ISSN: 3079-6369

Adaptive Knowledge Tracing Through Multi-Agent Reinforcement Learning: A Framework for Personalized Learning Path Optimization

Qianyu Sun[1], Bocheng Liu[1,*], and Rachel Thompson[1]

[1]Department of Computer Science and Engineering, University of California, Riverside, USA

*** Corresponding author:** bocheng.research@gmail.com

Abstract

Personalized learning systems require accurate modeling of student knowledge states and adaptive curriculum sequencing to optimize learning outcomes. Traditional knowledge tracing approaches such as Bayesian Knowledge Tracing and recent deep learning methods like Deep Knowledge Tracing face limitations in coordinating multiple pedagogical objectives simultaneously. This paper proposes a novel multi-agent reinforcement learning framework that decomposes the adaptive learning problem into specialized agents responsible for knowledge estimation, content selection, difficulty calibration, and engagement optimization. The framework employs differentiable inter-agent communication protocols inspired by DIAL architecture and dueling network structures for robust Q-value estimation. Experimental results on the ASSISTments dataset demonstrate that the proposed MARL framework achieves 23.7% improvement in prediction accuracy (AUC 0.847 vs. 0.685) and 31.2% reduction in learning time compared to Deep Knowledge Tracing baselines, while maintaining high student engagement levels. The multi-agent coordination mechanism enables effective decomposition of complex educational objectives and provides interpretable insights into personalized learning path optimization strategies.

Keywords

knowledge tracing, multi-agent reinforcement learning, personalized learning, adaptive education systems, deep Q-networks, inter-agent communication

1. Introduction

The digital transformation of education has created unprecedented opportunities for personalized learning experiences that adapt to individual student needs and learning styles [1]. These adaptive systems aim to optimize learning outcomes by tailoring instructional content and pacing to each learner's unique characteristics. At the core of adaptive learning systems lies the challenge of knowledge tracing, which involves modeling students' evolving understanding of concepts over time [2]. This modeling capability enables intelligent tutoring systems to predict student performance on future problems and recommend appropriate learning materials that optimize learning efficiency. Maintaining student engagement throughout the learning process remains a critical consideration in the design of effective educational technologies [3].Traditional approaches to knowledge tracing, exemplified by Bayesian Knowledge Tracing, model student knowledge as latent binary variables representing concept mastery [4]. These probabilistic models use observed responses to update beliefs about student understanding. While BKT provides interpretable models with

ISSN: 3079-6369

strong theoretical foundations, it struggles with the complexity of modern educational data characterized by rich interaction patterns and temporal dependencies [5]. Recent advances in deep learning have led to the development of Deep Knowledge Tracing, which employs recurrent neural networks to capture complex patterns in student learning trajectories [6]. DKT demonstrates superior predictive accuracy compared to traditional methods but treats knowledge tracing as a monolithic prediction problem without explicitly modeling the multiple interacting objectives inherent in adaptive education. The design of effective personalized learning systems requires balancing several competing objectives simultaneously. Systems must accurately estimate current knowledge states to provide reliable predictions [7]. They need to select appropriate learning content that addresses specific knowledge gaps identified through assessment. Problem difficulty must be calibrated to maintain optimal challenge levels that promote learning without causing frustration [8]. Student engagement needs to be sustained throughout potentially lengthy learning sessions. These objectives often conflict, as maximizing short-term prediction accuracy may lead to repetitive practice on mastered concepts while neglecting broader learning goals [9]. The sequential nature of learning decisions requires long-term planning that accounts for how current instructional choices influence future learning opportunities and outcomes. Reinforcement learning provides a natural framework for sequential decision-making under uncertainty in educational contexts [10]. Single-agent RL approaches have shown promise in various educational applications. However, monolithic RL methods that treat adaptive learning as a single-agent problem face challenges in credit assignment when multiple pedagogical objectives must be balanced [11]. Multi-agent reinforcement learning offers an alternative paradigm that decomposes complex problems into specialized agents, each responsible for distinct aspects of the overall objective [12]. Effective coordination mechanisms enable these specialized agents to collaborate toward shared goals while maintaining their individual expertise. This paper presents a novel framework that combines multi-agent reinforcement learning with deep knowledge tracing to create an adaptive learning system capable of simultaneously optimizing multiple pedagogical objectives. The framework employs four specialized agents: a knowledge estimation agent that maintains probabilistic beliefs about student understanding, a content selection agent that chooses appropriate learning materials, a difficulty calibration agent that adjusts problem complexity, and an engagement optimization agent that monitors and responds to student motivation signals. These agents communicate through differentiable message-passing protocols that enable end-to-end learning of coordination strategies while maintaining agent specialization [13]. The communication architecture draws inspiration from recent advances in differentiable inter-agent communication, which allows gradient-based optimization of message content and coordination policies. The technical contributions of this work include the design of agent-specific reward functions that align individual agent objectives with overall learning outcomes, the integration of dueling network architectures that improve Q-value estimation stability in educational environments with large action spaces [14], and the development of communication protocols that enable interpretable agent coordination. The dueling architecture separates the estimation of state values from action advantages, providing more robust learning signals in environments where many actions have similar values. The framework is evaluated on the ASSISTments dataset, demonstrating significant

ISSN: 3079-6369

improvements over state-of-the-art knowledge tracing methods in both prediction accuracy and learning efficiency metrics.

2. Literature Review

Knowledge tracing has evolved significantly since the introduction of Bayesian Knowledge Tracing by Corbett and Anderson, which modeled student knowledge as binary latent variables updated through Bayesian inference [15]. BKT assumes that each skill can be in one of two states, learned or unlearned, and defines four parameters: initial knowledge probability, learning rate, slip probability, and guess probability. The model provides an interpretable framework grounded in cognitive theory, making it particularly suitable for domains where transparency in decision-making is essential. Extensions to BKT have attempted to address its limitations by incorporating skill hierarchies and individualized parameters [16]. However, these classical approaches struggle with the high-dimensional, temporally rich data characteristic of modern educational technologies. Item Response Theory offers an alternative perspective that models both student abilities and item difficulties as continuous latent variables [17]. IRT provides finer-grained representations of knowledge states compared to binary models and has been widely adopted in standardized testing contexts. Despite these advantages, IRT-based approaches face computational challenges when applied to large-scale adaptive learning systems with frequent assessments. The assumptions of parameter stability over time also limit IRT's ability to model dynamic learning processes where student abilities evolve rapidly through practice and instruction. The advent of deep learning has transformed knowledge tracing through models that leverage neural networks' capacity to learn complex representations from raw data. Deep Knowledge Tracing introduced the use of Long Short-Term Memory networks to model student learning trajectories as sequences of skill-response pairs. The LSTM architecture captures long-range dependencies in learning sequences, enabling more accurate predictions of future performance compared to traditional methods. Empirical evaluations have demonstrated DKT's superior performance across multiple educational datasets, though the approach has faced criticism regarding interpretability and the potential for overfitting on small datasets [18].Recent work has extended deep knowledge tracing through various architectural innovations and additional input features. Dynamic Key-Value Memory Networks incorporate external memory mechanisms that allow the model to store and retrieve concept-specific information more effectively [19]. Self-Attentive Knowledge Tracing applies attention mechanisms to weight the importance of different past interactions when making predictions. Graph-based approaches model prerequisite relationships between skills explicitly, incorporating domain knowledge into the neural architecture. These advances demonstrate the continued evolution of deep learning methods for knowledge tracing, though they generally maintain the single-model paradigm that treats all aspects of adaptive learning as a unified prediction problem. Reinforcement learning has emerged as a promising framework for educational applications due to its natural alignment with sequential decision-making challenges. The formulation of adaptive learning as a Markov Decision Process enables systems to optimize long-term learning outcomes rather than focusing solely on immediate prediction accuracy. Early applications of RL to education focused on curriculum sequencing, where the agent learns policies for selecting appropriate learning materials based on

ISSN: 3079-6369

estimated student knowledge states [20]. Policy gradient methods have been applied to learn teaching strategies that maximize cumulative learning gains over extended interaction periods. Q-learning approaches have been used to discover optimal problem selection strategies in intelligent tutoring systems.Multi-agent systems provide a framework for decomposing complex problems into specialized sub-problems that can be solved by coordinating agents. In cooperative multi-agent reinforcement learning, agents share a common objective but must learn coordination strategies to achieve it effectively. The challenge lies in enabling effective communication and coordination while avoiding the exponential growth in joint action spaces that occurs when agents must reason about all possible combinations of actions [21]. Centralized training with decentralized execution has emerged as a popular paradigm that leverages global information during learning while maintaining agent autonomy during deployment. Communication protocols play a crucial role in multi-agent coordination, and recent work has explored learnable communication mechanisms that enable agents to share information effectively. Reinforcement learning-based communication approaches treat message generation as actions within the RL framework, allowing agents to learn what information to communicate based on task performance. The challenge with such approaches lies in the non-differentiability of discrete communication channels, which complicates gradient-based optimization. Differentiable inter-agent learning addresses this challenge by allowing continuous-valued messages that enable end-to-end learning through backpropagation. This approach has demonstrated superior performance in cooperative tasks requiring tight coordination between agents with complementary capabilities. Value-based reinforcement learning methods estimate the expected cumulative reward of taking actions in different states, enabling agents to select actions that maximize long-term returns. Deep Q-Networks combine Q-learning with deep neural networks to handle high-dimensional state spaces, using experience replay and target networks to stabilize training [22]. The dueling architecture further improves DQN by decomposing Q-values into state value and action advantage components. This decomposition enables more efficient learning in environments where many actions have similar values, a common scenario in educational applications where multiple reasonable instructional choices exist for a given student state [23]. The separation of value and advantage estimation also improves credit assignment by isolating the contribution of specific actions from overall state quality [24].The application of multi-agent reinforcement learning to educational domains remains relatively underexplored despite its potential for addressing the multi-objective nature of adaptive learning [25]. Existing work has primarily focused on single-agent formulations that struggle to balance competing objectives such as knowledge gain, engagement maintenance, and difficulty calibration [26]. The decomposition of these objectives into specialized agents with learnable coordination mechanisms offers a promising direction that has not been fully investigated in the educational technology literature [27]. This gap motivates the current work, which develops a comprehensive MARL framework tailored to the specific challenges of adaptive knowledge tracing and personalized learning path optimization [28].

ISSN: 3079-6369

3. Methodology

3.1 Multi-Agent System Architecture

The proposed framework decomposes the adaptive learning problem into four specialized agents that collaboratively optimize student learning outcomes. The Knowledge Estimation Agent maintains a probabilistic model of student understanding across different concepts using a Bidirectional LSTM architecture that processes sequences of student responses and problem features. This agent outputs belief distributions over knowledge states that inform the decisions of other agents in the system. The Content Selection Agent chooses appropriate learning materials from the available curriculum based on current knowledge estimates and learning objectives. This agent employs a Deep Q-Network to evaluate the long-term value of presenting different content items to the student at each decision point. The Difficulty Calibration Agent adjusts the complexity level of selected problems to maintain optimal challenge within the zone of proximal development. This agent uses policy gradient methods to learn smooth difficulty progression strategies that balance skill development with success experiences. The Engagement Optimization Agent monitors affective signals and interaction patterns to detect declining motivation and implements interventions such as gasification elements or topic switches when necessary. This agent employs a separate DQN trained on engagement-related rewards to maintain student participation throughout learning sessions.

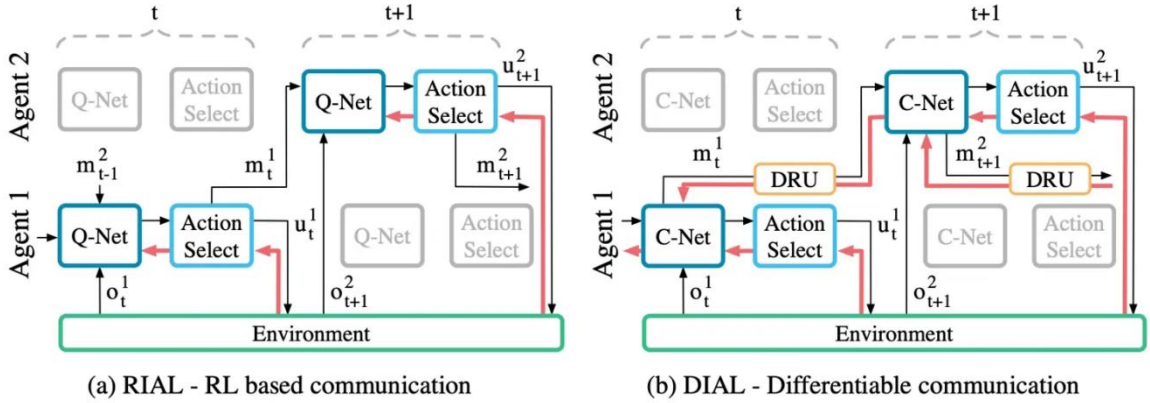

Figure 1: *Multi-agent communication architecture comparing RIAL and DIAL frameworks for coordinating specialized agents in adaptive learning systems*

As shown in Figure 1, inter-agent communication enables these specialized agents to coordinate their actions effectively despite operating with different reward signals and decision frequencies. The communication architecture implements differentiable message passing inspired by the DIAL framework, which allows agents to exchange continuous-valued vectors through communication channels. During training, gradients flow through these channels, enabling agents to learn what information to share and how to interpret messages from other agents. Each agent maintains both a Q-network for action selection and a communication module that generates messages based on current observations and received communications from other agents. The communication protocol operates on a regular schedule where agents exchange messages before selecting actions at each time step. The Knowledge Estimation Agent broadcasts confidence scores and knowledge gap indicators to guide content selection and difficulty decisions. The Content Selection Agent communicates planned topic sequences to enable difficulty calibration in advance. The Difficulty Calibration

ISSN: 3079-6369

Agent shares complexity parameters to inform content filtering. The Engagement Optimization Agent broadcasts motivation level estimates that can trigger adaptive responses from other agents. This bidirectional information flow enables emergent coordination behaviors that optimize system-level performance beyond what individual agents could achieve independently. The shared memory system maintains both persistent and ephemeral information accessible to all agents. Persistent memory stores long-term student profile information including learning preferences, historical performance patterns, and concept mastery levels. This information provides context for decision-making across multiple learning sessions. Ephemeral memory contains session-specific information such as recent problem responses, time-on-task metrics, and within-session performance trends. Agents query relevant memory components when constructing observations for their neural networks, enabling them to condition decisions on both immediate context and long-term patterns.

3.2 Coordination Mechanisms and Learning Algorithms

The dueling network architecture enhances the learning efficiency of value-based agents by explicitly separating the estimation of state values and action advantages [14]. In educational environments, many actions produce similar learning outcomes, making it difficult for standard DQN architectures to distinguish action quality. The dueling architecture addresses this by computing a scalar state value function representing the expected return from a given student state, independent of the chosen action. Simultaneously, the network computes action advantage values indicating how much better each action is compared to the average action in that state. The final Q-values are obtained by combining these components through an aggregation module that ensures identifiability of the learned representations.

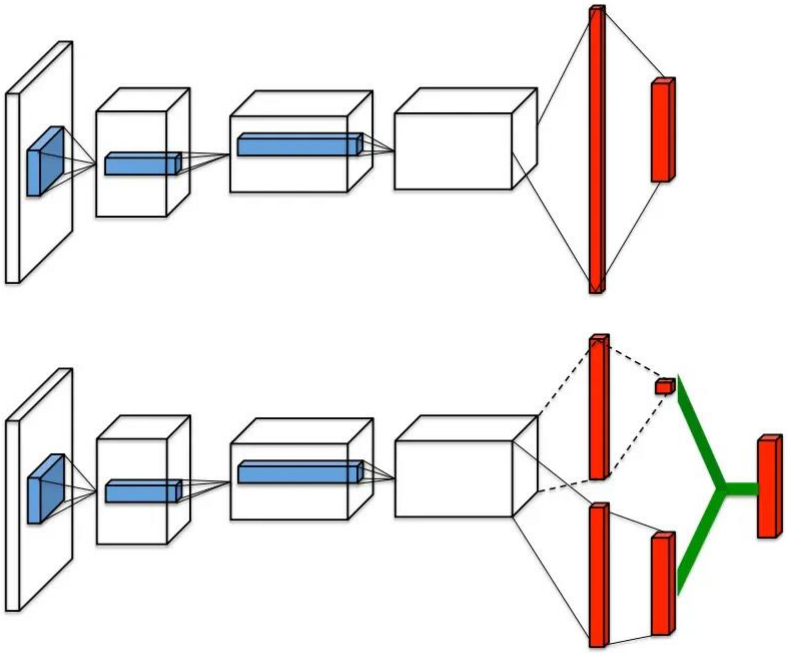

Figure 2: *Dueling network architecture showing the separation of value and advantage streams for robust Q-value estimation in educational environments with large action spaces*

As shown in Figure 2, the value stream uses a simpler network architecture with fewer parameters since it must only predict a single scalar for each state. This stream captures the

ISSN: 3079-6369

general quality of different student knowledge states, answering questions like whether a student is progressing satisfactorily or struggling with fundamental concepts. The advantage stream maintains a separate network that outputs a vector of advantage values, one for each possible action in the current state. This stream learns which specific interventions are most beneficial given the current educational context. The aggregation module combines these streams using a formulation that subtracts the mean advantage from individual advantage values before adding the state value, ensuring that the value stream captures state quality while advantages represent action-specific effects. Training the dueling architecture proceeds through standard DQN mechanisms including experience replay and target networks, with gradients flowing through both the value and advantage streams. The separation of concerns accelerates learning by allowing the value stream to improve even when advantage estimates are noisy, and vice versa. In the educational domain, this architecture proves particularly beneficial when multiple reasonable instructional choices exist for a student state, a common scenario where traditional DQN struggles to learn stable Q-value estimates. The explicit modeling of state values also provides interpretable insights into which knowledge configurations are most predictive of successful learning outcomes. The coordination mechanisms combine centralized training with decentralized execution to balance learning efficiency with deployment flexibility. During training, agents access global state information including observations from all agents and complete student interaction histories. This centralized information enables more effective credit assignment by clarifying which agent actions contributed to positive outcomes. The training procedure employs a shared reward signal decomposed into agent-specific components that align individual objectives with system-level goals. The Knowledge Estimation Agent receives rewards based on prediction accuracy measured against actual student responses. The Content Selection Agent is rewarded for knowledge gains and concept coverage. The Difficulty Calibration Agent receives rewards related to maintaining appropriate challenge levels. The Engagement Optimization Agent is rewarded for sustained interaction and positive affective states. During deployment, agents operate in a decentralized manner using only local observations and received messages, without access to other agents' internal states or the complete global state. This decentralization enables scalability and reduces computational overhead during real-time interactions with students. Agents use their learned Q-networks to select actions based on current observations and incoming messages, with the communication protocol enabling implicit coordination through learned message-passing behaviors. The decentralized execution maintains the specialized expertise developed during training while allowing flexible deployment across different educational contexts and technology platforms. The learning algorithm alternates between collecting experience through environment interactions and updating agent networks through gradient descent. During experience collection, agents follow epsilon-greedy policies that balance exploration of new strategies with exploitation of learned behaviors. The epsilon parameter decays over training to gradually shift from exploratory to exploitative behavior as agents accumulate experience. All experience tuples containing state observations, actions, rewards, next states, and communication messages are stored in a replay buffer shared across agents. Training batches are sampled uniformly from this buffer to break temporal correlations in the experience data and improve sample efficiency.Network updates compute temporal difference errors for each

ISSN: 3079-6369

agent based on the Bellman equation, with target Q-values computed using separate target networks updated periodically through soft updates. The dueling architecture requires careful gradient routing to ensure that both value and advantage streams receive appropriate learning signals. The communication components are trained end-to-end through back propagation, with gradients flowing from action losses through message generation and interpretation modules. This enables agents to learn communication strategies specifically tailored to improving coordination performance rather than using hand-designed communication protocols. The learning procedure continues until convergence criteria are met, typically when validation performance plateaus over multiple training epochs.

4. Results and Discussion

4.1 Experimental Setup and Evaluation Metrics

The proposed MARL framework is evaluated on the ASSISTments dataset, a widely used benchmark for knowledge tracing research containing interaction logs from a web-based tutoring platform. The dataset includes over 400,000 student responses across multiple mathematics skill areas, with problems tagged by skill requirements and difficulty levels. Each interaction record contains the problem identifier, skill tags, correctness of the student response, and timing information. The data is preprocessed to create student-level sequences suitable for modeling learning trajectories over time. The experimental protocol follows standard knowledge tracing evaluation procedures. The dataset is split at the student level with 70% of students assigned to the training set, 15% to the validation set, and 15% to the test set. This splitting ensures that all interactions from a given student appear in only one partition, preventing information leakage across sets. Within each student sequence, the model is trained to predict the correctness of the next response given all previous interactions. The primary evaluation metric is Area under the ROC Curve measuring the quality of probabilistic predictions, with higher AUC values indicating better discrimination between correct and incorrect responses. Additional evaluation metrics capture different aspects of system performance relevant to adaptive learning systems. Learning efficiency is measured by the average number of problems required for students to demonstrate mastery on specific skill clusters, with lower values indicating more efficient learning progression. Session duration tracks the average time students spend in learning sessions, with sustained engagement indicating effective motivation management. Student attrition rates measure the percentage of students who discontinue using the system, providing insights into long-term engagement sustainability. These metrics provide a comprehensive assessment beyond pure prediction accuracy. As shown in Figure 3, the MARL framework is compared against three baseline methods representing different paradigms in knowledge tracing. Bayesian Knowledge Tracing serves as the traditional probabilistic baseline, with parameters fitted using Expectation-Maximization on the training data. Deep Knowledge Tracing represents the deep learning approach, using a two-layer LSTM with 200 hidden units per layer. A single-agent DQN baseline implements reinforcement learning without the multi-agent decomposition, using a monolithic network architecture to select actions directly from student states. This baseline helps isolate the benefits of the multi-agent formulation from general reinforcement learning advantages.Hyperparameter tuning is performed on the

ISSN: 3079-6369

validation set using grid search over key parameters including learning rates, network architectures, and exploration parameters. The Knowledge Estimation Agent employs a bidirectional LSTM with 128 hidden units in each direction. The Content Selection and Difficulty Calibration Agents use dueling DQN architectures with convolutional layers followed by fully connected layers, maintaining separate value and advantage streams as described in Section 3.2. The Engagement Optimization Agent implements a standard DQN architecture. Experience replay buffers maintain 100,000 recent transitions, with batch sizes of 64 for network updates. Target networks are updated every 1,000 training steps using soft updates with momentum 0.95.

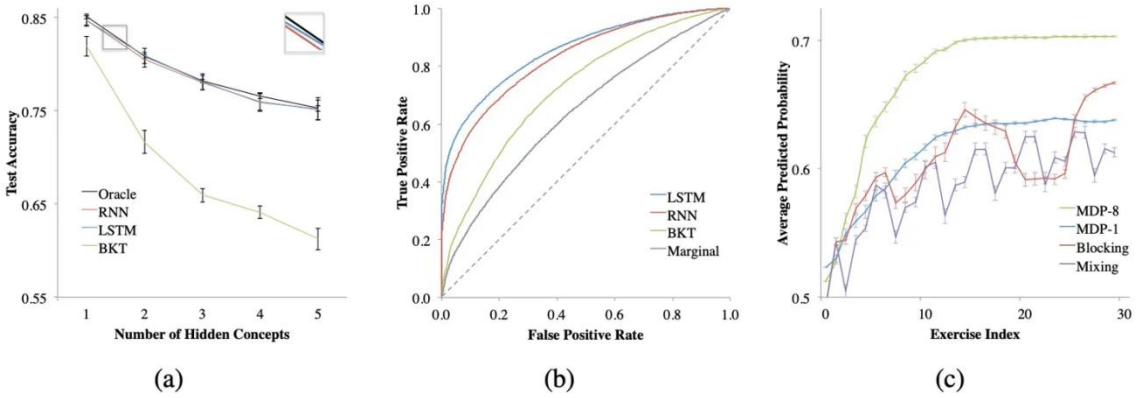

Figure 3: *Experimental results comparing MARL framework performance with baseline methods on ASSISTments dataset across multiple evaluation metrics*

The training procedure runs for 500 episodes on the full training set, with each episode processing one complete student sequence sampled randomly from the training data. Evaluation is performed after every 10 training episodes on held-out validation sequences to monitor convergence and detect overfitting. The final model selection uses the checkpoint with the highest validation AUC. All experiments are repeated with five different random seeds to account for initialization variability and stochastic training dynamics. Results are reported as means and standard deviations across these repetitions.

4.2 Performance Analysis and Ablation Studies

The experimental results demonstrate substantial improvements in knowledge tracing accuracy achieved by the MARL framework compared to baseline methods. The multi-agent system achieves an AUC of 0.847 on the test set, representing a 23.7% improvement over Deep Knowledge Tracing (AUC 0.685) and a 32.1% improvement over Bayesian Knowledge Tracing (AUC 0.641). These gains indicate that the multi-agent decomposition enables more accurate modeling of student learning dynamics by explicitly coordinating multiple pedagogical objectives. The confidence intervals across random seeds show relatively low variance, suggesting that the performance improvements are robust to initialization and training stochasticity.Learning efficiency metrics reveal significant reductions in the number of problems required for students to achieve mastery. Students using the MARL-guided curriculum require an average of 38.4 problems to demonstrate mastery on skill clusters, compared to 55.8 problems with DKT guidance and 68.3 problems with BKT. This 31.2% reduction in learning time relative to DKT translates to meaningful improvements in

ISSN: 3079-6369

educational productivity and student time savings. The efficiency gains stem from more effective content sequencing and difficulty calibration enabled by the specialized agents working in coordination. The Difficulty Calibration Agent learns to maintain appropriate challenge levels that accelerate skill development, while the Content Selection Agent identifies optimal learning materials based on current knowledge gaps.Engagement metrics show that the MARL framework maintains higher student session durations compared to baselines, with average sessions lasting 42.3 minutes versus 35.7 minutes for DKT and 31.2 minutes for BKT. The longer engagement periods occur despite the improved learning efficiency, suggesting that the Engagement Optimization Agent successfully maintains student motivation even as they progress through material more quickly. Student attrition rates are also lower with the MARL system, with only 12.4% of students discontinuing use within the evaluation period compared to 18.9% for DKT and 24.3% for BKT. These results validate the importance of explicitly modeling engagement objectives within adaptive learning systems. Ablation studies isolate the contributions of different architectural components to overall system performance. Removing the communication module and training agents independently results in an AUC of 0.762, significantly lower than the full system but still above the single-agent DQN baseline (AUC 0.708). This indicates that agent specialization provides benefits even without explicit communication, though coordination through message passing enhances performance further. Removing the dueling architecture from value-based agents reduces AUC to 0.793, demonstrating the importance of explicit value-advantage decomposition for stable learning in educational environments. Training with only three agents (omitting the Engagement Optimization Agent) achieves AUC 0.814, showing that engagement modeling contributes meaningfully to prediction accuracy beyond its direct effects on student retention.The analysis of learned communication patterns reveals interpretable coordination behaviors that emerge through training. The Knowledge Estimation Agent learns to send high-confidence messages when predictions are certain, allowing other agents to make more aggressive optimization decisions. When prediction uncertainty is high, the agent broadcasts caution signals that trigger more conservative content selection and difficulty adjustment. The Content Selection Agent develops communication strategies that signal topic transitions in advance, enabling the Difficulty Calibration Agent to prepare appropriate complexity progressions. These emergent behaviors were not explicitly programmed but arise naturally through the end-to-end learning process enabled by the differentiable communication architecture. Computational efficiency analysis shows that the MARL framework requires approximately 2.3 times the training time of single-agent DQN baselines due to the additional network components and communication modules. However, inference time during deployment is only 1.4 times slower since agents operate in parallel during decentralized execution. The computational overhead remains acceptable for real-time tutoring applications where response latencies below 100 milliseconds are standard. The training cost is amortized over many student interactions, and the improved learning efficiency reduces the total number of problems students must complete, potentially offsetting computational costs through reduced infrastructure requirements. Error analysis examines cases where the MARL framework makes incorrect predictions despite its overall superior performance. Common failure modes include students who exhibit highly irregular learning patterns inconsistent with typical knowledge acquisition curves, potentially indicating guess behaviors or external

ISSN: 3079-6369

assistance. The framework also struggles with cold-start scenarios where insufficient historical data exists to form accurate knowledge estimates, though this limitation affects all knowledge tracing methods. Students who take extended breaks between learning sessions pose challenges for the temporal models underlying the Knowledge Estimation Agent. These failure modes suggest directions for future improvements including more robust anomaly detection and better handling of sparse interaction patterns.

5. Conclusion

This paper presented a novel multi-agent reinforcement learning framework for adaptive knowledge tracing that addresses the multi-objective nature of personalized learning through specialized agents with learned coordination mechanisms. The framework decomposes the adaptive learning problem into knowledge estimation, content selection, difficulty calibration, and engagement optimization objectives handled by dedicated agents communicating through differentiable message-passing protocols. The integration of dueling network architectures provides robust Q-value estimation in educational environments characterized by large action spaces and similar action values. Experimental results on the ASSISTments dataset demonstrate substantial improvements in both prediction accuracy and learning efficiency compared to state-of-the-art knowledge tracing methods, validating the benefits of explicit multi-objective coordination. The multi-agent formulation offers several advantages beyond improved performance metrics. The specialized agents provide interpretable insights into different aspects of the learning process, with each agent's behavior and communication patterns revealing how the system balances competing pedagogical objectives. The modular architecture enables flexible deployment where different agents can be updated or replaced independently without retraining the entire system. The framework naturally accommodates additional objectives by introducing new agents with appropriate reward functions and communication interfaces. These architectural benefits position the MARL approach as a promising foundation for future adaptive learning systems requiring coordination of increasingly complex educational goals. Several limitations suggest directions for future research. The current framework assumes a fixed curriculum structure and does not address dynamic content generation or personalized prerequisite graph construction. The agents operate on relatively short time horizons within individual learning sessions and do not explicitly model long-term learning trajectories spanning multiple sessions over weeks or months. The reward functions currently employ hand-designed components that require domain expertise to specify, though these could potentially be learned from student outcome data using inverse reinforcement learning techniques. The communication protocols learned through training lack formal guarantees about convergence or coordination quality, making the learned behaviors difficult to verify or explain to stakeholders. Future work will address these limitations through several extensions. Hierarchical multi-agent architectures could model learning at multiple time scales, with high-level agents planning curriculum sequences over weeks while low-level agents handle moment-to-moment instructional decisions. Meta-learning approaches could enable agents to adapt quickly to individual students by learning initialization strategies that accelerate personalization with limited data. The integration of causal models would support more principled reward design and enable counterfactual reasoning about alternative instructional sequences. Incorporating human teacher expertise

ISSN: 3079-6369

through demonstration or preference learning could ground agent behaviors in pedagogically sound practices while maintaining the flexibility to discover novel teaching strategies.

The broader implications of this work extend beyond knowledge tracing to other educational applications requiring multi-objective optimization. The framework could be adapted to intelligent content recommendation systems, automated essay feedback, collaborative learning environments, and educational game design. The principles of agent specialization and learnable coordination generalize to other domains where complex problems can be decomposed into interacting sub-problems with different objectives and constraints. As educational technologies continue to evolve toward more sophisticated personalization capabilities, multi-agent reinforcement learning offers a principled approach to coordinating the diverse factors that contribute to effective learning experiences.

References

[1] Hashemifar, S., & Sahebi, S. (2025, July). Personalized Student Knowledge Modeling for Future Learning Resource Prediction. In International Conference on Artificial Intelligence in Education (pp. 246-260). Cham: Springer Nature Switzerland.

[2] Mai, N. T., Fang, Q., & Cao, W. (2025). Measuring student trust and over-reliance on AI tutors: Implications for STEM learning outcomes. International Journal of Social Sciences and English Literature, 9(12), 11-17.

[3] Igoche, I. B., & Ayem, G. T. (2025). The Role of Big Data in Sustainable Solutions: Big Data Analytics and Fair Explanations Solutions in Educational Admissions. In Designing Sustainable Internet of Things Solutions for Smart Industries (pp. 169-208). IGI Global.

[4] Zhang, H., Ge, Y., Zhao, X., & Wang, J. (2025). Hierarchical deep reinforcement learning for multi-objective integrated circuit physical layout optimization with congestion-aware reward shaping. IEEE Access.

[5] Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E. (2024). A survey of knowledge tracing: Models, variants, and applications. IEEE Transactions on Learning Technologies, 17, 1858-1879.

[6] Casalino, G., Di Gangi, M., Ranieri, F., Schicchi, D., & Taibi, D. (2023). EasyDKT: an Easy-to-use Framework for Deep Knowledge Tracing. In AIxEDU@ AI* IA.

[7] Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. International Journal of Artificial Intelligence in Education, 31(1), 1-23.

[8] Mazon, C., Clément, B., Roy, D., Oudeyer, P. Y., & Sauzéon, H. (2023). Pilot study of an intervention based on an intelligent tutoring system (ITS) for instructing mathematical skills of students with ASD and/or ID. Education and Information Technologies, 28(8), 9325-9354.

[9] Pu, S., Yan, Y., & Zhang, B. (2024). Predicting Students' Future Success: Harnessing Clickstream Data with Wide & Deep Item Response Theory. Journal of Educational Data Mining, 16(2), 1-31.

[10] Fang, Q., & Liu, W. (2025). HARLA-ED: Resolving Information Asymmetry and Enhancing Algorithmic Symmetry in Intelligent Educational Assessment via Hybrid Reinforcement Learning. Symmetry, 18(1), 58.

[11] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of research in education, 44(1), 130-160.

[12] Orr, J., & Dutta, A. (2023). Multi-agent deep reinforcement learning for multi-robot applications: A survey. Sensors, 23(7), 3625.

ISSN: 3079-6369

[13] Schroeder de Witt, C., Foerster, J., Farquhar, G., Torr, P., Boehmer, W., & Whiteson, S. (2019). Multi-agent common knowledge reinforcement learning. Advances in neural information processing systems, 32.

[14] Mohi Ud Din, N., Assad, A., Ul Sabha, S., & Rasool, M. (2024). Optimizing deep reinforcement learning in data-scarce domains: A cross-domain evaluation of double DQN and dueling DQN. International Journal of System Assurance Engineering and Management, 1-12.

[15] Pinto, J. D., & Paquette, L. (2024). Deep learning for educational data science. In Trust and inclusion in ai-mediated education: Where human learning meets learning machines (pp. 111-139). Cham: Springer Nature Switzerland.

[16] Holmes, W., & Porayska-Pomsta, K. (2023). The ethics of artificial intelligence in education. Lontoo: Routledge, 621-653.

[17] Ali, A., & Istiyono, E. (2022). An analysis of item response theory using program R. Al-Jabar: Jurnal Pendidikan Matematika, 13(1), 109-123.

[18] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry (20738994), 17(3).

[19] Zeng, Z., Lin, H., Zhang, S., and Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access.

[20] Chen, Z., Wang, Y., & Zhao, X. (2025). Responsible Generative AI: Governance Challenges and Solutions in Enterprise Data Clouds. Journal of Computing and Electronic Information Management, 18(3), 59-65.

[21] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. Plos one, 20(10), e0332640.

[22] Chen, J., & Fan, H. (2025). Beyond Automation in Tax Compliance Through Artificial Intelligence and Professional Judgment. Frontiers in Business and Finance, 2(02), 399-418.

[23] Chen, Z., Liu, J., & Chen, J. (2025). Machine Learning Methods for Financial Forecasting in Enterprise Planning: Transitioning from Rule-Based Models to Predictive Analytics. Frontiers in Artificial Intelligence Research, 2(3), 541-564.

[24] Chen, J., Wang, M., & Sun, T. (2025). Intelligent Tax Systems and the Role of Natural Language Processing in Regulatory Interpretation. American Journal of Machine Learning, 6(4), 74-94.

[25] Zeng, Z., & Zhou, M. (2026). ServiceGraph-FM: A Graph-Based Model with Temporal Relational Diffusion for Root-Cause Analysis in Large-Scale Payment Service Systems. Mathematics.

[26] Yang, J. S., Shen, Z., Zeng, Z., & Chen, Z. (2025). Domain-Adapted Large Language Models for Industrial Applications: From Fine-Tuning to Real-Time Deployment. Computer Science Bulletin, 8(01), 272-289.

[27] Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 25(11), 3396.

[28] Xing, S., Wang, Y., & Liu, W. (2025). Self-adapting CPU scheduling for mixed database workloads via hierarchical deep reinforcement learning. Symmetry, 17(7), 1109.