Offline Reinforcement Learning for Engagement and Correctness in EdNet

Aditya Khandelwal
Apurv Gupta

Abstract

In this project, we apply offline reinforcement learning (RL) techniques to improve recommendation policies in the EdNet dataset, a large-scale collection of student-system interactions from an AI tutoring platform. Traditionally, recommendation engines in Intelligent Tutoring Systems (ITS) have emphasized correctness as the sole optimization objective. However, we argue that engagement is also a critical component of effective learning, as it can foster deeper understanding and sustained effort over time. To this end, we design a custom reward function that balances engagement and correctness, and evaluate four state-of-the-art offline RL algorithms—Behavioral Cloning (BC), Conservative Q-Learning (CQL), Bootstrapping Q-Learning (BCQ), and Soft Actor-Critic (SAC) adapted to discrete actions—on a subset of EdNet.

Our results demonstrate that by carefully adjusting the relative weights of correctness and engagement, it is possible to produce policies that maintain strong performance on correctness while substantially increasing engagement. We provide a detailed analysis of the trade-offs between these objectives, show how different algorithms respond to varying reward structures, and discuss the implications for deploying such policies in real-world educational settings. We make the code publicly available at Offline RL for Education.

1 Introduction

As educational technologies become increasingly personalized, Intelligent Tutoring Systems (ITS) offer the promise of guiding students toward mastery through adaptive recommendations. Most systems, however, focus narrowly on correctness, ensuring that students answer a maximum number of questions correctly. While correctness is indeed a vital indicator of learning, it is not the entire story. Students who are engaged—spending quality time on tasks, reflecting on feedback, and exploring supplementary materials like lectures or explanations—may develop deeper conceptual understanding and long-term retention [1].

In this project, we investigate how to integrate engagement into the decision-making process of a recommendation policy. Specifically, we adopt an offline RL paradigm to learn policies from the EdNet dataset [3], one of the largest publicly available repositories of ITS interactions. Offline RL is well-suited for educational contexts because it allows policy improvement without active experimentation on learners, thereby avoiding the risk of negatively impacting their learning experiences [8].

Our contributions are as follows:

1. We propose a custom reward function that jointly optimizes correctness and engagement.
2. We apply four offline RL algorithms—BC, CQL, BCQ, and SAC (discrete)—and systematically vary the reward weights to explore the correctness-engagement trade-off.

3. We present empirical results showing that certain algorithms (notably BCQ and SAC) can effectively increase engagement with minimal loss in correctness, offering insights into algorithmic choices for educational recommendations.

2 Background and Related Work

Research on ITS often centers on knowledge tracing models [4, 11] that predict future correctness. Such models are primarily concerned with whether a student knows a particular concept, as reflected by accuracy. However, engagement, reflected through behaviors like time spent on tasks, re-reading explanations, and interacting with supplementary materials, is equally important for sustained learning and motivation [2].

Reinforcement Learning (RL) has been introduced into educational recommendation to dynamically adapt content [10], but few studies explicitly incorporate engagement as a reward signal. Offline RL, which learns from static logs without online exploration, is appealing in educational contexts due to ethical and practical considerations [9].

The offline RL methods considered here represent distinct paradigms:

- **Behavioral Cloning (BC):** Directly mimics historical data, acting as a baseline that reproduces existing policies.
- **Conservative Q-Learning (CQL) [7]:** Encourages conservative Q-value estimates to avoid overestimating actions that are not well-supported by data.
- **Bootstrapping Q-Learning (BCQ) [5]:** Restricts the policy's action set to those likely observed in the dataset, reducing the risk of distributional shift.
- **Soft Actor-Critic (SAC) [6] (Discrete variant):** Introduces entropy maximization to foster robust policies. We adapt SAC to the discrete action setting, ensuring that it can handle EdNet's discrete recommendation options.

3 Methods

3.1 Dataset and Preprocessing

We focus on the EdNet-KT4 dataset, which includes rich student interactions such as solving questions, watching lectures, and reading explanations. We filter for paid users who have full access to questions and lectures, ensuring a consistent and representative sample of engagement patterns. We define a maximum inter-action gap of 5 minutes to segment the data into sessions, each representing a coherent learning episode.

Following segmentation, we keep only those sessions that have atleast one organic interaction (sources like sprint, review, and diagnosis modes), as they reflect natural learning scenarios rather than artificially triggered events. Each question in a session is marked correct if the user's chosen answer matches the ground truth. Engagement features include time spent on questions and non-question activities, as well as the frequency with which learners access supplementary materials.

Further, we condense each session into a series of 3 actions:

- Question: We condense the traditional "enter","respond","submit" events into one "question" event, considering the answer selected in the last "respond" event as the answer chosen (as mentioned in the EdNet guidelines).

- Textual Explanation: We condense any "enter" and "quit" sequences with an item_id starting with "e" into a textual explanation event.
- Video Explanation: We condense any "play_audio", "pause_audio", "play_video" and "pause_video" events into one video explanation event.

We split the data into 70% training, 15% validation, and 15% test sets, ensuring that sessions are uniquely distributed so that no session appears in more than one set.

3.2 State and Action Representation

A state is derived from the session history, specifically the past 5 states the user was in. Each time-step has the following data points:

- **Event Type:** The type of event
- **Correctness History:** If the user interaction for that time was a question, whether or not they answered it correrctly (0 or 1).
- **Engagement metrics:** Time spent on the interaction, if it was a video/text explanation. This is normalized by dividing it by 30000.
- **Contextual information:** Platform type (mobile, web), session length, and question difficulty.
- **Source:** The source column retained from the initial dataset. This represents where the content was placed in the Santa UI.

The action space is discrete, representing different types of recommendations:

- **Question:** Presenting a new question.
- **Textual Explanation:** Suggesting a textual explanation.
- **Video Explanation:** Suggesting a video explanation.

3.3 Reward Function

We define correctness C as a normalized value between 0 and 1, representing the fraction of questions answered correctly in a small window. Engagement E is also normalized, reflecting session-level measures such as the ratio of time spent actively solving or studying to total session length.

The reward at each step is:

$$R = w_1 \cdot C + w_2 \cdot E, \quad w_1 + w_2 = 1.$$

We experiment with two weight configurations:

$$(w_1, w_2) \in \{(1.0, 0.0), (0.5, 0.5)\},$$

to observe how shifting emphasis from correctness to a blended representation of engagement and correctness influences policy behavior.

Further, we also normalize the engagement to be bounded between 0 and 1. We do this by first normalizing to the training data statistics, and then adding a min function:

$$\hat{X} = min(1.0, max(\frac{X - \bar{X}}{\sigma_X}, 0.0))$$

3.4 Offline RL Training Procedure

We implement each offline RL algorithm and run training for one iteration on the ttraining dataset (due to time and compute limitations).

During training, no online interactions occur. We rely entirely on historical data. This setup mirrors many real-world educational contexts, where experimenting directly on learners can be costly or unethical.

4 Results

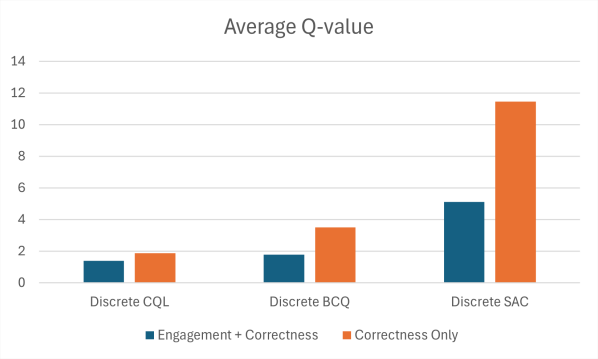
4.1 Overall Performance

Image 1 summarizes the Average Q-value per session. We see a clear trend:

$$Q_{SAC} > Q_{BCQ} > Q_{CQL}$$

.

Further, for all models, the average Q-value decreases when we introduce engagement into the rewards. This indicates that getting good engagement is a difficult task, and we argue thus is important to include as a target going forward.



.

The Discrete BC algorithm does not give an average Q-value (since it is not defined for this algorithm), but we note the action match accuracies in the table below. The Action Match Accuracy improves for the engagement + correctness reward, but this could be due to lack of training sufficiently.

Model	Reward Type	Action Match Accuracy
Discrete BC	Correctness Only	99.00%
Discrete BC	Engagement + Correctness	99.90%

Table 1: Action Match Accuracy for Different Reward Types

4.2 Policy Behavior and Analysis

We further analyse the specific instances of the trained RL Algorithm deviating from the existing sessions. We do this by sampling timestep 6 for various sessions, and using the previoud 5 time steps as input and noting the different in the actual action taken versus what our model recommended. We observe how this varies across our reward functions.

Here are a few examples:

Table 2: Event Sequences and Predictions for Each Algorithm

Algorithm	Step	Event Type	Correctness Reward	Engagement Reward	Next Action	Predicted Next Action
CQL (Combined)	1	Textual Explanation	0	0.35	Textual Explanation	Video Explanation
	2	Video Explanation	0	0.75		
	3	Video Explanation	0	0.19		
	4	Video Explanation	0	0.03		
	5	Video Explanation	0	0.15		
BCQ (Correctness)	1	Video Explanation	0	0.49	Video Explanation	Question
	2	Textual Explanation	0	0.87		
	3	Video Explanation	0	0.74		
	4	Question	1	0.00		
	5	Video Explanation	0	0.06		
BCQ (Combined)	1	Video Explanation	0	0.50	Video Explanation	Question
	2	Video Explanation	0	0.49		
	3	Question	1	0.00		
	4	Textual Explanation	0	0.57		
	5	Video Explanation	0	0.48		
SAC (Correctness)	1	Video Explanation	0	1.00	Textual Explanation	Question
	2	Video Explanation	0	0.02		
	3	Video Explanation	0	0.06		
	4	Video Explanation	0	1.00		
	5	Question	0	0.00		
SAC (Combined)	1	Question	0	0.00	Textual Explanation	Video Explanation
	2	Textual Explanation	0	0.72		
	3	Question	1	0.00		
	4	Textual Explanation	0	0.98		
	5	Question	1	0.00		

Specifically, on examining the policies' actions, we find that algorithms placing higher weight on engagement more frequently recommend additional explanations or lectures, as opposed to the algorithms trained with only correctness as a reward (most recommendations from the norm tended to be "Question"). Historically, students often skipped these suggestions. However, our offline RL policies identify contexts where such supplementary materials are likely to be well-received, potentially reducing skip rates and fostering more thorough concept exploration.

5 Discussion

These results highlight the potential of offline RL in educational contexts. By carefully designing the reward function to incorporate engagement, we can produce policies that balance short-term correctness with long-term educational value. Interestingly, while correctness-oriented policies mimic

traditional ITS strategies, engagement-oriented policies resemble a more holistic approach, encouraging sustained interaction, deeper processing, and concept mastery.

This study also underscores the value of comparing multiple offline RL methods. While BC provides a baseline that reproduces historical behavior, methods like BCQ and SAC can navigate the modified objective space more effectively. This suggests a need for careful algorithm selection based on desired outcomes.

6 Conclusion and Future Work

We have demonstrated that offline RL, when applied to EdNet and guided by a multi-objective reward function, can yield policies that improve engagement without severely undermining correctness. The insights gleaned here can inform designers of ITS to consider metrics beyond correctness and to utilize advanced RL techniques for more nuanced content recommendations.

Future work includes extending this approach to multi-step policy evaluations, conducting ablation studies on different engagement metrics, and integrating natural language processing to better understand which content fosters the most meaningful engagement. Ultimately, deploying these policies in a live environment and validating their impact on learners' long-term outcomes will be a critical next step.

Acknowledgments

We thank the EdNet team for making their dataset publicly available and acknowledge discussions and resources provided by the course staff.

References

References

[1] Anderson, J. R. (1985). *Cognitive Psychology and its Implications.* W.H. Freeman.

[2] Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education.*

[3] Choi, Y. et al. (2019). EdNet: A Large-Scale Hierarchical Dataset of Student Behavior. *arXiv preprint arXiv:1912.03072.*

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

[5] Fujimoto, S. et al. (2019). Off-Policy Deep Reinforcement Learning without Exploration. *ICML.*

[6] Haarnoja, T. et al. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *ICML.*

[7] Kumar, A. et al. (2020). Conservative Q-Learning for Offline Reinforcement Learning. *NeurIPS.*

[8] Levine, S. et al. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv:2005.01643.*

[9] Liu, L. et al. (2021). Reinforcement Learning for Educational Recommendations. *IJCAI Workshop on Educational AI.*

[10] Minn, S. et al. (2018). Deep Knowledge Tracing and Dynamic Student Classification for Knowledge State Reflection. *EDM*.

[11] Piech, C. et al. (2015). Deep Knowledge Tracing. *NIPS*.