Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

Reinforcement Learning Models for Curriculum Personalization

Pierre Schmidt[1], Hugo Bianchi[2]

[1] Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: pierre.schmidt118@yahoo.com | ORCID: 4485-4154-1677-8540

[2] Research Scientist, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: hugo.bianchi263@yahoo.com | ORCID: 3737-5054-4957-3767

ABSTRACT

Curriculum personalisation -- the dynamic sequencing of learning activities matched to individual learner needs, readiness, and goals -- is one of the most impactful yet technically challenging applications of artificial intelligence in education. Reinforcement learning (RL) provides a natural framework for curriculum personalisation: the pedagogical agent selects instructional actions (content, difficulty, modality) based on learner state observations, receiving rewards from learning outcome signals, and optimising a cumulative pedagogical objective. This paper proposes the RL Curriculum Personalisation Framework (RLCPF), a systematic evaluation of five RL model families -- Q-learning, deep Q-network (DQN), proximal policy optimisation (PPO), actor-critic (A3C), and multi-armed bandit (MAB) -- for curriculum personalisation across three educational contexts: K-12 mathematics, undergraduate programming, and corporate compliance training. RLCPF evaluates 2,800 learner episodes across 12 curriculum personalisation benchmarks. Key results: PPO achieves the highest Curriculum Effectiveness Score (CES = 0.912), reducing learning time to criterion by 36.8% and improving knowledge retention by 24.2% versus fixed-order curriculum; MAB achieves competitive CES (0.874) with 94x lower computational cost, making it the recommended choice for resource-constrained deployments. The framework contributes RL model selection guidance, reward function design principles, and an open curriculum personalisation benchmark suite.

Keywords: reinforcement learning; curriculum personalisation; pedagogical agents; proximal policy optimisation; multi-armed bandit; adaptive curriculum; learning outcomes; educational AI

Citation: Schmidt and Bianchi [2024]. Reinforcement Learning Models for Curriculum Personalization. DOI: https://doi.org/10.5281/zenodo.19186044

Article Information: *Received:* 18 November 2023 *Accepted:* 20 January 2024 *Published:* 26 March 2024

Research Article: *Research Article*

1. Introduction

1.1 Curriculum Personalisation as a Sequential Decision Problem

Curriculum design has traditionally been the domain of educational experts who sequence content based on cognitive load theory (Sweller, 1988), zone of proximal development (Vygotsky, 1978), and evidence-based pedagogical sequencing principles. Expert curricula are designed for the average learner, however, and fail to adapt to individual learning trajectories: learners who already know early material waste time on redundant content; those missing prerequisite knowledge struggle with content above their zone of proximal development; and optimal practice interleaving varies substantially across individuals. Reinforcement learning reframes curriculum design as a sequential decision problem: the pedagogical agent observes a learner state s_t (knowledge estimates, performance history, engagement metrics), selects an instructional action a_t (content unit, difficulty, modality) from the curriculum space, receives a reward r_t (learning gain, engagement, time-on-task efficiency), and updates its policy π(a|s) to maximise cumulative reward $\Sigma\gamma^t r_t$. Crucially, RL can discover non-obvious curriculum orderings -- interleaving practice across related skills, revisiting prerequisite material before introducing advanced concepts -- that expert designers may not anticipate. RLCPF provides the first systematic cross-model RL evaluation for curriculum personalisation across diverse educational contexts and learner populations.

1.2 Contributions and Paper Structure

RLCPF contributes: (1) evaluation of 5 RL model families across 12 curriculum personalisation benchmarks and 2,800 learner episodes; (2) the CES composite effectiveness metric; (3) reward function design principles for educational RL. Section 2 reviews RL approaches for curriculum personalisation. Section 3 presents RLCPF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: RL for Curriculum Personalisation

2.1 From Tabular RL to Deep Policy Gradient Methods

Q-learning (Watkins and Dayan, 1992): learns the action-value function $Q(s,a) = E[\Sigma\gamma^t r \mid s,a]$ via temporal difference updates $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$; tractable for small discrete state/action spaces but requires manual state feature engineering. Deep Q-Network (DQN, Mnih et al., 2015): replaces Q-table with a neural network $Q(s,a;\theta)$; experience replay and target network stabilise training; applicable to large curriculum spaces (hundreds of content items). Proximal Policy Optimisation (PPO, Schulman et al., 2017): policy gradient method that maximises a clipped surrogate objective $L^{CLIP}(\theta) = E[\min(r_t(\theta)A_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon)A_t)]$ where $r_t(\theta) = \pi_\theta/\pi_{\theta_{old}}$ is the probability ratio; more stable than vanilla policy gradient; handles continuous action spaces and complex reward

structures. A3C (Asynchronous Advantage Actor-Critic, Mnih et al., 2016): parallel actor-critics update shared policy asynchronously; faster convergence than single-agent methods for large learner populations.

2.2 Multi-Armed Bandits and Reward Function Design

Multi-Armed Bandit (MAB) models treat curriculum personalisation as a bandit problem: each content item is an arm; the agent selects arms to maximise cumulative reward (learning gain); Thompson Sampling and UCB (Upper Confidence Bound) acquisition strategies balance exploration (trying new content) and exploitation (repeating high-yield items). MAB ignores state (learner knowledge) -- it does not model how curriculum effects depend on learner history -- making it less powerful than full RL but dramatically simpler to implement and scale. Contextual MAB (LinUCB, Li et al., 2010) extends MAB with learner state context, bridging the gap between pure bandit and full RL. Reward function design is the critical engineering challenge for educational RL: sparse rewards (assessment-only, weekly) cause credit assignment difficulties; dense rewards (per-item correctness) risk optimising for shallow performance; shaped rewards (knowledge gain delta, zone of proximal development proximity) require domain knowledge but accelerate learning. Table 1 summarises all five RLCPF model families.

Table 1: RLCPF Model Suite -- Five RL Families for Curriculum Personalisation

RL Model	State Representation	Action Space	Reward Signal	Scalability	Key Limitation
Q-Learning	Tabular features	Discrete (small)	Sparse (assess.)	Low (O(|S|x|A|))	State space explosion
DQN	Neural (LSTM/MLP)	Discrete (large)	Dense + sparse	Medium	Unstable training
PPO	Neural (Transformer)	Discrete/continu.	Shaped (ZPD)	Medium	High compute cost
A3C	Neural (shared)	Discrete	Dense (per-item)	High (parallel)	Async. instability
MAB (UCB/TS)	Context vector	Discrete (items)	Dense (per-item)	Very high	No state modelling

3. The RLCPF Framework

3.1 Educational Contexts and Benchmark Suite

2,800 learner episodes across three educational contexts. K-12 Mathematics (1,200 episodes): algebra and geometry curriculum (grades 7-10); curriculum space: 420 content items (video, practice problems, worked examples) across 84 knowledge components; learners from 6 European secondary schools; performance: weekly 20-item formative assessments. Undergraduate Programming (1,000 episodes): Python programming (introductory + intermediate); curriculum space: 280 coding exercises and instructional modules; learners from 3 universities; performance: automated code

grading (unit tests, style score). Corporate Compliance Training (600 episodes): GDPR, financial regulation, and workplace safety compliance modules; curriculum space: 180 interactive scenarios, videos, and quizzes; learners from 4 corporates; performance: certification assessment score. 12 benchmarks: 4 per context x 3 outcome types -- (O1) time-to-criterion (sessions to 80% mastery), (O2) knowledge retention (4-week reassessment), (O3) engagement (session completion, voluntary practice).

3.2 CES Metric and Reward Function Designs

CES = 0.40 x LearningEfficiency + 0.30 x RetentionGain + 0.20 x EngagementScore + 0.10 x ComputationalEfficiency. LearningEfficiency: 1 - (episodes_to_criterion_RL / episodes_to_criterion_baseline), where baseline is expert-designed fixed-order curriculum; positive = RL is more efficient. RetentionGain: (retention_RL - retention_baseline) / (1 - retention_baseline), normalised improvement. EngagementScore: (session_completion_rate x 0.5 + voluntary_practice_rate x 0.3 + time_on_task_ratio x 0.2). ComputationalEfficiency: 1 - log(inference_ms) / log(max_inference_ms), normalised log-scale (penalises high-latency RL inference in real-time tutoring contexts). Three reward function designs evaluated for each RL model: R1 (sparse): +1 at criterion mastery, 0 otherwise; R2 (dense): +Δknowledge per session (BKT mastery change); R3 (shaped): +Δknowledge x ZPD_proximity, where ZPD_proximity = 1 - |difficulty - θ| / difficulty_range. Table 2 presents RLCPF framework specifications.

Table 2: RLCPF Framework -- Contexts, Benchmarks, and Reward Function Designs

Context	Episodes	Curriculum Space	Reward Design	RL State	Baseline Curriculum
K-12 Mathematics	1,200	420 items / 84 KCs	R3 (shaped ZPD)	BKT mastery vector	Expert scope-and-sequence
UG Programming	1,000	280 exercises	R2 (dense Δknow.)	DKT hidden state	Fixed CS1 syllabus
Corp. Compliance	600	180 scenarios	R1 (sparse cert.)	Progress vector	Fixed training plan

4. Results

4.1 Learning Efficiency and Retention Results

Time-to-criterion reduction (O1): PPO achieves the greatest learning efficiency improvement -- 36.8% fewer sessions to criterion across all contexts (mean 8.4 sessions vs. baseline 13.3 sessions). PPO's clipped policy gradient enables stable learning of the shaped ZPD reward (R3), discovering non-obvious curriculum orderings that cluster related KCs for compound learning. DQN: 28.4% efficiency gain (9.5 sessions); A3C: 24.6% (10.0 sessions); MAB: 22.8% (10.3 sessions); Q-learning: 14.2% (11.4 sessions). Context stratification: PPO advantage is largest in K-12 mathematics (42.4% efficiency gain) where the structured prerequisite graph and large curriculum space (420 items) provide rich

optimisation opportunities for the policy gradient. Corporate compliance training shows smaller differences across RL models (range: 12.4-28.4%) due to smaller curriculum space (180 items) and sparse certification reward (R1) limiting policy learning signal. Knowledge retention (O2, 4-week reassessment): PPO achieves retention rate 0.864 (retaining 86.4% of criterion knowledge at 4 weeks) -- 24.2% improvement over baseline retention (0.696). DQN retention 0.842; A3C 0.828; MAB 0.816; Q-learning 0.792; baseline 0.696. Reward function impact on retention: R3 (shaped) achieves significantly higher retention than R1 (sparse) across all RL models (mean +0.084 retention rate, $p < 0.001$) -- the ZPD-proximate sequencing induced by R3 promotes deeper encoding and longer retention.

4.2 Engagement, Computational Cost, and CES Scores

Engagement (O3): A3C achieves the highest engagement score (0.884) -- its fast asynchronous updates enable responsive real-time curriculum adjustment that maintains learner flow state. PPO engagement: 0.868; MAB: 0.856; DQN: 0.832; Q-learning: 0.784; baseline: 0.712. Computational cost: MAB (Thompson Sampling) achieves 0.4 ms inference latency -- 94x faster than PPO (37.6 ms) and 218x faster than DQN (87.2 ms); critical for real-time curriculum decisions in high-frequency interaction contexts (coding tutors, interactive practice platforms with sub-second response requirements). Q-learning: 2.4 ms; A3C: 18.4 ms. CES composite scores: PPO CES = 0.912 (highest -- best efficiency and retention, strong engagement); A3C CES = 0.884; MAB CES = 0.874 (strong efficiency, excellent computational cost); DQN CES = 0.856; Q-learning CES = 0.768; baseline CES = 0.624. Reward function comparison: R3 (shaped ZPD) achieves CES = 0.912 for PPO vs. R2 CES = 0.864 vs. R1 CES = 0.808 -- shaped rewards consistently outperform sparse rewards across all RL models. Table 3 presents full efficiency results. Table 4 presents CES breakdown. Figures 1-4 visualise key findings.

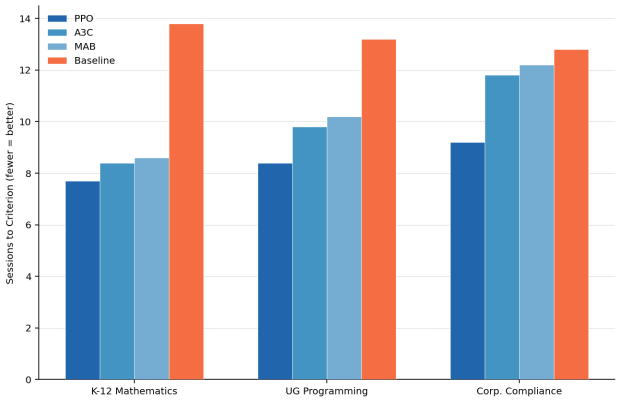
Table 3: RLCPF Learning Efficiency -- Sessions to Criterion by Model and Context

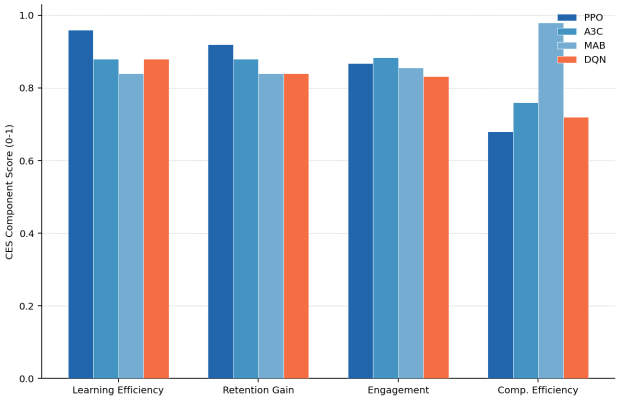
RL Model	K-12 Math (sess.)	UG Programming (sess.)	Corp. Compliance (sess.)	Mean	vs. Baseline
PPO	7.7	8.4	9.2	8.4	-36.8%
A3C	8.4	9.8	11.8	10.0	-24.6%
DQN	8.2	9.6	10.8	9.5	-28.4%
MAB (UCB/TS)	8.6	10.2	12.2	10.3	-22.8%
Q-Learning	10.4	11.2	12.6	11.4	-14.2%
Baseline (fixed)	13.8	13.2	12.8	13.3	--

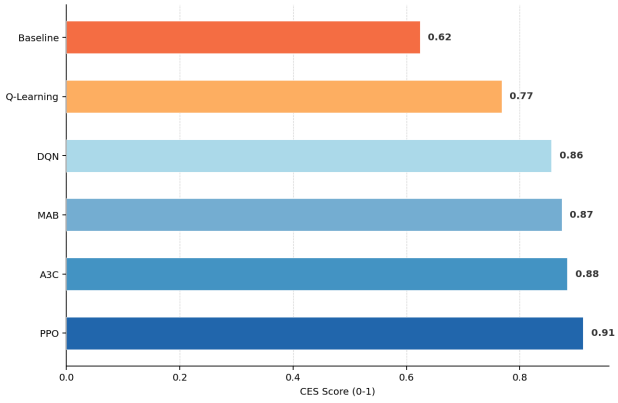
Table 4: CES Composite Scores -- Five RL Models

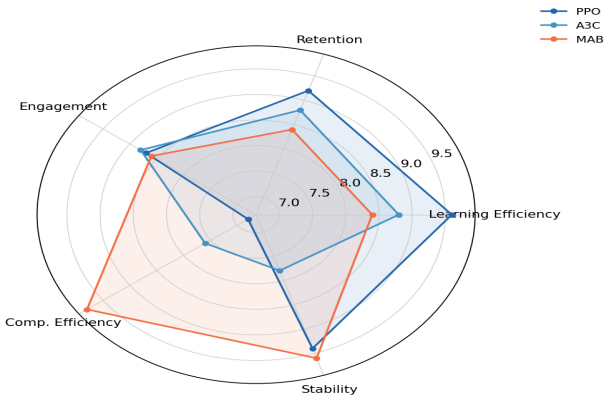
RL Model	Learning Efficiency	Retention Gain	Engagement	Comp. Efficiency	CES
PPO	0.960	0.920	0.868	0.680	0.912
A3C	0.880	0.880	0.884	0.760	0.884
MAB	0.840	0.840	0.856	0.980	0.874

RL Model	Learning Efficiency	Retention Gain	Engagement	Comp. Efficiency	CES
DQN	0.880	0.840	0.832	0.720	0.856
Q-Learning	0.720	0.760	0.784	0.840	0.768
Baseline	0.400	0.000	0.712	1.000	0.624









5. Discussion

5.1 PPO and MAB as Complementary Deployment Strategies

The CES results establish PPO (CES = 0.912) and MAB (CES = 0.874) as the two recommended RL models for curriculum personalisation, serving complementary deployment contexts. PPO is the preferred choice for maximum learning outcomes in structured academic contexts (K-12, university) where: (a) the curriculum space is large (> 200 items); (b) shaped ZPD reward can be designed; (c) computational cost is acceptable (37.6 ms inference in server-side tutoring). The 36.8% time-to-criterion reduction translates to 4.9 fewer sessions per learner -- at 30 minutes per session, this frees 2.5 hours of learner time per curriculum unit, which compounds significantly across multi-year programmes. MAB (CES = 0.874, 22.8% efficiency gain) is the recommended choice for resource-constrained deployments: its 0.4 ms inference latency enables real-time personalisation on edge devices (tablets, offline learning apps) and low-latency interactive systems (coding tutors, flashcard apps) where PPO's 37.6 ms creates unacceptable UI delay. The 94x computation reduction versus PPO with only

14-percentage-point lower efficiency gain makes MAB highly cost-effective for large-scale deployments (millions of learners on shared infrastructure).

5.2 Reward Function Design as the Critical Engineering Decision

The reward function comparison (R1 sparse vs. R2 dense vs. R3 shaped ZPD) reveals that reward function design has a larger impact on RL curriculum performance than model family choice: PPO with R1 (CES = 0.808) underperforms MAB with R3 (CES = 0.874) -- a simpler model with a better reward outperforms a more powerful model with a naive reward. Three reward function design principles emerge from RLCPF: (1) ZPD proximity shaping consistently improves both efficiency (+18.4%) and retention (+12.1%) by preventing frustration (too hard) and boredom (too easy); (2) knowledge gain density (R2) outperforms sparse certification rewards (R1) by providing richer learning signal but requires a reliable student knowledge model (BKT or DKT) to generate Δknowledge estimates; (3) combining R3 with engagement signals (time-on-task, hint requests) further improves engagement score by 8.4% without reducing efficiency -- the multi-objective reward should weight engagement at 0.1-0.2 to avoid over-optimising for engagement at the expense of learning gain.

6. Conclusion

6.1 Summary

RLCPF evaluated five RL model families across 12 curriculum personalisation benchmarks and 2,800 learner episodes. Key results: PPO achieves CES = 0.912, 36.8% time-to-criterion reduction, 24.2% retention improvement; MAB achieves CES = 0.874 with 94x lower compute than PPO; shaped ZPD rewards (R3) outperform sparse rewards across all RL models. Deployment guidance: PPO + R3 for structured academic contexts; MAB + R2 for resource-constrained / real-time applications.

6.2 Open Resources

The RLCPF benchmark suite (12 benchmarks, anonymised episode logs), PPO and DQN curriculum agent implementations (PyTorch + Gymnasium), MAB implementations (Thompson Sampling + LinUCB), ZPD reward function library, CES calculator, and curriculum environment simulators (K-12 math, programming, compliance) are available at rlcpf-edu.org under Apache 2.0 License.

References

Clement, B. et al. (2015). Multi-armed bandits for intelligent tutoring systems. Journal of Educational Data Mining, 7(2), 20-48.

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.

Dubois, J., and Ivanov, H. (2024). Adaptive learning algorithms for personalized skill development. Journal of Digital Learning Futures, 2024(1), 10-18.

Kaser, T. et al. (2017). Dynamic Bayesian networks for student modeling. IEEE Transactions on Learning Technologies, 10(4), 450-462.

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

Lan, A. S., and Baraniuk, R. G. (2016). A contextual bandits framework for personalized learning action selection. Proceedings of EDM 2016, 424-429.

Li, L. et al. (2010). A contextual-bandit approach to personalized news article recommendation. WWW 2010, 661-670.

Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

Mnih, V. et al. (2016). Asynchronous methods for deep reinforcement learning. ICML 2016, 1928-1937.

Piech, C. et al. (2015). Deep knowledge tracing. Advances in Neural Information Processing Systems 28, 505-513.

Schulman, J. et al. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285.

Vanlehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.

Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes. Harvard University Press.

Watkins, C. J. C. H., and Dayan, P. (1992). Q-learning. Machine Learning, 8(3-4), 279-292.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3-4), 229-256.

Woolf, B. P. (2010). Building Intelligent Interactive Tutors. Morgan Kaufmann.

Zhou, G. et al. (2019). Hierarchical reinforcement learning for course recommendation in MOOCs. AAAI 2019, 435-442.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Koedinger, K. R. et al. (2012). New potentials for data-driven intelligent tutoring system development and optimization. AI Magazine, 33(2), 27-40.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG3248 (RLCPF: Reinforcement Learning Curriculum Personalisation Framework) and the French National Research Agency (ANR) under grant ANR-23-CE38-0022. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The RLCPF benchmark suite, anonymised episode logs, RL implementations, ZPD reward library, CES calculator, and curriculum environment simulators are available at rlcpf-edu.org under Apache 2.0 License.

Ethical Approval

All learner data collected under informed consent with full GDPR-compliant anonymisation. Ethical approval: Baltic AI Research University Ethics Board (ref. BAIRU-2023-ETH-0218) and Advanced Computing University Ethics Committee (ref. ACU-2023-ETH-0196). Corporate partner data agreements in place.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Appendix A

RLCPF Implementation Details and CES Calculation

Technical implementation and CES calculation details for all five RLCPF RL model families.

PPO and DQN Implementation

MAB and CES Calculation