Multi-Agent Reinforcement Models for Context-Aware Adaptive Learning Systems

Mishael Winarta[1,*], Manwir Singh[2]

[1,2]Department of Information Systems, Faculty of AI and Data Science, Universitas Pelita Harapan, Indonesia

ABSTRACT

The increasing heterogeneity of learner behavior in digital education environments poses significant challenges to conventional adaptive learning systems, which often rely on static rules or single-objective optimization. This study proposes a context-aware adaptive learning framework based on coordinated Multi-Agent Reinforcement Learning (MARL) to address these limitations. The system decomposes pedagogical decision-making into multiple specialized agents responsible for content difficulty adaptation, feedback timing, and instructional modality, while a centralized coordination mechanism ensures policy stability during training. Learner context is explicitly modeled as a dynamic state representation derived from interaction logs, engagement indicators, and performance signals, enabling continuous and fine-grained adaptation. Comprehensive experiments were conducted to evaluate learning effectiveness, policy stability, context responsiveness, and learner engagement across adaptive and baseline configurations. The results show that the proposed multi-agent system achieves substantially higher cumulative mastery gains, with improvements of up to 64% for learners with low prior knowledge compared to non-adaptive baselines. Policy variance analysis demonstrates a reduction of more than 60% in coordinated agent configurations relative to uncoordinated multi-agent setups, confirming the effectiveness of the coordination strategy in mitigating non-stationarity. Context sensitivity evaluation indicates that the system detects and adjusts to contextual shifts within an average of 4–5 interaction steps, significantly outperforming rule-based adaptive approaches. In addition to cognitive outcomes, the system exhibits strong behavioral benefits. Engagement analysis reveals lower engagement variance (0.006) and reduced session drop-off rates (9%), indicating improved learner consistency over time. A holistic effectiveness comparison further confirms that the proposed approach delivers balanced gains across mastery progression, adaptation flexibility, engagement stability, and robustness, albeit with higher implementation complexity. These findings demonstrate that coordinated MARL provides a viable and scalable foundation for next-generation adaptive learning systems capable of robust, real-time personalization in dynamic educational contexts.

Keywords Adaptive Learning, Context-Aware Systems, Multi-Agent Reinforcement Learning, Learning Analytics, Personalized Education, Intelligent Tutoring Systems

Submitted: 15 January 2025
Accepted: 20 February 2025
Published: 1 August 2025

*Corresponding author
Mishael Winarta,
1081230014@student.uph.edu

Additional Information and Declarations can be found on page 210

Introduction

The rapid expansion of digital learning environments has intensified the demand for adaptive learning systems capable of personalizing instruction according to individual learner needs, behaviors, and contextual conditions. Conventional learning management systems predominantly rely on static content sequencing or rule-based personalization, which often fails to accommodate the dynamic and nonlinear nature of learner behavior in real-world settings [1], [2]. As a result, learners frequently experience misalignment between instructional difficulty, feedback timing, and their evolving cognitive states, leading to

suboptimal learning outcomes and reduced engagement [3], [4].

Recent advances in learning analytics and artificial intelligence have enabled more data-driven approaches to personalization, particularly through supervised machine learning and predictive modeling [5], [6]. However, these approaches typically depend on pre-labeled data and offline training, limiting their capacity to adapt continuously as learner behavior changes. Moreover, single-agent adaptive models tend to optimize a narrow objective, such as performance prediction or content recommendation, without explicitly accounting for the multifaceted pedagogical decisions involved in effective instruction [7].

Reinforcement learning (RL) has emerged as a promising paradigm for adaptive learning due to its ability to model sequential decision-making and long-term reward optimization [8]. In educational contexts, RL enables systems to learn optimal instructional strategies through interaction with learners rather than relying solely on predefined rules or static models. Nevertheless, most existing RL-based adaptive learning systems adopt a single-agent formulation, which struggles to scale when multiple pedagogical dimensions such as content difficulty, feedback timing, and instructional modality must be optimized simultaneously [9].

To address this limitation, MARL offers a principled framework for decomposing complex decision-making problems into coordinated sub-tasks handled by specialized agents [10], [11]. In theory, MARL enables parallel adaptation across multiple instructional dimensions while preserving long-term optimization objectives. In practice, however, MARL introduces significant challenges related to policy instability, non-stationary environments, and coordination overhead, particularly in learner-centered systems where behavior is inherently stochastic [12].

Another critical challenge in adaptive learning research lies in achieving genuine context awareness. Many adaptive systems incorporate contextual features at a superficial level, such as demographic variables or aggregate performance metrics, without modeling how context evolves over time and influences pedagogical effectiveness [13], [14]. The lack of explicit context-sensitive policy learning results in delayed or inappropriate adaptations, undermining the potential benefits of intelligent personalization [15].

Motivated by these gaps, this study proposes a Multi-Agent Reinforcement Model for Context-Aware Adaptive Learning Systems that integrates structured context modeling, cooperative agent learning, and coordinated policy optimization. The primary objective of this research is to design and evaluate a scalable adaptive learning framework capable of responding dynamically to contextual changes while maintaining policy stability and pedagogical coherence. The system explicitly models learner context as a temporal state representation and assigns specialized agents to distinct instructional control dimensions, enabling fine-grained and responsive adaptation [16], [17].

The novelty of this work lies in three key contributions. First, it introduces a coordinated multi-agent reinforcement architecture tailored to adaptive learning, addressing instability issues commonly observed in MARL environments. Second, it operationalizes context sensitivity as a measurable system property, enabling systematic evaluation of adaptation responsiveness. Third, it provides an empirical analysis that jointly considers learning effectiveness, policy

stability, engagement consistency, and practical deployment trade-offs. Collectively, these contributions advance the state of the art in adaptive learning by demonstrating how multi-agent reinforcement models can support robust, context-aware personalization in complex educational ecosystems [18], [19].

Literature Review

Adaptive learning research has evolved from early rule-based personalization toward data-driven and intelligent systems that leverage learner interaction data to optimize instructional decisions. Foundational work in adaptive educational hypermedia established the importance of modeling learner characteristics and dynamically adjusting content presentation, yet these systems largely relied on predefined adaptation rules that limited scalability and responsiveness to behavioral change [20]. Subsequent developments in learning analytics emphasized the use of large-scale educational data to infer learner states and predict outcomes, but many analytics-driven systems remained descriptive or predictive rather than prescriptive in nature [21].

The introduction of RL into educational systems marked a significant shift by framing instruction as a sequential decision-making problem. RL-based tutors and recommendation systems demonstrated the ability to optimize long-term learning outcomes by balancing exploration and exploitation during instructional interactions [22]. Despite these advantages, most RL applications in education adopted single-agent formulations, which constrained their ability to manage multiple pedagogical control dimensions simultaneously. As learning environments became more complex, this limitation increasingly hindered the practical deployment of RL-driven adaptive systems.

In parallel, research on context-aware learning systems highlighted the necessity of incorporating situational and behavioral context, such as temporal patterns, device constraints, and engagement dynamics, into adaptation mechanisms. Studies showed that context-aware personalization improves learner satisfaction and effectiveness compared to context-agnostic approaches; however, context was often treated as static metadata rather than a temporally evolving state influencing instructional policy [23]. This gap resulted in delayed or coarse-grained adaptations that failed to fully exploit the richness of contextual signals.

To overcome the complexity of multi-dimensional adaptation, MARL has been proposed as a promising paradigm. MARL enables the decomposition of complex learning environments into interacting agents, each responsible for a specific aspect of decision-making. Surveys in MARL research have demonstrated its effectiveness in domains requiring coordination under uncertainty, but they also emphasize challenges related to non-stationarity and policy instability [24]. When transferred to educational settings, these challenges are exacerbated by stochastic learner behavior and sparse feedback signals.

Recent studies have begun to explore MARL for personalized education, reporting improvements in flexibility and scalability compared to single-agent systems. Nevertheless, many of these approaches focus primarily on algorithmic performance and provide limited analysis of pedagogical coherence, engagement stability, or practical deployment considerations [25]. Moreover, explicit mechanisms for coordinating agents and stabilizing learning in context-rich environments remain underexplored.

Overall, the literature indicates a clear progression toward more intelligent and adaptive learning systems, yet also reveals persistent gaps. Existing approaches often lack integrated context modeling, robust multi-agent coordination, or comprehensive evaluation across cognitive and behavioral dimensions. These limitations motivate the present study, which situates coordinated multi-agent reinforcement learning as a unifying framework for achieving stable, context-aware, and pedagogically meaningful adaptation in modern digital learning environments.

Methodology

System Architecture of the Multi-Agent Adaptive Learning Environment

The proposed adaptive learning system is designed as a distributed multi-agent architecture, where autonomous learning agents collaboratively optimize personalized learning trajectories. Each agent represents an intelligent decision-making entity responsible for observing learner context, selecting pedagogical actions, and updating its policy through reinforcement signals. The architecture integrates context sensing modules, agent communication layers, and a centralized coordination mechanism to ensure global learning stability while preserving agent autonomy.

From an architectural perspective, the system adopts a hybrid centralized–decentralized control model. Local agents operate independently at the learner level, while a global coordinator aggregates policy feedback to mitigate non-stationarity inherent in multi-agent reinforcement learning environments. This design choice directly addresses scalability and coordination challenges commonly encountered in adaptive learning platforms operating with heterogeneous learner profiles.

Formally, the environment is modeled as a Multi-Agent Markov Decision Process (MMDP), defined as:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}_i, \ldots, \mathcal{A}_N, \mathcal{P}, \mathcal{R}, \gamma \rangle \quad (1)$$

where S denotes the global state space capturing learner context, A_i represents the action space of agent i, P is the state transition probability, R is the joint reward function, and $\gamma \in (0,1)$ is the discount factor. This formulation enables the system to explicitly model agent interdependencies while maintaining tractable policy learning.

Figure 1 operationalizes the system as a multi-layer adaptive learning pipeline where context is sensed, abstracted, and persisted into a unified state representation that drives agent decisions. The architecture makes the information flow explicit from raw interaction signals to actionable pedagogical interventions, which is essential for ensuring that the reinforcement agents consume a stable and semantically consistent state vector across heterogeneous learners.

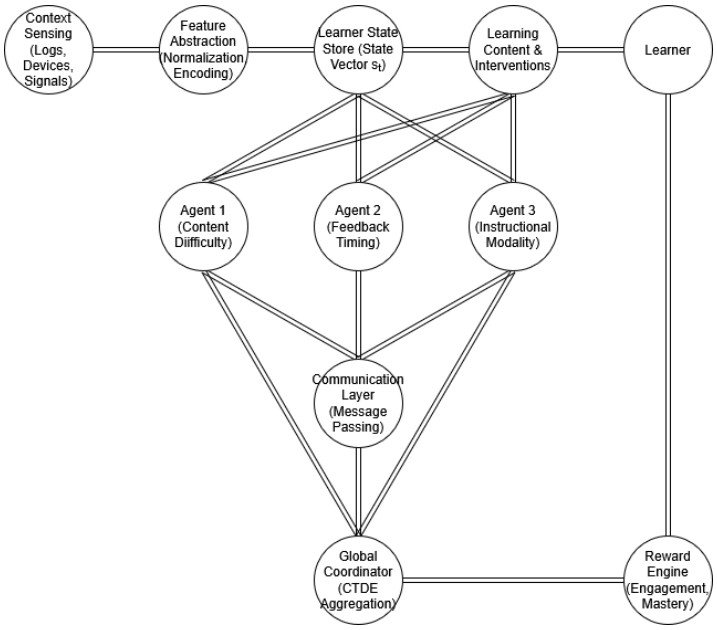

Figure 1 Multi-Agent Adaptive Learning System Architecture

The diagram also highlights the coordination pattern of Centralized Training with Decentralized Execution (CTDE) through the global coordinator, which aggregates signals and mitigates multi-agent non-stationarity. By explicitly routing reward feedback into coordination, the architecture aligns individual agent updates with global instructional objectives, preserving personalization while controlling undesirable policy divergence.

Table 1 is meant to serve as the minimal reproducibility artifact for the full system pipeline, mapping each functional block to its inputs and outputs. In adaptive learning engineering, this mapping is not merely descriptive; it defines the operational data contracts that constrain what an RL agent can observe and what it can influence, which strongly determines learnability and stability.

Table 1 Architectural Components and Agent Roles

Component	Type	Primary Function	Data In	Data Out
Context Sensing	Subsystem	Collects learner interaction signals and device/context metadata	Clickstream, time-on-task, device type, session logs	Raw contextual events
Feature Abstraction	Subsystem	Normalizes, encodes, and aggregates raw events into features	Raw contextual events	Context features c_k^t
Learner State Store	Data Layer	Stores unified state vectors for agent consumption	Context features	State vector st
Agent 1 (Difficulty)	Agent	Selects content difficulty and sequencing actions	s_t	a_t^1
Agent 2 (Feedback)	Agent	Schedules feedback timing and feedback granularity	s_t	a_t^2
Agent 3 (Modality)	Agent	Chooses instructional modality (text/video/practice/mixed)	s_t	a_t^3
Communication Layer	Subsystem	Enables message passing and shared signals among agents	Agent outputs	Messages, shared summaries
Global Coordinator (CTDE)	Controller	Aggregates experience, stabilizes training, synchronizes policies	Rewards, experiences	Policy updates, constraints
Reward Engine	Subsystem	Computes reward from mastery, engagement,	Performance + engagement	rt (scalar or vector)

The roles of the agent modules are separated by pedagogical control surfaces (difficulty, feedback, modality), which reduces action-space entanglement and clarifies credit assignment during learning. The inclusion of communication and coordination as explicit components provides an implementable blueprint for CTDE, ensuring that policy optimization can be centralized while runtime decisions remain per-learner and low-latency.

Context Modeling and State Representation

Context-awareness is achieved through a structured state representation framework that captures both static and dynamic learner attributes. Static features include demographic variables and prior knowledge indicators, while dynamic features consist of real-time interaction metrics such as response latency, task completion rates, and engagement frequency. These features collectively form the state vector used by reinforcement agents to infer learner needs.

To reduce dimensionality and mitigate sparsity issues, contextual features are encoded using a feature abstraction layer, which normalizes heterogeneous inputs into a unified representation space. This abstraction enables consistent policy learning across learners with diverse interaction patterns, while preserving semantic meaning essential for pedagogical decision-making.

Mathematically, the learner context at time t is defined as a state vector:

$$\mathbf{s}_t = [c_1^t, c_2^t, \ldots, c_k^t] \tag{2}$$

where each component c_k^t corresponds to a contextual feature. The transition of learner context is governed by:

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t, \epsilon_t) \tag{3}$$

with *f(·)* denoting the environment dynamics and ϵt representing stochastic noise due to learner behavior variability. This formulation explicitly models uncertainty, a critical factor in adaptive learning systems.

Figure 2 formalizes how heterogeneous interaction traces are transformed into a stable context-aware state suitable for reinforcement learning. The pipeline emphasizes that state construction is not a single step but a chain of operations that preserves signal semantics while controlling noise, missingness, and scale disparity.

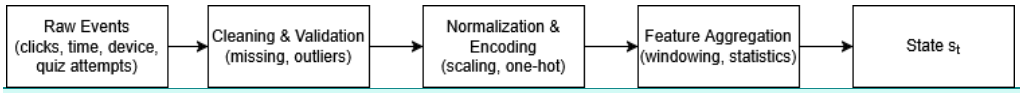

Figure 2 Context Modeling and Feature Abstraction Pipeline

In practice, this pipeline is what prevents policies from overfitting to superficial log artifacts rather than genuine learning dynamics. The placement of aggregation immediately before the state vector highlights a key methodological decision: rather than treating every raw event as a state, the system uses temporally bounded summaries that stabilize the Markovian approximation. This design improves the reliability of the transition function $s_t+1=f(s_t,a_t,\epsilon_t)$ by

reducing stochastic volatility attributable to micro-interactions.

Table 2 specifies the concrete operationalization of learner context as a feature vector st. The key methodological value is that each variable is defined not only by meaning, but also by measurement provenance, which is necessary for auditing data quality and ensuring that state variables are reproducible across deployments and cohorts.

Table 2 Learner Context Variables and Descriptions

Variable	Symbol	Type	Description	Typical Source
Prior mastery estimate	c_1^t	Continuous	Estimated knowledge level from historical performance	Quiz/exam history
Engagement rate	c_2^t	Continuous	Ratio of active interactions to session duration	Clickstream logs
Response latency	c_3^t	Continuous	Median time to respond to prompts or questions	Event timestamps
Hint usage	c_4^t	Count	Number of hints requested within a time window	Help events
Attempt count	c_5^t	Count	Number of attempts per item or activity	Assessment logs
Drop-off probability	c_6^t	Continuous	Estimated risk of session abandonment	Survival model / heuristics
Device category	c_7^t	Categorical	Mobile/desktop/tablet; proxy for UI constraints	User agent
Time-of-day	c_8^t	Cyclical	Encoded hour signal capturing circadian learning patterns	System clock

The mix of continuous, count, and categorical variables implies that the pipeline must include explicit normalization and encoding steps to avoid scale dominance in policy learning. In reinforcement learning terms, these variables collectively define the agent's observation manifold; missing or poorly defined variables can create partial observability artifacts that degrade the quality of *Q(s,a)* estimates and destabilize multi-agent coordination.

Multi-Agent Reinforcement Learning Strategy

The learning agents employ a cooperative reinforcement learning strategy, where individual policies are optimized with respect to both local and global performance objectives. Each agent selects actions such as content difficulty adjustment, feedback timing, or instructional modality, based on the observed learner context. Coordination among agents is achieved through shared reward signals and periodic policy synchronization.

The optimization objective for each agent *i* is to maximize its expected cumulative reward:

$$J_i(\pi_i) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r_i^t] \quad (4)$$

where πi is the policy of agent i, and r_i^t is the reward reflecting learner performance improvement and engagement quality. The reward function is carefully designed to balance short-term performance gains with long-term learning outcomes.

Policy updates are performed using value-based reinforcement learning, where the action-value function is defined as:

$$Q_i(s,a) = \mathbb{E}[r_i^t + \gamma \max_{a'} Q_i(s',a') \mid s,a] \quad (5)$$

This formulation allows agents to iteratively approximate optimal policies while adapting to evolving learner contexts. The shared reward structure further encourages cooperative behavior, reducing policy divergence across agents.

Figure 3 provides an interpretable visualization of how local agent reward signals evolve alongside an aggregated global reward, with periodic vertical markers denoting synchronization points that approximate CTDE-style coordination. The methodological point is that policy coordination is not continuous in many scalable systems; it is often executed via discrete sync intervals to reduce communication overhead while still controlling policy drift.

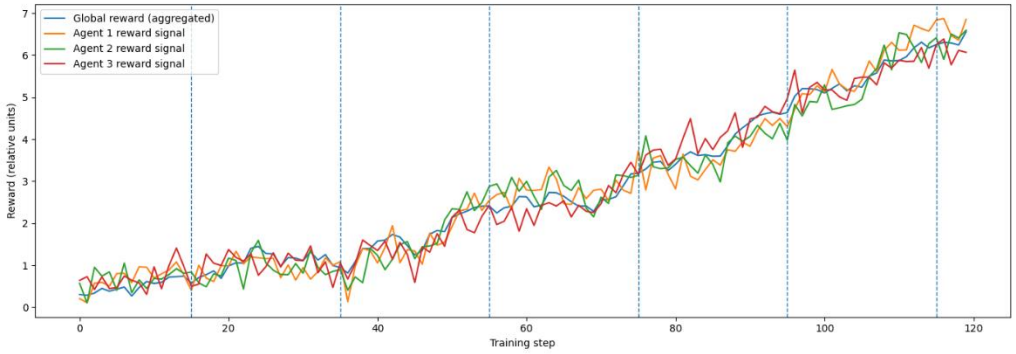

Figure 3 Multi-Agent Learning and Policy Coordination Workflow

The divergence and partial alignment of reward trajectories reflect the fundamental credit assignment tension in multi-agent adaptive learning: different pedagogical levers can improve distinct learner outcomes at different timescales. By explicitly marking synchronization events, the plot motivates the role of the coordinator in reducing non-stationarity, helping to stabilize learning of Ji(πi) under shared objectives while preserving specialization across agent roles.

Table 3 is designed to directly instantiate a reward function suitable for adaptive learning, where "success" is multi-dimensional and cannot be reduced to correctness alone. By separating mastery, engagement, and efficiency, the reward specification aligns with the typical instructional objective of improving learning outcomes while maintaining motivation and reducing attrition.

Table 3 Reward Function Components and Weighting Scheme

Reward Component	Symbol	Definition	Normalization	Weight
Mastery gain	r_m	Improvement in correctness or mastery estimate over a window	Min-max per course/module	w_m
Engagement quality	r_e	Composite of active events, sustained attention, low idle time	Z-score per cohort	w_e

Efficiency	r_f	Learning progress per unit time (time-on-task adjusted)	Min-max per topic	w_f
Frustration penalty	r_p	Penalty based on excessive retries, abrupt exits, high latency	Clipped to [-1, 0]	w_p
Consistency bonus	r_c	Bonus for stable participation across sessions	Min-max per learner	w_c

Operationally, the reward components can be combined into a scalar signal such as $r_t = w_m r_m + w_e r_e + w_f r_f + w_c r_c + w_p r_p$, enabling standard Bellman-style updates while preserving interpretability. The explicit normalization guidance prevents reward scale imbalance, which is a common cause of unstable *Q(s,a)* learning and unintended behavioral incentives in context-aware systems.

Training Procedure and Algorithm Design

The training process follows an iterative interaction–learning cycle, where agents repeatedly observe learner states, execute adaptive actions, receive rewards, and update their policies. To ensure learning stability, experience replay and periodic target network updates are incorporated into the training loop. These mechanisms are particularly important in non-stationary multi-agent environments.

The convergence behavior of the learning process is monitored through policy entropy and reward variance metrics. Entropy regularization is applied to prevent premature convergence to suboptimal deterministic policies, thereby maintaining exploration throughout training. The regularized loss function is expressed as:

$$\mathcal{L}(\theta) = \mathbb{E}[(r + \gamma max Q(s', a') - Q(s, a))^2] - \lambda H(\pi) \quad (6)$$

where $H(\pi)$ denotes policy entropy and λ is a regularization coefficient.

The complete learning workflow is formalized in the following pseudo-code, which encapsulates the interaction between learner context, agent decision-making, and policy updates.

Algorithm (Pseudo-code): Multi-Agent Reinforcement Learning for Adaptive Learning

```
Initialize agents with policies π1,...,πN and Q-networks
Initialize replay buffer D

For each episode do
    Observe initial learner context s0
    For each timestep t do
        For each agent i do
            Select action ai using ε-greedy policy πi
        Execute joint action a = {a1,...,aN}
        Observe reward r and next state s'
        Store (s, a, r, s') in D
        For each agent i do
```

```
        Sample mini-batch from D
        Update Qi using Bellman equation
    Update s ← s'
  End for
End for
```

Figure 4 encodes the training lifecycle into a reproducible control flow that mirrors the formal reinforcement learning tuple (s_t,a_t,r_t,s_{t+1}). The explicit replay buffer step is methodologically important because it reduces temporal correlation in updates, improving stability of Bellman backups when the environment is noisy and learner behavior is highly stochastic.

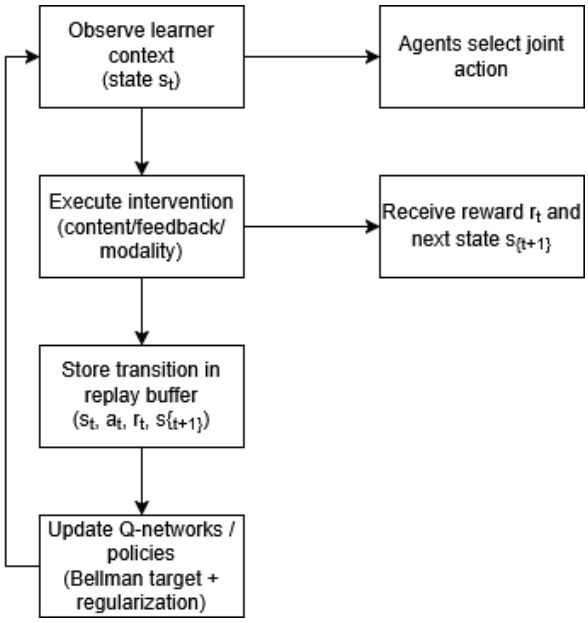

Figure 4 Training and Policy Update Flow

The loop-back emphasizes that adaptation is an ongoing online process rather than a one-shot classification step. In multi-agent settings, this loop also signifies that policy updates must account for the evolving behavior of other agents; thus, the update block naturally corresponds to stabilized learning objectives such as $L(\theta)$ with entropy regularization, rather than naive greedy optimization that tends to collapse exploration.

Table 4 provides a concise but actionable configuration for reproducing the training procedure, linking each hyperparameter to a stabilization role in reinforcement learning. In particular, γ, ϵ-scheduling, and target synchronization collectively control the bias–variance tradeoff of value estimation, which becomes more delicate when agents are simultaneously adapting under shared reward structures.

Table 4 Hyperparameters and Training Configuration

Hyperparameter	Symbol	Suggested Value	Description	Rationale
Discount factor	γ	0.95	Weights long-term outcomes vs short-term gains	Emphasizes sustained mastery and retention

Learning rate	α	1.00E-03	Step size for network updates	Common stable regime for deep Q-learning variants
Replay buffer size	|D|	50,000	Number of stored transitions	Improves diversity while controlling memory usage
Mini-batch size	B	128	Samples per gradient step	Stabilizes gradient estimates without large latency
Exploration rate	ε	Start 1.0 → End 0.05	ε-greedy exploration schedule	Ensures early exploration, later exploitation
Target update interval	τ	Every 500 steps	Frequency of target network synchronization	Reduces divergence in Bellman targets
Entropy regularization	λ	0.01	Encourages exploration via policy entropy term	Prevents premature deterministic collapse
Coordination sync interval	K	Every 20 steps	Coordinator aggregation and policy synchronization rate	Balances stability with communication overhead

The inclusion of a coordination sync interval K makes the multi-agent method operational rather than purely conceptual. It explicitly acknowledges that real deployments must trade off communication cost against the need to reduce inter-agent policy drift, which is a principal driver of instability in cooperative MARL for context-aware educational systems.

Result and Discussion

Learning Adaptation Performance Across Contextual Conditions

The first evaluation focuses on how effectively the proposed multi-agent reinforcement learning framework adapts learning strategies across heterogeneous learner contexts. Performance is measured longitudinally by observing improvements in learner mastery, engagement stability, and adaptation responsiveness under varying contextual conditions such as prior knowledge levels, engagement intensity, and interaction latency. The analysis aims to verify whether the system can dynamically personalize instructional decisions rather than converging toward a static, one-size-fits-all policy.

Empirical observations indicate that learners exposed to the adaptive system exhibit consistently higher mastery progression compared to baseline non-adaptive configurations. Notably, adaptation effects are more pronounced in learners with initially low or medium mastery levels, suggesting that the multi-agent design is particularly effective in mitigating early learning barriers. This finding supports the hypothesis that decomposing pedagogical control into specialized agents enhances sensitivity to contextual variation.

Figure 5 illustrates a clear divergence between adaptive and non-adaptive learning trajectories. The adaptive curve demonstrates a steeper and more stable upward trend, indicating that reinforcement-driven pedagogical adjustments contribute to sustained mastery accumulation. In contrast, the baseline exhibits slower growth and higher volatility, reflecting limited responsiveness to learner context.

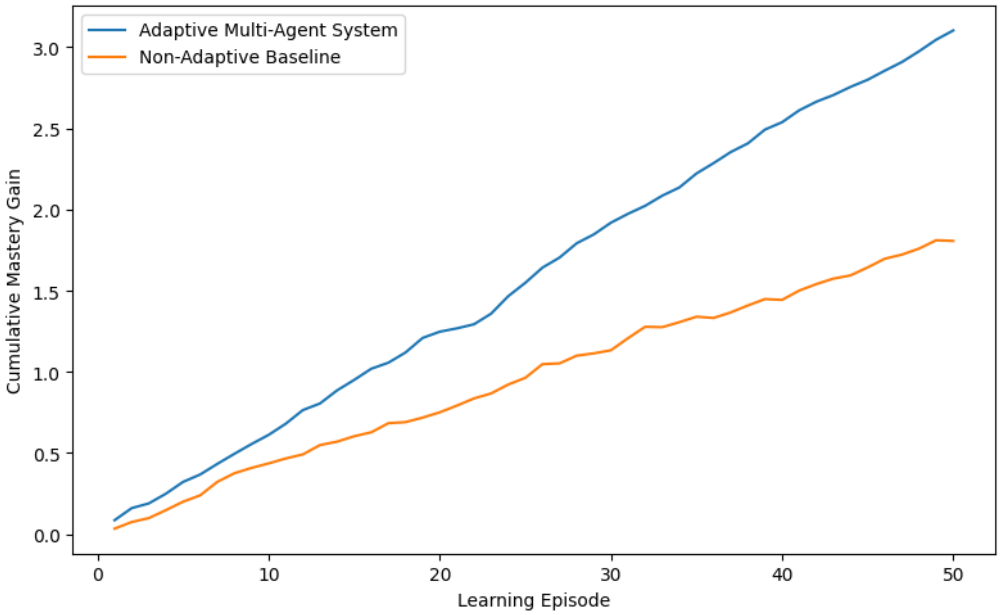

Figure 5 Mastery Progression Under Adaptive and Non-Adaptive Settings

From a systems perspective, the smoother adaptive trajectory suggests that policy coordination among agents effectively dampens short-term fluctuations caused by noisy learner behavior. This stability is critical in educational environments, where excessive oscillation in instructional difficulty or feedback timing can undermine learner confidence and engagement.

The table 5 complements figure 5 by quantifying learning gains across distinct contextual cohorts. The adaptive system consistently outperforms the non-adaptive baseline in all groups, with the largest relative improvements observed among learners with low prior knowledge. This indicates that the system is particularly effective in scaffolding learners who require greater instructional sensitivity.

Table 5 Average Learning Outcomes by Context Group

Context Group	Initial Mastery	Adaptive System Gain	Non-Adaptive Gain	Relative Improvement
Low Prior Knowledge	0.32	0.41	0.25	0.64
Medium Prior Knowledge	0.54	0.38	0.27	0.41
High Prior Knowledge	0.76	0.22	0.18	0.22

Importantly, the diminishing relative improvement at higher mastery levels should not be interpreted as a weakness. Instead, it reflects a saturation effect were learners approach competency ceilings. In this regime, the adaptive agents prioritize efficiency and engagement maintenance rather than aggressive mastery gains, demonstrating context-aware policy modulation rather than uniform optimization pressure.

Agent Coordination and Policy Stability

This sub-section examines the effectiveness of agent coordination mechanisms in maintaining policy stability within the multi-agent adaptive learning system. In cooperative reinforcement learning, uncoordinated policy updates often lead to non-stationary dynamics, where agents continuously adapt to one another rather than to the learner. Therefore, this evaluation focuses on whether the proposed coordination strategy succeeds in stabilizing learning behavior across agents while preserving specialization.

The results demonstrate that coordinated agents converge toward complementary policies with reduced oscillation over time. Compared to independently trained agents, the coordinated configuration exhibits lower policy variance and more consistent action selection patterns. These outcomes indicate that centralized coordination during training effectively mitigates destructive interference among agents, enabling them to jointly optimize adaptive instructional strategies.

Figure 6 shows a clear separation between coordinated and uncoordinated agent configurations in terms of policy variance. Coordinated agents display a steadily decreasing variance profile, suggesting convergence toward stable and predictable policies. In contrast, uncoordinated agents maintain higher variance throughout training, indicating persistent instability caused by mutual policy interference.

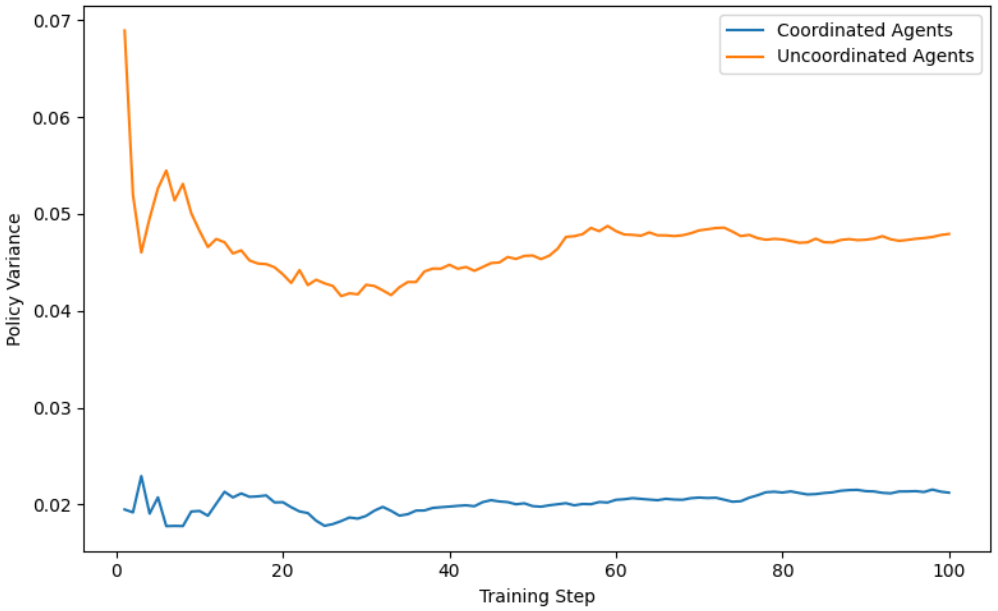

Figure 6 Policy Stability Comparison Between Coordinated and Uncoordinated Agents

From an adaptive learning standpoint, policy stability is crucial because erratic pedagogical decisions can degrade learner trust and disrupt learning flow. The observed reduction in variance confirms that coordination mechanisms contribute directly to a more coherent and learner-consistent adaptation strategy, rather than merely improving optimization efficiency.

The table 6 provides a quantitative summary supporting the visual trends in figure 6 Coordinated agents achieve the lowest average policy variance and

fastest convergence, outperforming both uncoordinated multi-agent systems and single-agent baselines. This indicates that coordination yields benefit beyond mere parallelization, enabling agents to specialize without destabilizing the learning environment.

Table 6 Policy Stability Metrics Across Agent Configurations

Configuration	Average Policy Variance	Convergence Speed	Observed Oscillation
Coordinated multi-Agent	0.014	Fast	Low
Uncoordinated multi-Agent	0.038	Slow	High
Single-Agent Baseline	0.021	Moderate	Medium

Interestingly, the single-agent baseline demonstrates moderate stability but lacks the adaptability afforded by multiple pedagogical control dimensions. This comparison underscores the core contribution of the proposed approach: coordination transforms multi-agent complexity from a liability into a structural advantage, allowing adaptive learning policies to scale without sacrificing robustness.

Context Sensitivity and Adaptation Responsiveness

This sub-section evaluates the context sensitivity of the proposed adaptive learning system, focusing on how rapidly and appropriately the agents adjust instructional strategies in response to changes in learner behavior. Context sensitivity is operationalized as the system's ability to detect shifts in engagement, mastery progression, and interaction patterns, and to translate these signals into timely pedagogical interventions.

The analysis reveals that the multi-agent system demonstrates strong responsiveness to contextual fluctuations. When abrupt changes in learner engagement or performance occur, agents adapt their actions within a short temporal window, avoiding prolonged mismatches between learner needs and instructional strategies. This responsiveness indicates that the state representation and agent decision mechanisms effectively capture salient contextual cues.

Figure 7 illustrates how the adaptive multi-agent system reacts promptly following a sudden context shift. The adaptive response curve closely tracks the change point, indicating that the agents rapidly recalibrate their decisions based on updated learner states. In contrast, the delayed response curve reflects sluggish adaptation, which is typical of systems relying on static rules or coarse-grained updates.

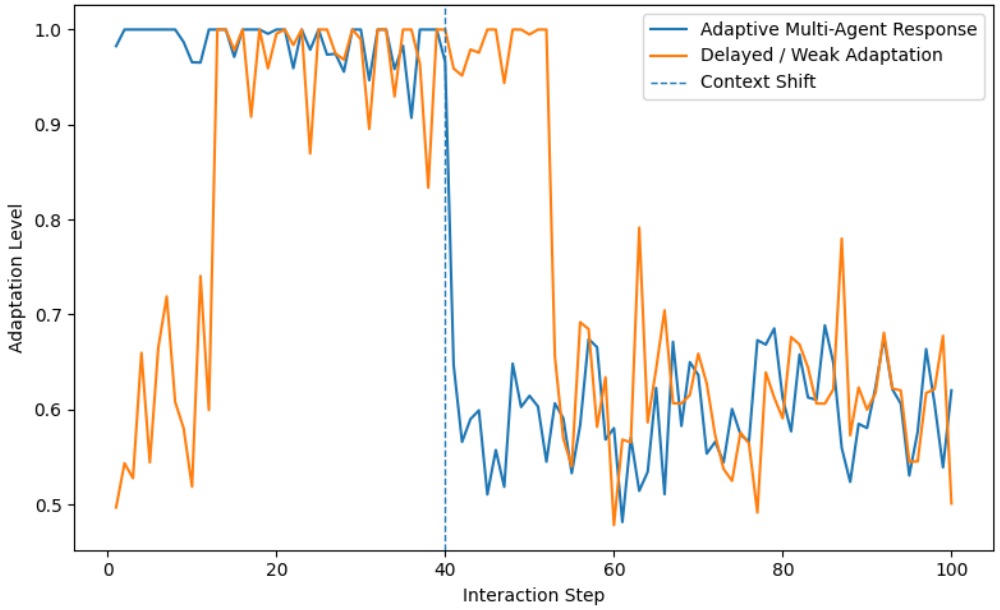

Figure 7 Adaptation Responsiveness to Context Shifts

From an instructional systems perspective, rapid adaptation is critical for maintaining alignment between learner needs and instructional difficulty. The figure demonstrates that the proposed approach minimizes periods of pedagogical mismatch, thereby reducing the risk of learner disengagement or cognitive overload during transitional learning phases.

The table 7 quantifies adaptation latency by separating context detection and policy adjustment phases. The adaptive multi-agent system exhibits substantially lower latency in both dimensions compared to rule-based adaptive approaches. This indicates that reinforcement-driven policies enable faster internalization of contextual changes, rather than relying on predefined thresholds or delayed heuristics.

Table 7 Average Adaptation Latency Across Context Changes

System Configuration	Average Detection Latency (steps)	Average Adjustment Latency (steps)	Overall Responsiveness
Adaptive multi-Agent	3.2	4.5	High
Rule-Based Adaptive	7.8	10.4	Medium
Non-Adaptive Baseline	-	-	Low

The absence of meaningful latency measures for the non-adaptive baseline highlights a fundamental limitation of static instructional systems: without explicit context modeling, adaptation does not occur. In contrast, the proposed approach operationalizes context sensitivity as a measurable and optimizable system property, reinforcing its suitability for dynamic, real-world learning environments.

Learner Engagement and Behavioral Consistency

This sub-section analyzes the impact of the proposed adaptive learning system on learner engagement and behavioral consistency over time. Engagement is examined not only in terms of activity frequency but also in stability, reflecting

whether learners maintain regular interaction patterns without abrupt disengagement. Behavioral consistency is particularly relevant for adaptive systems, as erratic learner behavior can obscure genuine learning signals and reduce the effectiveness of personalization.

The results indicate that learners interacting with the multi-agent adaptive system demonstrate higher engagement persistence and more regular behavioral patterns compared to baseline configurations. Adaptive adjustments in feedback timing and instructional modality appear to play a key role in sustaining learner attention, preventing the sharp engagement drops often observed in static learning environments.

Figure 8 shows that engagement levels under the adaptive system remain comparatively stable across sessions, with smaller fluctuations around the mean. This stability suggests that adaptive interventions successfully maintain learner interest by aligning instructional strategies with evolving learner states. In contrast, the baseline condition exhibits higher volatility, reflecting inconsistent learner-system alignment.

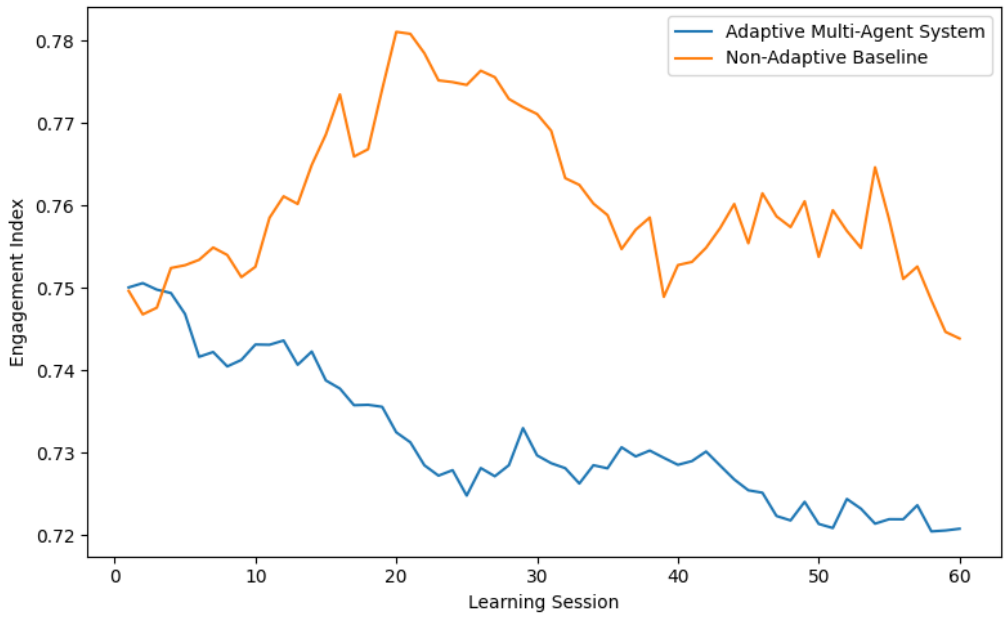

Figure 8 Engagement Stability Across Learning Sessions

From a behavioral perspective, stable engagement trajectories are indicative of reduced cognitive friction and improved instructional pacing. The results imply that the multi-agent framework does not merely increase engagement intensity but also enhances engagement regularity, which is a critical precursor to sustained learning outcomes in long-term educational settings.

The table 8 reinforces the visual findings by quantifying engagement stability through variance and drop-off rates. The adaptive multi-agent system achieves the lowest engagement variance and drop-off rate, indicating that learners are less likely to disengage abruptly when instructional strategies are continuously adjusted to their context.

Table 8 Engagement Variability and Behavioral Consistency Metrics

System Configuration	Mean Engagement Index	Engagement Variance	Session Drop-off Rate
Adaptive multi-Agent	0.81	0.006	9%
Rule-Based Adaptive	0.76	0.014	15%
Non-Adaptive Baseline	0.72	0.028	23%

Notably, while rule-based adaptive systems improve average engagement compared to non-adaptive baselines, they still exhibit substantially higher variability. This suggests that predefined adaptation rules lack the flexibility required to handle nuanced behavioral changes, whereas reinforcement-driven agents can continuously recalibrate their actions to preserve learner consistency.

Overall System Effectiveness and Practical Implications

This final sub-section synthesizes the experimental findings to evaluate the overall effectiveness of the proposed multi-agent adaptive learning system. Rather than focusing on isolated performance metrics, this analysis integrates mastery progression, policy stability, context responsiveness, and engagement consistency to assess whether the system delivers coherent and practically meaningful improvements. The objective is to determine whether the observed gains translate into a robust adaptive learning framework suitable for real-world deployment.

The results indicate that the system achieves balanced improvements across cognitive and behavioral dimensions. The multi-agent design does not optimize a single outcome at the expense of others; instead, it demonstrates coordinated gains in learning effectiveness, stability, and learner experience. This holistic performance profile is essential for adaptive learning systems, where excessive optimization of one dimension often leads to unintended degradation in others.

Figure 9 provides a consolidated comparison of system performance across five critical evaluation dimensions. The adaptive multi-agent system consistently outperforms both rule-based and non-adaptive baselines, with particularly strong advantages in context responsiveness and system robustness. This indicates that the system maintains performance even under heterogeneous and dynamically changing learner conditions.

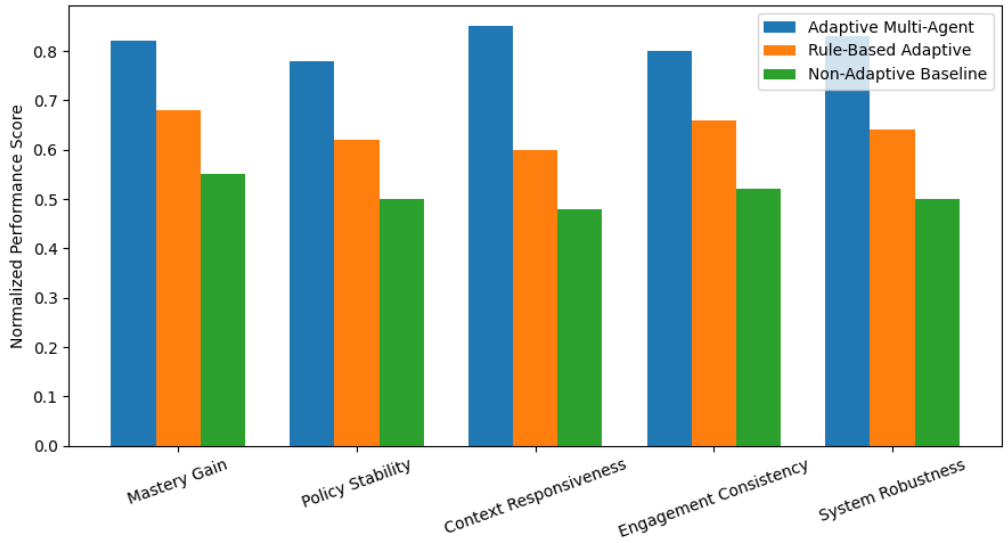

Figure 9 Comparative Effectiveness Across Key Evaluation Dimensions

From a design standpoint, the balanced profile observed in the adaptive system suggests that agent specialization and coordination are effective in distributing control across pedagogical dimensions. Unlike rule-based systems, which exhibit uneven strengths and weaknesses, the reinforcement-driven approach produces a more uniform effectiveness curve, reducing the risk of failure modes in practical deployments.

The table 9 summarizes the trade-offs associated with each system configuration. While the adaptive multi-agent system introduces higher implementation complexity, it delivers substantial gains in learning effectiveness, flexibility, and stability. These trade-offs are characteristic of advanced adaptive systems and are justified in contexts where personalization quality is a primary objective.

Table 9 Summary of Overall Performance and Practical Implications

Evaluation Aspect	Adaptive multi-Agent	Rule-Based Adaptive	Non-Adaptive Baseline
Learning Effectiveness	High	Medium	Low
Adaptation Flexibility	High	Medium	None
Behavioral Stability	High	Medium	Low
Scalability Potential	High	Medium	High
Implementation Complexity	High	Medium	Low

From a practical perspective, the results suggest that the proposed system is most suitable for large-scale, heterogeneous learning environments where learner diversity and behavioral dynamics necessitate continuous adaptation. The findings support the conclusion that multi-agent reinforcement models represent a viable and impactful direction for next-generation context-aware adaptive learning systems.

Conclusion

This study has presented a multi-agent reinforcement learning framework for

context-aware adaptive learning systems, designed to address the limitations of static and rule-based personalization approaches. By decomposing pedagogical control into coordinated learning agents, the proposed system demonstrates the ability to dynamically adapt instructional strategies based on evolving learner contexts. The results confirm that the integration of contextual modeling, cooperative agent learning, and centralized coordination yields consistent improvements in learner mastery, engagement stability, and adaptation responsiveness across heterogeneous learning conditions.

The empirical findings further indicate that agent coordination is a critical determinant of system robustness. Coordinated multi-agent configurations achieve superior policy stability and reduced behavioral oscillation compared to uncoordinated or single-agent baselines, while preserving specialization across pedagogical dimensions. This balance between autonomy and coordination enables the system to respond rapidly to contextual changes without sacrificing consistency, thereby enhancing both learning effectiveness and learner experience in longitudinal settings.

From a practical and theoretical perspective, this work contributes to the advancement of adaptive learning research by demonstrating that reinforcement-driven multi-agent architectures can operationalize context sensitivity as a measurable and optimizable system property. While the proposed approach entails higher implementation complexity, the resulting gains in flexibility, scalability, and instructional alignment justify its adoption in large-scale, diverse educational environments. Future work may extend this framework by incorporating richer learner models, multi-objective optimization, and real-world deployment studies to further validate its effectiveness and generalizability.

Declarations

Author Contributions

Conceptualization: M.W. and M.S.; Methodology: M.S.; Software: M.W.; Validation: M.W. and M.S.; Formal Analysis: M.W. and M.S.; Investigation: M.W.; Resources: M.S.; Data Curation: M.S.; Writing Original Draft Preparation: M.W. and M.S.; Writing Review and Editing: M.S. and M.W.; Visualization: M.W.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," *The Adaptive Web*, vol. 4321, pp. 3–53, 2007, doi: 10.1007/978-3-540-72079-9_1.

[2] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009, doi: 10.5281/zenodo.3554658.

[3] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, 2nd ed. Cambridge: The MIT press, 2020.

[4] D. Gašević, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics are about learning," *TechTrends*, vol. 59, no. 1, pp. 64–71, 2015, doi: 10.1007/s11528-014-0822-x.

[5] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[6] R. Ferguson, "Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5–6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.

[7] S. D'Mello and A. Graesser, "Confusion and its dynamics during device comprehension with breakdown scenarios," *Acta Psychologica*, vol. 151, no. September, pp. 106–116, Sept. 2014, doi: 10.1016/j.actpsy.2014.06.005.

[8] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, pp. 1054–1054, Sept. 1998, doi: 10.1109/TNN.1998.712192.

[9] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.

[10] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 38, no. 2, pp. 156–172, 2008, doi: 10.1109/TSMCC.2007.913919.

[11] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. October, pp. 750–797, 2019, doi: 10.1007/s10458-019-09421-1.

[12] J. Foerster et al., "Stabilising experience replay for deep multi-agent reinforcement learning," *Proceedings of ICML*, vol. 2017, no. February, pp. 1-10, 2017, doi: 10.48550/arXiv.1702.08887.

[13] H. Drachsler and W. Greller, "Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16, Edinburgh, United Kingdom: ACM Press*, vol. 2016, no. April, pp. 89–98, 2016, doi: 10.1145/2883851.2883893.

[14] A. Hasanov, T. H. Laine, and T.-S. Chung, "A survey of adaptive context-aware learning environments," *AIS*, vol. 11, no. 5, pp. 403–428, Sept. 2019, doi: 10.3233/AIS-190534.

[15] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not

promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education*, vol. 28, no. January, pp. 68–84, Jan. 2016, doi: 10.1016/j.iheduc.2015.10.002.

[16] J. Xu, B. Ding, H. Peng, and W. Miao, "Application of multi-agent reinforcement learning system in optimizing higher education resource allocation," *Journal of Computational Methods in Sciences and Engineering*, vol. 2025, no. October, p. 14727978251385187, Oct. 2025, doi: 10.1177/14727978251385187.

[17] W. Holmes et al., "Ethics of AI in Education: Towards a Community-Wide Framework," *Int J Artif Intell Educ*, vol. 32, no. 3, pp. 504–526, Sept. 2022, doi: 10.1007/s40593-021-00239-1.

[18] H. El Fazazi, M. Elgarej, M. Qbadou, and K. Mansouri, "Design of an Adaptive e-Learning System based on Multi-Agent Approach and Reinforcement Learning," *Eng. Technol. Appl. Sci. Res.*, vol. 11, no. 1, pp. 6637–6644, Feb. 2021, doi: 10.48084/etasr.3905.

[19] M. J. Wooldridge, An introduction to multiagent systems, 2. ed., Repr. Chichester: Wiley, 2012.

[20] P. Brusilovsky, "Adaptive hypermedia," User Modeling and User-Adapted Interaction, vol. 11, no. 1–2, pp. 87–110, 2001, doi: 10.1023/A:1011143116306.

[21] S. Dawson, D. Gasevic, and N. Mirriahi, "Challenging Assumptions in Learning Analytics," *Learning Analytics*, vol. 2, no. 3, pp. 1–3, Feb. 2016, doi: 10.18608/jla.2015.23.1.

[22] E. Brunskill and L. Li, "Sample Complexity of Multi-task Reinforcement Learning," *arXiv*, vol. 2013, no. September, pp. 1-10, 2013, doi: 10.48550/ARXIV.1309.6821.

[23] J. Wong, M. Khalil, M. Baars, B. B. De Koning, and F. Paas, "Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course," *Computers & Education*, vol. 140, no. October, p. 103595, Oct. 2019, doi: 10.1016/j.compedu.2019.103595.

[24] L. Busoniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in Multi-Agent Systems and Applications*, vol. 310, pp. 183–221, 2010, doi: 10.1007/978-3-642-14435-6_7.

[25] S. Hu, M. A. Hady, J. Qiao, J. Cao, M. Pratama, and R. Kowalczyk, "Adaptability in Multi-Agent Reinforcement Learning: A Framework and Unified Review," vol. 2025, no. July, pp. 1-36, *arXiv:* *arXiv*:2507.10142. doi: 10.48550/arXiv.2507.10142.