt to the evolution of students 'ability and the fluctuation of job demand. Although collaborative filtering and deep recommendation improve the degree of personalization, the description of continuous decision-making process is still insufficient. DQN can improve the dynamic decision-making ability to a certain extent, but the stability is slightly weak in complex action space. In contrast, the model based on PPO in this paper has the best performance in terms of training program matching, indicating that it can more effectively coordinate the relationship between students 'growth path and enterprise job demand. Figure 3 shows the comparison of matching degrees of training schemes under different decision models.

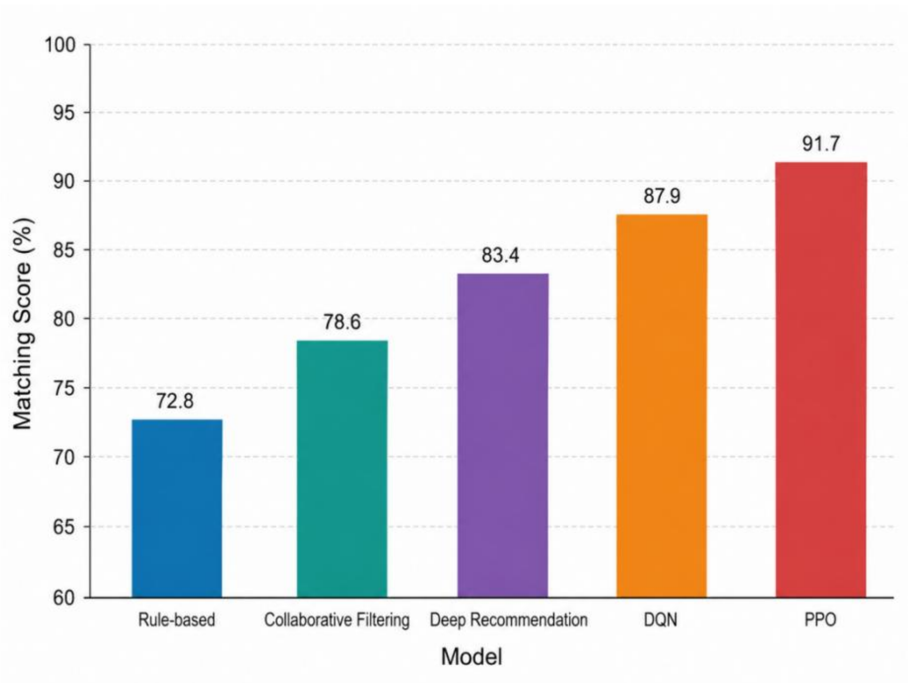

Figure 3: Bar charts for comparison of matching degree of cultivation schemes under different decision models

As can be seen from Figure 3, the matching degrees of training schemes of rule recommendation, collaborative filtering, deep recommendation and DQN models are 72.8%, 78.6%, 83.4% and 87.9%, respectively, while the proposed PPO model reaches 91.7%. Compared with rule recommendation, it was increased by 18.9 percentage points, and compared with DQN, it was further increased by 3.8 percentage points, indicating that the proposed method has better comprehensive effect in matching dynamic training path arrangement and job demand.

4.3 Analysis of convergence and reward change of reinforcement learning policy

In order to verify the training stability and optimization effectiveness of the proposed PPO strategy network in the dynamic decision-making of school-enterprise collaborative education, this paper analyzes the convergence trend of average reward, the change of network loss and value error, and the evolution of multi-objective reward composition. The experimental results show that with the increase of training rounds, the policy network can gradually learn a better training intervention plan, the reward output is more stable, the value estimation error continues to decrease, and the multi-objective reward structure also turns from the fluctuation state to a relatively balanced state, indicating that the model has better convergence and decision optimization ability. Figure 4 shows the convergence curve of training average reward for PPO policy.

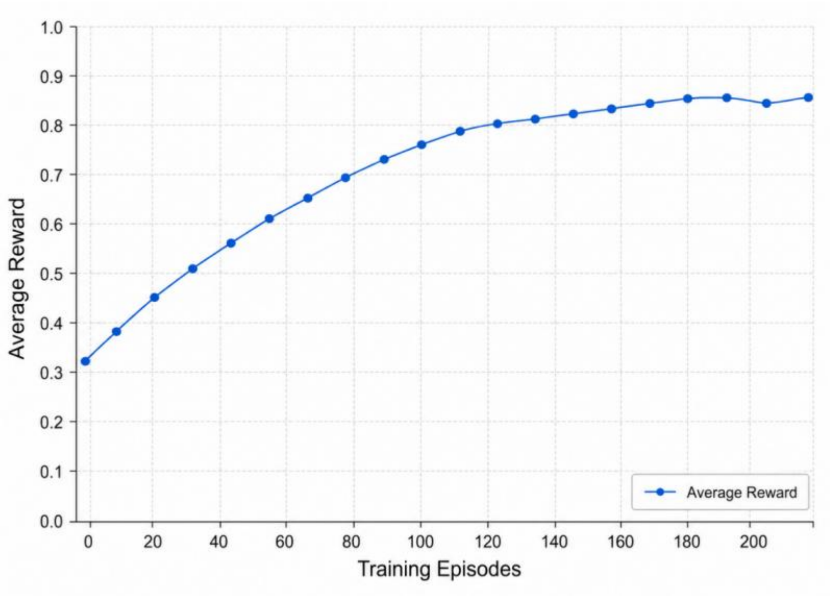

Figure 4: Convergence curve of average reward for training of PPO policy

As can be seen from Figure 4, the average reward increases rapidly in the early stage of model training, rising from 0.32 to 0.61 in the first 50 rounds, reaching 0.78 after 100 rounds, and stabilizing around 0.86 around 180 rounds with only a small fluctuation. This shows that PPO can constantly revise the training path decision in continuous interaction, and the policy output gradually transitions from the exploration phase to the stable optimization phase. Figure 5 shows the network loss and value error variation of PPO strategy.

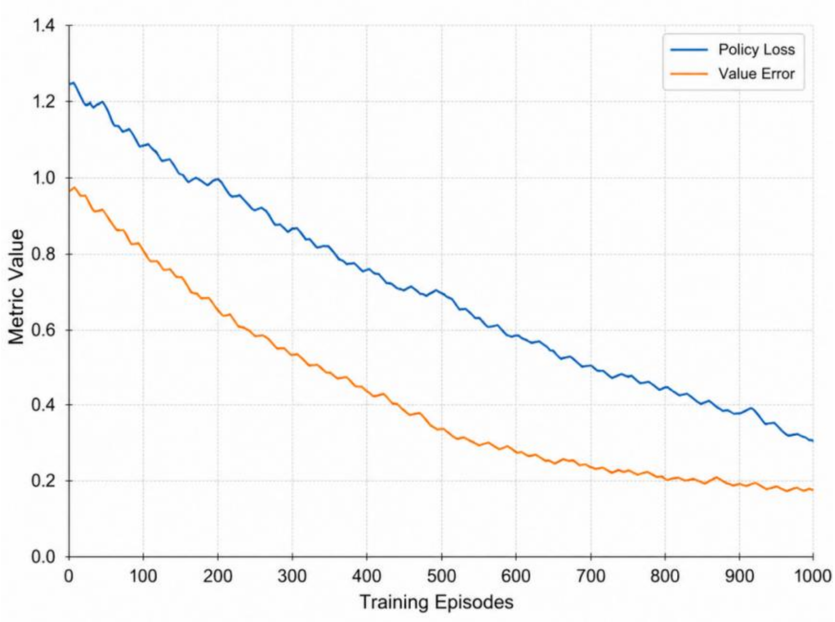

Figure 5: Graph of network loss and value error variation of PPO strategy

It can be seen from Figure 5 that the policy loss decreases from 1.24 to 0.31 at the beginning of training, and the value error decreases from 0.96 to 0.18, and tends to be stable in the later stage. The synchronous decline of the two indicators indicates that the direction of strategy update is stable, the estimation of long-term returns by the value network is gradually accurate, and there is no obvious oscillation or divergence in the model. Figure 6 shows how the multi-objective reward composition changes with training rounds.

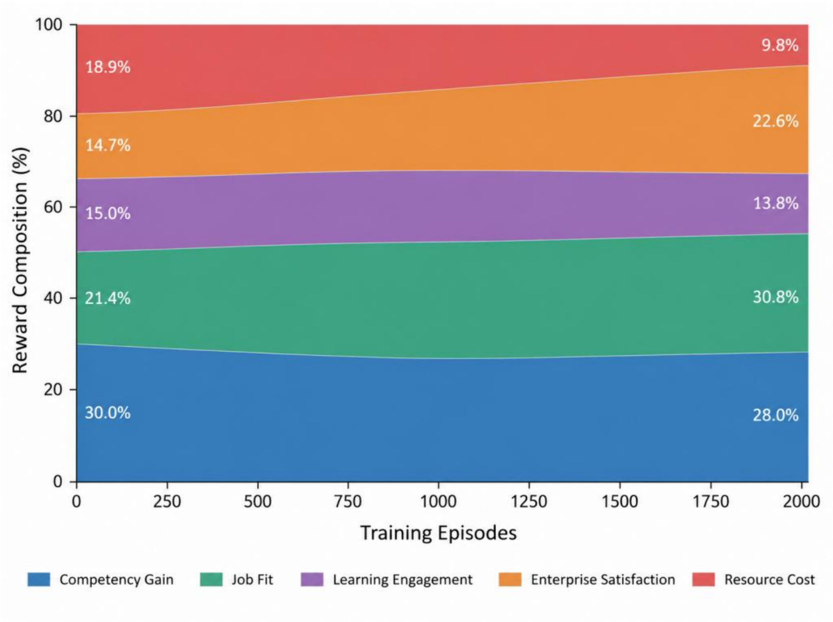

Figure 6: Stacked area plot of multi-objective reward composition as a function of training rounds

It can be seen from Figure 6 that the reward in the early stage of training is mainly driven by ability improvement and post adaptation, and the learning participation maintains a medium contribution level. After entering the middle and late stage, the contribution of job adaptation and enterprise satisfaction continued to increase, of which the contribution of job adaptation increased from 21.4% to 30.8%, enterprise satisfaction increased from 14.7% to 22.6%, and the proportion of resource cost decreased from 18.9% to 9.8%. This shows that the model gradually realizes the multi-objective optimization from the single learning benefit orientation to the ability improvement, job matching and collaborative efficiency in the process of strategy optimization.

4.4 Students' ability improvement and job suitability evaluation

In order to verify the education effect of the model intervention, this paper further evaluates from two dimensions of students 'ability improvement and job suitability. The evaluation indicators included knowledge mastery, practical operation ability, project completion ability and post suitability, and the statistical object was the stage results after the continuous intervention of the model. The experimental results show that PPO strategy network can dynamically adjust the training path according to the change of students 'ability and the needs of enterprises, so as to form a higher consistency between the improvement process of students 'comprehensive ability and the requirements of positions. Figure 7 shows the change trend of students' ability improvement and job suitability.

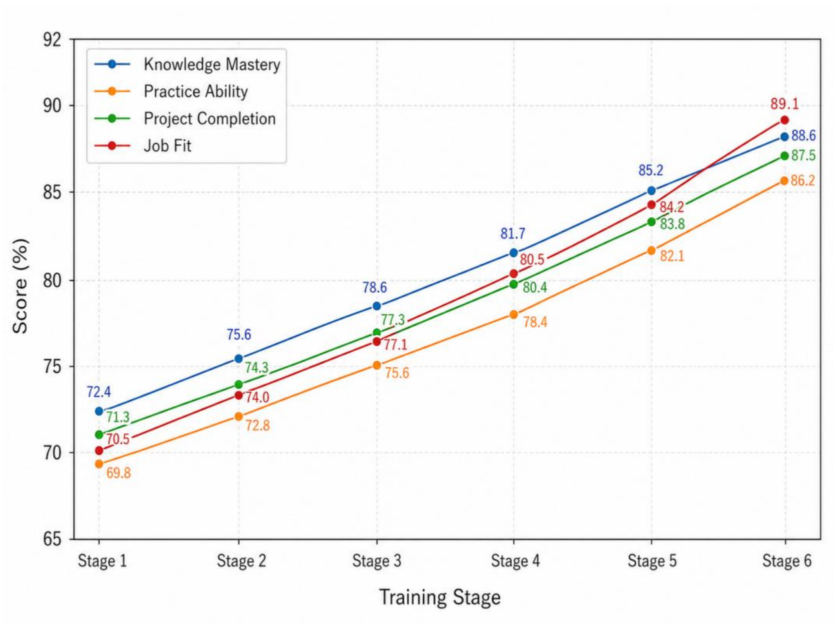

Figure 7: The change trend of students' ability improvement and job suitability

As can be seen from Figure 7, with the advancement of the training cycle, students 'knowledge mastery degree increased from 72.4% to 88.6%, practical ability increased from 69.8% to 86.2%, project completion ability increased from 71.3% to 87.5%, and job suitability increased from 70.5% to 89.1%. This shows that the proposed model can better promote the growth of students 'ability and improve the level of person-post matching in the scene of school-enterprise collaborative education.

4.5 Ablation experiment and model stability verification

In order to verify the actual contribution of each key module to the model performance, this paper designs an ablation experiment and further investigates the stability of the model under the condition of feedback disturbance. In the stability verification stage, 5%-20% delayed feedback, missing feedback and score fluctuation are added to the original feedback data to simulate the feedback disturbance in the real school-enterprise collaboration process. The experiment removed the state representation coding, action mask, multi-objective reward and online update modules respectively, and compared the differences between the complete model and each variant in the training scheme matching degree, post fitness degree, average reward and stability score. Figure 8 shows the results of ablation experiments and model stability verification.

Figure 8: Heat map of ablation experiments and model stability verification

It can be seen from Figure 8 that the complete model remains optimal in all indicators, with the matching degree of training program reaching 91.7%, the matching degree of position reaching 89.1%, the average reward being 0.86, and the stability score being 0.88. After removing the state representation encoding, the matching degree is reduced to 87.6%, the average reward is reduced to 0.79 after removing the action mask, the post fitness is reduced to 84.3% after removing the multi-objective reward, and the stability score is reduced to 0.81 after removing the online update. The results show that all modules have obvious supporting effects on the performance of the model, and online update and multi-objective reward have more significant effects on stability and comprehensive decision-making effect.

5 Discussion

The core value of the dynamic decision-making model of school-enterprise collaborative education driven by reinforcement learning constructed in this paper is to incorporate students 'ability growth, school curriculum supply, enterprise job demand and collaborative feedback into the same decision-making closed loop, so that the adjustment of training program changes from experience judgment to data-driven continuous optimization. Compared with rule recommendation, collaborative filtering and deep recommendation methods, PPO policy network can correct action selection according to reward feedback in multiple rounds of interaction, so it is more suitable for dealing with problems with long-term benefits such as training path choreography, practice task allocation and job adaptation. The experimental results show that the model in this paper has achieved good performance in training program matching degree, students 'ability improvement and job suitability, indicating that the reinforcement learning mechanism can effectively capture the relationship between students 'state changes and enterprise demand fluctuations.

From the perspective of technical implementation, the state representation encoding has a great impact on the performance of the model. Student ability is not a result that can be summarized by a single achievement index, but a dynamic feature composed of knowledge mastery, practical performance, project completion quality and learning behavior. Through the fusion of timing coding and job demand graph, this paper enables the model to identify the gap between students 'current ability and the target position, which provides more accurate input for subsequent action selection. Action space design also plays an important role. After course recommendation, practice task, tutor matching, internship position and path adjustment are uniformly coded, the model can select a training intervention strategy that is more suitable for the current state under resource constraints, reducing invalid recommendation and repeated training.

Multi-objective reward function is the key to achieve collaborative optimization of the model. School-enterprise collaborative education should not only focus on students 'academic performance, but also take into account job adaptation, learning participation, enterprise satisfaction and resource consumption. In the experiment, the contribution proportion of job adaptation and enterprise satisfaction increased, and the proportion of resource cost decreased in the later stage of training, indicating that the model gradually formed a decision-making tendency that was more in line with the goal of collaborative education in the process of strategy iteration. Ablation experiments further show that the stability and comprehensive effect of the model are decreased after removing multi-objective rewards or online updates, which indicates that the feedback loop is necessary for dynamic adjustment in real education scenarios.

Of course, the model in this paper still has some limitations. The experimental data mainly come from information technology related majors, and the job types and course structures have certain field characteristics. In the future, it is necessary to verify the generalization ability of the model in different professional scenarios such as manufacturing, nursing, and finance and economics. At the same time, the subjectivity and delay of enterprise feedback data may affect the accuracy of reward signals. In the future, federated learning, causal inference and interpretable reinforcement learning methods can be further introduced to improve the model interpretation ability and cross-scenario transfer ability on the basis of protecting the privacy of school and enterprise data, so that the dynamic decision-making model can better serve the long-term, stable and sustainable collaborative education practice.

6 Conclusion

Focusing on the problems of slow decision response, extensive adjustment of training path and insufficient utilization of enterprise feedback in school-enterprise collaborative education, this paper proposes a dynamic decision model driven by reinforcement learning. Through multi-source data fusion, state-aware modeling, student ability evolution coding and job demand graph construction, the model transformed scattered teaching data, practice data and enterprise evaluation into computable state representation. Through the hierarchical action space, action feasibility mask and PPO strategy network, the continuous optimization of course recommendation, practical task allocation, post ability reinforcement and training path adjustment was realized. Experimental results show that the proposed model is superior to rule recommendation, collaborative filtering, deep recommendation and DQN models in terms of training program matching degree, job suitability, reward convergence and stability.

The matching degree of PPO model reached 91.7%, which was 3.8 percentage points higher than that of DQN. The degree of knowledge mastery, practical ability, project completion ability and job suitability showed a continuous improvement trend. Ablation experiments further prove that the state encoding, multi-objective reward and online update mechanism have obvious supporting effects on the comprehensive performance of the model. In the future, federated learning, causal inference and interpretable reinforcement learning can be combined to further improve the generalization ability and deployment value of the model in multi-professional, multi-enterprise and cross-regional collaborative education scenarios.

Author's Profile

Wensheng Liu was born in Nanyang, Henan in August 1983. He obtained his Bachelor's degree in Management from Zhengzhou University in July 2007 and his Master's degree in Education from Henan University in July 2010. Currently, he serves as a dedicated lecturer at the School of Marxism, Yantai Nanshan University.

References

[1] Riedmann A, Schaper P, Lugrin B. Combining Human and Artificial Intelligence for Adaptive Learning Using Reinforcement Learning and Knowledge Tracing[J]. International Journal of Artificial Intelligence in Education, 2025, 35: 2669-2723. DOI: 10.1007/s40593-025-00494-6.

[2] Mon F A, Wasfi A, Hayajneh M, Slim A, Abu Ali N. Reinforcement Learning in Education: A Literature Review[J]. Informatics, 2023, 10(3): 74. DOI: 10.3390/informatics10030074.

[3] Amin A, Uddin M, Alarood A A, Mashwani W K, Alzahrani A, Alzahrani M. Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning[J]. IEEE Access, 2023, 11: 89769-89790. DOI: 10.1109/ACCESS.2023.3305584.

[4] El Fazazi H, Elgarej M, Qbadou M, Mansouri K. Design of an Adaptive E-Learning System Based on Multi-Agent Reinforcement Learning[J]. Engineering, Technology & Applied Science Research, 2021, 11(1): 6637-6644. DOI: 10.48084/etasr.3905.

[5] Afsar M M, Crump T, Far B. Reinforcement Learning Based Recommender Systems: A Survey[J]. ACM Computing Surveys, 2023, 55(7): Article 145. DOI: 10.1145/3543846.

[6] Kabudi T, Pappas I, Olsen D H. AI-Enabled Adaptive Learning Systems: A Systematic Mapping of the Literature[J]. Computers and Education: Artificial Intelligence, 2021, 2: 100017. DOI: 10.1016/j.caeai.2021.100017.

[7] Okonkwo C W, Ade-Ibijola A. Chatbots Applications in Education: A Systematic Review[J]. Computers and Education: Artificial Intelligence, 2021, 2: 100033. DOI: 10.1016/j.caeai.2021.100033.

[8] Lemay D J, Baek C, Doleck T. Comparison of Learning Analytics and Educational Data Mining: A Topic Modeling Approach[J]. Computers and Education: Artificial

Intelligence, 2021, 2: 100016. DOI: 10.1016/j.caeai.2021.100016.

[9] Tahiru F. AI in Education: A Systematic Literature Review[J]. Journal of Cases on Information Technology, 2021, 23(1): 1-20. DOI: 10.4018/JCIT.2021010101.

[10] Akgun S, Greenhow C. Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings[J]. AI and Ethics, 2022, 2(3): 431-440. DOI: 10.1007/s43681-021-00096-7.

[11] Bhutoria A. Personalized Education and Artificial Intelligence in the United States, China, and India: A Systematic Review Using a Human-in-the-Loop Model[J]. Computers and Education: Artificial Intelligence, 2022, 3: 100068. DOI: 10.1016/j.caeai.2022.100068.

[12] Holmes W, Tuomi I. State of the Art and Practice in AI in Education[J]. European Journal of Education, 2022, 57(4): 542-570. DOI: 10.1111/ejed.12533.

[13] Celik I, Dindar M, Muukkonen H, Järvelä S. The Promises and Challenges of Artificial Intelligence for Teachers: A Systematic Review of Research[J]. TechTrends, 2022, 66: 616-630. DOI: 10.1007/s11528-022-00715-y.

[14] Chounta I A, Bardone E, Raudsep A, Pedaste M. Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support Their Practice in Estonian K-12 Education[J]. International Journal of Artificial Intelligence in Education, 2022, 32: 725-755. DOI: 10.1007/s40593-021-00243-5.

[15] Molenaar I. The Concept of Hybrid Human-AI Regulation: Exemplifying How to Support Young Learners' Self-Regulated Learning[J]. Computers and Education: Artificial Intelligence, 2022, 3: 100070. DOI: 10.1016/j.caeai.2022.100070.

[16] Molenaar I. Towards Hybrid Human-AI Learning Technologies[J]. European Journal of Education, 2022, 57(4): 632-645. DOI: 10.1111/ejed.12527.

[17] Bearman M, Ryan J, Ajjawi R. Discourses of Artificial Intelligence in Higher Education: A Critical Literature Review[J]. Higher Education, 2023, 86: 369-385. DOI: 10.1007/s10734-022-00937-2.

[18] Crompton H, Burke D. Artificial Intelligence in Higher Education: The State of the Field[J]. International Journal of Educational Technology in Higher Education, 2023, 20: Article 22. DOI: 10.1186/s41239-023-00392-8.

[19] Gašević D, Siemens G, Sadiq S. Empowering Learners for the Age of Artificial Intelligence[J]. Computers and Education: Artificial Intelligence, 2023, 4: 100130. DOI: 10.1016/j.caeai.2023.100130.

[20] Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education[J]. Learning and Individual Differences, 2023, 103: 102274. DOI: 10.1016/j.lindif.2023.102274.

[21] Major L, Francis G A, Tsapali M. The Effectiveness of Technology-Supported Personalised Learning in Low- and Middle-Income Countries: A Meta-Analysis[J]. British Journal of Educational Technology, 2021, 52(5): 1935-1964. DOI: 10.1111/bjet.13116.

[22] Park E, Ifenthaler D, Clariana R. Adaptive or Adapted to: Sequence and Reflexive Thematic Analysis to Understand Learners' Self-Regulated Learning in an Adaptive Learning Analytics Dashboard[J]. British Journal of Educational Technology, 2023, 54(1): 98-125. DOI: 10.1111/bjet.13287.

[23] Hemmler Y M, Rasch T, Ifenthaler D. A Classification of Facets of Adaptive Learning Systems[J]. TechTrends, 2023, 67: 98-111. DOI: 10.1007/s11528-022-00777-y.

[24] Hiim H. How Can Collaboration between Schools and Workplaces Contribute to Relevant Vocational Education? Results of an Action Research Project in the School-Based Part of Norwegian Vocational Education and Training[J]. Vocations and Learning, 2023, 16: 1-21. DOI: 10.1007/s12186-022-09300-z.

[25] Jackson D, Dean B A. The Contribution of Different Types of Work-Integrated Learning to Graduate Employability[J]. Higher Education Research & Development, 2023, 42(1): 93-110. DOI: 10.1080/07294360.2022.2048638.

[26] Jackson D, Rowe A. Impact of Work-Integrated Learning and Co-Curricular Activities on Graduate Labour Force Outcomes[J]. Studies in Higher Education, 2023, 48(3): 490-506. DOI: 10.1080/03075079.2022.2145465.

[27] Tushar H, Sooraksa N. Global Employability Skills in the 21st Century Workplace: A Semi-Systematic Literature Review[J]. Heliyon, 2023, 9(11): e21023. DOI: 10.1016/j.heliyon.2023.e21023.