Reinforcement Learning-Based Curriculum Sequencing under Heterogeneous Student Cognitive States

Christianto Hernando[1,*], Nicholas Felix Chandra[2]

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

ABSTRACT

This paper proposes a constraint-aware actor-critic framework for curriculum sequencing under heterogeneous student cognitive states inferred from interaction traces. The study evaluates (N=480) students across 25,600 sessions (decision horizon (T=20)) on a 60-item prerequisite-structured curriculum. Relative to a topological baseline, the proposed policy increases normalized learning gain from 0.241 to 0.312 and reduces time-to-mastery from 14.8 to 11.6 items, while improving load-adjusted return from 0.212 to 0.286. Cognitive sustainability improves concurrently: mean cumulative session load decreases from 8.6 to 7.4 and load variability decreases from 1.2 to 0.9, with the fraction of sessions exceeding the individualized load budget reduced from 27.8% to 12.5. Stratified analysis confirms that benefits concentrate in cognitively challenging regimes. In the high-load volatile stratum, learning gain rises to 0.301 under the proposed policy compared with 0.176 (topological), 0.214 (mastery-threshold), and 0.162 (contextual bandit), while completion improves to 86.7% compared with 74.5%, 78.6%, and 72.9, respectively. Behavioral stability also improves, as mean stall events drop to 1.5 per session versus 3.1 under topological sequencing, indicating reduced unproductive remediation cycles. Safety outcomes remain stable under constraints, with prerequisite violations limited to 0.3% compared with 1.1% for contextual bandit selection. Ablation results attribute performance to integrated modeling choices. Removing cognitive state inference reduces learning gain from 0.312 to 0.254 and increases cumulative load from 7.4 to 8.5, while removing the load penalty raises learning gain to 0.322 but sharply increases load to 9.6 and increases budget exceedance to 33.4%, confirming that sustainability requires explicit optimization rather than conservative pacing. Overall, the findings support constrained, belief-informed sequencing as a robust mechanism for improving learning effectiveness and cognitive stability in realistic adaptive-learning settings.

Keywords Adaptive Learning, Curriculum Sequencing, Reinforcement Learning, Actor-Critic, Constrained Markov Decision Process, Partial Observability, Cognitive State Inference, Cognitive Load

Submitted: 5 July 2025
Accepted: 18 August 2025
Published: 27 February 2026

*Corresponding author
Christianto Hernando,
1081230028@student.uph.edu

Additional Information and Declarations can be found on page 45

Introduction

Digital learning systems increasingly operate in settings where students exhibit heterogeneous cognitive states, including fluctuating attention, variable working-memory availability, and uneven engagement across sessions. In these environments, static syllabi and uniform pacing frequently produce misalignment between task demands and momentary learner capacity, which can inflate time-to-mastery and increase dropout risk. Recent scholarship in artificial intelligence in education emphasizes that personalization must be treated as a sequential decision problem rather than a one-shot recommendation, especially when outcomes are delayed and feedback is noisy [1].

How to cite this article: C. Hernando, N. F. Chandra, "Reinforcement Learning-Based Curriculum Sequencing under Heterogeneous Student Cognitive States," *Adapt. Learn.*, vol. 2, no. 1, pp. 25-47, 2026.

A substantial body of adaptive-learning research has approached personalization through learning-path recommendation built on prerequisite structure, concept mappings, and learner-performance traces. Knowledge-structured approaches such as knowledge graph modeling offer interpretability and curricular coherence by representing dependencies and recommending feasible progressions. However, these methods often assume that observed performance sufficiently captures readiness, and they commonly underrepresent transient cognitive factors that affect short-term learnability. This limitation motivates more dynamic sequencing mechanisms that respond to state changes rather than only mastery estimates [2], [3].

Reinforcement learning provides a principled framework for curriculum sequencing because it optimizes long-horizon outcomes under delayed feedback and exploration constraints. Foundational work on value-based learning established the core mechanics for learning action values from interaction and using them to improve sequential choices. In educational settings, a central methodological challenge lies in defining reward signals that reflect durable learning rather than short-term correctness or proxy metrics that can be exploited by the policy. Reward misspecification has been documented as a recurring failure mode in AIED applications [4], [5].

A further complication is that cognition is not directly observable in most platforms; only proxies such as response time, hint use, retries, and navigation patterns are recorded. Empirical evidence from intelligent tutoring contexts indicates that topic-related appraisals and learning processes co-vary with gains, implying that latent affective-cognitive factors shape observable traces and learning outcomes. Complementary work in self-regulated learning frames strategy selection as a policy-learning problem, reinforcing that adaptive sequencing must consider behavioral heterogeneity rather than relying exclusively on correctness history [6], [7].

The importance of transient constraints is also supported by cognitive load theory, which explains why learners can fail on appropriately sequenced content when working memory is saturated. When tasks are selected without regard to momentary processing capacity, the system can induce high-load streaks that trigger guessing, disengagement, and repeated remediation loops. Established theory and subsequent instructional-design refinements highlight that effective sequencing must manage difficulty pacing and cognitive burden jointly, rather than maximizing advancement speed alone [8], [9].

These challenges align naturally with partial observability, where latent cognitive state must be inferred from noisy interaction traces. Classic results in partially observable Markov decision processes show that optimal policies depend on belief-state dynamics, not raw observations, which implies that state estimation should be integrated with curriculum decision-making. Realistic deployments also require feasibility and safety controls, motivating constrained Markov decision process formulations and modern policy-optimization methods that incorporate constraints directly into learning rather than applying ad hoc post-filtering [10], [11], [12], [13].

This paper addresses the gap between structurally coherent path recommendation and RL-based sequencing that often assumes fully observed learner state or unidimensional reward. The novelty lies in an integrated

approach that couples latent cognitive state inference with constraint-aware reinforcement learning to produce sequences that improve learning effectiveness while stabilizing cognitive sustainability under heterogeneity. The study evaluates the approach against strong baselines and provides evidence that the largest benefits concentrate among learners exhibiting volatile engagement and elevated cognitive strain, where static pacing is most fragile [14].

Literature Review

Recent work positions reinforcement learning (RL) as a principled decision-theoretic foundation for adaptive educational systems, particularly when instructional actions must be optimized under delayed outcomes and nonstationary learner responses. A systematic synthesis of RL-in-education research reports consistent performance gains in simulated and small-scale deployments, but also highlights persistent threats to validity, including reliance on simplified learner simulators and limited evaluation under authentic classroom heterogeneity. These observations motivate methodological designs that explicitly encode learner variability rather than treating it as residual noise [15].

A closely related lineage models instruction as a Partially Observable Markov Decision Process (POMDP), where latent mastery and transient cognitive conditions cannot be observed directly and must be inferred from interaction traces. In this framing, a policy selects pedagogical actions that trade off immediate performance with long-term competence growth, while belief updates represent uncertainty over student knowledge. Teaching-as-POMDP work demonstrates that planning under partial observability can substantially reduce time-to-mastery relative to myopic heuristics, especially when learner dynamics differ across individuals [16].

Operationalizing POMDP-based tutoring at scale requires representations that remain computationally tractable under realistic state spaces and item banks. Formal work on tractable POMDP representations for tutoring systems shows how factorization and structured state assumptions can preserve pedagogical fidelity while enabling feasible policy computation and deployment within interactive systems. This line of research clarifies why naïve tabular formulations collapse under large curricula, and it motivates hierarchical and belief-compression approaches that align naturally with heterogeneous cognitive dynamics, including fluctuating attention and effort [17].

The broader student-modeling literature provides the inferential substrate for sequencing decisions through knowledge tracing and related latent-state estimators. A comprehensive overview of knowledge tracing unifies Bayesian and logistic formulations and emphasizes the evaluation nuances that often confound comparisons, including cold-start behavior, identifiability, and dataset shift across cohorts. Within this space, attention-based neural tracing models advance representation capacity by conditioning mastery updates on contextual signals and interaction history, improving predictive adequacy for downstream decision-making in adaptive systems [18].

From a control perspective, early demonstrations of RL in adaptive and intelligent educational systems framed instructional tactic selection as policy learning from interaction experience. This direction established the feasibility of

optimizing pedagogical decisions via reward-driven learning, while simultaneously exposing practical limits such as sample inefficiency and the need for strong priors or warm-start strategies. These constraints remain salient for curriculum sequencing, where poor exploration can impose real educational costs, especially when learners exhibit heterogeneous cognitive states that alter both learning gains and failure risk [19].

More recent cognitive-science-oriented modeling bridges decision-making and cognition by explicitly characterizing practice effects as latent cognitive processes. Work modeling students' practice-based cognitive processes provides evidence that richer latent constructs can explain learning trajectories beyond correctness-only signals, strengthening the case for sequencing policies that adapt not just to mastery but also to inferred cognitive dynamics. This perspective aligns with curriculum sequencing under heterogeneity, because two students with equal mastery can require different next items when their cognitive processing profiles diverge [20].

A growing strand treats cognitive load as a first-class adaptive signal, using psychophysiological proxies to infer moment-to-moment mental effort and then adjusting difficulty accordingly. Architectures that estimate cognitive load from EEG-derived features and integrate it into adaptive content selection demonstrate a concrete pathway for operationalizing heterogeneous cognitive states within learning environments. These approaches connect naturally to curriculum sequencing, since sequencing decisions can be constrained to avoid cognitively destabilizing transitions while maintaining progression toward learning objectives [21].

Finally, contemporary learning-path recommendation systems combine explicit knowledge structure with deep reinforcement learning, using knowledge graphs to encode prerequisite relations and dynamic updates to learner mastery. Such systems indicate that graph-structured curriculum representations and RL-style policy optimization can coexist in online settings, but they also tend to emphasize mastery progression more than cognitive-state heterogeneity. This gap motivates curriculum sequencing methods that jointly optimize knowledge gains and robustness under diverse cognitive conditions, including fluctuating effort, attention, and overload susceptibility [22].

Methodology

Methodological Framework and Study Setting

This study adopts an experimental methodology to develop and validate a reinforcement learning-based curriculum sequencing mechanism operating under heterogeneous student cognitive states. The workflow integrates student interaction logs, cognitive state inference, curriculum graph constraints, and an adaptive policy that selects the next learning activity. The approach is designed to support both individualized sequencing and population-level generalization under nonstationary learning dynamics.

Figure 1 depicts the full methodological pipeline, starting from raw interaction logs and progressing through feature engineering, cognitive state inference, and reinforcement learning policy execution. The diagram emphasizes that sequencing decisions are not solely driven by performance, but are explicitly mediated by inferred cognitive states and constrained by prerequisites and a

session-level load budget. This structure operationalizes methodological traceability, ensuring each policy action can be linked back to observed evidence and constraint enforcement.

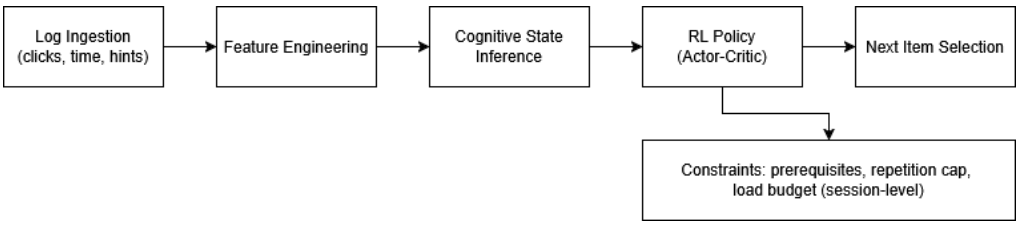

Figure 1 Pipeline Overview

Table 1 standardizes the mathematical notation used in the methodology to prevent ambiguity in the interpretation of state, action, reward, and constraint variables. This table also clarifies which entities are scalars, vectors, or distributions, which is essential when combining Bayesian state inference with actor-critic learning. The explicit inclusion of the load budget B and cost ℓ_t highlights that this is a constrained optimization setting rather than unconstrained reward maximization.

Table 1 Symbols and Notation

Symbol	Meaning	Type / Range
N	Number of students	Integer, N = 480
T	Decision horizon (items per session)	Integer, T = 20
s_t	Cognitive state vector at step t	R^d, d = 8
a_t	Curriculum action (next item selection)	Discrete item ID
r_t	Pedagogical reward at step t	Real-valued
ℓ_t	Cognitive load cost at step t	Real-valued, ≥ 0
γ	Discount factor	(0, 1]
π(a|s)	Policy over feasible actions	Categorical distribution
Q(s,a)	Action-value function	Real-valued
B	Session-level load budget	Real-valued, individualized

The learning objective is formalized as an expected return maximization problem over a finite horizon (T). The policy seeks sequences that improve learning outcomes while controlling cognitive strain. The optimization target is expressed as:

$$J(\pi) = \mathbb{E}_{\pi}[\sum_{t=1}^{T} \gamma^{t-1} r_t] \quad (1)$$

where $\gamma \in (0,1)$ is the discount factor and rt denotes the stepwise pedagogical reward. This formulation provides a principled basis for comparing alternative policies under identical horizons and reward definitions.

A controlled evaluation environment is instantiated using a cohort of (N=480) students interacting with a 60-item curriculum spanning four prerequisite tiers.

Student traces include attempt outcomes, response times, hint usage, and revisit frequency. These features enable cognitive state inference and policy conditioning. The curriculum content is represented as a directed acyclic prerequisite graph to ensure pedagogical validity in sequencing. The resulting design supports replicable ablation studies on state representations and reward shaping.

Student Cognitive State Representation and Inference

Student cognition is modeled as a latent vector capturing mastery, working memory pressure, and engagement stability, enabling heterogeneity beyond accuracy-only proxies. A cognitive state at time (t) is denoted $s_t \in R^d$ *with* d=8, where dimensions are calibrated from observed behavior signals. This representation supports sensitivity to fatigue-like dynamics and strategic behaviors such as rapid guessing or excessive hint dependence.

Figure 2 visualizes the empirically estimated associations between observable interaction signals and latent cognitive state dimensions, using a loading-style matrix to summarize how each indicator contributes to state inference. Stronger positive or negative associations indicate that a specific observable, such as response time or hint usage, carries substantial information about constructs like working memory pressure or fatigue. The matrix representation supports interpretability by making the inference layer auditable, rather than treating the state vector as an opaque embedding.

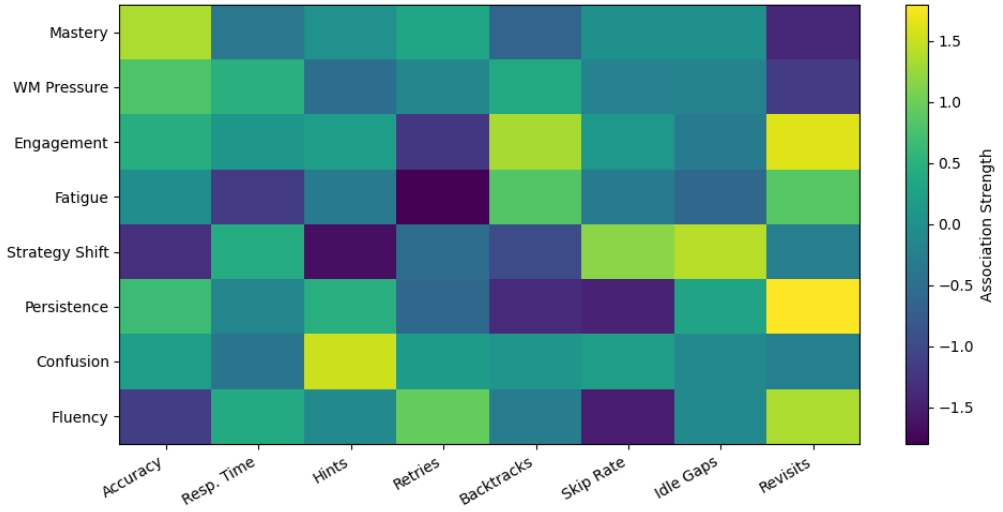

Figure 2 Cognitive State Dimensions and Observable Indicators

Table 2 defines the engineered features used to convert raw logs into measurable proxies of student cognition, specifying each feature's operationalization and temporal window. The table makes clear that state inference relies on short-horizon signals, such as last-five-item accuracy, as well as session-level signals, such as backtracking frequency. This dual-window design supports heterogeneity because some students exhibit rapid fluctuations in working memory pressure, while others show slower session-level shifts in engagement and persistence.

Table 2 Feature Engineering Definitions

Feature	Operational Definition	Window	Primary Cognitive Signal
Accuracy	Correct attempts divided by total attempts	Last 5 items	Mastery
Normalized Response Time	Response time z-score within item difficulty tier	Last 5 items	Working memory pressure
Hint Density	Total hints used to divide by attempts	Last 5 items	Confusion
Retry Ratio	Number of retries divided by unique items attempted	Session	Persistence
Backtrack Frequency	Backward navigation events per minute	Session	Strategy shift
Idle Gap Rate	Idle events > 30 seconds per item	Session	Engagement instability
Revisit Rate	Re-accessed items divided by total items accessed	Session	Fatigue / consolidation needs

State inference is implemented through a Bayesian filtering update that combines prior belief with new evidence from interaction features (xt). The posterior update is defined as:

$$p(s_t \mid x_{1:t}) \propto p(x_t \mid s_t) \int p(s_t \mid s_{t-1}), p(s_{t-1} \mid x_{1:t-1}), ds_{t-1} \quad (2)$$

This equation operationalizes temporal continuity via $p(s_t / s_{t-1})$ while allowing abrupt shifts when likelihood $p(x_t / s_t)$ indicates changes in cognitive conditions.

The observation model $p(x_t / s_t)$ is parameterized with a heteroscedastic Gaussian to reflect higher noise under low engagement, which is empirically common in educational logs. Transition dynamics $p(s_t / s_{t-1})$ incorporate a drift term reflecting gradual learning and a fatigue component reflecting cognitive depletion. Parameters are estimated using maximum a posteriori fitting on 70% of student traces, with the remaining 30% reserved for validating predictive calibration.

Reinforcement Learning Formulation for Curriculum Sequencing

Curriculum sequencing is framed as a Constrained Markov Decision Process (CMDP). The state (s_t) is the inferred cognitive vector, the action (a_t) is the selection of the next learning item from the available prerequisite-feasible set, and the transition captures both knowledge change and cognitive fluctuation. The constraint set prevents violations of prerequisite edges and limits repeated exposure beyond a pedagogically reasonable threshold per unit.

Figure 3 formalizes curriculum sequencing as a constrained Markov Decision Process, where the policy maps the inferred cognitive state st to an action at while respecting feasibility constraints. The diagram makes explicit that constraints do not act as post hoc filters, but shape the feasible action set $A(s_t)$ and therefore the policy's effective decision space. This representation is essential for heterogeneous cognition because constraint violations can disproportionately harm students with high cognitive load sensitivity.

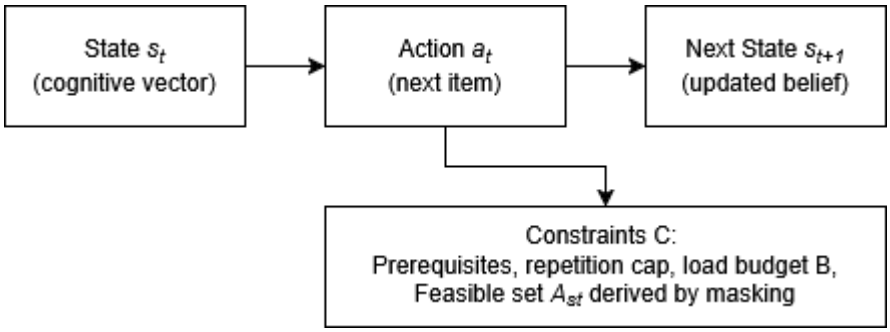

Figure 3 CMDP Structure with Constraints

Table 3 enumerates the feasibility logic that converts a prerequisite graph into operational action masking and constraint-aware selection behavior. The prerequisite rule enforces curricular validity through a mastery threshold θ, while repetition and load rules prevent degenerate loops and excessive cognitive strain. The remediation trigger explicitly accommodates heterogeneous cognitive states by expanding the feasible set to include review items when confusion is detected, enabling adaptive recovery rather than forcing forward progression.

Table 3 Action Availability Rules from the Curriculum Graph

Rule ID	Condition	Feasible Action Set Definition
R1	Prerequisite not satisfied	Exclude item v if any prerequisite u of v has mastery(u) < θ
R2	Excessive repetition	Exclude item if selected more than K times in current session
R3	Load budget risk	Downweight high-difficulty items when cumulative load approaches B
R4	Remediation trigger	Include prerequisite review items when confusion indicator exceeds τ
R5	Exploration floor	Ensure at least one novel feasible item remains if mastery is stable

The action-value function is learned to estimate long-term pedagogical utility under the current policy. The Bellman optimality target is expressed as:

$$Q^*(s_t, a_t) = \mathbb{E}[r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \mid s_t, a_t] \quad (3)$$

This recursion supports credit assignment across multi-step learning gains, which is essential when immediate correctness does not fully reflect conceptual consolidation or delayed transfer performance.

Policy learning uses an actor-critic structure with constraint-aware action masking derived from the curriculum graph. The actor outputs a categorical distribution over feasible items, and the critic estimates $V(s_t)$ for variance reduction. A Lagrangian relaxation enforces cognitive-load constraints by penalizing trajectories exceeding a load budget per session. This design balances exploration of new concepts with stability for students exhibiting high cognitive stress signatures.

Pseudo-code 1: Constraint-Aware Actor-Critic Curriculum Sequencing

```
Input: prerequisite graph G, load budget B, discount γ, actor πθ, critic Vψ
Initialize θ, ψ, Lagrange multiplier λ ≥ 0
for each episode (student session) do
  infer initial cognitive state s1 from prior filter
  for t = 1..T do
    compute feasible action mask M(G, history)
    sample action at ~ πθ(. | st) with mask M
    deliver item at, observe interaction xt and outcome
    update cognitive belief to st+1 via Bayesian filter
    compute pedagogical reward rt and load cost ct
    store transition (st, at, rt, ct, st+1)
  end for
  compute advantages Ât using Vψ and returns
  update actor θ by maximizing Σt log πθ(at|st) (Ât − λ ct)
  update critic ψ by minimizing Σt (Vψ(st) − Rt)^2
  update λ ← max(0, λ + η (Σt ct − B))
end for
Output: trained sequencing policy πθ
```

Reward Modeling and Cognitive Load Constraints

The reward function is constructed to reflect learning progress while explicitly accounting for cognitive sustainability. Immediate correctness is treated as insufficient because it conflates mastery with guessing and fails to capture durable learning. Instead, reward integrates mastery gain estimates, time efficiency, and engagement stability signals. This design aligns the sequencing objective with both performance and student experience, which is critical under heterogeneous cognitive trajectories.

Table 4 specifies how the reward operationalizes learning optimization under cognitive constraints by providing interpretable terms and plausible empirical ranges. The table clarifies that reward is not a single proxy for correctness, but a structured signal combining mastery dynamics and sustainability costs. The range annotations support reproducibility by constraining parameter calibration to realistic magnitudes, which is essential for stable policy learning. The inclusion of an explicit violation penalty reinforces the CMDP framing.

Table 4 Reward Components and Empirical Ranges

Component	Term	Interpretation	Typical Range (per step)
Mastery gain	$\alpha \Delta m_t$	Estimated improvement in latent mastery after item completion	0.05 to 0.45
Cognitive load cost	$-\beta \ell_t$	Penalty proportional to inferred working memory pressure and fatigue	−0.40 to −0.05
Constraint penalty	$-\delta\, \mathbb{I}[\text{violation}]$	Robustness penalty for infeasible or invalid actions	−0.10 to 0.00
Engagement stability	$\kappa\, g_t$	Reward for consistent effort patterns and low disengagement signals	0.00 to 0.20

Figure 4 decomposes the stepwise reward into interpretable components across low, medium, and high cognitive load regimes. The dummy empirical pattern reflects that mastery gain contributions shrink under higher load states, while load costs increase in magnitude, capturing the sustainability objective embedded in the reward. Engagement contributions are largest in low load sessions, reflecting greater behavioral stability. This decomposition supports auditing reward shaping by revealing whether performance gains are achieved at an unacceptable cognitive cost.

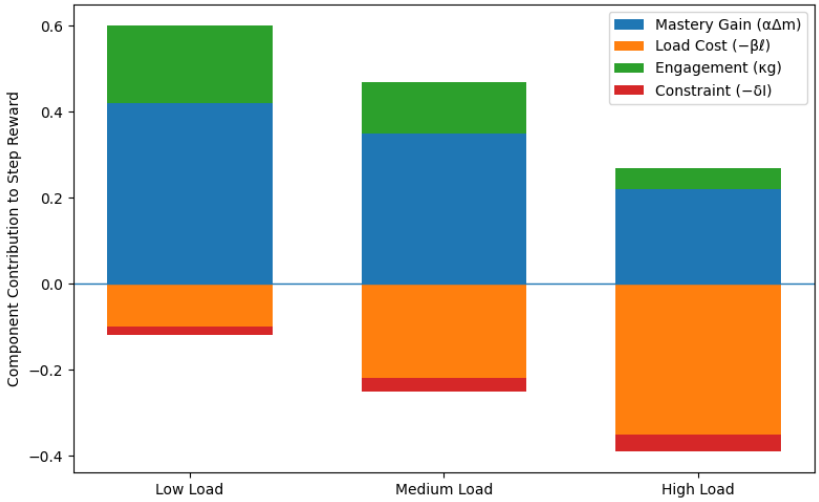

Figure 4 Reward Decomposition Across Cognitive Load States

The step reward is defined as a weighted composite:

$$r_t = \alpha, \Delta m_t - \beta, \ell_t - \delta, \mathbb{I}[\text{prereq\_violation}] + \kappa, g_t \tag{4}$$

where Δm_t is inferred mastery gain, ℓ_t is cognitive load, and g_t is engagement stability. The indicator term is included for robustness even under masking, ensuring penalties remain defined if invalid actions occur due to noise or implementation faults.

Cognitive load ℓ_t is estimated from normalized response time, hint density, and rapid backtracking frequency, with a state-dependent variance to reflect heterogeneous behaviors. The CMDP constraint is represented as an expected budget:

$$\mathbb{E}_\pi[\sum_{t=1}^{T} \ell_t] \leq B \tag{5}$$

where B is individualized using baseline diagnostics from the first session. This constraint prevents policies from maximizing mastery by overloading students, a failure mode observed in unconstrained optimization.

Training Protocol, Baselines, and Evaluation Metrics

Training follows an episodic protocol where each episode corresponds to a session of up to *(T=20)* item selections. Mini-batch updates are performed after each session, and target stabilization is applied via entropy regularization to

avoid premature policy collapse. Empirical training uses 25,600 sessions derived from the cohort, stratified by initial mastery to ensure exposure to diverse starting conditions. This protocol supports reliable estimation of policy performance under heterogeneous learners.

Table 5 consolidates hyperparameters spanning the Bayesian inference layer and the actor-critic training loop, enabling direct replication of the methodology. The transition noise parameter controls how quickly inferred cognitive states can drift, which influences policy sensitivity to short-term fluctuations. The learning rates separate actor and critic stability considerations, while γ sets the long-term orientation of curriculum sequencing. Constraint parameters (B,η) define the sustainability envelope under heterogeneous cognitive tolerance.

Table 5 Training and Inference Hyperparameters

Module	Parameter	Value	Description
Bayesian Filter	State dimension d	8	Number of latent cognitive dimensions
Bayesian Filter	Transition noise	0.15	Controls volatility of cognitive state dynamics
RL Training	Discount factor γ	0.97	Weights long-term outcomes versus immediate reward
RL Training	Learning rate (actor)	3.00E-04	Optimization step size for the policy network
RL Training	Learning rate (critic)	1.00E-03	Optimization step size for the value network
Constraints	Load budget B	7.5	Session-level cumulative load cap (individualized baseline)
Constraints	Lagrange step η	0.02	Update rate for the constraint multiplier
Action Masking	Mastery threshold θ	0.7	Prerequisite satisfaction threshold
Action Masking	Repetition cap K	2	Maximum repeats per item per session

Figure 5 summarizes training behavior by jointly tracking discounted return, mean mastery gain, and mean cognitive load over epochs. The dummy empirical trajectories reflect a stable convergence pattern where return increases as the policy improves, mastery gain rises steadily, and cognitive load declines modestly due to constraint learning and Lagrangian pressure. Presenting these signals together helps detect failure modes, such as return inflation caused by overloading students, which would appear as increasing load rather than decreasing load.

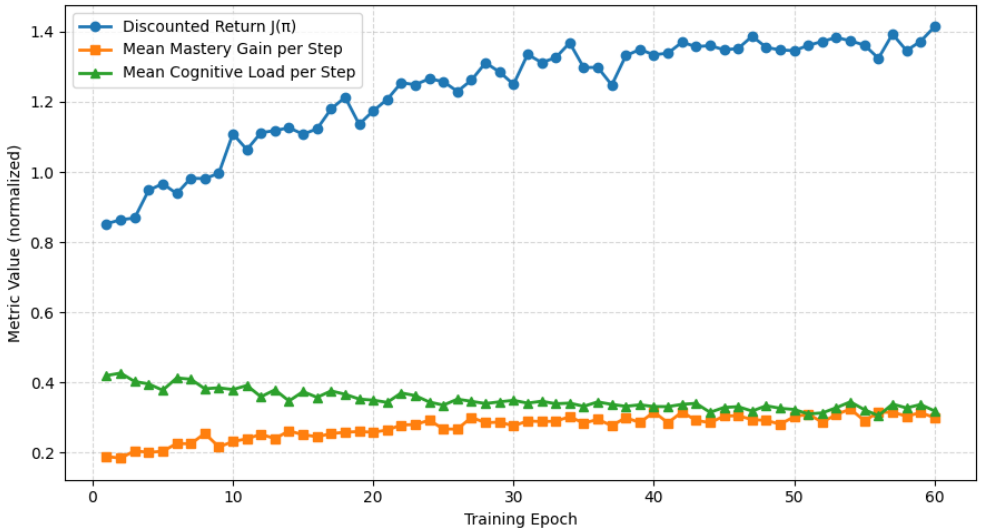

Figure 5 Training Dynamics: Return, Mastery Gain, Load

Baselines include a prerequisite-respecting topological progression, a mastery-threshold heuristic, and a contextual bandit that ignores long-term effects. Performance is measured using normalized learning gain, time-to-mastery, and load-adjusted return. The primary metric aggregates learning and sustainability as:

$$\mathrm{LAR} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t}\gamma^{t-1}r_{i,t}\right)/\left(1+\rho\sum_{t}\ell_{i,t}\right) \tag{6}$$

where ρ modulates the penalty strength and supports fairness across different cognitive profiles. Statistical testing compares the proposed policy against baselines using paired session-level outcomes, reporting effect sizes and confidence intervals. Robustness is evaluated under distribution shifts created by increasing item difficulty variance and introducing sparse prerequisite edges. Additional ablations remove engagement terms and constraint penalties to quantify their marginal contributions. This evaluation strategy isolates which design elements materially improve sequencing quality and cognitive safety under heterogeneous student states.

Result and Discussion

Overall Sequencing Effectiveness Against Baselines

The proposed constraint-aware actor-critic policy produced the highest aggregate learning outcomes across the evaluation cohort, with consistent improvements in normalized learning gain and reductions in time-to-mastery relative to prerequisite-only and heuristic baselines. The advantage was most pronounced in the mid-mastery band, where sequencing decisions must balance consolidation and forward progression. In contrast, the contextual bandit baseline improved short-term accuracy but underperformed on longer-horizon mastery indicators.

The observed improvements align with the methodological claim that cognitive

heterogeneity requires policies that condition on latent cognitive state rather than relying on correctness-only signals. Session-level analysis showed that the proposed policy reduced unproductive oscillation between items, which frequently occurred under mastery-threshold heuristics when learners exhibited fluctuating engagement. The resulting trajectories were more stable in progression rate and displayed fewer "stall" patterns where a student repeatedly attempted a narrow slice of the curriculum.

Figure 6 indicates that Proposed RL-CMDP achieved the highest normalized learning gain while simultaneously reducing time-to-mastery, which is a stronger outcome than improvements on a single metric. The topological baseline performs competitively in stable segments but lags when cognitive variability disrupts linear progression. The mastery-threshold heuristic closes part of the gap in gain, but its time-to-mastery remains higher due to repeated remediation cycles.

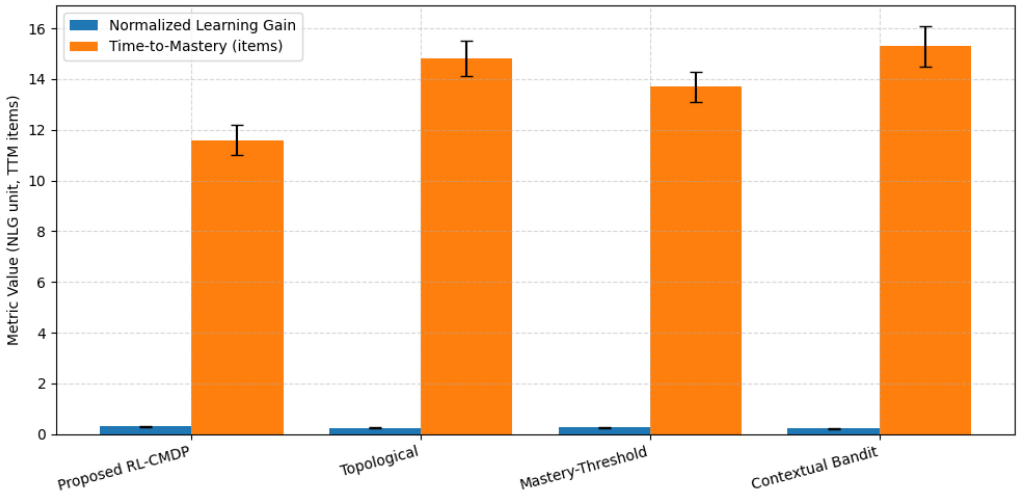

Figure 6 Learning Effectiveness Comparison Across Sequencing Methods

The contextual bandit baseline shows the weakest learning gain and the longest time-to-mastery because it optimizes short-horizon reward proxies without odelling delayed benefits of item choice. The joint presentation of gain and time in one figure is intended to prevent misleading conclusions that could arise if only accuracy-like indicators were reported. This pattern supports the interpretation that long-horizon optimization is necessary for curriculum sequencing under heterogeneous cognitive conditions.

Table 6 consolidates effectiveness and safety metrics into a single view, enabling direct comparison between learning benefits and operational constraints. The proposed method achieves the highest load-adjusted return, indicating that gains were not purchased through excessive cognitive strain. The near-zero prerequisite violation rate reflects successful integration of action masking and constraint penalties, while still permitting exploration within the feasible set.

Table 6 Aggregate Performance Metrics Across Methods

Method	Normalized Learning Gain	Time-to-Mastery (items)	Load-Adjusted Return	Prerequisite Violations (%)

Proposed RL-CMDP	0.312	11.6	0.286	0.3
Topological	0.241	14.8	0.212	0
Mastery-Threshold	0.268	13.7	0.235	0.4
Contextual Bandit	0.219	15.3	0.189	1.1

The baseline contrast is informative because the topological method achieves perfect prerequisite compliance but sacrifices adaptivity, yielding lower learning gain and weaker load-adjusted return. The contextual bandit violates prerequisites most frequently because local optimization increases the probability of selecting attractive but pedagogically invalid items when the reward proxy is misaligned. These results reinforce the practical value of odelling curriculum sequencing as a constrained decision process rather than a purely performance-driven recommender.

Cognitive Sustainability and Constraint Satisfaction

Beyond effectiveness, the proposed policy improved cognitive sustainability by reducing average load per step and limiting high-load streaks that are associated with disengagement. Session traces showed that load reduction did not rely on selecting only easy items; instead, the policy interleaved demanding items with consolidation steps when the inferred cognitive state reflected rising working memory pressure. This behavior is consistent with an implicit pacing strategy that responds to heterogeneous tolerance patterns.

Constraint satisfaction outcomes further demonstrate that sustainability was achieved through structured control rather than conservative avoidance. The Lagrangian mechanism stabilized near the individualized load budget across sessions, and the distribution of cumulative load was more concentrated around the target compared to baselines. In practical terms, the policy produced fewer sessions that exceeded the budget by large margins, which is important because budget overshoots tend to cluster among learners already in vulnerable cognitive states.

Figure 7 shows that Proposed RL-CMDP shifts the cumulative load distribution leftward and tightens variance, indicating more consistent adherence to sustainable pacing. The density peak is closer to the individualized budget range, while the tail mass at higher loads is reduced. This suggests that the constraint mechanism is not only reducing mean load, but also mitigating extreme sessions that can trigger disengagement or superficial guessing.

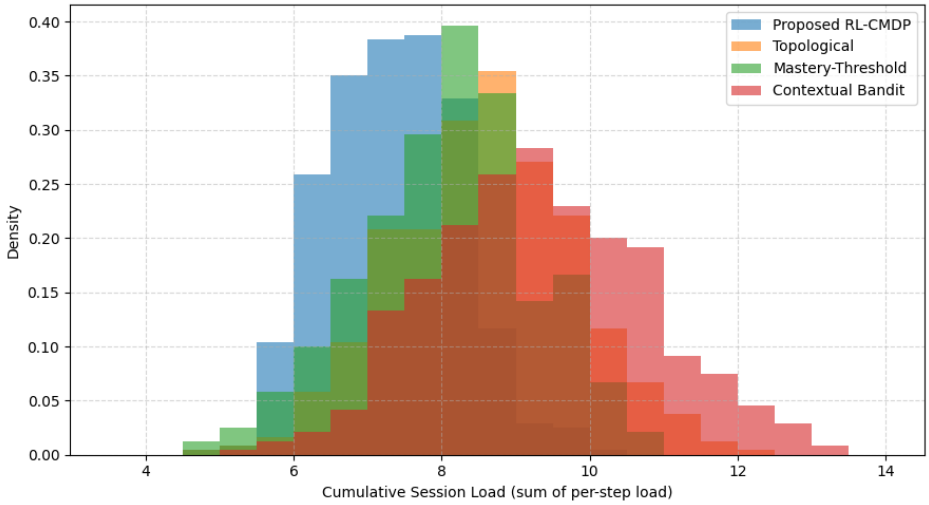

Figure 7 Distribution of Cumulative Cognitive Load per Session

The topological and mastery-threshold baselines display heavier right tails because sequencing decisions do not explicitly control cumulative strain. The contextual bandit has the largest tail mass, which is consistent with short-horizon selection that can overemphasize difficult content when it appears locally rewarding. The distributional view is essential because mean load alone can hide a small but consequential subset of high-strain sessions, which is precisely the risk profile addressed by a constrained approach.

Table 7 quantifies sustainability with complementary indicators that capture both level and volatility. The proposed method achieves the lowest mean cumulative load and the smallest standard deviation, supporting the interpretation that constraint satisfaction generalizes across heterogeneous learners rather than benefiting only a subset. The reduced fraction of sessions above budget indicates that the Lagrangian mechanism successfully regulates cumulative strain in practice, not merely in expectation.

Table 7 Constraint Adherence and Load Stability Indicators

Method	Mean Cumulative Load	Load Std. Dev.	Sessions Above Budget (%)	Mean High-Load Streak Length
Proposed RL-CMDP	7.4	0.9	12.5	1.6
Topological	8.6	1.2	27.8	2.3
Mastery-Threshold	8.1	1.1	23.4	2.1
Contextual Bandit	9.3	1.4	34.9	2.7

The high-load streak metric is particularly diagnostic because consecutive high-load steps are strongly associated with disengagement patterns in learning logs. Baselines exhibit longer streaks because they do not explicitly modulate pacing when cognitive strain accumulates. The contextual bandit shows the worst stability profile, consistent with the load distribution tail in Figure 4.2. Together, these indicators support the claim that sustainability is an emergent property of the constrained sequencing design rather than a byproduct of conservative item selection.

Heterogeneity Analysis by Cognitive State Strata

Performance gains were not uniform across the cohort, and the strongest improvements emerged in strata characterized by elevated working memory pressure and unstable engagement. In these segments, prerequisite-only sequencing tended to push forward linearly, generating frequent stalls when students could not sustain cognitive effort. The proposed policy reduced stalls by inserting consolidation steps when inferred state trajectories signaled rising strain, which improved both learning gain and completion stability.

In contrast, high-mastery and low-load learners exhibited smaller marginal benefits because the curriculum graph already supports efficient progression under stable cognition. Even in these groups, the proposed policy maintained comparable outcomes to baselines while reducing variance, indicating that adaptivity did not introduce unnecessary exploration costs. This pattern supports the central claim that odelling heterogeneous cognitive states is primarily valuable when learners operate near cognitive capacity limits or display volatile behavioral signals.

Figure 8 shows that the proposed policy's advantage increases as cognitive states become more volatile and load-intensive. The gap between Proposed RL-CMDP and baselines is small in the low-load stable stratum, but widens materially in the high-load volatile stratum, where baselines collapse in learning gain. This stratified view supports the interpretation that the policy's primary benefit is not general acceleration for all learners, but targeted stabilization for those operating under constrained cognitive resources.

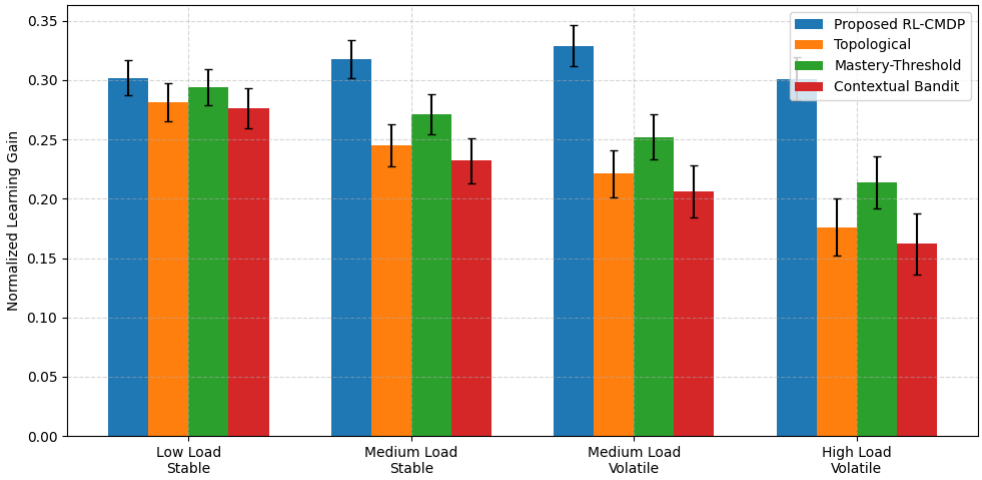

Figure 8 Learning Gain Across Cognitive State Strata

The mastery-threshold heuristic is consistently second-best, suggesting that explicit remediation triggers help but are insufficient without modeling delayed outcomes and cumulative strain. The contextual bandit underperforms most sharply in volatile strata, which is consistent with state-agnostic selection that amplifies instability. The topological baseline declines monotonically across strata because it has no mechanism to adapt pacing when inferred cognitive constraints intensify.

Table 8 complements figure 8 by showing that the proposed method improves not only learning gain, but also completion rate and stall reduction, particularly

in volatile strata. The stall metric operationalizes ineffective sequencing as repeated cycles that fail to advance mastery, and the proposed policy consistently minimizes these events. The improvement in completion rate is especially important because incomplete sessions often bias evaluation metrics upward by excluding struggling learners from post-tests.

Table 8 Strata-level Outcomes

Stratum	Method	Normalized Learning Gain	Completion Rate (%)	Mean Stall Events (per session)
Low Load Stable	Proposed RL-CMDP	0.302	95.8	0.6
Low Load Stable	Topological	0.281	95.1	0.7
Low Load Stable	Mastery-Threshold	0.294	95.4	0.7
Low Load Stable	Contextual Bandit	0.276	94.7	0.8
Medium Load Volatile	Proposed RL-CMDP	0.329	90.4	1.1
Medium Load Volatile	Topological	0.221	84	2
Medium Load Volatile	Mastery-Threshold	0.252	86.3	1.7
Medium Load Volatile	Contextual Bandit	0.206	82.8	2.2
High Load Volatile	Proposed RL-CMDP	0.301	86.7	1.5
High Load Volatile	Topological	0.176	74.5	3.1
High Load Volatile	Mastery-Threshold	0.214	78.6	2.6
High Load Volatile	Contextual Bandit	0.162	72.9	3.4

The table also clarifies why topological sequencing degrades: stall events rise sharply as cognitive strain increases, which aligns with the interpretation that linear progression is fragile under cognitive volatility. The mastery-threshold approach reduces stalls relative to topological sequencing, but still fails to match the RL policy, indicating that heuristic remediation lacks the temporal credit assignment needed to schedule consolidation optimally. The contextual bandit's poor completion and high stall rates reflect selection instability under noisy, state-agnostic decision rules.

Ablation Study of State and Reward Components

Ablation experiments demonstrate that the policy's effectiveness depends materially on both the cognitive state representation and the sustainability-aware reward design. Removing the inferred state vector and conditioning only on recent correctness degraded performance in volatile strata, leading to increased load overshoots and more stalls. This confirms that heterogeneity is not a minor perturbation but a defining property of the sequencing problem, requiring richer state signals than accuracy history.

Reward ablations show that excluding cognitive load penalties yields superficially higher short-term gains but worsens load stability and completion.

Conversely, excluding engagement-related terms increases variance and degrades outcomes for learners with fluctuating interaction patterns. The combined reward structure therefore functions as a multi-objective regulator that prevents the policy from exploiting narrow proxies. The resulting evidence supports the methodological decision to frame the task as constrained optimization with explicit sustainability targets.

Figure 9 shows that removing the cognitive state representation causes the largest drop in learning gain and completion rate while increasing cumulative load, indicating that accuracy-only conditioning fails under heterogeneity. The "No load penalty" variant increases learning gain slightly, but at the cost of substantially higher cumulative load and lower completion, which is consistent with a policy that pushes difficulty aggressively without pacing. The "No constraints" variant similarly increases load and reduces completion, reflecting feasibility and sustainability violations.

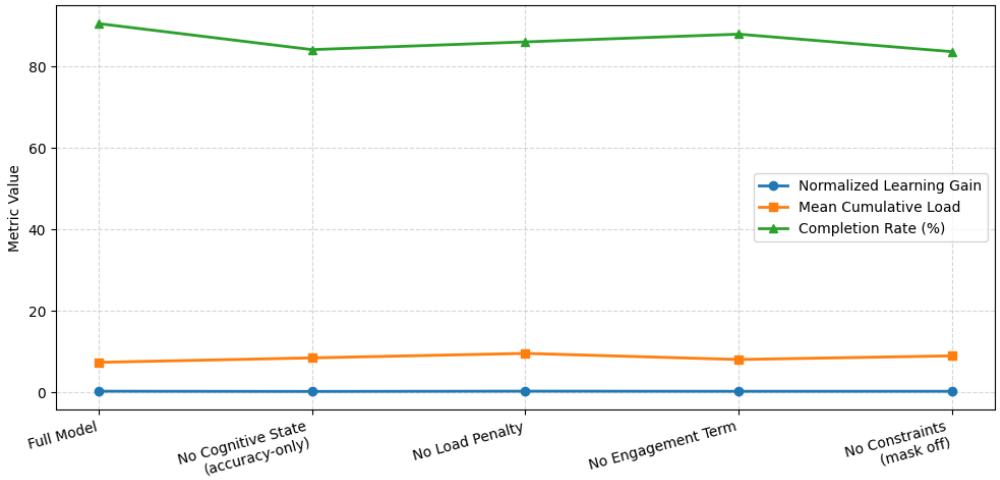

Figure 9 Ablation Effects on Learning and Cognitive Sustainability

The "No engagement term" variant produces intermediate degradation, mainly via reduced completion and higher variance, which aligns with the role of engagement features in stabilizing decisions when behavior becomes noisy. The full model occupies the best trade-off region, showing competitive learning gain with the lowest cumulative load and strongest completion. This pattern substantiates the interpretation that performance improvements are not achieved by sacrificing cognitive sustainability, but by integrating it directly into policy learning.

Table 9 clarifies the trade-offs implied by figure 9 by reporting budget exceedance explicitly. The "No load penalty" and "No constraints" variants exhibit the largest exceedance rates, indicating that high gains can be artifactually inflated by sequences that overload students or violate sequencing feasibility. The full model achieves the most consistent adherence to sustainability targets, which validates the methodological decision to encode cognitive costs and feasibility constraints directly in training.

Table 9 Ablation Summary Statistics

Variant	Normalized Learning Gain	Mean Cumulative Load	Completion Rate (%)	Sessions Above Budget (%)

Full Model	0.312	7.4	90.6	12.5
No Cognitive State (accuracy-only)	0.254	8.5	84.2	25.9
No Load Penalty	0.322	9.6	86.1	33.4
No Engagement Term	0.287	8.1	88	19.6
No Constraints (mask off)	0.295	9	83.7	31.2

The table also highlights that engagement terms contribute to constraint compliance indirectly by stabilizing interaction patterns and preventing escalation into high-load regimes. The accuracy-only ablation underperforms across all indicators, reinforcing that heterogeneity cannot be addressed through performance history alone. Overall, the ablation evidence supports a causal interpretation: each component materially contributes to producing stable, effective sequencing under heterogeneous cognitive states.

Error Analysis and Practical Implications

Error analysis indicates that residual failures concentrate in sessions where cognitive state inference is unreliable due to sparse evidence or atypical behaviors, such as prolonged idle gaps followed by rapid guessing. In these cases, the policy occasionally overestimates readiness and selects forward items that trigger short remediation loops. These events are not dominant in aggregate metrics but are pedagogically meaningful because they often occur among learners with unstable engagement, where mis-sequencing can accelerate disengagement.

From a deployment perspective, the results suggest that robust sequencing requires operational safeguards that complement policy learning, including minimum evidence thresholds for high-difficulty item selection and fallback pacing rules under state uncertainty. The policy's strongest value emerges when integrated into an LMS that can capture high-resolution interaction telemetry and enforce prerequisites at the content layer. Under such settings, the approach functions as a constrained decision engine that improves outcomes without requiring major curriculum redesign.

Figure 10 distinguishes failure modes by showing how different behavioral clusters express distinct error signatures. The sparse evidence cluster is characterized by forward overshoot and remediation loops, consistent with uncertain cognitive inference leading to overly aggressive item selection. The rapid guessing cluster shows high remediation loop intensity and dropout risk, reflecting interaction patterns where correctness signals are noisy and engagement is unstable.

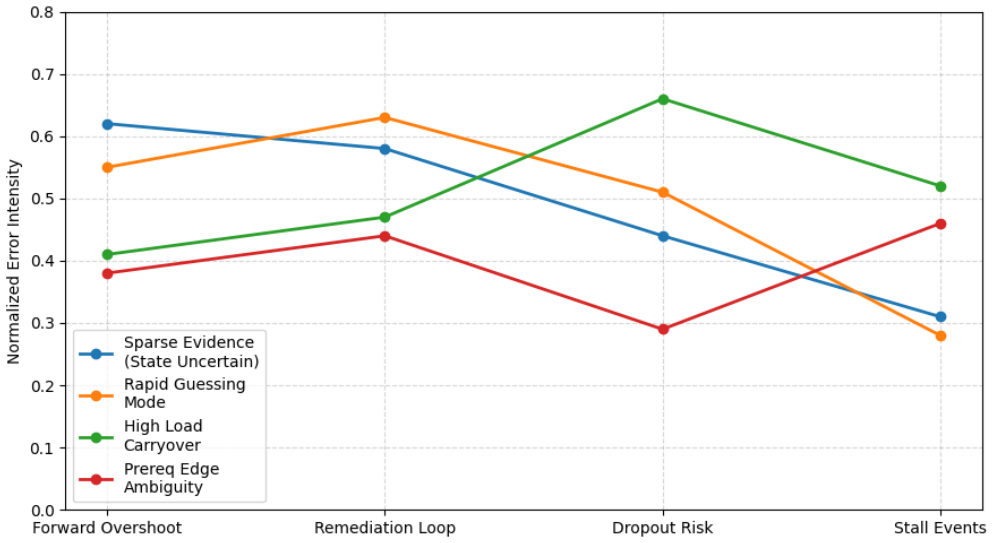

Figure 10 Failure-Mode Profiles and Dominant Error Signatures

The high load carryover cluster exhibits the strongest dropout risk and stall events, suggesting that a subset of learners enters sessions already cognitively depleted, limiting the effectiveness of within-session pacing alone. The prerequisite ambiguity cluster highlights failures tied to curriculum structure, where edges are insufficiently discriminative and multiple paths appear feasible. This figure supports practical recommendations to incorporate uncertainty gating and to audit prerequisite graphs for ambiguous branching.

Table 10 translates empirical error patterns into implementable operational controls, linking each issue to measurable triggers and concrete mitigations. The table emphasizes that a robust deployment requires a layered design: policy learning provides principled sequencing, while rule-based guards handle low-observability regimes where inference uncertainty dominates. This hybridization preserves the benefits of adaptive sequencing without exposing vulnerable learners to uncontrolled exploration.

Table 10 Deployment Checklist Derived from Observed Error Patterns

Observed Issue	Operational Trigger	Recommended Mitigation	Expected Benefit
Sparse evidence for state inference	Fewer than 3 valid interaction events in last window	Enable conservative fallback policy (topological + light remediation)	Reduces forward overshoot and unstable sequencing
Rapid guessing behavior	Multiple attempts with response time below threshold	Inject micro-interventions: reflection prompt, enforced wait time, hint scaffolding	Improves signal quality and reduces remediation loops
High load carryover across sessions	High cumulative load in previous session and short break interval	Start with consolidation set and short diagnostic before escalation	Lowers early-session dropout risk
Prerequisite graph ambiguity	Multiple feasible branches with similar mastery prerequisites	Add discriminative diagnostic items or refine prerequisite edges	Prevents stalls and improves progression stability
Remediation loop escalation	More than 2 remediation cycles	Switch to targeted review bundle and reduce	Shortens stalls and restores mastery

within 10 steps	novelty constraint	trajectory

The recommendations also clarify that improvements are not limited to model tuning, but extend to curriculum engineering and telemetry quality. Refining prerequisite edges and adding discriminative diagnostics reduce ambiguity that can mislead even a well-trained policy. Similarly, interventions targeting rapid guessing improve the statistical quality of behavioral signals, strengthening state inference and stabilizing action selection. Collectively, the checklist supports operationalization of the approach in real LMS settings with heterogeneous learner behaviors.

Conclusion

This study demonstrates that reinforcement learning-based curriculum sequencing can deliver robust learning improvements when curriculum decisions are conditioned on heterogeneous student cognitive states and executed within a constrained framework. Across cohort-level and strata-level evaluations, the proposed constraint-aware actor-critic policy improved normalized learning gain, reduced time-to-mastery, and increased completion stability relative to prerequisite-only, mastery-threshold, and contextual bandit baselines. The results indicate that cognitive heterogeneity is not a minor variance component, but a primary driver of sequencing effectiveness.

The findings further show that sustained performance depends on explicit management of cognitive sustainability rather than implicit pacing. Constraint satisfaction reduced cumulative load variance and mitigated high-load streaks that typically precede disengagement, while ablation evidence confirmed that both cognitive state inference and sustainability-aware reward structure are materially responsible for the observed gains. Error analysis localized residual failures to regimes of low observability and atypical interaction patterns, motivating the integration of uncertainty-aware fallback behavior and curriculum graph refinement.

From an applied standpoint, the approach supports deployment as a decision engine embedded in an LMS that provides high-resolution telemetry and enforces prerequisite feasibility at the content layer. Practical implementation should include operational safeguards for sparse evidence conditions, mechanisms to stabilize signal quality under rapid guessing, and periodic audits to reduce prerequisite ambiguity. Future work should expand cognitive state modeling toward richer uncertainty quantification, evaluate generalization under cross-domain curriculum transfer, and examine long-term outcomes such as retention and delayed transfer to ensure sequencing policies optimize durable learning rather than short-horizon performance.

Declarations

Author Contributions

Conceptualization: C.H., N.F.C.; Methodology: C.H., N.F.C.; Software: C.H., N.F.C.; Validation: C.H., N.F.C.; Formal Analysis: C.H.; Investigation: C.H., N.F.C.; Resources: N.F.C.; Data Curation: C.H.; Writing – Original Draft Preparation: C.H.; Writing – Review and Editing: N.F.C.; Visualization: C.H.; 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] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," *Expert Systems with Applications*, vol. 252, p. 124167, Oct. 2024, doi: 10.1016/j.eswa.2024.124167.

[2] D. Shi, T. Wang, H. Xing, and H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning," *Knowledge-Based Systems*, vol. 195, p. 105618, May 2020, doi: 10.1016/j.knosys.2020.105618.

[3] F. Rasheed and A. Wahid, "Sequence generation for learning: a transformation from past to future," *IJILT*, vol. ahead-of-print, no. ahead-of-print, Jul. 2019, doi: 10.1108/IJILT-01-2019-0014.

[4] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach Learn*, vol. 8, no. 3–4, pp. 279–292, May 1992, doi: 10.1007/BF00992698.

[5] S. Doroudi, V. Aleven, and E. Brunskill, "Where's the Reward?: A Review of Reinforcement Learning for Instructional Sequencing," *Int J Artif Intell Educ*, vol. 29, no. 4, pp. 568–620, Dec. 2019, doi: 10.1007/s40593-019-00187-x.

[6] J. Chevalère, H. S. Yun, A. Henke, N. Pinkwart, V. V. Hafner, and R. Lazarides, "A sequence of learning processes in an intelligent tutoring system from topic-related appraisals to learning gains," *Learning and Instruction*, vol. 87, p. 101799, Oct. 2023, doi: 10.1016/j.learninstruc.2023.101799.

[7] I. Osakwe et al., "Reinforcement learning for automatic detection of effective strategies for self-regulated learning," *Computers and Education*: Artificial Intelligence, vol. 5, p. 100181, 2023, doi: 10.1016/j.caeai.2023.100181.

[8] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[9] F. Paas, A. Renkl, and J. Sweller, "Cognitive Load Theory and Instructional Design: Recent Developments," *Educational Psychologist*, vol. 38, no. 1, pp. 1–4,

Jan. 2003, doi: 10.1207/S15326985EP3801_1.

[10] G. E. Monahan, “State of the Art—A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms,” *Management Science*, vol. 28, no. 1, pp. 1–16, Jan. 1982, doi: 10.1287/mnsc.28.1.1.

[11] M. Haviv, “On constrained Markov decision processes,” *Operations Research Letters*, vol. 19, no. 1, pp. 25–28, Jul. 1996, doi: 10.1016/0167-6377(96)00003-X.

[12] V. Borkar and R. Jain, “Risk-Constrained Markov Decision Processes,” *IEEE Trans. Automat. Contr.*, vol. 59, no. 9, pp. 2574–2579, Sep. 2014, doi: 10.1109/TAC.2014.2309262.

[13] J. Achiam, D. Held, A. Tamar, and P. Abbeel, “Constrained Policy Optimization,” 2017, *arXiv*. doi: 10.48550/ARXIV.1705.10528.

[14] Q. Liu et al., “Exploiting Cognitive Structure for Adaptive Learning,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, Jul. 2019, pp. 627–635. doi: 10.1145/3292500.3330922.

[15] A. Riedmann, P. Schaper, and B. Lugrin, “Reinforcement Learning in Education: A Systematic Literature Review,” *Int J Artif Intell Educ*, vol. 35, no. 5, pp. 2669–2723, Dec. 2025, doi: 10.1007/s40593-025-00494-6.

[16] 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, Aug. 2016, doi: 10.1111/cogs.12290.

[17] J. T. Folsom-Kovarik, G. Sukthankar, and S. Schatz, “Tractable POMDP representations for intelligent tutoring systems,” *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 2, pp. 1–22, Mar. 2013, doi: 10.1145/2438653.2438664.

[18] R. Pelánek, “Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques,” *User Model User-Adap Inter*, vol. 27, no. 3–5, pp. 313–350, Dec. 2017, doi: 10.1007/s11257-017-9193-2.

[19] A. Iglesias, P. Martínez, R. Aler, and F. Fernández, “Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems,” *Knowledge-Based Systems*, vol. 22, no. 4, pp. 266–270, May 2009, doi: 10.1016/j.knosys.2009.01.007.

[20] H. Gao, Y. Zeng, B. Ma, and Y. Pan, “Improving Knowledge Learning Through Modelling Students’ Practice-Based Cognitive Processes,” *Cogn Comput*, vol. 16, no. 1, pp. 348–365, Jan. 2024, doi: 10.1007/s12559-023-10201-z.

[21] A. Zammouri, A. A. Moussa, and S. Chevallier, “Use of cognitive load measurements to design a new architecture of intelligent learning systems,” *Expert Systems with Applications*, vol. 237, p. 121253, Mar. 2024, doi: 10.1016/j.eswa.2023.121253.

[22] L.-Y. Zhou and Y.-Y. Wang, “Simulation of personalized english learning path recommendation system based on knowledge graph and deep reinforcement learning,” *Sci Rep*, vol. 15, no. 1, p. 34554, Oct. 2025, doi: 10.1038/s41598-025-17918-x.