Adaptive Pessimism with Normalised Return Targets for Educational Offline Reinforcement Learning

by Manajemen S1

Submission date: 07-Jul-2026 12:39PM (UTC+0700)
Submission ID: 2857809231
File name: yusrilhelmi_paper3jict_AdaptivePessimism.pdf (785.88K)
Word count: 6502
Character count: 37023

JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Katuk, N., Brown, P., & Yusof, Y. (202X). The title of your manuscript: Read and follow this guideline to avoid desk rejection. *Journal of Information and Communication Technology, XX*(X), XX-XXX. https://doi.org/10.32890/jict202X.XX.X.X

Adaptive Pessimism with Normalised Return Targets for Educational Offline Reinforcement Learning

[1]Muhammad Yusril Helmi Setyawan, [2]Yudi Priyadi & [3]Patah Herwanto
[1] Applied Bachelor Program of Informatics Engineering, School of Information Technology, Universitas Logistik dan Bisnis Internasional, Indonesia
[2]Center of Excellence of Artificial Intelligence for Learning and Optimization, Telkom University, Indonesia
[3]Department of Informatics, Faculty of Information Technology, Universitas Ekuitas, Indonesia

*[1]yusrilhelmi@ulbi.ac.id
[2]whyphi @telkomuniversity.ac.id
[3]pherwanto@ekuitas.ac.id
*Corresponding author

ABSTRACT

Historical learner-system logs offer a practical route for studying adaptive educational policies without exposing learners to online trial-and-error. The difficulty is that offline reinforcement learning must learn from a fixed behaviour dataset, where distribution shift and Q-value overestimation can make unsupported actions appear better than they are. Using EdNet-KT4, this study examines a tabular question part recommendation setting and adapts the target-pessimism idea behind APTQ by driving the conservative coefficient toward a normalised return-to-go target. Across experiments with 50,000, 100,000, and 200,000 transitions, Conservative Q-Learning remains stronger in Bellman consistency and fitted Q evaluation (FQE) value. The normalised target variant learns a policy closer to logged learner behaviour, improving action match on the 200,000-transition subset while keeping Bellman and FQE fit competitive. The evidence points to a targeted role for normalised target pessimism: it stabilises behaviour-aligned educational offline RL, while CQL retains the stronger value-fit profile. The contribution is a reproducible EdNet-KT4 proxy benchmark and systematic evidence for the trade-off between conservative value learning and behaviour alignment in educational logs.

Keywords: Adaptive learning, conservative Q-learning, EdNet-KT4, fitted Q evaluation, offline reinforcement learning.

INTRODUCTION

Adaptive learning systems increasingly leave behind detailed traces of learner activity: responses, attempts, pauses, review actions, and other forms of interaction with learning content. These traces are valuable because they describe how learners and systems actually interacted over time, not only as isolated prediction labels. A natural question follows from this setting: can such historical logs be used to learn adaptive policies without experimenting on learners online? Offline reinforcement learning is appealing here because it learns from a fixed dataset and does not require further interaction with the environment during training (Prudencio et al., 2024; Chen et al., 2024).

The main obstacle is not only computational. In offline RL, the learned policy may assign high values to actions that were rarely, or never, observed for similar learner states. In a simulator this is already a problem; in education it is more sensitive because an unsupported recommendation could mean unsuitable learning material, not merely a lower benchmark score. For this reason, educational offline-RL evidence is more informative when value estimates are reported together with behaviour alignment and support coverage.

EdNet is a useful public setting for this question because it contains large-scale learner-system interactions from the Santa learning platform (Choi et al., 2020). The KT4 level is especially relevant because it includes richer activity information than question-answer correctness alone. Earlier work has already used EdNet to study reinforcement learning for question sequencing and to examine how state representation affects policy quality (Azhar et al., 2022). In this study, EdNet-KT4 serves as a public proxy for educational policy learning, allowing the offline-RL mechanism to be examined before any controlled deployment or longitudinal validation.

CQL is a natural conservative baseline in this setting because it penalises unsupported high Q-values and is designed to reduce overestimation under distribution shift (Kumar et al., 2020). APTQ adds another idea: instead of keeping pessimism fixed, it adapts the penalty strength toward a target Q-value (Liu et al., 2024). The transfer question is whether this target-adaptive idea remains stable on educational logs. Raw return-to-go values in EdNet-KT4 can be much larger than early tabular Q-values, so a direct target can make the adaptive penalty collapse or drift.

This paper focuses on one specific form of that question: whether a normalised return-to-go target can make an APTQ-style mechanism stable enough for an EdNet-KT4 educational proxy task. The analysis compares that behaviour-aligned target design with CQL under Bellman consistency, behaviour support, action match, and fitted Q evaluation (FQE).

Three questions guide the study, moving from problem formulation to target stability and finally to the policy trade-off:

RQ1. Which tabular Markov Decision Process (MDP) design is most stable for EdNet-KT4 offline reinforcement learning experiments?

RQ2. Does normalised return-to-go targeting stabilise APTQ compared with reward-based or raw return-to-go targets?

RQ3. How does APTQ with normalised return-to-go targets compare with CQL under Bellman consistency, behaviour alignment, and FQE?

The contribution is organised around four points. First, the study builds a reproducible tabular offline-RL formulation for EdNet-KT4 using state-action-reward-next-state transitions. Second, it adapts target

pessimism to an educational proxy task through a normalised return-to-go target. Third, it reports a multi-step empirical study covering MDP design, target-source ablation, scaling validation, and FQE-based offline policy evaluation. Fourth, it reports the trade-off directly: normalised target pessimism improves behaviour alignment relative to CQL, while CQL remains stronger on Bellman mean squared error (MSE) and FQE estimated value.

The article follows this sequence. The method section makes the EdNet-KT4 tabular construction and normalised target update explicit; the results section tests those choices through the 200,000-transition comparison, scaling, ablation, and FQE; and the discussion interprets the outcome as a trade-off between conservative value fit and behaviour alignment.

RELATED WORKS

Educational Data Mining and EdNet

Large-scale educational data mining depends on datasets that preserve the order and context of learner behaviour. EdNet was introduced as a hierarchical dataset of interactions with an intelligent tutoring service (Choi et al., 2020). Its levels allow researchers to move from question-response records to richer traces of learning activity. This paper uses EdNet-KT4 because those richer traces are closer to the sequential decision-making view required by offline reinforcement learning.

Recent EDM surveys place knowledge tracing, learner behaviour modelling, performance prediction, and personalised recommendation among the central tasks of learning analytics (Lin et al., 2025; Batool et al., 2023). This literature supports the use of rich learner logs while underlining that policy-behaviour evidence ultimately requires controlled learning-effect validation. Reviews of reinforcement learning for instructional sequencing add a more specific warning: reward design and offline evaluation are recurring difficulties when RL is applied to educational data (Doroudi et al., 2019; Memarian & Doleck, 2024).

Prior reinforcement learning work on EdNet has examined question sequencing and the role of state representation (Azhar et al., 2022). It is relevant here because it shows how state design and data support can shape policy quality. We retain that concern with support while shifting the methodological focus to conservative and target-adaptive tabular offline-RL policies, with CQL and an APTQ-style variant as the main contrast.

Offline Reinforcement Learning and Conservative Value Learning

Offline reinforcement learning learns a policy from a static dataset collected by another behaviour policy. The central difficulty is familiar but severe: once the learned policy chooses actions outside the logged support, standard Q-learning can repeatedly maximise values for state-action pairs that were not sufficiently observed. This is why reviews of offline RL emphasise support constraints, pessimism, and reliable off-policy evaluation (Prudencio et al., 2024). The same caution appears in real-world RL studies, where limited interaction, safety, evaluation, and distribution shift are treated as deployment obstacles rather than secondary implementation issues (Dulac-Arnold et al., 2021).

CQL addresses this risk by learning conservative Q-values (Kumar et al., 2020). It reduces values for unsupported actions while preserving values for observed state-action pairs. This makes it a strong baseline for educational policy learning, where unsupported actions should be treated cautiously rather than rewarded by extrapolation.

CQL is not the only family of offline-RL methods. Policy-constraint approaches such as batch-

constrained Q-learning restrict the learned policy to actions close to the data (Fujimoto et al., 2019), TD3+BC adds an explicit behaviour-cloning term to the policy objective (Fujimoto & Gu, 2021), implicit Q-learning avoids querying out-of-sample actions altogether (Kostrikov et al., 2022), and Decision Transformer treats offline RL as return-conditioned sequence modelling (Chen et al., 2021). CQL is used as the main conservative baseline here because its penalty acts directly on Q-values, which transfers naturally to a tabular educational setting; a comparison with these other families is left for future work.

Adaptive Pessimism and Off-Policy Evaluation

APTQ adapts the pessimism coefficient toward a target Q-value rather than fixing it in advance (Liu et al., 2024). The attraction is clear: a single fixed penalty can be too weak in one setting and too strong in another. The difficulty is target design. If the target is on the wrong scale, the adaptive coefficient may collapse or drift, especially in tabular educational settings where rewards, returns, and learned Q-values do not naturally occupy the same range.

Offline policy evaluation is needed because online evaluation is often unavailable during educational system development. FQE estimates the value of a fixed target policy by fitting an evaluation Q-function on logged transitions, drawing on batch-RL and off-policy evaluation literature (Ernst et al., 2005; Le et al., 2019; Kallus & Uehara, 2020; Wang et al., 2024). In this study, FQE complements Bellman MSE, support, and action match by adding an offline value-estimation view.

This way of reading the evidence is consistent with the intelligent tutoring systems literature, where controlled or longitudinal validation remains central when log-based adaptation is interpreted as learning-effect evidence (Kulik & Fletcher, 2016; Letourneau et al., 2025).

Comparative Analysis and Research Gap

Table 1 brings these strands together by comparing each study's dataset, method, and point of contact with the present work. EdNet studies make the learner-log setting concrete, especially state design and question sequencing. Offline-RL studies, in turn, motivate conservative value learning, target-adaptive pessimism, and off-policy evaluation mainly in benchmark or logged-decision domains. What remains to be examined is how these ideas behave in an EdNet-KT4 educational proxy, where behaviour alignment, Bellman consistency, and FQE can be read side by side.

Table 1

Comparison of Related Studies and the Present Work

Study	Dataset or domain	Main method	Relevance and remaining gap
Choi et al. (2020)	EdNet learner logs	Large-scale dataset construction	Provides the public learner-system interaction source used in this study.
Lin et al. (2025)	Educational data mining	Deep learning survey for EDM	Positions knowledge tracing and personalised recommendation as central EDM tasks related to sequential learner modelling.
Azhar et al. (2022)	EdNet	RL-based question sequencing and state representation analysis	Shows the relevance of EdNet for educational RL and motivates a closer look at conservative and target-adaptive offline policies.
Doroudi et al. (2019)	Instructional sequencing	Review of RL for instructional sequencing	Identifies reward design and offline evaluation as recurring difficulties, which this study addresses with a proxy reward and FQE.
Prudencio et al. (2024)	Offline RL	Taxonomy and review of	Supports the framing of support

Study	Dataset or domain	Main method	Relevance and remaining gap
	benchmarks	offline RL	constraints, pessimism, and offline evaluation as core offline-RL issues.
Kumar et al. (2020)	General offline RL benchmarks	Conservative Q-Learning	Provides the conservative baseline used to control unsupported action overestimation.
Liu et al. (2024)	General offline RL benchmarks	Adaptive pessimism via target Q-value	Motivates target-adaptive pessimism for transfer to an EdNet-KT4 educational setting.
Wang et al. (2024)	Logged decision data	Reliable off-policy evaluation	Supports uncertainty-aware evaluation before deploying policies learned from fixed historical logs.
Present study	EdNet-KT4	Tabular CQL and APTQ-style normalised target pessimism with FQE	Evaluates the behaviour-alignment and value-estimation trade-off in an educational offline-RL proxy task.

THE PROPOSED METHOD

MDP Construction

The logged EdNet-KT4 interactions are converted into a fixed transition dataset, written as

$$D = \{(s_t, a_t, r_t, s_{t+1}, d_t)\}_{t=1}^{N}. \quad (1)$$

In Equation 1, s_t is the learner state, a_t is the selected action, $r_t$is the reward, s_{t+1} is the next learner state, and dt marks the end of a trajectory. Every policy in the study learns from these logged transitions, so the experiment is framed as an offline educational proxy. The state is the discrete tuple

$$s_t = \left(b_t^{att},\ b_t^{acc},\ p_{t-1},\ c_{t-1},\ b_t^{gap}\right). \quad (2)$$

where b_t^{att} is the attempt-count bin, b_t^{acc} the running-accuracy bin, p_{t-1} the previous question part, c_{t-1} the previous correctness, and b_t^{gap} the time-gap bin. This discretisation yields the 626 unique states reported in the results. The tabular formulation keeps the analysis focused on conservative and adaptive pessimism before introducing neural function approximation.

The action is the question part, encoded as seven discrete choices. A richer action space, for example a part-source combination, would carry more semantic detail, but it would also make the logged support thinner. In an offline setting, that sparsity directly weakens value estimation. The seven-action formulation is therefore a practical compromise between educational meaning and sufficient observations per state-action pair.

The reward is a correctness proxy used to compare policies on the same logged interaction signal. Let $c_t \in \{0,1\}$ denote whether the learner answered correctly. The reward and return-to-go are defined as

$$r_t = 2c_t - 1, \qquad RTG_t = \sum_{k=t}^{T} \gamma_{rtg}^{k-t}\, r_k, \qquad \gamma_{rtg} = 0.99 \quad (3)$$

Equation 3 keeps the signal simple and reproducible: a correct response gives +1, an incorrect response gives −1, and the discounted return-to-go summarises later correctness within the same learner trajectory. Return-to-go is also used as a conditioning signal in sequence-modelling offline RL (Chen et al., 2021); here it serves instead as the source of the adaptive pessimism target. This return scale becomes important later because the raw return-to-go distribution can be misaligned with early tabular

Q-values.

Baseline Policies

The value-learning backbone is a one-step Bellman backup shared by the tabular policies:

$$y_t = r_t + \gamma(1 - d_t)\max_{a' \in A} Q(s_{t+1}, a'), \qquad Q(s_t, a_t) \leftarrow (1 - \lambda_q)Q(s_t, a_t) + \lambda_q y_t \tag{4}$$

In Equation 4, A is the seven-action question-part space and λq is the tabular learning rate. Offline Q-learning is included as an optimistic diagnostic baseline because it uses this update without conservative regularisation. If this baseline produces high estimated values while moving far away from logged behaviour, the result is a warning sign for overestimation and support risk.

CQL is the main conservative baseline. Following the CQL(H) regulariser of Kumar et al. (2020), the tabular objective combines the Bellman error with a penalty that pushes down a soft maximum over Q-values while pushing up the values of logged state-action pairs:

$$\mathcal{L}(Q) = \alpha \mathbb{E}_{s \sim \mathcal{D}} \left[\log \sum_{a \in A} \exp Q\,(s, a) - \mathbb{E}_{a \sim \hat{\pi}_\beta(\cdot|s)} Q(s, a) \right] + \frac{1}{2} \mathbb{E}_{\mathcal{D}}[(Q(s_t, a_t) - y_t)^2] \tag{5}$$

In Equation 5, $\hat{\pi}\beta$ is the empirical behaviour policy in the dataset and α controls the penalty strength. The tabular implementation applies the corresponding gradient step with learning rate λq, so unsupported actions receive a softmax-weighted downward correction while logged actions are preserved. The selected setting uses $\alpha = 0.20$, based on the hyperparameter ablation, because it is the strongest CQL setting observed in the present EdNet-KT4 validation.

Normalised Target Pessimism

The APTQ-style method changes the conservative penalty during training instead of keeping it fixed. Let α_t denote the pessimism coefficient at epoch t, and let $\bar{Q}_D$ denote the mean Q-value over dataset state-action pairs. The update rule is shown in Equation 6.

$$\alpha_{t+1} = \max\left(0, \alpha_t + \eta(\bar{Q}_D - \tau)\right) \tag{6}$$

In Equation 6, η is the adaptation learning rate and τ is the target Q-value. We evaluate three target sources: a reward quantile (0.60), a raw return-to-go quantile, and a normalised return-to-go quantile. The raw return-to-go target is unstable in this setting because its scale is too large for early tabular Q-values. In practice, it pushes α toward zero and makes the method behave like the optimistic offline Q-learning baseline.

The target used in the final configuration is therefore a scaled return-to-go quantile, as shown in Equation 7.

$$\tau = q_{0.80}(0.10 \times RTG) \tag{7}$$

In Equation 7, $q_{0.80}$ denotes the 0.80 quantile of the scaled return-to-go distribution. The final configuration uses $\eta = 0.02$ and initial $\alpha = 0.20$. This target acts as a practical stabiliser: it keeps τ in a range where the penalty remains active and the learned policy stays conservative.

The three target sources differ mainly in where they drive the pessimism coefficient. Under the raw return-to-go target, τ sits far above early tabular Q-values, so the update in Equation 6 drives α to zero and the method degenerates toward optimistic offline Q-learning. Under the reward-quantile target, τ sits below the learned dataset Q-values throughout training, so α grows without a natural stopping point; the resulting over-pessimism shows up later as the weakest Bellman fit among the conservative variants. The normalised return-to-go target is retained precisely because it keeps $\bar{Q}_D - \tau$ small, so α stays in a moderate range near the tuned CQL setting. The ablation findings quantify these three regimes.

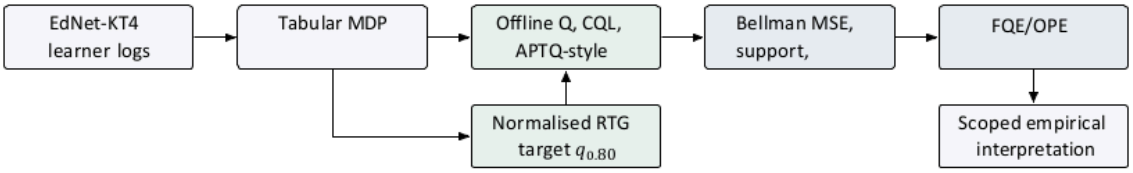
Algorithmic Procedure

Algorithm 1. APTQ-style normalised target pessimism

Input: offline transition dataset D, discount factor γ, learning rate, initial α, target quantile 0.80, target scale 0.10.
Step 1: Construct tabular states, actions, rewards, next states, and trajectory termination flags from EdNet-KT4 learner sequences.
Step 2: Compute discounted return-to-go using Equation 3.
Step 3: Set τ using Equation 7.
Step 4: Initialise tabular Q-values and α = 0.20.
Step 5: Update Q-values using the Bellman backup in Equation 4 and the conservative regularisation in Equation 5.
Step 6: Estimate $\bar{Q}D$ over observed dataset state-action pairs.
Step 7: Update α using Equation 6.
Step 8: Repeat until the training epochs are completed.
Output: greedy target policy and evaluation metrics on held-out learner split.

Figure 1 summarises the workflow as an evidence chain: EdNet-KT4 logs are converted into tabular transitions, conservative and target-adaptive policies are learned, and the resulting policies are judged with both value-based and behaviour-based metrics. The workflow keeps the interpretation focused on an empirical offline-RL proxy study.

Figure 1
Evidence Workflow for APTQ-Style Normalised Target Pessimism on EdNet-KT4



Evaluation Metrics

The evaluation uses three groups of metrics. Bellman MSE measures Bellman consistency on a held-out learner split. Support checks whether the selected greedy action remains in the training support for the corresponding state. All learned policies act greedily with respect to their Q-values, and support is measured on the held-out split:

$$\pi(s) = \arg\max_{a \in A} Q(s,a), \qquad Support(\pi) = \frac{1}{|D_{test}|} \sum_{(s,a) \in D_{test}} \mathbf{1}\left[\pi(s) \in A_D(s)\right] \tag{8}$$

where $A_D(s)$ is the set of actions observed for state s in the training transitions. FQE value and FQE

temporal-difference MSE provide the offline policy-evaluation view. Dataset Q is the mean learned Q-value on logged state-action pairs and is used only as a scale diagnostic. To make the value-based and behaviour-based readings explicit, FQE and action match are written as

$$y_t^{\pi} = r_t + \gamma(1 - d_t)Q_{\phi}\left(s_{t+1}, \pi(s_{t+1})\right), \quad Match(\pi) = \frac{1}{|D_{test}|}\sum_{(s,a)\in D_{test}} \mathbf{1}\left[\pi(s) = a\right] \quad (9)$$

In Equation 9, Q_{ϕ} is the fitted evaluation Q-function for a fixed policy π, while action match measures how often the learned greedy policy agrees with the logged behaviour action on the held-out split. These metrics are read together because each one captures a different aspect of offline policy quality. High FQE value with low action match suggests off-policy risk, while high action match with poor Bellman fit suggests behaviour imitation with weak value consistency. The results are therefore interpreted as a trade-off across complementary metrics.

One caution is needed when reading action match. Pure behaviour cloning would maximise this metric by construction, since it directly imitates the logged actions; offline methods that constrain or regularise the policy toward logged behaviour make the same connection explicit (Fujimoto et al., 2019; Fujimoto & Gu, 2021). Action match is therefore never used here as a stand-alone quality measure: it is informative only for policies that also learn a value function, because such policies can support conservative improvement and FQE-based evaluation, which behaviour cloning cannot. A policy whose action match rises while its Bellman fit and FQE value deteriorate is drifting toward imitation; this reading is applied to the reward-target variant in the results.

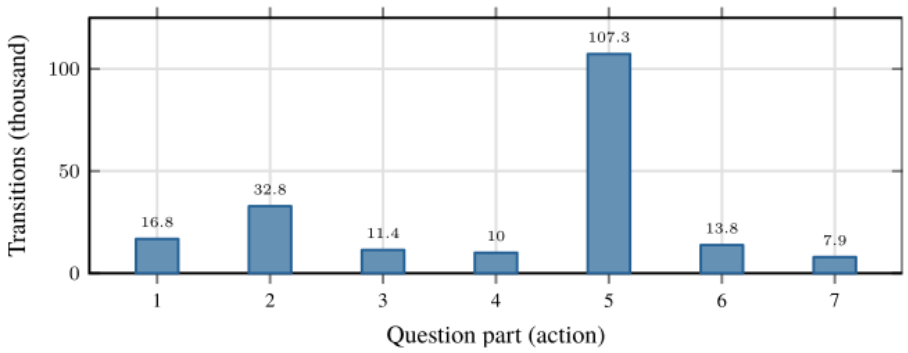
EVALUATION AND RESULTS

Dataset and Experimental Design

The main validation subset contains 200,000 EdNet-KT4 transitions from 2,665 learners, with 626 unique tabular states and seven actions. Figure 2 shows that the action distribution is strongly skewed: question part 5 accounts for 107,329 of the 200,000 transitions (53.7%), while part 7 contributes only 7,867 (3.9%). The largest activity sources in this subset are sprint (144,493), diagnosis (22,390), my_note (15,472), adaptive_offer (8,940), review_quiz (6,752), tutor (1,237), and review (716). This imbalance is one reason support-aware conservative learning matters here: state-action pairs outside the dominant parts are thinly observed, which is exactly where unconstrained Q-learning is prone to overestimation.

Figure 2

Action Distribution of the 200,000-Transition EdNet-KT4 Subset



The experiments were organised as a chain of checks. We first compared MDP and action granularities,

then tuned CQL and APTQ hyperparameters, tested raw and normalised return-to-go targets, validated the selected design at 100,000 and 200,000 transitions, and finally added FQE-based offline policy evaluation. All main results use three random seeds: 2026, 2027, and 2028.

All policies are trained for 20 epochs over the transition dataset with tabular learning rate $\lambda q = 0.08$, Bellman discount factor $\gamma = 0.95$, and the APTQ adaptation rate $\eta = 0.02$; the return-to-go preprocessing uses $\gamma_{rtg} = 0.99$ as in Equation 3. Evaluation uses a held-out learner split with 20% of learners held out, so no learner contributes transitions to both training and evaluation; in the 200,000-transition subset this gives 164,627 training and 35,373 evaluation transitions. FQE is implemented in the same tabular state-action space: the evaluation Q-function Q_ϕ is fitted on the logged transitions with learning rate 0.05, using the target policy's greedy action in the bootstrap term of Equation 9. Bellman MSE and FQE value are reported as mean ± standard deviation over the three seeds; the remaining metrics are three-seed means.

Main 200,000-transition Results

Table 2 gives the central 200,000-transition result. CQL has the lowest Bellman MSE and the highest FQE value among the conservative policies. APTQ normalised return-to-go shows its advantage on behaviour alignment: action match rises from 0.7500 to 0.8569 while support remains nearly identical.

The reward-target variant deserves a direct comparison because it reaches the highest action match (0.9001) and the lowest FQE TD MSE. It is nevertheless not selected as the main APTQ configuration for two reasons. First, its pessimism coefficient grows steadily to 1.2052 without approaching a stable level (Table 3), which is the over-pessimism regime anticipated in the method section: among the conservative variants it has the worst Bellman MSE (1.9928) and the lowest FQE value (3.0147). Second, its rising action match combined with weakening value fit matches the imitation-drift pattern flagged in the metrics section, so the extra alignment carries little value-learning content. The normalised return-to-go variant is preferred because it dominates the reward-target variant on both value-centred metrics while keeping α in a moderate range (0.4102) near the tuned CQL setting.

Table 2

Main Results on the 200,000-transition EdNet-KT4 Subset

Policy	Bellman MSE	Action match	In support	Dataset Q	FQE value	FQE TD MSE
Offline Q	2.5114 ± 0.1767	0.1879	0.9986	9.6644	8.5125 ± 0.5309	2.1080
CQL $\alpha = 0.20$	1.7225 ± 0.0803	0.7500	0.9982	3.6066	3.8703 ± 0.1703	1.5972
APTQ normalised RTG	1.7942 ± 0.0904	0.8569	0.9983	3.5887	3.3866 ± 0.2725	1.5691
APTQ reward-target	1.9928 ± 0.0959	0.9001	0.9987	3.6655	3.0147 ± 0.3216	1.3592

Figure 3 makes the same result easier to read as a policy trade-off. The dashed lines mark the CQL point. Points to the right are more behaviour-aligned, while points lower on the vertical axis lose FQE value.

Offline Q reaches the highest FQE value, but its action match is only 0.1879 and its dataset Q is much larger than that of the conservative methods. This pattern is a warning sign for unconstrained offline Q-learning: the value estimate looks strong, while the policy moves far away from logged behaviour. Offline Q therefore serves as a diagnostic rather than a candidate policy, motivating the conservative baselines used for subsequent comparison.

Scaling Validation

Table 3 reports how the pattern changes as the transition subset grows from 50,000 to 200,000. At 50,000 and 100,000 transitions, APTQ normalised return-to-go stays close to CQL in Bellman MSE. At 200,000 transitions, the MSE gap becomes larger, but the action-match gain also becomes clearer. The pattern positions APTQ normalised return-to-go as a behaviour-aligned conservative policy, with CQL retaining the stronger value fit.

Figure 3

Main Policy Trade-Off on the 200,000-Transition EdNet-KT4 Subset

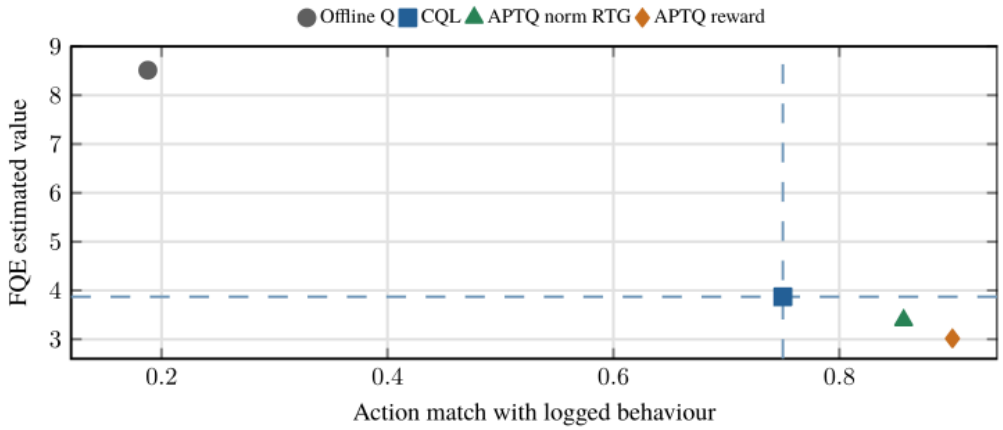


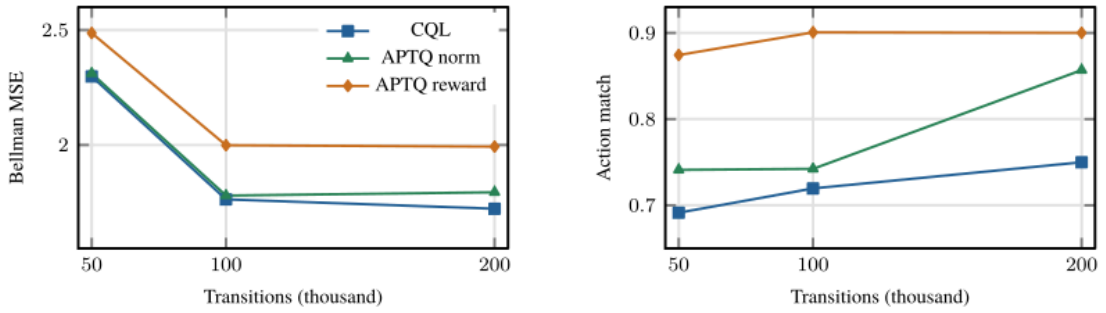
Table 3

Scaling Results from 50,000 to 200,000 Transitions

Subset	Policy	Bellman MSE	Action match	Final α	Delta MSE vs CQL
50k	CQL $\alpha = 0.20$	2.2983 ± 0.4064	0.6914	0.2000	0.0000
50k	APTQ normalised RTG	2.3109 ± 0.3926	0.7411	0.2902	0.0126
50k	APTQ reward-target	2.4870 ± 0.3935	0.8744	1.0594	0.1887
100k	CQL $\alpha = 0.20$	1.7633 ± 0.1046	0.7196	0.2000	0.0000
100k	APTQ normalised RTG	1.7798 ± 0.0892	0.7423	0.2498	0.0165
100k	APTQ reward-target	1.9988 ± 0.1178	0.9007	1.0285	0.2355
200k	CQL $\alpha = 0.20$	1.7225 ± 0.0803	0.7500	0.2000	0.0000
200k	APTQ normalised RTG	1.7942 ± 0.0904	0.8569	0.4102	0.0717
200k	APTQ reward-target	1.9928 ± 0.0959	0.9001	1.2052	0.2703

Figure 4

Scaling Validation Across EdNet-KT4 Transition Subsets



(a) Bellman consistency

(b) Behaviour alignment

Figure 4 shows the same trade-off visually: normalised RTG remains close to CQL in Bellman MSE while improving action match at 200k transitions.

Ablation Findings

The ablation study explains why the final configuration was retained, and it answers RQ1 directly. On the 50,000-transition subset, the seven-action question-part formulation reaches a CQL Bellman MSE of 2.4634, whereas the richer part-source action space degrades it to 3.4439; the sparser support of the larger action space directly weakens value estimation, so the seven-action design is retained. CQL with $\alpha = 0.20$ is the strongest CQL configuration under Bellman MSE among the tested penalty strengths. For the target source (RQ2), the raw return-to-go target collapses the adaptive penalty to $\alpha = 0.0000$ with a Bellman MSE of 3.2890, while the reward-quantile target inflates α to 1.0594 on the same subset. Among the tested target designs, normalised return-to-go with quantile 0.80 and scale 0.10 gives the most usable APTQ-style configuration (Bellman MSE 2.3109, closest to CQL); the quantile and scale were selected from this ablation rather than fixed in advance.

Figure 5

Target-Source Ablation on the 50,000-Transition Subset

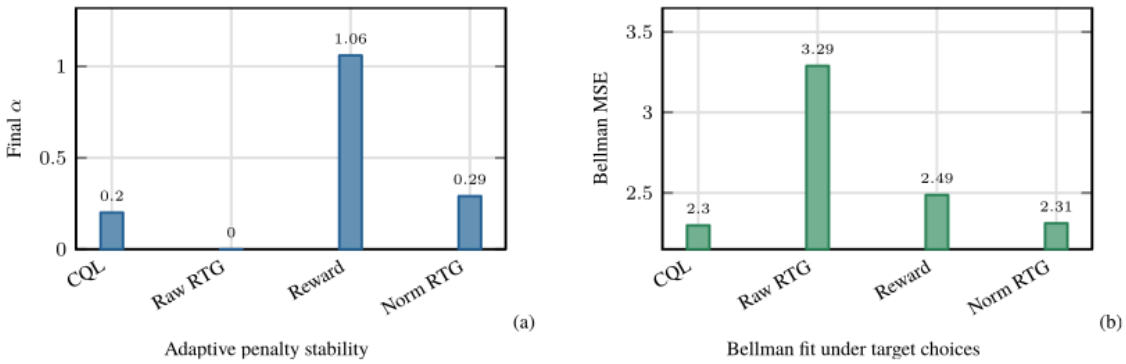
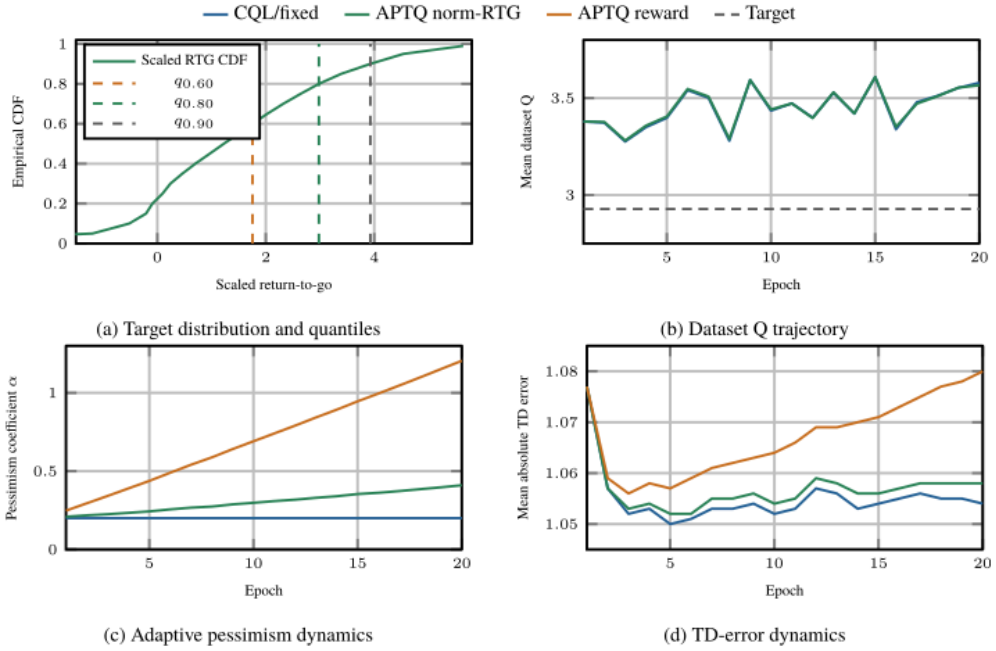


Figure 5 clarifies why the normalised RTG target was retained. The raw RTG target drives the adaptive coefficient toward collapse, whereas the normalised RTG target keeps the penalty active and close to the CQL setting.

To examine whether this target choice remains stable in the larger experiment, Figure 6 reports the 200,000-transition diagnostics. Panel (a) marks the 0.60, 0.80, and 0.90 quantiles of the scaled return-to-go distribution: the neighbouring quantiles are shown to make clear that the selected $q_{0.80}$ target is not a boundary choice but sits inside the well-populated part of the distribution. The dataset-Q trajectory remains near the target range, and the adaptive coefficient grows more moderately than the reward-target variant.

Figure 6

Normalised Return-to-Go Target Diagnostics



(a) Target distribution and quantiles

(b) Dataset Q trajectory

(c) Adaptive pessimism dynamics

(d) TD-error dynamics

Taken together, these diagnostics show the role of target design in educational offline reinforcement learning. APTQ becomes stable in this setting when the target scale is aligned with the learned Q-value range.

FQE Findings

FQE supports a conservative reading of the results. Relative to CQL, APTQ normalised return-to-go has an average FQE value difference of −0.4836. The same paired comparison, however, shows a +0.1069 average action-match improvement and a small −0.0281 average FQE TD MSE difference.

The FQE result sharpens the interpretation. APTQ normalised return-to-go improves behaviour alignment while maintaining competitive FQE fit, whereas CQL remains stronger on estimated value. This distinction matters in educational technology, where policy trustworthiness depends on more than a single estimated return.

DISCUSSION

The main finding is a trade-off between conservative value estimation and behaviour alignment. CQL is preferable when the priority is Bellman consistency and FQE estimated value. APTQ normalised return-to-go is preferable when the priority is to stay closer to logged educational behaviour while avoiding the instability of raw return-to-go targeting. The result identifies normalised target pessimism as a useful design component for behaviour-aligned educational offline RL.

This trade-off matters in education. A policy with a high estimated value but very low agreement with logged behaviour is difficult to trust without online validation. A more behaviour-aligned policy offers a more plausible starting point for cautious simulation, teacher-facing decision support, or a later controlled deployment. In this role, behaviour alignment functions as a safety-oriented signal within the

offline evaluation suite.

The results also show why target design is central to target-adaptive pessimism. Reward-target APTQ produces the highest action match but weaker Bellman and FQE value. Raw return-to-go targeting collapses the adaptive penalty. Normalised return-to-go gives a more balanced target scale, allowing APTQ to remain conservative without collapsing. The practical implication is direct: when target-adaptive pessimism moves from general offline-RL benchmarks to educational data, the scale of returns and Q-values needs to be checked explicitly.

Two caveats bound this interpretation. First, the correctness-based reward carries a known incentive risk in educational RL: a policy can raise its observed reward by steering learners toward question parts they are already likely to answer correctly, which is not the same as promoting learning gain. In this study the reward is therefore used strictly as a shared comparison signal across policies on the same logs, not as a learning-outcome measure; reward designs that credit knowledge growth rather than immediate correctness, for example signals derived from knowledge-tracing models (Piech et al., 2015), are a necessary step before any deployment-oriented claim; this reward-design difficulty is a long-standing theme in instructional-sequencing research (Doroudi et al., 2019). Second, the behaviour-alignment argument would be sharpened by an explicit behaviour-cloning baseline. Behaviour cloning would provide the action-match ceiling and make the value-learning contribution of APTQ directly visible; its absence here means the alignment claim rests on the joint reading of action match with Bellman and FQE fit rather than on a direct imitation reference point.

The dataset implication is equally important. EdNet-KT4 is large enough to support a meaningful tabular proxy study because it records learner-system activity from a public educational platform at scale. The resulting evidence is best read as offline-RL policy behaviour on educational logs and as groundwork for later intervention-oriented validation.

CONCLUSION

This paper examined adaptive pessimism for tabular educational offline reinforcement learning on EdNet-KT4. Across MDP design, target-source ablation, scaling validation, and FQE experiments, CQL remained the stronger conservative baseline under Bellman MSE and FQE estimated value. APTQ with normalised return-to-go targets produced a more behaviour-aligned policy and avoided the collapse observed with raw return-to-go targets.

These findings support a focused conclusion: normalised target pessimism is useful when the policy-learning goal includes closeness to observed learner behaviour as well as estimated value. In this sense, the paper contributes an EdNet-KT4 tabular proxy benchmark and a measured account of the trade-off between conservative value fit and behaviour alignment in educational logs.

Several limitations remain. The experiments are tabular, which improves interpretability while leaving neural function approximation for future work. The correctness-based reward is a comparison proxy and can favour recommending already-mastered content, so it should not be read as a learning-gain signal. A behaviour-cloning baseline is not included, so the action-match ceiling is not directly observed. The 200,000-transition subset is substantial for an initial empirical study, and full KT4 or a second dataset would strengthen external validation. FQE improves the offline evidence, while online or randomised validation remains the next step for learning-effect assessment. Future work should extend the method to neural CQL and APTQ, add behaviour-cloning and learning-gain-oriented reward baselines, evaluate larger EdNet-KT4 subsets, add a second educational dataset, and test whether the behaviour-alignment advantage remains under richer state representations.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Azhar, A. Z., Segal, A., & Gal, K. (2022). Optimizing representations and policies for question sequencing using reinforcement learning. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853123

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34, 15084–15097. https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html

Chen, X., Wang, S., McAuley, J., Jannach, D., & Yao, L. (2024). On the opportunities and challenges of offline reinforcement learning for recommender systems. *ACM Transactions on Information Systems*, 42(6), Article 150, 1–26. https://doi.org/10.1145/3661996

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). EdNet: A large-scale hierarchical dataset in education. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millan (Eds.), *Artificial Intelligence in Education* (pp. 69–73). Springer. https://doi.org/10.1007/978-3-030-52240-7_13

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? A review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9), 2419–2468. https://doi.org/10.1007/s10994-021-05961-4

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18), 503–556. https://jmlr.csail.mit.edu/papers/v6/ernst05a.html

Fujimoto, S., & Gu, S. S. (2021). A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 20132–20145. https://proceedings.neurips.cc/paper/2021/hash/a8166da05c5a094f7dc03724b41886e5-Abstract.html

Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2052–2062). PMLR. https://proceedings.mlr.press/v97/fujimoto19a.html

Kallus, N., & Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167), 1–63. https://www.jmlr.org/papers/v21/19-827.html

Kostrikov, I., Nair, A., & Levine, S. (2022). Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=68n2s9ZJWF8

Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 1179–1191. https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. https://doi.org/10.3102/0034654315581420

Le, H. M., Voloshin, C., & Yue, Y. (2019). Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 3703–3712). PMLR. https://proceedings.mlr.press/v97/le19a.html

Letourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Leger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10, Article 29. https://doi.org/10.1038/s41539-025-00320-7

Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z., & Liu, Y. (2025). A comprehensive survey on deep learning techniques in educational data mining. *Data Science and Engineering*, 10, 564–590. https://doi.org/10.1007/s41019-025-00303-z

Liu, J., Zhang, Y., Li, C., Yang, Y., Liu, Y., & Ouyang, W. (2024). Adaptive pessimism via target Q-value for offline reinforcement learning. *Neural Networks*, 180, Article 106588. https://doi.org/10.1016/j.neunet.2024.106588

Memarian, B., & Doleck, T. (2024). A scoping review of reinforcement learning in education. *Computers and Education Open*, 6, Article 100175. https://doi.org/10.1016/j.caeo.2024.100175

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28, 505–513. https://papers.nips.cc/paper/5654-deep-knowledge-tracing

Figueiredo Prudencio, R., Maximo, M. R. O. A., & Colombini, E. L. (2024). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8), 10237–10257. https://doi.org/10.1109/TNNLS.2023.3250269

Wang, J., Gao, R., & Zha, H. (2024). Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 72(2), 699–716. https://doi.org/10.1287/opre.2022.2382

Adaptive Pessimism with Normalised Return Targets for Educational Offline Reinforcement Lear

ORIGINALITY REPORT

8%	5%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of the Philippines Diliman Student Paper	
2	files01.core.ac.uk Internet Source	
3	arxiv.org Internet Source	
4	Submitted to Imperial College of Science, Technology and Medicine Student Paper	
5	jutif.if.unsoed.ac.id Internet Source	
6	Fengying Li, Yuqiang Li, Xianyi Wu. "Minimax weight learning for absorbing MDPs", Statistical Papers, 2024 Publication	
7	Tao Wang, Shaorong Xie, Mingke Gao, Xue Chen, Zhenyu Zhang, Hang Yu. "Offline Reinforcement Learning via Policy Regularization and Ensemble Q-Functions", 2022 IEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), 2022 Publication	
8	Tao Wang, Xiangfeng Luo, Zhenyu Zhang, Shaorong Xie. "Efficient offline-to-online reinforcement learning with pre-reduced out-of-distribution Q-values", Neural Network 2025 Publication	
9	proceedings.mlr.press Internet Source	
10	Ekarsi Lodh, Shalini Majumder, Tapan Chowdhury, Manashi De. "NetPolicy-RL: network informed offline reinforcement learning for pharmacogenomic drug prioritization", Journal of Computer-Aided Molecular Design, 2026 Publication	
11	Saepudin Nirwan, Rolly Maulana Awangga, Naufal Fachrudin Nirwan. "Comparative Performance Analysis of Several Python Libraries Utilizing the Least Significant Bit Method", Jurnal Online Informatika, 2026 Publication	

12 Shengbo Eben Li. "Reinforcement Learning for Sequential Decision and Optimal Contro
Springer Science and Business Media LLC, 2023
Publication

13 Yesen Chen, Teng Zhang, Tao Li, Jiangcheng Chen, Dongyun Wang. "Dynamic-based representation inconsistency and implicit constraints for offline reinforcement learning
Neural Networks, 2026
Publication

14 export.arxiv.org
Internet Source

15 proceedings.neurips.cc
Internet Source

16 Mehrdad Moghimi, Hyejin Ku. "Risk-sensitive Actor-Critic with Static Spectral Risk Measures for Online and Offline Reinforcement Learning", Expert Systems with Applications, 2026
Publication

17 Niranjani Prasad, Aishwarya Mandyam, Corey Chivers, Michael Draugelis et al. "Guiding Efficient, Effective, and Patient-Oriented Electrolyte Replacement in Critical Care: An Artificial Intelligence Reinforcement Learning Approach", Journal of Personalized Medicine, 2022
Publication

18 Zijian Wang, Bin Wang, Hongbo Dou, Zhongyuan Liu. "Windows deep transformer Q-networks: an extended variance reduction architecture for partially observable reinforcement learning", Applied Intelligence, 2024
Publication

19 jestp.com
Internet Source

20 raw.githubusercontent.com
Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		