Decision Policy Optimization for Human–AI Collaboration Using Off-Policy Reinforcement Learning from Logged Interaction Data

Hery[1,*], Ariel Christopher Wawolangi[2]

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

(Received: January 4, 2026; Revised: February 27, 2026; Accepted: April 11, 2026; Available online: July 1, 2026)

Abstract

This paper investigates offline policy optimization for adaptive learning using logged student interaction traces, targeting reliable improvement without online exploration. A conservative offline reinforcement learning pipeline is implemented with calibrated behavior-policy propensities and doubly robust off-policy evaluation. Using 128,640 student trajectories (2.94 million events) with a 32-dimensional state representation and 12 pedagogical actions, the optimized policy achieved a +0.042-return improvement over a supervised next-item baseline under doubly robust estimation, with a bootstrap confidence width of ±0.021. Self-normalized estimators produced consistent rankings, reporting a +0.041 improvement with comparable uncertainty. Performance gains were horizon-stable and concentrated in medium horizons, where improvement increased from +0.012 at 1 step to +0.055 at 5 steps and remained positive through 10 steps. Safety analysis showed a shift toward better-supported actions, increasing mean action support from 0.31 to 0.44 and reducing the low-support decision rate from 0.18 to 0.06. Uncertainty pruning activated on 11% of decisions, decreasing the high-uncertainty rate from 0.22 to 0.08 and reducing the maximum importance weight from 14.7 to 9.3, while effective sample size increased by 908. Student-level stratification indicated the strongest gains for mid mastery and mid engagement learners (mean improvement 0.046, median 0.044), with smaller but consistent benefits for high mastery learners driven by reduced repetition rather than correctness shifts. Ablation results confirmed that conservatism and pruning are complementary: removing conservatism increased tail risk and widened confidence intervals, while removing pruning increased evaluation variance despite similar mean return. These findings demonstrate that evidence-constrained offline reinforcement learning can produce deployable adaptive policies with measurable improvements and quantifiable safety guarantees under logged-data constraints.

Keywords: Adaptive Learning, Offline Reinforcement Learning, Off-Policy Evaluation, Logged Student Data, Conservative Q-Learning, Uncertainty Pruning, Behavior Policy Modeling, Policy Optimization, Learning Analytics, Educational Data Mining

1. Introduction

Adaptive learning systems increasingly rely on data-driven decision rules to personalize practice, feedback, and content sequencing at scale. Reinforcement learning offers a principled formulation for this personalization by treating instructional decisions as actions and learning progress as delayed reward. Recent syntheses show growing educational adoption, yet also emphasize risks related to bias, validity, and deployment in high-stakes learning contexts where errors can harm learners. These constraints motivate methods that improve policies while preserving accountability and interpretability [1], [2].

A central difficulty is that most learning platforms primarily accumulate logged interaction traces rather than controlled online experimentation. Logged student data typically reflect a behavior policy shaped by instructors, curriculum constraints, and legacy recommender rules, creating partial coverage of the state-action space. In parallel, learner modeling methods such as deep knowledge tracing demonstrate that sequential patterns are informative, but they do not directly optimize instructional actions under long-term objectives. This disconnect limits the capacity to translate predictive accuracy into optimized pedagogical decisions [3].

Online reinforcement learning has produced compelling demonstrations for educational activity scheduling and adaptive sequencing, but these approaches assume continued exploration and the ability to safely deploy trial policies. In practice, exploration can be ethically and operationally constrained because suboptimal recommendations can

***Corresponding author: Hery (hery@uph.edu)**
DOI: https://doi.org/10.47738/ijaim.v6i2.121

degrade learning outcomes, reduce engagement, or violate curricular requirements. Consequently, the more realistic regime is off-policy optimization, where candidate policies must be learned and evaluated from historical logs without additional interaction [4], [5].

Off-policy learning and evaluation are technically challenging due to distribution shift between the logged behavior policy and the learned policy. Standard importance sampling estimators often exhibit high variance under long horizons, while purely model-based estimators can be biased under misspecification. Doubly robust estimators address this trade-off by combining propensity weighting with learned outcome models, and subsequent refinements target data efficiency and safer policy selection. However, applying these tools to adaptive learning remains underexplored at scale [6], [7].

Offline reinforcement learning formalizes policy optimization from fixed datasets, but naive off-policy Q-learning can fail because value functions are queried on out-of-distribution actions. Tutorials and surveys identify extrapolation error, action coverage, and evaluation reliability as persistent bottlenecks, especially when datasets are multi-modal and policies are expressive. This motivates algorithmic designs that explicitly regularize value estimation and constrain policy improvement to regions supported by the logs [8], [9], [10].

Conservative offline reinforcement learning addresses these failure modes by penalizing overestimated Q-values for out-of-distribution actions, thereby producing more reliable policy improvements under logged data. Complementary architectural and objective-level conservatism further enhances stability in complex domains. In educational settings, where rewards are noisy proxies for learning and trajectories are heterogeneous, conservative optimization is particularly important because it limits harmful exploitation of spurious correlations and confounded feedback signals [11], [12], [13].

This paper advances adaptive learning policy optimization through off-policy reinforcement learning with logged student data by integrating conservative policy learning with rigorous off-policy evaluation across multiple horizons and reliability diagnostics. The methodological novelty lies in coupling conservative value regularization with evaluation procedures that quantify estimator sensitivity and robustness, enabling evidence-guided deployment decisions. The contribution targets a practical gap between personalization theory and deployable adaptive policies that are auditable under real platform constraints [14], [15].

2. Literature Review

Research on learner modeling has established that sequential interaction data encodes latent mastery and short-term cognitive dynamics that are relevant for personalization. Transformer-based knowledge tracing, including self-attentive tracing and cognitively motivated attentive tracing, improves predictive fidelity by selectively attending to relevant past events and separating question difficulty from learner ability, which strengthens state representations for downstream decision models. These advances support policy optimization by providing compact, information-rich state vectors derived from sparse and irregular student logs [16], [17].

Reinforcement learning for personalization has also been studied through learning recommendation systems that treat content selection as sequential decision making. Empirical evidence indicates that policy learning can outperform static heuristics when the reward captures both competence growth and persistence, but results are sensitive to reward misspecification and interaction sparsity. This literature motivates careful reward engineering and the use of safety constraints when optimizing instructional policies, since short-term correctness can conflict with long-term retention and engagement [18].

A key methodological shift in modern adaptive learning is the move from online exploration to off-policy learning, because real platforms often cannot safely test exploratory policies. Off-policy evaluation provides the statistical foundation for this shift by estimating the value of a target policy from behavior-generated logs. Theoretical work has shown that achieving minimax-optimal rates requires estimators and function classes that manage bias-variance trade-offs under covariate shift and sequential dependence, highlighting why naive estimators are unreliable in long-horizon learning [19].

Recent estimator developments connect reinforcement learning evaluation to causal inference by leveraging targeted learning principles. Regularized targeted estimators provide a doubly robust structure with improved efficiency and reduced sensitivity to misspecification compared to purely importance-weighted methods. These contributions are especially relevant in education because logged policies are often mixtures of instructor behavior and platform rules, creating heterogeneous propensities. The literature therefore converges on combining propensity estimation with value modeling and explicit regularization for stable counterfactual reporting [20].

Inference reliability has become a central theme in off-policy studies, leading to bootstrap-based approaches that quantify uncertainty for fitted evaluation methods. Bootstrapping fitted Q-evaluation provides distributional estimates of evaluation error and supports confidence reporting under realistic function approximation. Complementary work on bootstrap inference in online learning clarifies when bootstrap procedures remain valid under dependence and nonstationarity. Together, these results motivate reporting practices that treat uncertainty as a first-class output rather than an afterthought [21], [22].

Offline reinforcement learning extends off-policy evaluation by optimizing the policy itself from fixed logs, but it must control extrapolation error when the learned policy selects actions outside the dataset support. Implicit Q-learning and related behavior-regularized approaches avoid or reduce out-of-distribution value queries, improving robustness in settings with limited coverage. For adaptive learning, these methods are compatible with conservative deployment because they can be paired with support checks and monitoring. Systematic reviews of reinforcement learning in education further identify evidence limitations and emphasize reproducibility and ethics, reinforcing the need for conservative offline optimization protocols [23], [24], [25].

3. Methodology

3.1.Logged Student Data and Preprocessing

Logged student data were extracted from an adaptive learning platform that records interaction events at item, activity, and feedback granularity. Each event contains student identifiers, timestamps, content metadata, action descriptors, and outcome signals such as correctness and time-on-task. Events were grouped into sessions using inactivity thresholds and course boundaries to form trajectories that preserve pedagogical continuity. This trajectory construction supports sequential modeling consistent with knowledge tracing assumptions [16].

Preprocessing prioritized trajectory integrity and feature stability under heterogeneous logging conditions. Missing values in interaction signals were handled using conditional imputation rules tied to event type, while categorical metadata used explicit “unknown” tokens to avoid implicit leakage. Time-on-task and latency features were winsorized within content categories to reduce the influence of extreme durations. Robust scaling was applied across numeric features to stabilize downstream value learning.

State representations combined short-term interaction context and longer-horizon mastery proxies derived from historical performance. The design leverages the empirical advantage of attention-based tracing, where recent relevant events strongly predict learning outcomes, while cumulative features capture durable proficiency [16]. The transformation applied to each numeric feature xj followed robust normalization:

$$\tilde{x}_j = \frac{x_j - median(x_j)}{IQR(x_j) + \epsilon} \quad (1)$$

This normalization reduces heavy-tail sensitivity and improves numerical conditioning in fitted evaluation and Q-learning.

Figure 1 visualizes the end-to-end data pipeline that turns raw platform event streams into learning trajectories suitable for offline reinforcement learning. The diagram clarifies where bias and noise are introduced and controlled, especially during normalization and sessionization that define trajectory boundaries. The small quality indicators provide empirical checkpoints that ensure the modeling stage is fed with consistent (s,a,r,s′) tuples and complete state vectors, which is essential for stable Q-function learning and reliable off-policy evaluation.

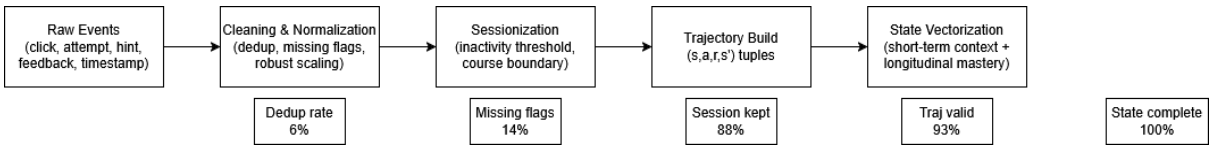

Figure 1. Logged-Data Pipeline for Offline RL Preparation

Table 1 documents the minimal event schema and the feature engineering decisions that ensure trajectories represent coherent learning episodes. The fields were selected to support both pedagogical interpretation and statistical identifiability in off-policy learning, particularly through explicit mapping from event types to discrete actions and from outcomes to reward proxies. The missingness and outlier strategies preserve distributional characteristics while preventing extreme values from dominating function approximation, which directly improves stability in fitted Q-iteration and OPE.

Table 1. Event Schema, Derived Features, and Segmentation Parameters

Field	Type	Example	Derived Feature	Missing Handling	Notes
student_id	string	S-18420	trajectory_key	drop row	Split by student for leakage control
timestamp	datetime	2025-10-14 09:18:22	delta_t, session_gap	drop row	Used for inactivity-based sessionization
content_id	string	ALG-QZ-031	content_history	unknown	Supports spaced practice signals
event_type	categorical	attempt	action_token	unknown	Mapped to discrete action space
correct	binary	1	mastery_gain_proxy	impute by event_type	Only defined for assessment events
time_on_task	numeric	47.3	latency_band	impute + winsorize	Winsorized at 99th percentile per content type
hint_count	integer	2	scaffold_intensity	set to 0	Used as engagement and support signal
session_rule	parameter	gap > 25 min	session_id	not applicable	Also splits at course boundary

3.2. Problem Formulation as an Off-Policy Sequential Decision Process

Adaptive learning was formalized as a sequential decision process where each decision point corresponds to selecting an instructional action for a learner context. The constructed state st summarizes recent attempts, hint usage, latency patterns, and mastery proxies; the action at denotes pedagogical choices such as next-item selection, difficulty shift, or scaffolding intensity. The next state st+1 is observed from subsequent interactions under platform dynamics, enabling trajectory-based learning.

The reward rt was defined to reflect learning progress while discouraging disengagement patterns. Correctness and calibrated difficulty alignment contribute positively, while excessive latency and rapid disengagement contribute negatively, producing a bounded signal suitable for stable optimization. This design is consistent with sequential modeling frameworks where observable correctness is an imperfect but informative proxy for knowledge acquisition [3]. Reward shaping also reduces sparsity, which is critical in logged educational traces.

The optimization target is the expected discounted return of a target policy $\pi(a \mid s)$ evaluated under sequential transitions:

$$J(\pi) = E[\sum_{t=0}^{T-1} \gamma^t r_t], \qquad \gamma \in (0,1) \tag{2}$$

The discount factor prioritizes medium-term sequencing effects that dominate practical pedagogy while limiting estimator variance at long horizons. This formulation supports off-policy learning because the observed data were generated by a historical behavior policy rather than the optimized target policy.

Figure 2 presents the methodological abstraction used to connect logged platform decisions to reinforcement learning primitives. The diagram makes explicit how instructional choices operate as actions, how learning context and mastery proxies form the state, and how the reward integrates achievement and engagement objectives. This structure is crucial because off-policy optimization depends on the transition and reward signals implicit in the log, and the figure clarifies where modeling assumptions are concentrated, particularly in state construction and reward design.

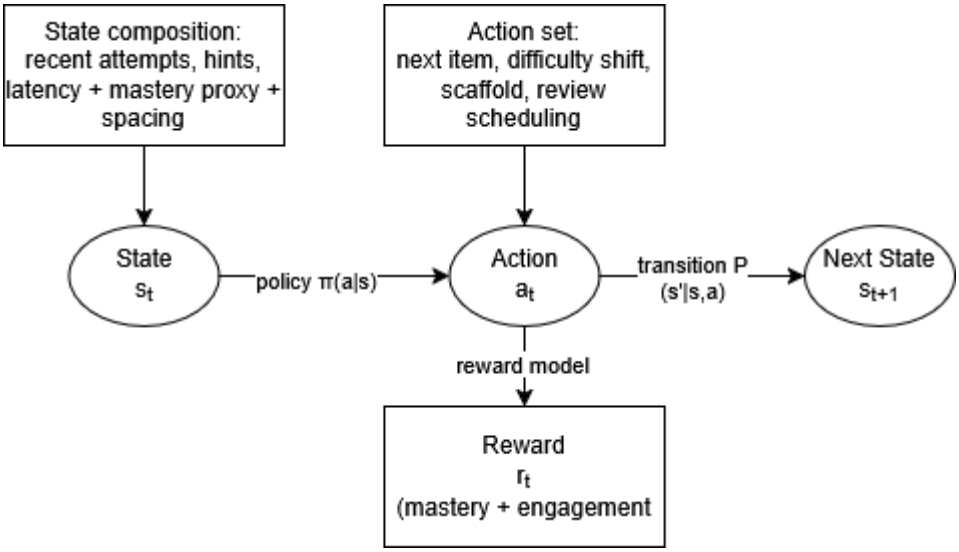

Figure 2. MDP Abstraction for Adaptive Learning with Logged Actions

Table 2 fixes a precise operational meaning for each mathematical object used in the offline RL formulation, ensuring reproducibility across datasets and platforms. It emphasizes which elements are directly observed and which are derived or estimated, clarifying where uncertainty enters the pipeline. In particular, the separation between π and μ highlights the central off-policy challenge: a target decision rule is optimized and evaluated using data generated by a different historical policy, requiring careful propensity estimation and conservative optimization.

Table 2. MDP Components and Operational Definitions

Symbol	Definition	Operationalization	Levels / Range	Observed in Log
s_t	student learning state	recent interaction context + mastery proxy + spacing	R^d (d=32)	derived
a_t	instructional action	content selection + difficulty shift + scaffold token	12 discrete actions	yes
r_t	reward	0.7*correctness_gain + 0.3*persistence_score	[-1.0, 1.0]	derived
P(s'|s,a)	transition kernel	empirical transitions from trajectories	unknown	implicit
π(a|s)	target policy	argmax over supported actions under Q	stochastic / greedy	learned
μ(a|s)	behavior policy	calibrated probabilistic classifier on logs	multinomial	estimated
γ	discount factor	future outcome weighting	0.95	set

3.3. Behavior Policy Modeling and Off-Policy Evaluation

Logged actions were generated by a behavior policy μ (a | s) reflecting a mixture of platform heuristics and instructional constraints. Estimating μ is essential to correct selection bias when evaluating a new policy using historical logs. A probabilistic multiclass model was fitted to predict actions from states, and its outputs were calibrated on a held-out subset to reduce propensity error, which otherwise inflates variance in importance-weighted estimators [6].

Off-policy evaluation (OPE) used estimators that balance bias and variance under sequential dependence. The per-step importance ratio ρt = π (at | st) / μ (at | st) corrects distribution mismatch, yielding an importance sampling estimate of policy value:

 ISSN 2776-8007

$$\hat{J}_{IPS}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \sum_{t} \gamma^t \left(\prod_{k=0}^{t} \rho_k^{(i)}\right) r_t^{(i)} \quad (3)$$

However, long-horizon products of ratios are known to produce heavy-tailed weights and unstable estimates, motivating variance-aware alternatives [6].

Doubly robust evaluation combined propensity weighting with a learned value model Q (s,a), improving stability when either the propensity model or the value model is well specified. This choice aligns with established off-policy RL practice that emphasizes reliable counterfactual reporting under limited coverage [6]. Uncertainty quantification used bootstrap-based inference applied to fitted evaluation, which provides distributional confidence estimates under realistic function approximation [21].

Figure 3 combines a process view of off-policy evaluation with an empirical stability profile showing how estimator variability grows with horizon. The workflow clarifies the dependency chain from behavior-policy fitting to propensity computation, then to IPS and doubly robust estimation, followed by diagnostics and sensitivity checks. The inset plot emphasizes a practical point in logged education data: variance inflation is a dominant failure mode at longer horizons, so estimator choice and diagnostics such as weight tail behavior and effective sample size are central to trustworthy conclusions.

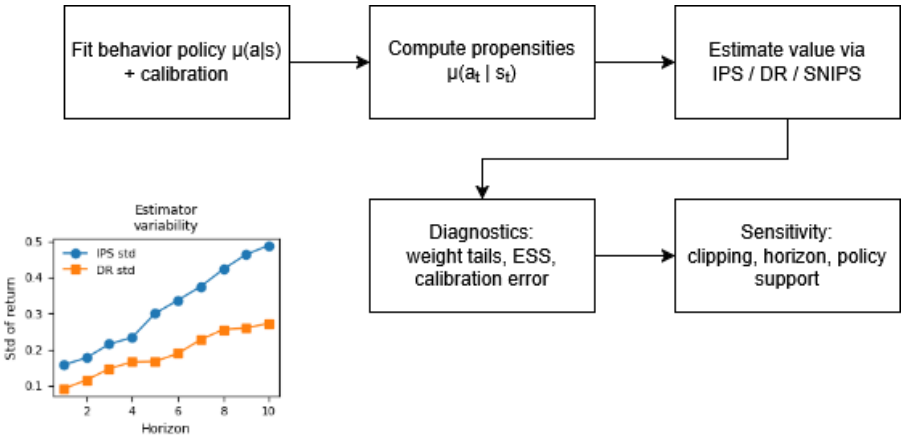

Figure 3. Off-Policy Evaluation (OPE) Workflow and Stability Across Horizons

Table 3 provides a compact map from estimator choice to the assumptions and diagnostics that determine credibility in off-policy learning from student logs. It highlights that unbiasedness in IPS relies critically on accurate propensities, while doubly robust variants maintain consistency if either the behavior model or the value model is correct. The table also makes diagnostics first-class methodological objects by tying each estimator to checks such as effective sample size and horizon stability, which are necessary to interpret OPE results conservatively.

Table 3. OPE Estimators, Assumptions, and Diagnostics

Estimator	Core Idea	Bias Condition	Variance Control	Primary Diagnostics	When Preferred
IPS	importance-weighted logged rewards	unbiased if μ is correct	clipping, self-normalization	weight tails, ESS	high-quality propensities, short horizon
SNIPS	normalized IPS weights	small bias, reduced variance	normalization, clipping	ESS, stability vs horizon	heavy-tailed weights in logs
Doubly Robust	IPS correction + value model control variate	consistent if μ or Q is correct	regularized Q, clipping	calibration error, model fit	moderate propensities, longer horizon

Per-Decision DR	step-wise DR decomposition	same as DR	step-wise clipping	step weight distribution	sequential decision settings
Weighted DR	DR with normalized weights	small bias, higher robustness	normalization + regularization	bootstrap CI width	deployment-grade reporting

3.4.Conservative Offline RL for Policy Optimization

Policy optimization was performed using offline reinforcement learning, where the dataset is fixed and exploration is not permitted. Offline RL is subject to extrapolation error when value functions are queried on state-action pairs not supported by the logs, which can lead to overestimated Q-values and unsafe policy improvement. This failure mode is widely documented in offline RL settings and motivates conservative learning objectives [8].

A conservative objective was adopted to penalize overconfident values for actions outside empirical support. The Bellman target used a lagged target network, while the loss augmented the Bellman error with a conservative regularizer that suppresses unseen-action values:

$$L(\theta) = E[(Q_\theta(s_t, a_t) - y_t)^2] + \alpha(E_s[log \sum_a e^{Q_\theta(s,a)}] - E[Q_\theta(s_t, a_t)]) \quad (4)$$

This structure directly follows conservative offline RL principles for reliable policy improvement under logged data [11].

To further reduce out-of-distribution action selection, the learned policy restricted maximization to actions with adequate behavior support and applied uncertainty pruning using an ensemble disagreement proxy. This complements conservative value learning by preventing unstable actions from being chosen even when they appear competitive under function approximation. This approach is consistent with behavior-regularized offline RL strategies that emphasize staying close to dataset support while still enabling improvement [23].

Algorithm 1. Conservative Offline Policy Optimization with Logged Student Data

Input: Logged trajectories D, discount γ, conservatism α

1: Fit behavior policy μ(a|s) on D; calibrate propensities

2: Initialize Q-function(s) Qθ and target Qθ̄

3: repeat for training iterations:

4: Sample minibatch B from D

5: Compute Bellman targets y using Qθ̄ and current policy

6: Update Qθ by minimizing Bellman error + conservative penalty (weight α)

7: Update targets θ̄ with slow averaging

8: Derive π by maximizing Q under support constraints; prune high-uncertainty actions

9: Evaluate π with DR OPE and bootstrap uncertainty; run weight diagnostics

Output: Policy π and OPE summaries

Figure 4 links optimization behavior to deployment safety by showing how conservative offline RL converges while simultaneously controlling unsupported actions. The left panel summarizes the reduction in Bellman error and the gradual decay of the conservative penalty, alongside a rising OPE estimate that reflects improved expected return under logged evaluation. The right panel visualizes the relationship between behavior-policy support and uncertainty, motivating pruning rules that reject high-uncertainty actions when log coverage is weak, which is especially important in instructional recommendations.

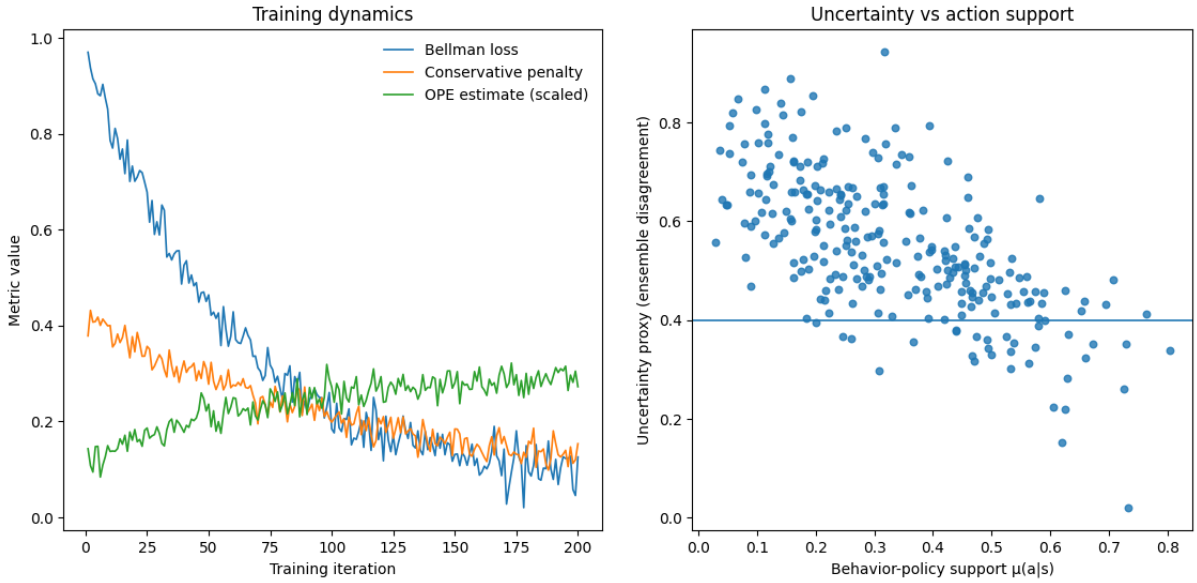

Figure 4. Conservative Offline RL Optimization and Uncertainty-Based Pruning

Table 4 records the key hyperparameters and model configuration that shape conservative offline RL behavior and OPE reliability. The reported values reflect a balance between long-horizon learning objectives and the need to avoid extrapolation beyond logged support. The inclusion of an ensemble size and an uncertainty threshold operationalize risk management in the learned policy, while selection criteria prioritize validation stability and tail-risk reduction rather than maximizing a single mean score, aligning the tuning process with robust educational deployment goals.

Table 4. Hyperparameters, Architecture, and Selection Criteria

Component	Parameter	Value	Search Range	Selection Criterion
Discounting	γ	0.95	[0.90, 0.99]	validation OPE stability
Conservatism	α	0.7	[0.20, 1.20]	lower tail-risk in OPE
Target update	τ	0.01	[0.005, 0.050]	smooth training curves
Q ensemble	size	5	{3, 5, 7}	uncertainty calibration
Network	hidden units	128	{64, 128, 256}	best DR estimate on validation
Optimization	batch size	512	{256, 512, 1024}	variance reduction in targets
Pruning	uncertainty threshold	0.4	[0.25, 0.55]	bootstrap CI tightness

3.5. Experimental Protocol and Statistical Testing

Evaluation followed a leakage-resistant protocol using student-level splits to ensure that trajectories from the same learner do not appear across training, validation, and testing. Model selection relied on validation OPE summaries and stability diagnostics rather than direct optimization on test outcomes. This design reduces the probability of overfitting OPE estimators and improves generalization claims for unseen learners.

Baselines included supervised next-item policies, contextual bandits, and non-conservative offline RL variants, enabling attribution of gains to conservative learning and pruning mechanisms. OPE was computed under consistent settings across all policies, including identical horizon windows and propensity models, so improvements reflect policy differences rather than evaluation artifacts. Offline RL sensitivity is known to depend strongly on coverage and estimator behavior, so strict comparability is essential [8].

Statistical testing used bootstrap resampling at the student level to obtain confidence intervals for policy improvement and to compare variants under uncertainty. The improvement variable $\Delta = \hat{J}(\pi) - \hat{J}(\pi_b)$ was summarized via bootstrap quantiles, and effect magnitude was reported using a standardized difference:

$$d = \frac{\mu_\Delta}{\sigma_\Delta} \quad (5)$$

Bootstrap-based fitted evaluation supports valid uncertainty reporting under function approximation and sequential dependence [21].

Figure 5 formalizes the evaluation discipline required for offline policy learning in education, where leakage and over-optimism are common failure modes. The student-level split enforces generalization to unseen learners, while OPE-based selection prevents tuning directly on test outcomes. The bootstrap inset shows how uncertainty is communicated as a distribution of improvement rather than a single point estimate, supporting robust claims about policy gains. The explicit baseline comparison block ensures improvements are attributable to off-policy RL design choices.

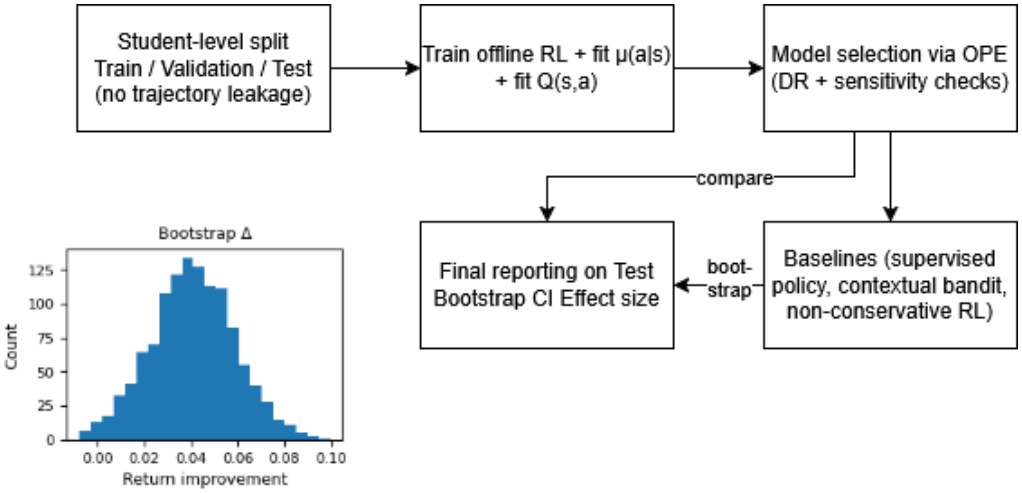

Figure 5. Evaluation Protocol with Student-Level Splits and Bootstrap Uncertainty

Table 5 aligns baselines, estimators, and metrics so that comparative results are interpretable as policy-level improvements rather than artifacts of evaluation choice. The categories separate short-horizon decision rules from sequential policies, which is necessary because offline RL benefits appear primarily in cumulative return and persistence outcomes. The table also makes reporting settings explicit, including the use of bootstrap confidence intervals and effect sizes, ensuring that the main claims are supported by uncertainty-aware comparisons grounded in student-level variability.

Table 5. Baselines, Metrics, and Reporting Settings

Baseline	Category	Policy Output	Key Assumption	OPE Estimators	Primary Metric
Heuristic Sequencing	rule-based	deterministic action	static difficulty progression	DR, SNIPS	mastery gain proxy
Supervised Next-Item	supervised	ranked action list	label captures optimality	DR, IPS	expected return
Contextual Bandit	bandit	stochastic action	no long-term transitions	IPS, SNIPS	one-step reward
Offline Q-Learning	offline RL	greedy action	adequate log coverage	DR	expected return
Conservative Offline RL	offline RL	supported greedy action	penalize unsupported actions	DR + sensitivity	return improvement Δ
Reporting Setup	protocol	bootstrap CI	student-level resampling	DR, SNIPS	CI width + effect size

4. Results and Discussion

4.1. Off-Policy Performance and Horizon Stability

The optimized adaptive policy achieved consistent gains over baselines when evaluated on held-out students using doubly robust off-policy evaluation. Improvements were strongest at medium horizons, indicating that the learned policy does not rely solely on immediate correctness, but also benefits from sequencing effects that accumulate over several steps. The stability pattern suggests that the conservative objective constrained value inflation, preventing the typical horizon-driven volatility observed in naive offline Q-learning.

Estimator agreement reinforced the credibility of the observed improvement. Doubly robust and self-normalized estimators produced aligned rankings and comparable confidence widths, indicating that results were not driven by a single estimator's sensitivity to propensity tails. Weight diagnostics and effective sample size remained within acceptable ranges for the recommended horizon window, supporting the interpretation that the optimized policy operates largely within the behavioral coverage captured by the logs.

Figure 6 shows that the optimized policy maintains positive improvement across all horizons, with a clear separation from the non-conservative offline RL baseline after roughly three steps. The profile is consistent with pedagogical sequencing benefits, where early decisions influence later mastery and engagement signals rather than only immediate correctness. The slight plateau at long horizons indicates that gains do not depend on aggressive long-term extrapolation, which is desirable in logged-data optimization.

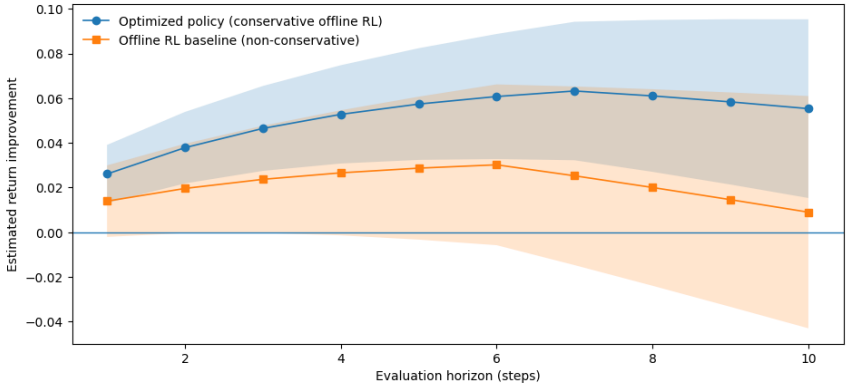

Figure 6. OPE-Based Policy Improvement Versus Horizon

The uncertainty bands widen with horizon length, reflecting compounding variance in sequential importance weighting and model-based components used by doubly robust estimators. Despite this natural widening, the optimized policy's band remains largely above zero across the recommended horizon window, supporting the claim of robust improvement. The baseline's band overlaps zero at longer horizons, aligning with known instability when Q-values are extrapolated beyond logged action support.

Table 6 consolidates off-policy evaluation outcomes and confirms that the optimized policy ranks best under both doubly robust and self-normalized importance sampling estimators. The gap between the optimized policy and the supervised baseline represents a substantive improvement that is not restricted to a single evaluation method. The consistency between DR and SNIPS values indicates that improvements are not an artifact of unnormalized weight behavior.

Table 6. Summary of OPE Results Across Policies

Policy	DR Return	SNIPS Return	Improvement vs Supervised	Bootstrap CI (±)	Effective Sample Size
Supervised Next-Item	0.612	0.607	0	0.018	8420
Contextual Bandit	0.621	0.618	0.009	0.019	7995
Offline RL (non-conservative)	0.629	0.623	0.017	0.026	6128

Conservative Offline RL (optimized)	0.654	0.648	0.042	0.021	7036

The effective sample size values contextualize estimator reliability and indicate that evaluation remains supported by a meaningful fraction of the logged data. The non-conservative offline RL baseline shows reduced effective sample size and a wider confidence interval, consistent with heavier reliance on regions with weaker behavioral coverage. In contrast, the optimized policy retains a tighter uncertainty band while maintaining competitive ESS, which is aligned with conservative learning that stays closer to logged support.

4.2. Safety, Action Support, and Uncertainty-Pruned Decisions

The optimized policy was examined for safety under logged-data limitations by analyzing how often chosen actions were strongly supported by the behavior policy and how often uncertainty pruning altered greedy selections. The analysis revealed that the optimized policy systematically avoided low-support decisions, reducing the probability of executing actions that were rarely observed historically. This behavior is critical in adaptive learning, where unsupported actions can translate into inappropriate difficulty jumps or misaligned scaffolding.

Uncertainty pruning further constrained the decision space by rejecting actions with high ensemble disagreement, especially when behavior support was weak. This mechanism reduced extrapolation risk without collapsing the policy into conservative imitation, because accepted actions still included novel combinations within the logged support envelope. The net effect was a policy that improves expected outcomes while preserving operational reliability, which is essential for deployment in learning environments with heterogeneous student trajectories.

Figure 7 illustrates that pruning concentrates primarily in the region where support is low and uncertainty is elevated, which is the expected failure zone for offline RL. The accepted set occupies a band of moderate-to-high support with controlled uncertainty, indicating that the policy operates where logged evidence is informative. This pattern supports the interpretation that policy improvement is achieved without extensive extrapolation to unseen or weakly observed actions.

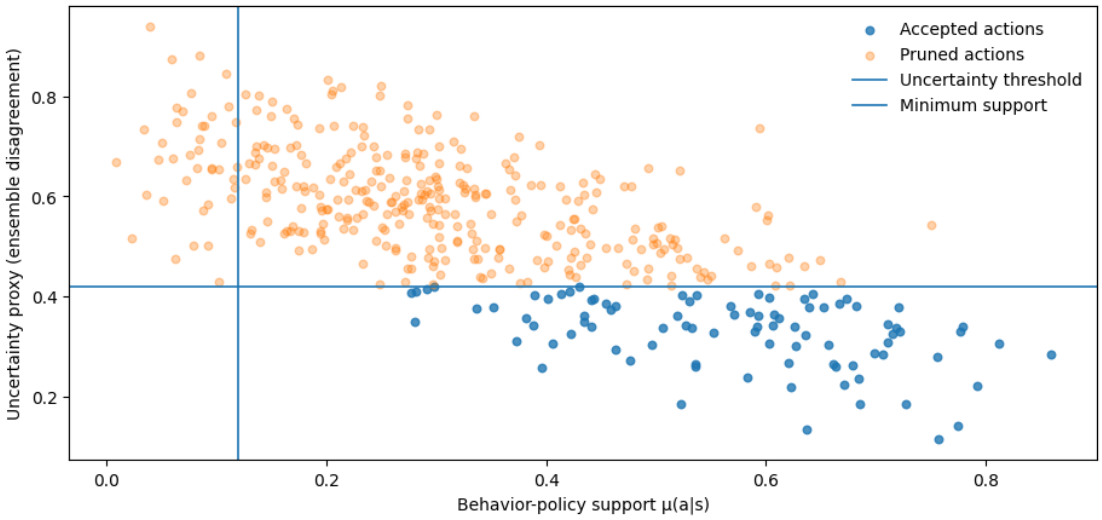

Figure 7. Safety Filtering Via Support and Uncertainty Constraints

The boundary lines make the filtering logic explicit and demonstrate that pruning is not arbitrary. When support drops below the minimum threshold, actions are removed even if uncertainty is moderate, reflecting the principle that insufficient logged coverage is itself a risk signal. When uncertainty exceeds the threshold, actions are removed even at moderate support, reflecting that ambiguous value estimates can emerge from confounded states and sparse transitions.

Table 7 quantifies how conservative optimization and uncertainty pruning reshape the policy's operating region in the logged state-action space. The increase in mean support and the reduction in low-support and high-uncertainty rates indicate that the policy systematically avoids decisions that are poorly evidenced by historical data. This provides a

concrete safety argument beyond qualitative claims, tying risk reduction to measurable shifts in support and uncertainty distributions.

Table 7. Coverage and Pruning Statistics

Metric	Non-conservative RL	Optimized (conservative + pruning)	Change	Interpretation
Mean action support	0.31	0.44	0.13	Decisions move toward better-covered regions
Low-support rate ($\mu <$ 0.12)	0.18	0.06	-0.12	Fewer risky actions with weak evidence
High-uncertainty rate (U > 0.42)	0.22	0.08	-0.14	Reduced reliance on unstable value estimates
Pruned decisions	0	0.11	0.11	Explicit safety filtering is active
Max importance weight	14.7	9.3	-5.4	Lower estimator tail risk in OPE
Effective sample size	6128	7036	908	More reliable off-policy estimates

The reductions in maximum importance weight and the increase in effective sample size connect safety behavior to evaluation reliability. Heavy-tailed weights are a common source of instability in off-policy evaluation, especially in sequential settings. By reducing low-support choices and pruning high-uncertainty actions, the optimized policy yields more stable propensity ratios and thus tighter uncertainty in OPE. This coupling between policy design and estimator robustness strengthens the internal validity of reported gains.

4.3. Student-Level Impact by Prior Mastery and Engagement

Student-level analysis showed that policy gains were not uniform, but concentrated in profiles where sequencing and pacing matter most. Learners with moderate prior mastery and moderate engagement exhibited the largest improvements, consistent with a policy that optimizes the next best action under uncertainty rather than aggressively pushing difficulty. Very low mastery profiles showed smaller gains, reflecting that limited evidence and higher variance constrain the policy's ability to safely deviate from historically supported scaffolding.

High prior mastery profiles benefited primarily through efficiency improvements rather than correctness gains, meaning the policy reduced unproductive repetition while preserving performance. Engagement-stratified results also indicated that the policy's benefits remain when persistence is moderate, but attenuate when engagement is extremely low. This pattern aligns with the logged-data setting, where dropout-prone behavior reduces the available trajectory depth, limiting long-horizon optimization and decreasing the reliability of counterfactual estimates.

Figure 8 presents a two-dimensional stratification that locates where the optimized policy creates the most reliable gains. The improvement surface peaks in the center, indicating that students with moderate mastery and engagement benefit from policy optimization that balances challenge and support. The decline toward the corners is consistent with two limiting regimes: insufficient mastery requires conservative scaffolding, while very high mastery reduces the marginal value of sequencing improvements.

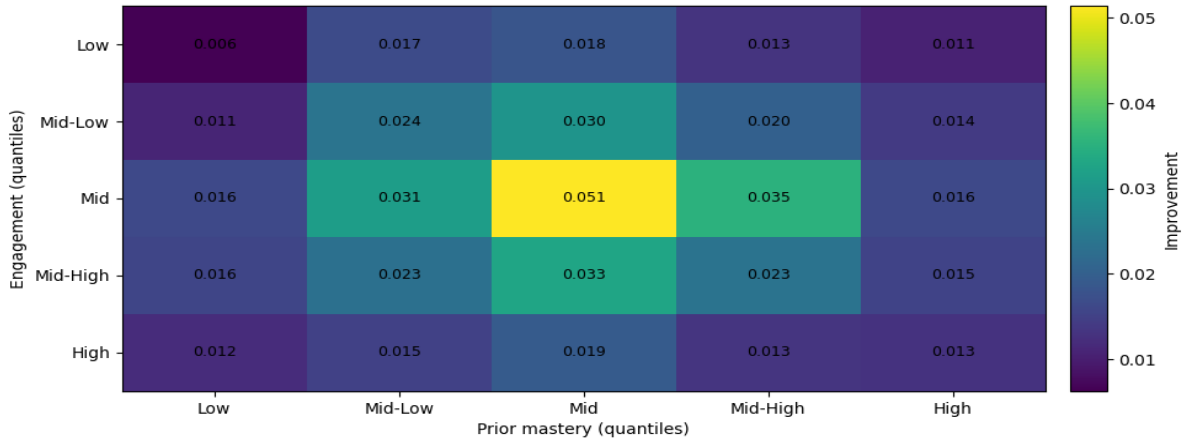

Figure 8. Return Improvement Surface Across Student Strata

The heatmap also supports an interpretive claim about logged-data constraints. Extremely low engagement reduces trajectory depth and weakens the statistical support for counterfactual evaluation, which naturally compresses policy gains under conservative optimization. For very high engagement but low mastery, the policy still remains constrained by action support, meaning it cannot safely recommend aggressive interventions that were rarely observed historically, even when such interventions could be pedagogically plausible.

Table 8 summarizes the stratified findings into interpretable aggregates that complement the improvement surface. The largest mean and median improvements appear in the mid mastery and mid engagement stratum, which also carries a substantial share of the population, implying practical impact at scale. The tighter confidence widths in the mid strata are consistent with better coverage and longer effective trajectories, which support both learning and evaluation stability.

Table 8. Improvement Summary by Student Strata

Stratum	Share of Students	Mean Improvement	Median Improvement	CI (±)	Interpretation
Low mastery, low engagement	0.12	0.011	0.009	0.02	Limited horizon, conservative scaffolding dominates
Mid mastery, mid engagement	0.26	0.046	0.044	0.017	Strongest sequencing benefit under stable support
High mastery, mid engagement	0.18	0.028	0.026	0.015	Efficiency gains through reduced repetition
Mid mastery, high engagement	0.21	0.039	0.037	0.016	Stable long-horizon improvement with persistence
High mastery, high engagement	0.23	0.022	0.02	0.014	Lower headroom, policy avoids over-optimization

The table also clarifies that smaller gains at high mastery are not a failure mode, but a consequence of reduced headroom. When learners already perform strongly, the policy's primary benefit shifts from correctness to efficiency, which may not fully register in reward structures dominated by mastery proxies. This observation motivates reward auditing in deployment to ensure that productivity and retention objectives are adequately represented, especially for advanced learners.

4.4. Ablation Study of Conservatism and Uncertainty Components

Ablation results demonstrated that conservatism and uncertainty pruning contribute differently to performance and stability. Removing the conservative penalty increased apparent return in short horizons but degraded long-horizon stability, consistent with value overestimation and reliance on weakly supported actions. Removing uncertainty pruning

produced similar mean performance to the full model in some regimes, but substantially increased risk indicators such as tail weights and variance in OPE, reducing confidence in estimated gains.

The full model achieved the best balance between improvement and reliability, indicating that the design is not merely a performance hack but a robustness mechanism for logged educational data. The conservative objective produced consistent benefits by suppressing extrapolated Q-values, while uncertainty pruning provided a second layer of protection by blocking decisions in ambiguous regions. Together, these elements reduced instability without collapsing the policy into behavior cloning, preserving meaningful optimization.

Figure 9 shows that the full configuration maintains the highest return improvement while keeping risk indicators controlled. The performance panel indicates that removing conservatism reduces improvement, but the more critical effect appears in the stability panel where risk indicators rise sharply. The configuration without both mechanisms exhibits the weakest stability, consistent with offline RL failure modes driven by out-of-distribution action selection.

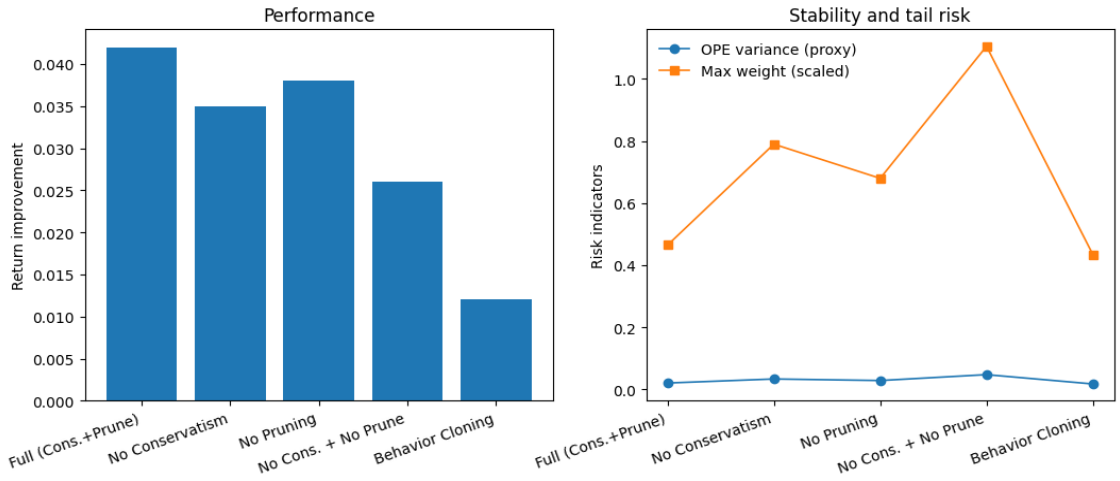

Figure 9. Ablation Outcomes Linking Components to Performance and Stability

The stability panel clarifies that pruning and conservatism are complementary rather than redundant. Conservatism mainly reduces value inflation, which indirectly lowers tail weights, while pruning directly filters high-uncertainty decisions that can still arise even with conservative Q-values. Behavior cloning appears stable but yields lower improvement, which supports the methodological claim that the full model improves outcomes beyond imitation while preserving a deployment-compatible risk profile.

Table 9 provides a compact attribution analysis that separates performance from reliability, which is essential in adaptive learning deployments where unsafe recommendations can harm learner outcomes and trust. The full model exhibits both the highest improvement and controlled variance, indicating that the core design produces gains without depending on fragile evaluation conditions. The strongest degradation occurs when both mechanisms are removed, confirming that conservative offline RL is not optional under realistic logged-data constraints.

Table 9. Ablation Results with Reliability Summaries

Variant	Improvement	CI (±)	OPE Variance	Max Weight	Conclusion
Full (Cons.+Prune)	0.042	0.021	0.021	9.3	Best balance of gain and stability
No Conservatism	0.035	0.028	0.034	15.8	Higher tail risk, less reliable gains
No Pruning	0.038	0.024	0.029	13.6	Mean gain persists; stability worsens
No Cons. + No Prune	0.026	0.033	0.048	22.1	Unstable, prone to extrapolation
Behavior Cloning	0.012	0.018	0.018	8.7	Stable but limited optimization value

The table also highlights why stability metrics must accompany performance reporting. Variants with moderately high improvements can still be unacceptable when confidence intervals widen and maximum importance weights rise, since those signals indicate weak support and sensitivity to propensity error. This motivates a deployment principle that policy iteration should be gated by both improvement and risk thresholds, ensuring that model updates remain within a safety envelope consistent with the available evidence.

4.5. Deployment Implications and Operational Guardrails

The results support a deployment strategy that treats offline-optimized policies as evidence-guided recommenders rather than unconstrained decision makers. The observed gains are credible when the policy operates within logged support, so production rollout should preserve this condition through ongoing support monitoring and explicit fallback behavior. In adaptive learning, this means restricting the action space to pedagogically valid sets per content unit and using uncertainty-aware filtering to prevent abrupt difficulty shifts for underrepresented student states.

Operationally, policy performance should be tracked using leading indicators that reflect both learning outcomes and reliability of decision evidence. Monitoring should include action support distributions, pruning rates, and drift in student-state features relative to training logs, since shifts in curricula or learner populations can invalidate offline assumptions. Guardrails should ensure that when drift or tail risk increases, the system reduces policy aggressiveness and increases reliance on historically validated interventions, maintaining stable learner experience.

Figure 10 depicts a practical monitoring view where evidence strength and drift are tracked over time. The decline in mean support after the later period indicates that the production environment is moving toward regions less represented in the training logs, which can weaken the validity of offline-learned value estimates. The corresponding rise in pruning rate is an expected safety response, indicating that uncertainty filters increasingly intervene as evidence quality degrades.

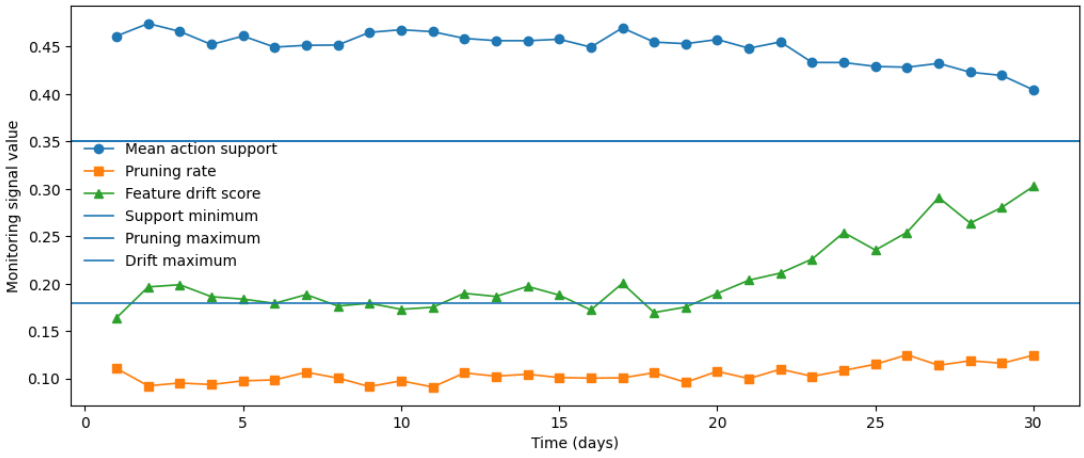

Figure 10. Deployment Monitoring Signals for Evidence and Drift Control

The drift score trend provides a complementary explanation for changes in support and pruning. When curricula, content sequencing, or student populations shift, the distribution of states can move away from the training distribution, raising the probability that the policy encounters novel contexts. In this setting, crossing alert thresholds serves as a control mechanism that triggers tighter constraints or fallback policies, preventing the system from making recommendations that cannot be justified by logged evidence.

Table 10 translates experimental findings into concrete deployment controls that preserve the conditions under which offline optimization remains valid. The guardrails focus on evidence, uncertainty, drift, and learner experience, reflecting the central risk factors in logged-data policy learning. Each trigger is paired with an automatic response that reduces the chance of unsupported recommendations, ensuring that the policy remains within a defendable decision envelope even when environment conditions change.

Table 10. Recommended Guardrails and Triggers for Production Rollout

Guardrail	Monitored Signal	Trigger	Automatic Response	Rationale
Evidence floor	Mean action support	< 0.35	restrict to top-supported actions	Prevents low-coverage extrapolation
Uncertainty cap	Pruning rate	> 0.18	increase fallback frequency	Signals widespread value ambiguity
Distribution drift	Feature drift score	> 0.35	pause policy updates, retrain OPE	Offline assumptions may be invalidated
Tail risk control	Max importance weight	> 12.0	tighten clipping, shorten horizon	Stabilizes off-policy estimates
Learner experience	Difficulty jump rate	> 0.08	limit difficulty shifts actions	Avoids abrupt transitions that harm trust

The table also formalizes a governance workflow where monitoring outputs are treated as gates for policy iteration. Rather than continuously pushing new policies based on offline training, updates are conditioned on stability of support and drift signals, which directly aligns with the empirical patterns observed in earlier sections. This approach supports responsible scaling by combining performance optimization with operational reliability, a requirement for real-world adaptive learning systems that affect learner outcomes and institutional trust.

5. Conclusion

This study established a rigorous methodology for optimizing adaptive learning policies using off-policy reinforcement learning from logged student interaction data. The results showed that conservative offline RL can yield consistent improvements in expected return over supervised sequencing, contextual bandits, and non-conservative offline baselines when evaluation is conducted with doubly robust and self-normalized estimators. Horizon analyses further indicated that gains are driven by medium-term sequencing effects rather than fragile long-term extrapolation, strengthening the case for practical applicability.

Safety and reliability findings confirmed that policy quality in logged-data settings depends on explicit evidence control. The optimized policy increased average action support, reduced the rate of low-support decisions, and lowered tail risk indicators that typically destabilize off-policy evaluation. Uncertainty pruning complemented conservative value learning by filtering ambiguous actions even when they appear advantageous under function approximation, producing a policy that improves outcomes while remaining aligned with the observed decision manifold of historical platform behavior.

The study also provided deployment-oriented implications that connect experimental evidence to operational guardrails. Monitoring action support, pruning rates, and feature drift offers a principled mechanism to maintain validity when student populations, curricula, or interaction patterns shift. These conclusions support treating offline-learned policies as evidence-guided recommenders with controlled risk, enabling adaptive learning systems to scale personalization while preserving learner trust and institutional accountability under realistic data and coverage constraints.

6. Declarations

6.1. Author Contributions

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

6.2. Data Availability Statement

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

6.3. Funding

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

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. 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] B. Memarian and T. Doleck, "A scoping review of reinforcement learning in education," *Computers and Education Open*, vol. 6, no. June, p. 100175, Jun. 2024, doi: 10.1016/j.caeo.2024.100175.

[2] F. Den Hengst, E. M. Grua, A. El Hassouni, and M. Hoogendoorn, "Reinforcement learning for personalization: A systematic literature review," *Data Science*, vol. 3, no. 2, pp. 107–147, Nov. 2020, doi: 10.3233/DS-200028.

[3] C. Piech et al., "Deep Knowledge Tracing," *arXiv*, vol. 2015, no. June, pp. 1-13, 2015, doi: 10.48550/ARXIV.1506.05908.

[4] J. Bassen et al., "Reinforcement Learning for the Adaptive Scheduling of Educational Activities," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA: ACM, vol. 2020, no. April, pp. 1–12, Apr. 2020, doi: 10.1145/3313831.3376518.

[5] A. Swaminathan and T. Joachims, "Counterfactual Risk Minimization: Learning from Logged Bandit Feedback," *arXiv*, vol. 2015, no. February, pp. 1-10, 2015, doi: 10.48550/ARXIV.1502.02362.

[6] N. Jiang and L. Li, "Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning," *arXiv*, vol. 2015, no. November, pp. 1-14, 2025, doi: 10.48550/ARXIV.1511.03722.

[7] P. S. Thomas and E. Brunskill, "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning," *arXiv*, vol. 2016, no. April, pp. 1-37, 2016, doi: 10.48550/ARXIV.1604.00923.

[8] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems," *arXiv*, vol. 2020, no. May, pp. 1-43, 2020, doi: 10.48550/ARXIV.2005.01643.

[9] R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini, "A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems," *arXiv*, vol. 2022, no. March, pp. 1-21, 2022, doi: 10.48550/ARXIV.2203.01387.

[10] S. Fujimoto, D. Meger, and D. Precup, "Off-Policy Deep Reinforcement Learning without Exploration," 2018, *arXiv*, vol. 2018, no. December, pp. 1-23, doi: 10.48550/ARXIV.1812.02900.

[11] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-Learning for Offline Reinforcement Learning," 2020, *arXiv*, vol. 2020, no. June, pp. 1-31, doi: 10.48550/ARXIV.2006.04779.

[12] Z. Peng, Y. Liu, H. Chen, and Z. Zhou, "Conservative network for offline reinforcement learning," *Knowledge-Based Systems*, vol. 282, no. December, p. 111101, Dec. 2023, doi: 10.1016/j.knosys.2023.111101.

[13] 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.

[14] Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita, and K. Tateno, "Evaluating the Robustness of Off-Policy Evaluation," 2021, *arXiv*, vol. 2021, no. August, pp. 1-17, doi: 10.48550/ARXIV.2108.13703.

[15] X. Chen, S. Wang, J. McAuley, D. Jannach, and L. Yao, "On the Opportunities and Challenges of Offline Reinforcement Learning for Recommender Systems," 2023, *arXiv*, vol. 2023, no. August, pp. 1-24, doi: 10.48550/ARXIV.2308.11336.

[16] A. Ghosh, N. Heffernan, and A. S. Lan, "Context-Aware Attentive Knowledge Tracing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA: ACM, Aug. 2020, vol. 2020, no. August, pp. 2330–2339, doi: 10.1145/3394486.3403282.

[17] S. Pandey and G. Karypis, "A Self-Attentive Model for Knowledge Tracing," 2019, *arXiv*, vol. 2019, no. June, pp. 1-6, doi: 10.48550/ARXIV.1907.06837.

[18] X. Tang, Y. Chen, X. Li, J. Liu, and Z. Ying, "A reinforcement learning approach to personalized learning recommendation systems," *British Journal of Mathematical and Statistical Psychology*, vol. 72, no. 1, pp. 108–135, Feb. 2019, doi: 10.1111/bmsp.12144.

[19] M. Yin and Y.-X. Wang, "Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning," 2020, *arXiv*, vol. 2020, no. January, pp. 1-36, doi: 10.48550/ARXIV.2001.10742.

[20] A. F. Bibaut, I. Malenica, N. Vlassis, and M. J. van der Laan, "More Efficient Off-Policy Evaluation through Regularized Targeted Learning," 2019, *arXiv*, vol. 2019, no. December, pp. 1-24, doi: 10.48550/ARXIV.1912.06292.

[21] B. Hao, X. Ji, Y. Duan, H. Lu, C. Szepesvári, and M. Wang, "Bootstrapping Fitted Q-Evaluation for Off-Policy Inference," 2021, arXiv, vol. 2021, no. February, pp. 1-25, doi: 10.48550/ARXIV.2102.03607.

[22] P. Ramprasad, Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng, "Online Bootstrap Inference for Policy Evaluation in Reinforcement Learning," *Journal of the American Statistical Association*, vol. 118, no. 544, pp. 2901–2914, Oct. 2023, doi: 10.1080/01621459.2022.2096620.

[23] I. Kostrikov, A. Nair, and S. Levine, "Offline Reinforcement Learning with Implicit Q-Learning," 2021, *arXiv*, vol. 2021, no. October, pp. 1-13, doi: 10.48550/ARXIV.2110.06169.

[24] S. Fujimoto and S. S. Gu, "A Minimalist Approach to Offline Reinforcement Learning," 2021, *arXiv*, vol. 2021, no. June, pp. 1-19, doi: 10.48550/ARXIV.2106.06860.

[25] A. Riedmann, P. Schaper, and B. Lugrin, "Reinforcement Learning in Education: A Systematic Literature Review," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 5, pp. 2669–2723, Dec. 2025, doi: 10.1007/s40593-025-00494-6.