2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #51416

PedagoReLearn: A Reinforcement Learning Framework for Adaptive Instructional Sequencing

Mr. Thomas Franklin Hallmark, Texas A&M University

Thomas F. Hallmark is a doctoral student in Curriculum and Instruction (Engineering Education) in the Department of Teaching, Learning, and Culture at Texas A&M University. He holds degrees in Nuclear Engineering, Legal Studies, and Business Administration (MBA) and brings more than 30 years of experience in the nuclear power and energy industries. His research focuses on the integration of artificial intelligence and reinforcement learning in engineering and STEM education, with particular emphasis on adaptive tutoring systems, veteran transitions, and cross-cultural learning. He is the developer of Reflexive Reciprocity Theory (RRT), a diagnostic framework for analyzing how instructional design shapes emergent behavior in AI-supported learning systems. His work bridges pedagogical theory and computational modeling to design human-centered AI learning environments that support personalized, data-informed, and pedagogically grounded instruction.

PedagoReLearn: Reinforcement Learning as a Diagnostic Instrument for Instructional Design

Abstract

Adaptive tutoring systems increasingly employ reinforcement learning (RL) to personalize instruction, yet many such systems exhibit behaviors that optimize short-term performance at the expense of pedagogically meaningful learning. This work-in-progress paper investigates how instructional behaviors emerge when RL is applied to adaptive tutoring without explicit modeling of knowledge stability and forgetting. Using a simulated tutoring environment framed as a Markov Decision Process (MDP), a tabular State–Action–Reward–State–Action (SARSA) agent selects among three pedagogical actions — teach, quiz, and review — to guide learner progression. The system is evaluated against fixed-sequence and random baselines across 10 independent random seeds using cumulative reward, steps-to-mastery, and instructional action distribution metrics.

Results indicate that while the RL agent converges faster and outperforms both baselines on episodic efficiency — reducing steps-to-mastery from 37.33 (fixed and random baselines) to 25.25 ± 1.13 (aggregated-state) and 26.14 ± 2.02 (full-state) — these efficiency gains reflect optimization of the reward structure rather than pedagogically meaningful learning. The agent converges on a quiz-dominant instructional policy. Across converged policies, quiz actions comprised 36.1–38.6% of selections, with teach and review actions falling below equal distribution. This non-uniform action distribution reveals a structural feature of the instructional environment: in the absence of explicit stability and decay dynamics, quiz actions are reinforced disproportionately by the immediate reward structure (+1 per correct quiz response; +500 terminal mastery bonus; −1 per step), encouraging short-term optimization rather than pedagogically balanced sequencing. Interpreted through Reflexive Reciprocity Theory (RRT), these findings position RRT as a diagnostic framework that uses emergent agent behavior to reveal which pedagogical constructs are operationalized, underspecified, or absent within adaptive instructional designs. The paper contributes design insights for engineering and STEM education by identifying design conditions within this environment under which RL-based tutors can support pedagogically meaningful sequencing, informing the redesign of adaptive systems that balance episodic efficiency with long-term retention.

Introduction

Engineering and STEM education stand at an inflection point where advances in artificial intelligence intersect with long-standing principles of human learning. Adaptive tutoring systems promise personalized instruction at scale, yet their effectiveness depends not only on algorithmic sophistication but on how pedagogical assumptions are encoded within instructional environments. Reinforcement learning (RL), in particular, offers a compelling framework for discovering instructional policies through interaction and long-term optimization [1], and has been increasingly explored for instructional sequencing and adaptive tutoring in intelligent tutoring systems (ITS) research (e.g., [2], [3], [4], [5]). These approaches formalize teaching as a sequential decision-making problem in which instructional actions are selected based on anticipated long-term learning gains. However, when pedagogical structure is underspecified, RL agents may converge on behaviors that optimize immediate outcomes while diverging from established learning science [4], [6].

This paper examines a fundamental question relevant to engineering education researchers and instructional designers: ***What instructional behaviors emerge when RL-based tutors operate without explicit models of knowledge stability and forgetting?*** Rather than treating unexpected agent behavior as algorithmic failure, this study adopts a design-oriented perspective in which agent policies are interpreted as reflections of the pedagogical affordances made available by the environment. As an early application of Reflexive Reciprocity Theory (RRT) — which conceptualizes learning as emerging from reciprocal adaptation between learners and instructional systems — this work investigates how RL agents can reveal hidden pedagogical constraints embedded within instructional design. The study focuses on simulated tutoring environments and does not evaluate learner outcomes directly. Findings demonstrate that while the RL agent achieved superior episodic efficiency relative to baseline strategies — reducing steps-to-mastery from 37.33 to 25.25 ± 1.13 (aggregated-state) and 26.14 ± 2.02 (full-state) — it converged on a quiz-dominant instructional policy (36.1–38.6% quiz selections), contrary to established principles of durable learning [4], [6], [7].

Two research questions frame the inquiry:

RQ1: What instructional behaviors emerge when reinforcement learning is applied to adaptive tutoring without explicit modeling of knowledge stability and forgetting?

RQ2: What representational and reward design conditions are required for RL agents to exhibit pedagogically meaningful instructional sequencing?

By addressing these questions, the study contributes diagnostic insight into the design of adaptive tutoring systems for engineering and STEM contexts, where durable learning and principled sequencing are critical [4], [6].

Theoretical Foundations

This work integrates three complementary theoretical perspectives: experiential learning, cognitive models of forgetting, and reinforcement learning for instructional decision-making. Together, these perspectives motivate the treatment of adaptive tutoring as a sequential, state-dependent process shaped by interaction and feedback.

Experiential learning theory emphasizes that learning arises through interaction and reflective adjustment rather than passive transmission. Dewey [8] argued that the quality of experience — particularly its continuity and its capacity to inform subsequent action — is the central criterion of educational value. Schön [9] extended this perspective by framing professional competence as a function of reflective practice, in which practitioners adjust their strategies in response to feedback from the environment. From this view, instructional quality depends on how experiences are sequenced and how feedback structures subsequent action — a dynamic directly instantiated in reinforcement learning–based tutoring environments.

Cognitive research on memory further demonstrates that learning is inherently unstable: without reinforcement, knowledge decays predictably over time [7]. Ebbinghaus [7] established the foundational decay curve characterizing forgetting as an exponential function of elapsed time, a finding replicated and extended across decades of memory research. The spacing effect and optimal practice scheduling literature underscore the importance of timing, review, and retrieval practice for long-term retention [4], [6]. These findings establish a theoretical baseline against which the instructional policies of RL agents can be evaluated: a pedagogically sound tutoring system should exhibit sensitivity to temporal dynamics, not merely immediate correctness.

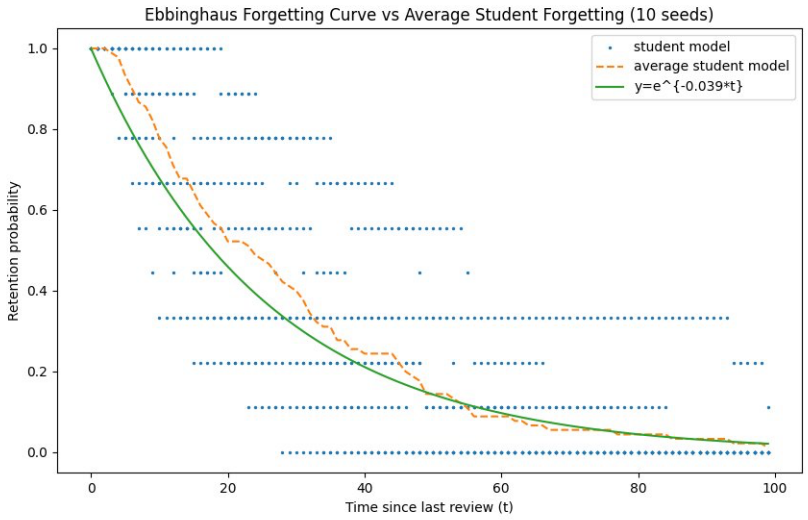

Figure 1. Ebbinghaus forgetting curve ($y = e^{-0.039 \cdot t}$) overlaid with simulated student retention trajectories across 10 seeds. The average student model (dashed) tracks the theoretical curve through the early training episodes, after which individual trajectories stabilize — reflecting the absence of explicit decay dynamics in the current environment specification and motivating future redesign.

Reinforcement learning provides a computational framework for operationalizing these theoretical principles. A Markov Decision Process (MDP) formalizes this as a sequential decision problem in which an agent selects actions based on the current state of the environment, receives rewards, and transitions to new states — enabling instructional actions to be evaluated in terms of their long-term consequences rather than immediate correctness alone [1]. This paper instantiates that formalism in a simulated tutoring environment to examine what happens when the MDP's state and reward representations omit explicit cognitive dynamics. However, RL agents optimize only what the environment makes salient. When temporal dynamics such as knowledge stability and decay are omitted from state representations and reward structures, the

resulting policies may align with immediate reward availability rather than with the pedagogical intent established by experiential and cognitive learning theory.

Reflexive Reciprocity Theory (RRT) provides the integrative interpretive framework binding these perspectives [10]. RRT posits that instructional systems and learners co-adapt through interaction, and that system-level behavior can be analyzed pedagogically without attributing human-like cognition to the system. In this study, RL agents function as structural reflections of instructional design — revealing which pedagogical assumptions are operationalized within the environment and which are absent. It should be noted that RRT is applied here as an interpretive and diagnostic framework positioned at an early stage of theoretical development, rather than as a fully validated or exhaustive theory of instructional learning.

Methodology

The adaptive tutoring environment examined in this study is formalized as a Markov Decision Process (MDP) defined by the tuple ⟨S, A, P, R, γ⟩, where S represents learner mastery states, A the instructional action space, P the transition function, R the reward structure, and γ the discount factor [1]. A Markov Decision Process formalizes instructional decision-making as a sequential problem in which an agent selects actions based on the current state of the environment, receives rewards, and transitions to new states — enabling instructional actions to be evaluated in terms of their long-term consequences rather than immediate correctness alone. The full environment and training specification is provided in Table I.

Table I
ENVIRONMENT AND TRAINING SPECIFICATION

Component	Specification
State Representation (S)	Concept mastery $m_i \in \{0, 1, 2, 3\}$; terminal mastery = 3
Action Space (A)	{Teach, Quiz, Review}
Transition Function (P)	Deterministic state advancement; no decay, regression, or recency effects
Episode Initialization	All concepts initialized to Mastery Level 0
Episode Termination	All concepts reach Mastery Level 3 **or** 100-step limit
Reward Function (R)	−1 per step; +1 correct quiz; −0.20 incorrect quiz; +500 terminal bonus
Learning Algorithm	Tabular SARSA (on-policy)
Exploration Policy	ε-greedy (ε = 0.1)
Learning Rate (α)	0.2
Discount Factor (γ)	0.9
Training Episodes	2,000 per seed
Random Seeds	10 (reporting); 50 (convergence visualization)
Performance Metric	Mean over final 100 episodes per seed

The state representation encodes learner mastery across four discrete levels per concept: $m_i \in \{0, 1, 2, 3\}$, where Mastery Level 0 represents an unlearned concept and Mastery Level 3 represents terminal mastery. Two state encoding configurations are evaluated: an aggregated-state representation that summarizes global mastery progress across concepts, and a full-state representation that encodes individual concept mastery levels independently. This design choice enables a comparative assessment of how representational granularity influences emergent instructional policy structure.

The action space consists of three pedagogical actions drawn from established instructional practice: teach (introduction of new material), quiz (retrieval-based assessment), and review (reinforcement of prior content). These actions reflect the core decision space available to a human instructor and are treated as functionally distinct within the environment specification, though — as the results demonstrate — their behavioral differentiation by the RL agent depends critically on how they are represented in the reward structure.

Transition dynamics are deterministic: successful instructional interaction advances learner mastery state without probabilistic decay, regression, or recency effects. Each episode initializes all concepts at Mastery Level 0 and terminates either when all concepts reach Mastery Level 3 or when a maximum of 100 instructional steps is reached. No temporal stability parameters, forgetting rates, or recency weighting are incorporated into the transition logic — an intentional design choice that isolates the effect of pedagogical underspecification on emergent agent behavior.

Rewards are assigned according to immediate performance indicators exclusively. A correct quiz response yields +1; an incorrect quiz response incurs a penalty of −0.20; each instructional step carries a cost of −1; and successful episode completion upon full mastery yields a terminal bonus of +500. No delayed reward, recency penalty, or stability bonus is included. This structure incentivizes rapid mastery achievement without differentiating between instructional actions based on their long-term pedagogical consequences — making the diagnostic signal of emergent policy behavior interpretable under RRT [10].

A tabular SARSA agent interacts with the environment using an ε-greedy exploration strategy ($\varepsilon = 0.1$, $\alpha = 0.2$, $\gamma = 0.9$) [1]. SARSA was selected for its interpretability and on-policy learning dynamics, enabling direct analysis of how agent behavior evolves in response to environmental structure. Two baseline agents are included for comparison: a random agent selecting instructional actions uniformly at random, and a fixed-sequence agent following a deterministic teach–quiz–review script representative of traditional linear curricula [11]. These baselines isolate the contribution of learned policy structure beyond stochastic or scripted instructional sequencing.

Each configuration was trained for 2,000 episodes across 10 independent random seeds. Final performance metrics were computed as the mean over the last 100 episodes per seed to reduce stochastic fluctuation. Across seeds, convergence patterns and action distributions exhibited minimal variance, indicating stable optimization under the specified reward structure. For convergence visualization, a separate 50-seed experiment was conducted to ensure stability of trajectory patterns and to illustrate training dynamics independent of final performance reporting.

Performance is evaluated using three metrics: cumulative reward, steps-to-mastery, and instructional action distribution, each aggregated across runs. Cumulative reward and steps-to-mastery reflect episodic efficiency; action distribution provides a diagnostic view of how the agent differentiates among pedagogical actions within the environment — the primary metric of interest for the RRT-based analysis. Together, these metrics enable the study to distinguish between surface-level instructional efficiency and the deeper pedagogical structure of the learned policy, establishing the empirical foundation for the diagnostic analysis developed in the Results and Discussion sections.

Results

Table II reports final cumulative reward and steps-to-mastery (mean ± SD across 10 seeds) for all configurations, computed as the mean over the last 100 episodes per seed to reduce stochastic fluctuation.

Table II

FINAL PERFORMANCE ACROSS 10 SEEDS (2000 EPISODES; LAST 100 AVERAGED). STATISTICAL COMPARISONS (WELCH'S T-TEST) INDICATE BOTH SARSA CONFIGURATIONS SIGNIFICANTLY OUTPERFORM FIXED BASELINE IN STEPS-TO-MASTERY ($P < .001$).

Agent / Configuration	Final Cumulative Reward (Mean ± SD)	Final Steps to Mastery ↓ (Mean ± SD)
SARSA (Aggregated)	123.28 ± 4.18	25.25 ± 1.13
SARSA (Full State)	125.55 ± 7.97	26.14 ± 2.02
Fixed Baseline	—	37.33 ± 0.00
Random Baseline	—	37.33 ± 0.00

Across 10 independent random seeds, both SARSA configurations converged to substantially lower steps-to-mastery than either baseline. The aggregated-state configuration achieved 25.25 ± 1.13 steps-to-mastery, and the full-state configuration achieved 26.14 ± 2.02 steps, compared to 37.33 ± 0.00 steps for both the fixed and random baselines. Variance across seeds remained modest for both of the SARSA variants, indicating stable convergence rather than stochastic fluctuation. Cumulative reward is not reported for baseline agents, as their deterministic and non-optimizing policies do not accumulate reward in a manner comparable to the learned SARSA policies.

To evaluate whether observed differences in steps-to-mastery were statistically significant, independent-samples comparisons were conducted between each SARSA configuration and the fixed baseline. Given the zero-variance observed in the fixed baseline condition, Welch's t-test was applied to accommodate unequal variances, though the zero-variance baseline represents a boundary condition that should be interpreted with appropriate caution. The aggregated-state SARSA configuration achieved significantly lower steps-to-mastery than the fixed baseline, $t(9) = 10.69$, $p < .001$, $d = 10.69$. The full-state SARSA configuration similarly outperformed the fixed baseline, $t(9) = 5.54$, $p < .001$, $d = 5.54$. Both effect sizes are very large, reflecting strong separation between learned and scripted policies under the specified reward structure.

To illustrate the training dynamics underlying these final performance statistics, Figure 2 presents the median steps-to-mastery trajectory across 50 independent random seeds.

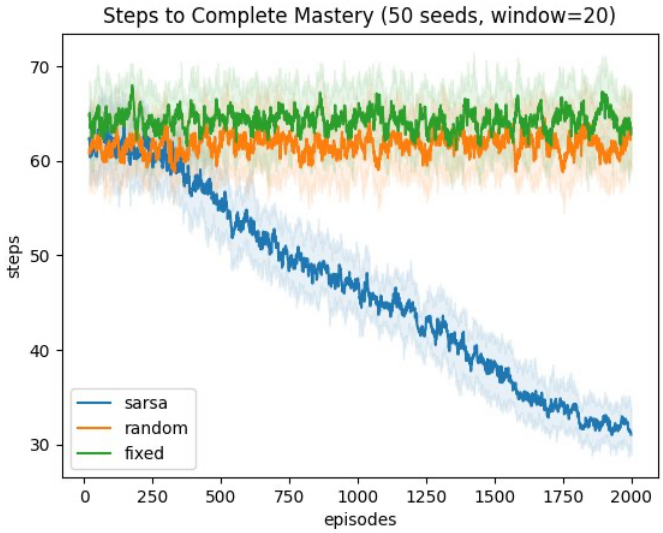

Fig. 2. *Median steps-to-mastery across 50 independent random seeds (moving average window = 20). Both SARSA configurations converge to substantially lower steady-state step counts relative to fixed and random baselines.*

As shown in Figure 2, the SARSA agent exhibits a steady, monotonic reduction in steps-to-mastery over training, whereas fixed and random baselines remain comparatively stable throughout. This gradual improvement indicates structural policy learning rather than episodic fluctuation. Figure 3 presents the corresponding median cumulative reward trajectories across 50 seeds, confirming that SARSA reward accumulation increases monotonically while baseline agents plateau.

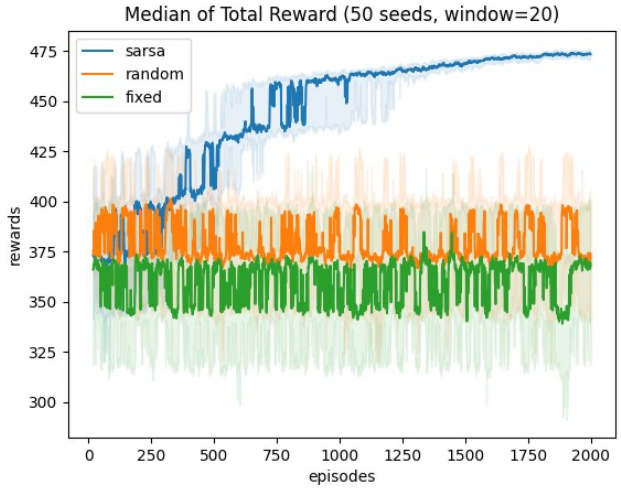

Fig. 3. *Median cumulative reward across 50 independent random seeds (moving average window = 20). The SARSA agent accumulates reward monotonically across training while random and fixed baselines remain flat, confirming structural policy learning.*

Figure 4 presents the 10-seed comparison across all four configurations, illustrating the modest but consistent separation between the full-state and aggregated-state SARSA variants, with SARSA (Aggregated) showing slightly faster convergence in later episodes.

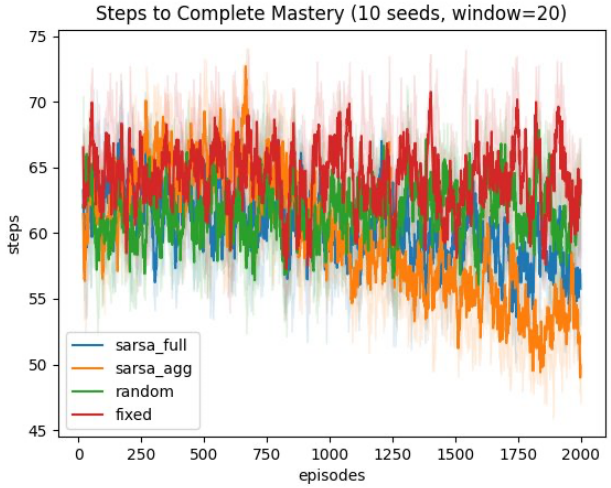

Fig. 4. *Steps to complete mastery across 10 seeds (moving average window = 20) for all four agent configurations. Both SARSA variants converge toward lower step counts relative to random and fixed baselines, with SARSA (Aggregated) showing slightly faster convergence in later episodes.*

These efficiency gains, however, do not tell the complete story. The similarity in steps-to-mastery between the aggregated-state and full-state configurations indicates that increased representational granularity alone does not produce pedagogically differentiated behavior when critical cognitive dynamics — knowledge stability and forgetting — are absent from the environment specification. To examine whether the learned policies are pedagogically meaningful beyond their efficiency advantage, Table III presents the instructional action distributions for all configurations.

Table III

INSTRUCTIONAL ACTION DISTRIBUTION

(FINAL 100 EPISODES AVERAGED ACROSS 10 SEEDS)

Agent	Quiz (%)	Teach (%)	Review (%)
SARSA (Aggregated)	36.1	33.2	30.7
SARSA (Full State)	38.6	31.9	29.5
Fixed Baseline	33.3	33.3	33.3
Random Baseline	~33.3	~33.3	~33.3

As shown in Table III, both SARSA configurations converged on non-uniform instructional policies that diverged meaningfully from the approximately equal distributions produced by baseline agents. The full-state SARSA policy selected quiz actions most frequently (38.6%), followed by teach (31.9%) and review (29.5%). The aggregated-state SARSA policy maintained quiz dominance at 36.1% while shifting modestly toward teach (33.2%) and review (30.7%),

indicating that reduced state granularity slightly altered action weighting without fundamentally changing the quiz-dominant policy structure. These distributions remained stable across seeds and training durations, indicating structural convergence rather than transient exploration behavior.

Notably, the quiz-dominant pattern emerges directly from the confirmed reward structure: quiz actions yielding correct responses produce an immediate +1 reward, while teach actions advance mastery deterministically but carry no immediate positive bonus beyond per-step cost recovery (−1 per step), and review actions offer no distinct advantage in the absence of decay or recency modeling. The agent therefore learns to favor the action most immediately reinforced by the environment — a rational optimization outcome given the encoded reward structure, but one that contradicts established findings on the importance of balanced retrieval practice and spaced review for long-term retention [4], [6], [7].

These findings demonstrate that the observed policy structure is not a failure of the learning algorithm, but a faithful reflection of the pedagogical assumptions embedded within the instructional environment. The results characterize emergent instructional policies under specific modeling constraints and do not constitute evidence of improved learner outcomes.

Figure 5 presents alpha-blended heatmaps of agent behavior across 2,000 episodes and 50 seeds, providing within-episode resolution of the mastery progression, recency logging, cumulative reward, and action frequency patterns that underlie the aggregate findings reported in Tables II and III.

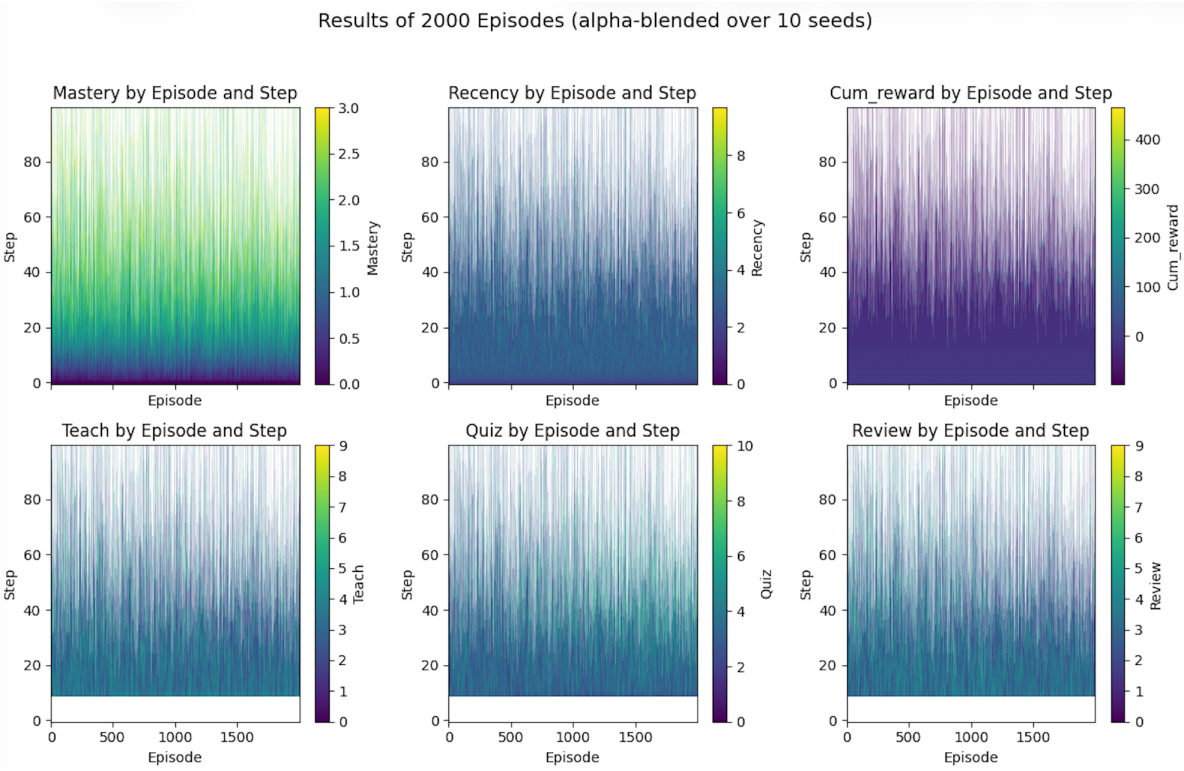

Fig. 5. *Alpha-blended heatmaps of SARSA agent behavior across 2,000 episodes and 50 seeds. Top row: mastery progression, action recency log (recorded for diagnostic purposes; recency is not modeled in the environment), and cumulative reward by episode and step. Bottom row: Teach, Quiz, and Review action frequencies. Increasing episode number (x-axis) reflects the agent learning to reach mastery in fewer steps, while action panels reveal the non-uniform distribution that emerges in the converged policy.*

Discussion

Interpreted through Reflexive Reciprocity Theory (RRT) [10], the observed results highlight the distributed nature of instructional agency in adaptive systems. The RL agent does not independently reason about pedagogy; rather, it enacts the instructional logic made possible by the environment. When pedagogical distinctions between actions are underspecified at the representational level — as in the absence of knowledge stability, temporal decay, and recency modeling confirmed in this study's environment specification — the system converges on behavior that appears efficient but conflicts with established learning science [4], [6]. Reinforcement learning, under this framing, functions not as a surrogate instructor but as a structured probe of instructional design: the learned policy is empirical evidence of the pedagogical assumptions embedded within the environment.

Figure 6 summarizes this diagnostic relationship, illustrating how instructional design choices — state representation, action space, and reward logic — give rise to emergent agent behavior that can be interpreted pedagogically under RRT.

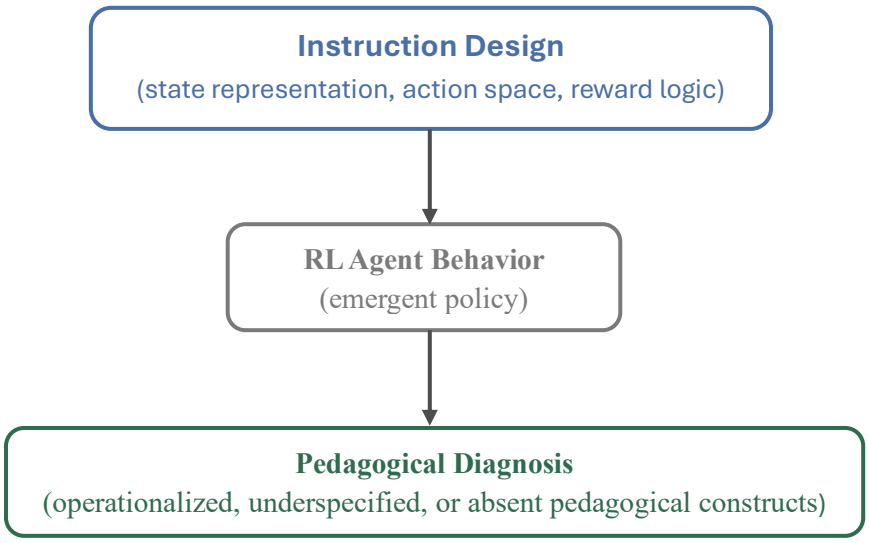

Fig. 6. *Reinforcement learning as a diagnostic instrument under Reflexive Reciprocity Theory (RRT). The figure illustrates the three-stage diagnostic relationship between instructional design choices (state representation, action space, and reward structure), emergent RL agent behavior (quiz-dominant policy; non-uniform action distribution), and RRT-based interpretation (identification of operationalized, underspecified, and absent pedagogical constructs). Arrows indicate the direction of diagnostic inference from environment specification to policy analysis.*

Operational Criteria for RRT-Based Diagnostic Analysis

RRT provides not only an interpretive lens but an operational framework for analyzing and redesigning adaptive instructional systems. Three functional criteria define its diagnostic logic. First, RRT conceptualizes instructional design elements — state representation, reward structure, and action space — as co-adaptive components that jointly shape system behavior. Second, it treats emergent policy structure as diagnostic output: the learned strategy constitutes evidence of which pedagogical constructs are effectively encoded within the environment. Third, RRT provides criteria for identifying underspecified constructs; when theoretically important distinctions — such as knowledge stability, temporal decay, or recency effects — fail to

influence policy differentiation despite repeated training, their absence becomes empirically observable.

An instructional construct is considered operationalized if systematic variation in its representational encoding produces differentiated policy behavior under RL optimization. Conversely, a construct is considered underspecified or absent if theoretically meaningful distinctions fail to influence learned policy structure across randomized seeds. RRT therefore generates testable diagnostic predictions: if knowledge stability, recency effects, or temporal decay are meaningfully represented in the state space or reward structure, RL agents should exhibit differentiated sequencing behaviors — such as increased review frequency or spacing-sensitive policies. If such constructs are omitted, policy convergence should collapse toward short-horizon efficiency behaviors, as observed here — with quiz-dominant strategies (36.1–38.6% quiz selections) driven by immediate reward availability at the expense of pedagogically balanced sequencing.

Replicating an RRT-based diagnostic analysis involves three procedural steps: (1) explicit specification of state, action, and reward representations; (2) training RL agents across a sufficient number of randomized seeds — a minimum of 10 for performance reporting and 50 for convergence visualization is recommended based on the present study; and (3) comparing emergent policy structure under systematically varied representational conditions.

Principled Redesign Under RRT

The quiz-dominant policy — alongside the underrepresentation of review actions (29.5–30.7% across configurations) — indicates specific design gaps that RRT makes empirically visible. Redesign of the PedagoReLearn environment would require four targeted modifications. First, the state representation should be augmented to include temporal variables such as time-since-last-retrieval or stability confidence levels, replacing reliance on static mastery states alone. Second, transition dynamics should incorporate probabilistic decay functions to model forgetting over time, consistent with established Ebbinghaus decay modeling [7]. Third, reward logic should be modified to include delayed bonuses for maintained mastery or penalties for regression, thereby differentiating short-term correctness from long-term retention and aligning the reward structure with the spacing and retrieval practice literature [4], [6]. Fourth, recency weighting should be introduced to adjust action valuation in favor of spaced retrieval over repeated teaching sequences.

These modifications would allow the agent's policy to become sensitive to temporal and stability dynamics rather than optimizing exclusively for immediate mastery advancement — shifting the system from episodic efficiency toward pedagogically grounded sequencing.

From a design perspective, these findings underscore a critical implication for engineering and STEM education: adaptive tutoring systems must explicitly encode cognitive principles such as knowledge stability, forgetting, and temporal spacing if they are to support durable learning [7]. Reward shaping alone is insufficient to induce pedagogically meaningful sequencing when underlying state representations lack temporal sensitivity. More broadly, the results position reinforcement learning as a structured diagnostic instrument for instructional design — one through which researchers can systematically evaluate whether theoretical learning principles are meaningfully operationalized within adaptive systems.

Implications for Learning Analytics and Classroom Telemetry

The findings in this study carry direct implications for how learning analytics pipelines are designed and evaluated. Classroom telemetry — including response accuracy, response latency, and action history — can be formally mapped to instructional state representations in adaptive systems [2]. In practical analytics workflows, learner state may be derived from Learning Management System (LMS) mastery metrics or Bayesian Knowledge Tracing (BKT) posterior probabilities that estimate current knowledge levels [12]. Reward signals can be computed from assessment correctness, optionally weighted by response latency to capture efficiency and fluency. Recency variables emerge naturally from timestamped interaction logs, enabling systems to model temporal decay or knowledge stability. Within this framework, the teach action corresponds to logged content exposure events (e.g., lecture views, module completions), while the review action maps to retrieval practice events such as spaced practice modules or low-stakes assessments. Episode termination can be operationalized using course-level competency thresholds or mastery benchmarks defined within the LMS.

Critically, the observed quiz-dominant policy illustrates that when analytics pipelines omit constructs such as knowledge stability or temporal decay, adaptive systems will optimize toward whichever action the immediate reward structure most consistently reinforces — in this case, quiz actions yielding correct-response rewards — at the expense of pedagogically balanced sequencing. This positions reinforcement learning not only as an optimization mechanism but as a diagnostic tool for evaluating the pedagogical adequacy of learning data pipelines and the completeness of the constructs they encode. Standards such as xAPI and IMS Caliper provide infrastructure pathways for operationalizing the temporal telemetry streams that would support more complete environment specifications in production adaptive systems.

These findings are specific to the instructional representations and reward structures examined in this simulated environment, and they should be interpreted as diagnostic rather than generalizable across all adaptive tutoring systems or RL algorithms.

Conclusion and Future Work

This work-in-progress examines how RL-based tutoring agents behave when instructional environments omit explicit representations of knowledge stability and forgetting. Rather than evaluating learner outcomes directly, the study treats emergent agent behavior as a diagnostic signal of the pedagogical structures encoded in the environment — demonstrating that the RL agent reduced steps-to-mastery from 37.33 to 25.25 ± 1.13 (aggregated-state) and 26.14 ± 2.02 (full-state) while converging on a quiz-dominant instructional policy (36.1–38.6% quiz selections) inconsistent with established principles of durable learning [4], [6], [7].

Interpreted through RRT [10], these results underscore a central insight: reinforcement learning optimizes the pedagogy the environment makes possible. When temporal stability and decay are not formally represented, instructional actions become functionally indistinguishable, and the agent defaults to short-term efficiency over pedagogically grounded sequencing. This reframes the challenge of adaptive tutoring design. The critical question is not whether RL can discover effective teaching strategies, but whether instructional environments encode the cognitive structures necessary for pedagogically meaningful sequencing to emerge — a condition this study demonstrates is not satisfied by immediate reward structures alone.

RQ1 asked what instructional behaviors emerge when RL is applied to adaptive tutoring without explicit modeling of knowledge stability and forgetting. The answer is a quiz-dominant policy driven by immediate reward availability, with teach and review actions distributed below equal frequency — stable across seeds, configurations, and training durations.

RQ2 asked what representational and reward design conditions are required for RL agents to exhibit pedagogically meaningful instructional sequencing. The findings indicate that static mastery-level state representations and immediate-correctness reward structures are insufficient. Meaningful sequencing requires explicit temporal variables, probabilistic decay dynamics, and differentiated reward consequences that distinguish short-term correctness from long-term retention.

Future work will incorporate explicit stability modeling and decay dynamics into the PedagoReLearn environment, enabling differentiated instructional actions with long-term consequences [6], [7]. Subsequent studies will examine whether these refinements produce sequencing strategies aligned with spaced retrieval and durable learning principles across alternative RL algorithms — including Q-learning and actor–critic methods — and assess whether algorithmic choice influences instructional emergence independently of environment specification. The diagnostic framework established here will also be extended to cross-cultural engineering and STEM contexts, including veteran learner populations navigating technical transitions in U.S. and international university settings, where adaptive sequencing and durable retention are particularly critical.

Bridging theory and practice, full deployment of cognitively grounded adaptive tutoring systems will require LMS-level instrumentation capable of capturing not only correctness but temporal engagement patterns. As noted in the Discussion, standards such as xAPI and IMS Caliper provide established infrastructure pathways for operationalizing these telemetry streams within production adaptive systems. By integrating RRT-grounded RL environments with real-world learning analytics infrastructure, adaptive tutoring systems can move beyond episodic efficiency toward designs that intentionally support long-term retention — a foundational criterion of educational effectiveness in engineering and STEM learning contexts.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction,* 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[2] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto, "Faster teaching via POMDP planning," *Cognitive Science,* vol. 40, no. 6, pp. 1290–1332, 2016.

[3] T. Mandel, Y. Liu, E. Brunskill, and Z. Popović, "Offline policy evaluation across representations with applications to educational games," in *Proc. 13th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS),* Paris, France, 2014, pp. 1429–1436.

[4] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," *Psychological Bulletin,* vol. 132, no. 3, pp. 354–380, 2006.

[5] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive Science*, vol. 36, no. 5, pp. 757–798, 2012.

[6] P. I. Pavlik Jr. and J. R. Anderson, "Using a model to compute the optimal schedule of practice," *Journal of Experimental Psychology: Applied,* vol. 14, no. 2, pp. 101–117, 2008.

[7] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology,* H. A. Ruger and C. E. Bussenius, Trans. New York, NY, USA: Teachers College Press, 1913. (Original work published 1885.)

[8] J. Dewey, *Experience and Education.* New York, NY, USA: Macmillan, 1938.

[9] D. A. Schön, *The Reflective Practitioner.* New York, NY, USA: Basic Books, 1983.

[10] T. F. Hallmark, "Reflexive Reciprocity Theory: Toward a Framework for Durable Pedagogical Change in AI-Supported Learning Environments," *EdArXiv,* 2025. doi: 10.35542/osf.io/zkyht_v1.

[11] E. Brunskill and E. L. Brunskill, "Towards understanding pedagogical decision-making in reinforcement learning," in *Proc. 10th Int. Conf. Educational Data Mining (EDM)*, Wuhan, China, 2017, pp. 1–10.

[12] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction,* vol. 4, no. 4, pp. 253–278, 1994.