When Offline RL Cannot Evaluate Teaching: A Diagnostic Case Study

Ozan Bayiz
CS 185/285, UC Berkeley
ozanbayiz@berkeley.edu

Narges Norouzi*
UC Berkeley
norouzi@berkeley.edu

Abstract

We imagine an intelligent tutoring system as a lightweight pedagogical policy steering a foundation-model interaction layer. This paper focuses on the high-level policy. We train it as a Decision Transformer over a Deep Knowledge Tracing world model and put two learning-sciences-grounded reward shapes through a four-diagnostic screening protocol: a cumulative-return correlation precondition, a DT-vs-empirical-teacher action-overlap check, a pairwise inter-policy diagnostic, and a cross-dataset replication check. The screening is backed by a 20-seed FQE bootstrap matched to policy-class variance and three OPE estimators (FQE, WIS, DualDICE) rank-validated on a tractable HMM toy environment before being applied. The two reward shapes are a preparedness-gain reward over a sibling neighborhood (drawn from Preparation for Future Learning) and a boundary-of-capability reward (drawn from Productive Failure, included as a deliberate test case). The diagnostics caught every comparison between a pedagogically-grounded policy and the immediate baseline, on both XES3G5M and Junyi Academy — for reasons ranging from algebraic degeneracy in the reward formula to dependence on the world model's calibration regime. Two follow-up isolation experiments converged on action-space flatness as the load-bearing structural cause across both datasets. The paper offers no positive claim about either reward shape. Its contribution is the screening protocol itself, paired with a specification of the data infrastructure a faithful evaluation would require: a directed prerequisite graph, a transfer-aligned action structure, a post-intervention transfer assessment, and a behavioral anchor outside synthetic trajectories. None of these are currently available; the paper names them as the prerequisites for the next attempt.

1 Introduction

Atkinson [1] framed instructional sequencing — choosing what topic to teach a student next — as an optimal-control problem and derived a learning-rate-aware policy for second-language vocabulary acquisition. The half-century since has produced a literature in which positive offline results recurrently fail more careful scrutiny. Doroudi et al. [9] survey the pattern in "Where's the Reward?", concluding that pedagogically-grounded reward design must be paired with robust offline analyses before any claim of policy quality is made. The recommendation is what this paper takes as its operating discipline.

Existing RL-for-tutoring work [3] typically uses immediate correctness as the reward — the student's predicted probability of answering the taught skill correctly. The learning sciences have argued for thirty years that immediate correctness is the wrong target. Preparation for Future Learning [5, 21] holds that the value of a teaching event is its effect on the student's readiness to learn future material, not its effect on immediate performance. Productive Failure [11] and the desirable-difficulty literature

*Faculty advisor.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

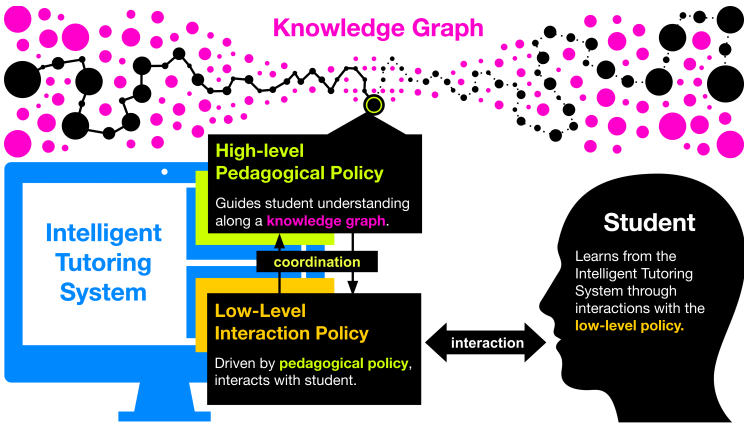

Figure 1: A two-policy intelligent tutoring system. A high-level pedagogical policy selects which knowledge-graph node to teach next. A low-level interaction policy executes the teaching turn over text or speech, coordinating with the high-level policy on the chosen node and interacting directly with the student. The high-level policy is the focus of this paper: we train it as a Decision Transformer under learning-sciences-grounded reward shapes and screen those shapes with the four-diagnostic protocol of subsection 4.6. The low-level policy is the foundation-model component the broader program (section 7) returns to; this paper does not build it. The knowledge graph is the structured curriculum object both policies operate over — deployed systems would extract it from a course's class assets, while our case study uses curated benchmark trees as a stand-in.

[4] hold that students learn most when they engage at the boundary of their current capability. From PFL we instantiate a preparedness-gain reward (PFL-GAIN) over a sibling neighborhood; from the PF / desirable-difficulty lineage we instantiate a boundary-of-capability reward (PF-DIFF at $p^\star = 0.5$). Both are trained as Decision Transformer [7] policies on a Deep Knowledge Tracing world model [19]. The offline data is two primary-school mathematics interaction logs: XES3G5M [15] and Junyi Academy [6].

The deployment constraint moves the load-bearing question from training to evaluation. Reinforcement learning ordinarily assumes the agent gathers new data through interaction; running policy-improvement loops on real students requires ethical review, school partnerships, and long-horizon timelines. We work in the offline setting, where the only measure of a policy's teaching quality comes through an off-policy estimator — a procedure that takes the policy and the logged data and returns a scalar meant to approximate the policy's true expected return. Estimators have known failure modes; the same number that licenses a published claim of teaching quality can also be the artifact of an out-of-distribution action, a saturated importance weight, or a world model that hallucinates outside the logged support. To pair the comparison with the robust offline analyses Doroudi et al. [9] recommend, we wrap it in four diagnostics. A cumulative-return correlation precondition flags reward shapes whose return-conditioned policy is mechanically tied to the baseline. A DT-vs-empirical-teacher action-overlap check catches the circular-Q-bootstrapping pathology [13] that inflates FQE on policies whose actions are out-of-distribution under the logged data. A pairwise inter-policy overlap diagnostic covers the case where multiple reward shapes produce identical policies. A cross-dataset check requires comparisons to survive a dataset switch before being treated as evidence about the policies. Three off-policy estimators (Fitted Q Evaluation [14], weighted importance sampling [20], and DualDICE [17]) are cross-checked so that no single estimator's failure mode silently propagates, and the diagnostics are backed by a 20-seed FQE bootstrap matched to policy-class variance.

Each of the four diagnostics fires on at least one comparison across the four reward shapes (Figure 2), and each firing has a specific data-infrastructure cause. The precondition catches SIB-MEAN as mechanically identical to the baseline at correlation $+0.9956$ — a property of how the XES3G5M curriculum tree was curated, by similarity rather than direction — and PF-DIFF at $p^\star = 0.5$ as algebraically anti-correlated with the baseline under myopic-greedy behavior on a flat 865-way action space. The action-overlap check flags PF-DIFF's high FQE on XES3G5M as Q-bootstrapping inflation on actions the logged data does not cover. The cross-dataset check inverts the PF-DIFF-

versus-IMM ranking on Junyi's lower-AUC world model. On Junyi we expand the comparison to five reward shapes by adding three further learning-sciences-grounded variants (subsection 5.6); the pairwise diagnostic collapses all five DTs at $\lambda = 2.0$ to a single underlying policy. The differences reduce to properties of the reward formula, the action space, the world model's calibration regime, or the synthetic-trajectory bridge between them; the four together name what the longer program needs as data infrastructure — a directed prerequisite graph, a transfer-aligned action structure, a post-intervention transfer assessment, and a behavioral anchor outside synthetic trajectories.

Contributions. We instantiate two pedagogically-grounded reward shapes drawn from distinct learning-sciences traditions: a preparedness-gain reward (PFL-GAIN) from Preparation for Future Learning, and a boundary-of-capability reward (PF-DIFF at $p^\star = 0.5$) from the Productive Failure / desirable-difficulty lineage. Both are trained as Decision Transformer policies over a DKT world model. We pair the comparison with four diagnostics: a cumulative-return precondition, an action-overlap check, a pairwise inter-policy diagnostic, and a cross-dataset check. A 20-seed FQE bootstrap matched to policy-class variance backs each one. The case study runs on XES3G5M and Junyi Academy. Every comparison between a pedagogically-grounded policy and the immediate-correctness baseline is caught by at least one diagnostic, and each catch traces to a specific data-infrastructure gap. The paper offers no positive claim about either reward shape. The contribution is the verified screening protocol and the catalogue of catches as a specification of the data infrastructure a faithful pedagogically-grounded reward needs before it can be evaluated on its merits.

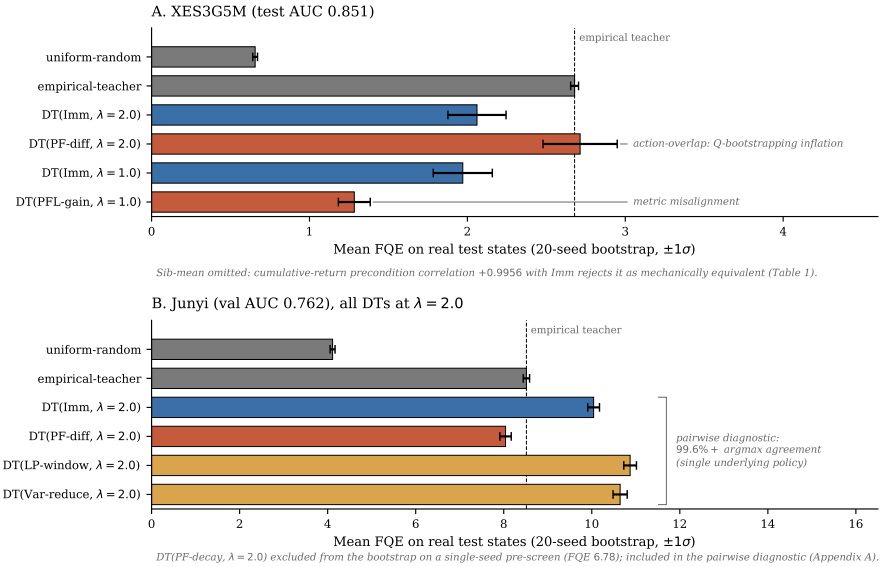

Figure 2: Two-panel summary of the case study; mean FQE on real test states with 20-seed bootstrap error bars ($\pm 1\sigma$). **A.** XES3G5M (test AUC 0.851). DT(PF-DIFF, $\lambda = 2.0$) sits above the empirical teacher; the action-overlap diagnostic catches this as Q-bootstrap inflation rather than teaching quality. DT(PFL-GAIN, $\lambda = 1.0$) scores reliably below DT(IMM, $\lambda = 1.0$) on FQE-immediate, a metric orthogonal to its training objective. Sibling-mean is omitted: the cumulative-return precondition catches it as mechanically equivalent to IMM at $+0.9956$ correlation (Table 1). **B.** Junyi (val AUC 0.762), every DT trained at $\lambda = 2.0$. The four bootstrap-included DTs span FQE values from 8.0 to 10.9. The pairwise inter-policy diagnostic shows $99.6\%+$ argmax agreement across all ten pairs of the five DTs (IMM, PF-DIFF, PF-DECAY, LP-WINDOW, VARIANCE-REDUCE), so the bootstrap measures FQE evaluation noise on a single underlying policy. The FQE rank between DT(PF-DIFF) and DT(IMM) inverts vs. Panel A; the cross-dataset check identifies this as a calibration-regime artifact.

2 Background

Deep Knowledge Tracing (DKT). DKT [19] is a recurrent neural network that processes a student's history of (topic, correctness) pairs and outputs, for every topic, a predicted probability that the student would answer correctly next. We use DKT both as the *student model* (the simulator we roll policies out against) and as the *belief encoder* whose hidden state becomes the input to the policy.

Decision Transformer (DT). A Decision Transformer [7] casts offline RL as supervised next-token prediction: given a history of states, return-to-go values, and actions, predict the next action. At inference, the user supplies a target return-to-go R_0 and the DT generates an action sequence aiming to achieve that return. The architecture is appealing for offline RL because it sidesteps both the value-function estimation and the policy-improvement loop — the two destabilizers in many offline-RL alternatives.

MOPO pessimism penalty. Model-based offline policy optimization (MOPO; 25) is a reward shaping for offline model-based RL: at training time, replace the dataset reward $r(s_t, a_t)$ with the penalized $r'(s_t, a_t) := r(s_t, a_t) - \lambda \cdot u_t$, where u_t is the model-uncertainty estimate at step t (the ensemble disagreement scalar from subsection 4.1). The penalty discourages the policy from exploiting regions where the world model is unreliable. We sweep $\lambda \in \{0.5, 1.0, 2.0\}$ for every DT configuration.

Off-policy evaluation (OPE). OPE estimates a policy's expected return without deploying it. Formally: estimate $V(\pi) := \mathbb{E}_{\tau\sim\pi}[G_0(\tau)]$ from a dataset $\mathcal{D}$ of trajectories collected under a different (behavior) policy μ. We use three estimators. *Fitted Q Evaluation* (FQE; 14) trains a parametric $\hat{Q}^\pi$ on logged transitions $(s, a, r, s') \in \mathcal{D}$ to satisfy the Bellman recursion

$$Q^\pi(s,a) \;=\; r(s,a) \;+\; \gamma\, \mathbb{E}_{s'\sim P,\, a'\sim\pi(\cdot|s')}\big[\, Q^\pi(s',a')\,\big], \tag{1}$$

and returns $\hat{V}_{\mathrm{FQE}}(\pi) := \mathbb{E}_{s_0\sim\mathcal{D},\, a_0\sim\pi}[\hat{Q}^\pi(s_0, a_0)]$. *Per-decision weighted importance sampling* (WIS; 20) reweights each logged transition by the importance ratio $w_t := \pi(a_t \mid s_t)/\mu(a_t \mid s_t)$ and self-normalizes the cumulative weight $\prod_{t'\le t} w_{t'}$ at each step. *DualDICE* [17] estimates the state-action density ratio $\rho(s,a) := d^\pi(s,a)/d^{\mathcal{D}}(s,a)$ via a convex dual objective and recovers $\hat{V}_{\mathrm{DICE}}(\pi) := \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\mathcal{D}}[\rho(s,a)\, r(s,a)]$.

Preparation for Future Learning (PFL). Schwartz and Bransford's framework [5, 21–23] reframes the assessment of teaching: measure how well students are prepared to learn from future instruction on related material. Empirically, activities that look unproductive under traditional sequestered-problem-solving assessments—such as having students invent procedures, generate predictions, or contrast cases before being shown the canonical method—substantially outperform direct instruction when measured by future-learning gains. PFL motivates our *preparedness-gain* reward, which maps onto its transfer-assessment metric: a teaching action's value is the predicted-correctness change it induces on a topic's neighborhood.

Productive Failure (PF) and desirable difficulty. *Productive Failure* [11] is a distinct learning-sciences framework whose central empirical claim is that students who first work on novel problems they cannot yet solve, before receiving structured instruction, learn more deeply than students given direct instruction first. The mechanism is engagement at the boundary of current capability: difficulty calibrated such that the student's intuitions are activated and surfaced as variant solution attempts before the canonical method is introduced. *Desirable difficulty* [4] makes a related claim from cognitive psychology: retrieval and effortful encoding under appropriate difficulty improve long-term retention. Both frameworks motivate our *boundary-of-capability* reward, which targets the moderate-difficulty band ($p^\star = 0.5$) where productive struggle is most plausibly induced. PFL by itself does not motivate this reward shape; PF and desirable difficulty do, and we credit them accordingly.

3 Related Work

Reinforcement learning for tutoring. The application of RL to instructional sequencing has a long history, surveyed in part by Doroudi et al. [9]. Modern offline-RL applications have explored

deep policy classes for tutoring [3], with most existing systems framing the reward as immediate correctness or learning-curve slope. Our work follows in this tradition but proposes pedagogically-grounded reward shapes drawn from learning-sciences theory. It emphasizes the off-policy evaluation methodology required to compare reward shapes defensibly.

Offline-RL methodology and OPE. The offline-RL methodology literature has matured substantially in the past five years. Kumar et al. [13] (CQL), Fujimoto et al. [10] (BCQ), and Kostrikov et al. [12] (IQL) are the standing methods we benchmark against. The three OPE estimators we cross-check across—Le et al. [14], Precup et al. [20], and Nachum et al. [17]—each have known failure modes documented in the broader literature; we cite them as estimators we use, and cross-checking three is one of several ways to hedge against single-estimator artifacts. The Decision Transformer architecture itself [7] is one instance of return-conditioned offline learning; the precondition we propose in section 4.6 applies to any return-conditioned policy class.

Two distinct learning-sciences lineages. The two reward shapes in this paper draw on two distinct empirical frameworks. *Preparation for Future Learning* [5, 21, 22] emphasizes transfer to future learning as the assessment metric and motivates our preparedness-gain reward (PFL-GAIN). *Productive Failure* [11] and *desirable difficulty* [4] emphasize engagement at the boundary of current capability and motivate our boundary-of-capability reward (PF-DIFF at $p^\star = 0.5$). The two lineages are adjacent but neither subsumes the other; PFL alone does not motivate difficulty targeting, and PF / desirable difficulty alone do not motivate transfer-of-learning measurement. The contribution is the screening protocol itself, paired with a specification of the data infrastructure a faithful evaluation would require (section 6).

4 Method

4.1 Notation

We adopt standard offline-RL notation [7, 13]. The environment is a finite-horizon Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, T)$ with state space $\mathcal{S} \subseteq \mathbb{R}^{129}$; action space $\mathcal{A}$ has one action per leaf knowledge component (KC; $|\mathcal{A}| = 865$ KCs on XES3G5M, 1,326 exercises on Junyi); the index 0 is reserved as a sequence-padding token and is masked out of action selection. The MDP also has a transition kernel $P(s_{t+1} \mid s_t, a_t)$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount $\gamma \in (0, 1]$, and horizon $T = 50$.

A trajectory is $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$. The return-to-go from step t is $G_t(\tau) = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'}$; we use $\gamma = 1$ throughout training and FQE evaluation, and report DualDICE at $\gamma = 0.99$ as a discounted approximation. A policy is $\pi(a \mid s)$. The behavior policy is $\mu(a \mid s)$; we use μ_ε for the specific ε-myopic-greedy policy that generates our offline data (Equation 4.3). The offline dataset is $\mathcal{D} = \{\tau^{(n)}\}_{n=1}^{N_{\text{traj}}}$.

The student model is a $K = 5$ ensemble of independently-trained DKTs [19]; ensemble member $i \in \{1, \ldots, K\}$ outputs the predicted-correctness function

$$\hat{p}_t^{(i)}(a) := \Pr\big[\text{student would answer KC } a \text{ correctly at step } t \mid h_t^{(i)}\big], \tag{2}$$

where $h_t^{(i)} \in \mathbb{R}^{128}$ is member i's LSTM hidden state after observing the prefix $(a_0, y_0), \ldots, (a_{t-1}, y_{t-1})$ of teaching actions a and observed correctness labels $y \in \{0, 1\}$. The ensemble mean is $\bar{p}_t(a) := \frac{1}{K} \sum_{i=1}^{K} \hat{p}_t^{(i)}(a)$ and the per-step ensemble disagreement is $u_t := \operatorname{std}_i \hat{p}_t^{(i)}(a_t^{\text{ref}})$ where $a_t^{\text{ref}} \in \arg\max_a \hat{p}_t^{(\text{ref})}(a)$ is the reference DKT's argmax action at step t. Throughout, i indexes ensemble members and k indexes KCs.

For $a \in \mathcal{A}$, the KC-neighborhood $\mathcal{N}(a) \subseteq \mathcal{A}$ is the set of leaves sharing a's parent in the topic hierarchy (one tree-hop laterally). For only-child leaves (131 of 865 on XES3G5M; 4 of 1,326 on Junyi), $\mathcal{N}(a) = \{a\}$.

Off-policy evaluation estimators are written $\hat{V}_{\text{FQE}}(\pi)$, $\hat{V}_{\text{WIS}}(\pi)$, $\hat{V}_{\text{DICE}}(\pi)$ for the procedure-returned scalar. We refer to the procedures themselves as FQE, WIS, and DualDICE.

4.2 Datasets and student model

We evaluate on two primary-school mathematics interaction logs. XES3G5M [15] is the primary dataset for the headline analysis: 18,066 students, 1,175 KCs organized into a four-level topic hierarchy, and 5.1M (student, KC, correctness) interactions. We use the pyKT preprocessing folds and a 50/50 validation/test split partitioned by student, so no student appears in more than one split. Junyi Academy [6] is the cross-dataset reproduction target (subsection 5.5): 16.2M interactions over 1,326 unique exercises (the ucid action space), also organized into a four-level topic hierarchy (level_1 $\rightarrow$ level_4 $\rightarrow$ exercise).

The belief encoder is a single-layer LSTM with hidden size 128 and embedding size 64 (308K parameters), trained with cross-entropy on the per-step correctness label. AdamW (learning rate 10^{-3}, weight decay 0.01), three-epoch linear warmup followed by cosine annealing to 10^{-5}, and early stopping on validation AUC with patience 5. The same architecture and training recipe are used on both datasets; only the input vocabulary changes. On XES3G5M the model reaches test AUC 0.851 at epoch 26; on Junyi it reaches validation AUC 0.762, an ~ 0.09 gap that becomes load-bearing in the cross-dataset analysis. A 4.4$\times$-larger 2-layer LSTM with hidden size 256 trained under the same regularization reaches test AUC 0.8497 on XES3G5M, slightly worse than the 1-layer model; the input feature set (topic ID and correctness) determines the ceiling, not model capacity. A K=5 ensemble of independently-seeded DKTs trained on the same configuration provides per-step uncertainty as the standard deviation of per-KC predictions across members.

4.3 World-model environment and behavior policy

The DKT ensemble is wrapped as a Gym-like environment. The state at step t concatenates the reference DKT's hidden vector and the ensemble-disagreement scalar:

$$s_t := \left[\, h_t^{(\text{ref})},\ u_t \,\right] \in \mathbb{R}^{129}, \tag{3}$$

where $h_t^{(\text{ref})} \in \mathbb{R}^{128}$ is the seed-42 ("reference") ensemble member's hidden vector and $u_t \in \mathbb{R}$ is the disagreement defined in subsection 4.1. Given action a_t, transitions sample a synthetic correctness label $\tilde{y}_t \sim \text{Bernoulli}(\bar{p}_t(a_t))$ and feed each ensemble member the tuple $(a_t, \tilde{y}_t)$ to advance every $h_t^{(i)} \rightarrow h_{t+1}^{(i)}$.

Trajectories for offline DT training are generated by rolling out an ε-myopic-greedy behavior policy μ_ε in the world model:

$$\mu_\varepsilon(a \mid s_t) \;=\; (1-\varepsilon)\,\mathbb{1}\!\left[a \in \arg\max_{a'} \hat{p}_t^{(\text{ref})}(a')\right] \;+\; \frac{\varepsilon}{|\mathcal{A}|}, \tag{4}$$

with $\varepsilon = 0.5$. We sample $N_{\text{traj}} = 10{,}000$ trajectories of horizon $T = 50$ to form $\mathcal{D}$. This concentrates the offline data on competent-teacher rollouts while preserving enough exploration to give the DT a usable conditioning signal. The pipeline chains real students $\rightarrow$ DKT student model $\rightarrow$ DKT-as-environment $\rightarrow$ myopic-greedy synthetic trajectories $\rightarrow$ DT trained on those trajectories $\rightarrow$ off-policy evaluation of the DT on real test states; each link introduces shift, and the choice of myopic-greedy biases offline data toward high-success-probability actions, which interacts with the boundary-of-capability reward as we discuss in subsection 5.3.

4.4 Reward shapes

We define four reward functions $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ computable from the world model, using the predicted-correctness shorthand $\hat{p}_t^{(i)}(a)$ and ensemble mean $\bar{p}_t(a)$ from subsection 4.1. The first two are baselines (IMM and SIB-MEAN). The last two are pedagogically-grounded shapes drawn from learning-sciences traditions: PFL-GAIN (PFL-grounded) and PF-DIFF at $p^\star = 0.5$ (PF / desirable-difficulty grounded).

Immediate (IMM).

$$r^{\text{IMM}}(s_t, a_t) := \bar{p}_t(a_t) = \frac{1}{K}\sum_{i=1}^{K} \hat{p}_t^{(i)}(a_t). \tag{5}$$

Maximizing this rewards "teach what the student is already most likely to know" — the standard formulation in the RL-for-tutoring literature, and the formulation PFL, PF, and desirable difficulty identify as pedagogically misaligned.

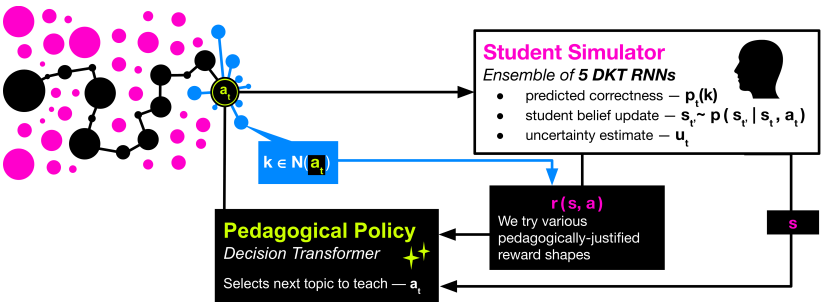

Figure 3: The world-model training loop. The pedagogical policy (a Decision Transformer) selects a topic a_t to teach. The student simulator (a $K = 5$ DKT ensemble) returns predicted correctness $p_t(k)$ over the neighborhood $\mathcal{N}(a_t)$, a stochastic belief update $s'_t \sim p(s'_t \mid s_t, a_t)$, and an ensemble-disagreement uncertainty u_t. The reward $r(s, a)$ is computed under one of the four shapes defined in subsection 4.4; the state s and reward r close the loop back to the policy. The DKT ensemble is fit on real student data (subsection 4.2); the highlighted node a_t and its neighbors visualize the locality structure the sibling-mean and PFL-gain reward shapes exploit.

Sibling-mean (SIB-MEAN).

$$r^{\text{SIB-MEAN}}(s_t, a_t) := \frac{1}{|\mathcal{N}(a_t)|} \sum_{k \in \mathcal{N}(a_t)} \bar{p}_t(k), \tag{6}$$

the ensemble-mean predicted correctness averaged over the KC-neighborhood $\mathcal{N}(a_t)$ defined in subsection 4.1. For only-child leaves, $\mathcal{N}(a_t) = \{a_t\}$ and Equation 6 reduces to Equation 5. The sibling neighborhood is a stand-in for the downstream signal faithful PFL would compute on prerequisite edges; subsection 5.1 reports what this proxy buys, and section 6 returns to the data-infrastructure gap that motivates the proxy in the first place.

PF boundary-of-capability (PF-DIFF, ours).

$$r^{\text{PF-DIFF}}(s_t, a_t) := -\left| r^{\text{IMM}}(s_t, a_t) - p^\star \right|. \tag{7}$$

Maximized when the predicted-correctness on the taught KC matches a target $p^\star$ at the boundary of capability. The moderate-difficulty band is motivated by Productive Failure and the desirable-difficulty literature, which place the productive-engagement zone near the boundary where the student's intuitions are surfaced as variant solution attempts before structured help. We report at $p^\star = 0.5$. The choice produces an algebraically near-perfect anti-correlation with IMM under any behavior policy whose p_t concentrates on one side of 0.5, the myopic-greedy $\varepsilon = 0.5$ behavior of Equation 4.3 included. We did not anticipate the anti-correlation when proposing the operationalization; the precondition diagnostic (section 4.6) caught it before any training was run, and the correlation value is reported in Table 1.

PFL preparedness-gain (PFL-GAIN, ours).

$$r^{\text{PFL-GAIN}}(s_t, a_t, s_{t+1}) := \frac{1}{|\mathcal{N}(a_t)|} \sum_{k \in \mathcal{N}(a_t)} \left[\bar{p}_{t+1}(k) - \bar{p}_t(k) \right]. \tag{8}$$

The change in ensemble-mean predicted correctness on the KC-neighborhood from before to after the teaching action. Operationalizes "learning happened" as a measurable belief shift; depends on the next state s_{t+1} in addition to (s_t, a_t). The neighborhood $\mathcal{N}(a_t)$ is the sibling set defined in subsection 4.1; section 6 discusses why this is a sibling proxy for PFL's downstream-belief-change construct rather than a faithful operationalization.

4.5 Decision Transformer with MOPO pessimism

The policy is a Decision Transformer [7] with 4 layers, 4 heads, $d_{\text{model}} = 128$ (~1.13M parameters), trained on (state, return-to-go, action) sequences via supervised cross-entropy on the action conditional on its prefix tokens. Optimizer Adam, learning rate 10^{-4}, batch size 128, 50 epochs. MOPO

pessimism is applied at training time with $\lambda \in \{0.5, 1.0, 2.0\}$ for every reward variant. At inference, each DT is queried at its own training-distribution 95th percentile target return-to-go: on XES3G5M, IMM $R_0 = 46.5$, SIB-MEAN 42.8, PF-DIFF ($p^\star = 0.5$) -17.0, PFL-GAIN 0.9; on Junyi, DT(IMM, $\lambda = 2.0$) $R_0 = 40.3$ and DT(IMM, $\lambda = 0.5$) $R_0 = 41.7$. subsection 4.6 explains why per-policy target RtGs are needed.

4.6 Evaluation infrastructure

We assemble four diagnostics around the offline-RL reward-shape comparison: a cumulative-return correlation precondition, a DT-vs-empirical-teacher action-overlap check, a pairwise inter-policy diagnostic, and a cross-dataset replication check (introduced in subsection 5.5). Each is a reapplication of standard offline-RL or empirical practice — correlation analysis, behavior-policy distribution comparison, replication on a second dataset — applied as a screening protocol at the reward-design stage rather than after results are reported. Two pieces of underlying machinery support the diagnostics: a three-estimator off-policy evaluation protocol (FQE, WIS, DualDICE) on test-split trajectories, and an out-of-distribution guard on the inference-time return-to-go. A multi-seed FQE bootstrap (20 random-seed re-runs per policy, Welch two-sample t-test on seed-mean values) scales statistical significance to the DT class's seed variance; a 5-seed pilot returned $t = 1.34$ on a comparison the 20-seed bootstrap returned $t = +5.41$ on, motivating the budget.

Three-estimator off-policy evaluation. A single OPE estimator's failure mode can silently propagate into a finding; using three lets agreement strengthen one and disagreement bound it. We score every policy on $N = 1{,}445$ real XES3G5M test-split trajectories (and the analogous Junyi test split). Test states are recovered by replaying each student's (a, y) sequence through the reference DKT to materialize the per-step hidden vector and ensemble disagreement. We use FQE, WIS, and DualDICE; agreement across all three strengthens a finding, and disagreement bounds it. FQE uses a 2-hidden-layer Q network with hidden size 128, $\gamma = 1$, soft target update $\tau = 0.005$, MSE on Bellman targets, 50 epochs. WIS uses a state-bucketed empirical behavior policy (k-means with $k = 64$, Laplace-smoothed action counts per bucket). We refer to this bucketed-empirical behavior policy — the WIS-fitted estimate of the teacher distribution induced by the logged interactions — as the *empirical teacher*, and use it as a comparison anchor for FQE and the action-overlap diagnostics throughout section 5. DualDICE uses $\gamma = 0.99$ as a discounted approximation, which makes its absolute scale incomparable with FQE/WIS but preserves cross-policy ranks within its column. Each estimator is rank-validated on an 8-state HMM toy environment with computable Monte Carlo ground truth before its values are reported: FQE underestimates absolute values by $\sim 2\times$ but its ranking is preserved on the (oracle, ε-oracle, uniform) policy triple, and DualDICE preserves rank at $\gamma = 0.95$.

Out-of-distribution evaluation guard. A return-conditioned policy must be queried at a target return-to-go inside its training distribution. An earlier comparison evaluated all DTs at a uniform $R_0 = 37.5$; a sibling-contrast DT trained at returns $R_0 \sim \mathcal{N}(-4.13, 1.53)$ was queried at a 27-σ extrapolation that produced nominal but pathologically extrapolative behavior, which was briefly mistaken for an interesting empirical finding. Per-policy target RtGs (each DT at its own training-distribution 95th percentile) raised zero-context single-step argmax agreement from 76% to 97% and dissolved the artifact.

Cumulative-return correlation precondition. A candidate reward whose return-to-go ranks trajectories the same way as immediate is mechanically equivalent to the immediate baseline; a return-conditioned policy trained on it learns the same policy as DT(IMM), regardless of the reward's formula. The precondition catches this before training. For $N = 1{,}000$ trajectories $\tau \sim \mu_\varepsilon$, compute the per-$(\tau, t)$ Pearson correlation between the candidate reward's return-to-go $G_t^{\text{cand}}(\tau) := \sum_{t' \geq t} r^{\text{cand}}(s_{t'}, a_{t'})$ and the immediate reward's return-to-go $G_t^{\text{IMM}}(\tau)$. A correlation > 0.95 in absolute value indicates the equivalence (up to a sign flip on the negative side).

Action-overlap diagnostic. A policy whose actions are out-of-distribution under the logged data will have high FQE despite teaching nothing the data covers, because the FQE Q-net inherits the same overestimation — the circular-Q-bootstrapping pathology. The action-overlap diagnostic catches this. For a policy π and a bucketed-empirical behavior policy μ, we sample 500 real test states and report (a) per-state argmax agreement $\Pr_s[\arg\max \pi(\cdot \mid s) = \arg\max \mu(\cdot \mid s)]$, (b) top-5 set overlap, and (c) the empirical-teacher probability at π's argmax versus uniform $1/|\mathcal{A}|$. We use the diagnostic in

Candidate reward	Per-(traj, step) corr.	Verdict
SIB-MEAN	0.9956	Fails (positive collapse)
PF-DIFF ($p^\star = 0.5$)	-0.9959	Fails (negative collapse)
PF-DIFF ($p^\star = 0.85$)	-0.9325	Borderline
PFL-GAIN	-0.1305	Passes (orthogonal)

Table 1: Cumulative-return correlation precondition: Pearson correlation between each candidate reward's cumulative return and the immediate reward's cumulative return, computed at the per-(trajectory, step) level over $N = 1{,}000$ behavior-policy trajectories. The threshold for "passing" the precondition is $|\text{corr}| < 0.95$. SIB-MEAN fails on the positive side—a return-conditioned policy trained on it will be mechanically equivalent to the immediate-reward baseline. PF-DIFF at $p^\star = 0.5$ fails on the negative side—the same mechanical equivalence holds up to a sign flip, so the trained policy is approximately "minimize immediate reward." Only PFL-GAIN passes the precondition strictly.

two variants. The DT-vs-empirical-teacher form catches policies whose argmax distribution is far from the behavior data; the pairwise inter-policy form catches the case where multiple reward shapes produce policies that are individually OOD but mutually identical (subsection 5.6).

Cross-dataset check. A comparison's signal might depend on the calibration regime of the world model that mediates it. We run the same comparison protocol on a second dataset (Junyi Academy) wherever the first dataset's result would otherwise stand alone, and require the conclusion to survive the dataset switch before treating it as evidence about the policies rather than the world model. subsection 5.5 reports the cross-dataset run.

5 Results

5.1 The cumulative-return correlation diagnostic

We run the cumulative-return correlation precondition (subsection 4.6) on the four reward shapes before training any DT, to test whether each candidate reward is mechanically distinguishable from the immediate baseline.

Table 1 reports the precondition for each candidate reward shape. SIB-MEAN's correlation with immediate is $+0.9956$ on XES3G5M — a property of how the curriculum tree was curated. Siblings in the XES3G5M four-level hierarchy are grouped by similarity, so the mean predicted correctness over $\mathcal{N}(a)$ is close to the value at a itself, and a return-conditioned policy trained on SIB-MEAN cannot distinguish the two return signals. The precondition catches the equivalence in five minutes, before training; an independently-trained SIB-MEAN DT subsequently agreed with the IMM DT on 99% of argmax actions across $5{,}000$ evaluation states. The precondition saved a day of compute on this shape that would otherwise have been spent re-discovering the mechanical equivalence.

PF-DIFF at $p^\star = 0.5$ has correlation -0.9959. The mechanism is algebraic: under a behavior policy whose per-step predicted correctness clusters above $p^\star$, the absolute value in $r^{\text{PF-DIFF}} = -|\bar{p}_t(a_t) - 0.5|$ resolves with a fixed sign, and $r^{\text{PF-DIFF}} \approx 0.5 - r^{\text{IMM}}$ across the regime the data covers. The cumulative-return pair $(G_t^{\text{PF-DIFF}}, G_t^{\text{IMM}})$ is therefore a near-perfect line with negative slope — not a coincidence the precondition discovers, but a consequence of the formula and the behavior distribution that the precondition computes in five minutes. A DT conditioned to maximize cumulative PF-DIFF should, in the limit, learn the inverse of the immediate-reward DT. The FQE result reported in subsection 5.2 comes in above the immediate baseline; subsection 5.3 runs the action-overlap diagnostic and resolves the gap. PF-DIFF at $p^\star = 0.85$ correlates with IMM at -0.9325, borderline against the threshold; we do not develop it as a separate result.

PFL-GAIN correlates with immediate at -0.1305. The low correlation is not a discovered orthogonality so much as a consequence of comparing a delta-based reward (per-step belief change on a sibling neighborhood) with a level-based reward (predicted correctness on the taught skill); cumulative deltas and cumulative levels routinely correlate weakly. PFL-GAIN clears the $|\text{corr}| < 0.95$ threshold the precondition uses, so a return-conditioned policy trained on it will not be mechanically tied to the immediate baseline.

Policy	Mean FQE	Std	Min	Max
`uniform-random`	0.656	0.015	0.628	0.685
`empirical-teacher`	2.679	0.025	2.624	2.733
DT(IMM,$\lambda = 2.0$)	2.061	0.184	1.6	2.4
DT(PF-DIFF,$\lambda = 2.0$)	**2.714**	0.235	2.240	3.182
DT(IMM,$\lambda = 1.0$)	1.971	0.187	1.593	2.240
DT(PFL-GAIN,$\lambda = 1.0$)	1.284	0.101	1.057	1.497

Table 2: 20-seed FQE bootstrap on the XES3G5M four-policy portfolio. Each row reports the seed-mean and seed-std of FQE training over random seeds 42–61.

Comparison	Mean diff	Pooled SE	Welch t
DT(PF-DIFF,$\lambda = 2.0$) vs. DT(IMM,$\lambda = 2.0$)	$+0.653$	0.066	$+\mathbf{9.86}$
DT(PF-DIFF,$\lambda = 2.0$) vs. `empirical-teacher`	$+0.036$	0.053	$+0.68$
DT(PFL-GAIN,$\lambda = 1.0$) vs. DT(IMM,$\lambda = 1.0$)	-0.687	0.047	$-\mathbf{14.48}$
DT(PFL-GAIN,$\lambda = 1.0$) vs. `empirical-teacher`	-1.394	0.023	-59.95

Table 3: Pairwise Welch two-sample t-statistics from the XES3G5M 20-seed FQE bootstrap.

5.2 XES3G5M boundary-of-capability bootstrap

We test whether DT(PF-DIFF, $\lambda = 2.0$) and DT(PFL-GAIN, $\lambda = 1.0$) score differently from their immediate baselines on FQE-immediate. The 20-seed bootstrap follows subsection 4.6: 20 random-seed FQE re-runs per policy, Welch two-sample t-test on the seed-mean values. Table 2 reports the per-policy seed statistics for the four-policy headline portfolio and Table 3 the pairwise comparisons.

DT(PF-DIFF, $\lambda = 2.0$) achieves mean FQE 2.714 ± 0.235; DT(IMM, $\lambda = 2.0$) achieves 2.061 ± 0.184. The $+0.653$ mean difference yields a Welch $t = +9.86$ ($p \ll 10^{-5}$). Against the empirical teacher's 2.679 ± 0.025, the $+0.036$ mean difference yields $t = +0.68$: at $n = 20$ seeds the minimum detectable effect at 80% power and $\alpha = 0.05$ is approximately 0.10, so $+0.036$ falls below detection. Numerically, the boundary-of-capability DT lands inside the empirical teacher's $\pm 1\sigma$ FQE band. We initially read this as a positive empirical claim about teaching quality. subsection 5.3 reports the action-overlap diagnostic that revisits that reading.

5.3 The PF-diff teacher-band FQE is a Q-bootstrapping artifact

The precondition diagnostic predicted, in advance, that DT(PF-DIFF, $\lambda = 2.0$) should not reach the empirical teacher's FQE band on the merits of teacher-like behavior. PF-diff at $p^\star = 0.5$ has cumulative-return correlation with immediate of -0.9959 (Table 1)—strongly anti-correlated under myopic-greedy behavior, the regime that drives synthetic-trajectory generation. A DT conditioned to maximize cumulative PF-DIFF should, in the limit, learn the inverse of the immediate-reward DT. That its FQE comes in above the immediate-reward baseline—and on top of the empirical teacher—is exactly the circular-bootstrapping signature: a policy whose Q-values are high under one Q network can be high under a second Q network trained on the same logged data even when the policy never visits states the data covers.

We ran the action-overlap diagnostic to test this. At 500 sampled real-test states, with $k = 5$ for the top-5 set, DT(PF-DIFF, $\lambda = 2.0$) at target RtG -17 against the empirical teacher returns the numbers in Table 4.

These numbers are worse than CQL's, the canonical example of a policy whose FQE is inflated by Q-bootstrapping. The teacher-band FQE on PF-diff is the same pathology, on a fresh policy class, detected by the same diagnostic.

The action-overlap diagnostic overturns the framing of DT(PF-DIFF, $\lambda = 2.0$) reaching the empirical teacher's FQE band as evidence of teacher-like teaching. The narrower factual claim that survives — DT(PF-DIFF, $\lambda = 2.0$) achieves higher FQE than DT(IMM, $\lambda = 2.0$) trained at the same MOPO penalty on the same data — is a relative-ranking statement on a single estimator, and the action-overlap evidence shows the ranking is driven by the Q-net's overestimation on out-of-distribution

Metric	DT(PF-DIFF, λ=2.0)	CQL	Uniform
Zero-overlap fraction (top-5)	100%	70%	—
Argmax agreement with empirical teacher	0%	—	$1/865 \approx 0.12\%$
Empirical-teacher prob. at argmax (median)	0.0005	0.0008	$1/865 \approx 0.00116$

Table 4: Action-overlap diagnostic on 500 real test states. DT(PF-DIFF, $\lambda = 2.0$) at target RtG -17 against the empirical teacher, with CQL as an offline-RL comparator (DiscreteCQL via `d3rlpy`, two seeds: FQE $4.50/11.94$, WIS NaN from importance-weight overflow). The DT chooses actions the empirical teacher places below uniform mass on; CQL exhibits the same Q-bootstrap pathology on this evaluation infrastructure.

actions rather than by either policy's teaching quality. The $+0.653$ mean difference and $t = +9.86$ in Table 3 are real arithmetic on the seed-level FQE values; their interpretation is agnostic to whether either DT teaches well.

Q-bootstrapping inflation under offline-RL FQE is the pathology CQL motivates pessimism against [13]; the action-overlap diagnostic catches the same pathology on a fresh policy class in five minutes on the same evaluation infrastructure. What the catch traces back to, on the data-infrastructure side, is sharp. The XES3G5M action space is a flat 865-way list of KCs, with no hierarchical structure separating "stay on this topic" from "advance to a new one," and the offline data is sparsely-covered $\varepsilon = 0.5$ myopic-greedy rollouts in the world model rather than logged behavior from a teacher of any kind. A Q network trained on those rollouts has no signal at most of the action space, and a return-conditioned policy that argmaxes into the uncovered region inherits an FQE that reflects Q-net optimism rather than population return. Two pieces of data infrastructure would give offline-RL evaluation of pedagogically-grounded rewards a tractable surface to work against on this dataset. The first is a transfer-aligned action structure — a hierarchical action that separates within-node from between-node decisions, so the policy is not forced to argmax over 865 alternatives at every step. The second is a behavioral anchor outside synthetic trajectories — logs from a real teacher, or a faithful behavior cloning of one — so that FQE's $\hat{Q}^\pi$ is fit on transitions that bound the policy's action distribution rather than on ε-greedy noise. Neither exists in the current pipeline; the action-overlap diagnostic catches the result that would have shipped without them.

Isolation experiment. The two-piece prescription above predicts that anchoring the behavior data on a higher-return slice should partially close the Q-bootstrap gap. We tested this directly. We re-trained DT(PF-DIFF, $\lambda = 2.0$) on the top-10% slice of the original myopic-greedy trajectories (1,000 trajectories retained, kept return range $[-17.531, -12.451]$ vs. source $[-23.664, -12.451]$) and re-ran the action-overlap diagnostic. The argmax agreement with the empirical teacher remains 0%; the top-5 set overlap moves marginally from $0.00/5$ to $0.06/5$; the median empirical-teacher probability at the DT's argmax stays at 0.00051. The pairwise argmax agreement between the new DT and the baseline DT(PF-DIFF) is 99.2% on 500 real test states. Filtering the synthetic trajectories does not move the policy meaningfully. The dominant mechanism is action-space flatness, not the synthetic-trajectory bridge per se; a higher-quality slice of the same flat-action-space data is not equivalent to a behavioral anchor outside the model.

5.4 PFL preparedness-gain on XES3G5M

DT(PFL-GAIN, $\lambda = 1.0$) achieves mean FQE 1.284 ± 0.101; DT(IMM, $\lambda = 1.0$) achieves 1.971 ± 0.187 (Table 2). The -0.687 mean difference yields $t = -14.48$ (Table 3). The DT trained on PFL-GAIN scores reliably lower than the immediate baseline on FQE-immediate, which follows trivially: the training objective is a delta on sibling beliefs and the evaluation metric is a level on the taught skill, and the two are weakly correlated (Table 1). The non-trivial observation is that on XES3G5M the DT class is responsive to reward signal at all — a property the Junyi pairwise diagnostic in subsection 5.6 rules out at $\lambda = 2.0$, and a contrast we discuss in section 8. The result speaks to a sibling-aggregation reward framed under the PFL motivation, not to a faithful PFL operationalization; we discuss the proxy gap in section 6. The diagnostic protocol refutes the hypothesis that PFL-GAIN would produce a policy that teaches better on FQE-immediate. The data show "different reward, different policy," nothing more; reframing PFL-GAIN as succeeding under a

Policy	Mean FQE	Std	Min	Max
`uniform-random`	4.112	0.056	4.016	4.235
`empirical-teacher`	8.515	0.072	8.388	8.705
DT(IMM,$\lambda = 2.0$)	**10.042**	0.133	9.818	10.332
DT(PF-DIFF,$\lambda = 2.0$)	8.041	0.129	7.821	8.308

Table 5: 20-seed FQE bootstrap on Junyi Academy. The relative ranking between DT(IMM, $\lambda = 2.0$) and DT(PF-DIFF, $\lambda = 2.0$) inverts compared to the XES3G5M result in Table 2.

Comparison	Mean diff	Pooled SE	Welch t
DT(PF-DIFF,$\lambda = 2.0$) vs. DT(IMM,$\lambda = 2.0$)	-2.001	0.041	$\mathbf{-48.30}$
DT(PF-DIFF,$\lambda = 2.0$) vs. `empirical-teacher`	-0.474	0.033	-14.33
DT(IMM,$\lambda = 2.0$) vs. `empirical-teacher`	$+1.527$	0.034	$+45.21$

Table 6: Pairwise Welch two-sample t-statistics from the Junyi 20-seed FQE bootstrap. The first row is the cross-dataset reproduction of Table 3's headline comparison; the sign reverses and the magnitude is approximately five times larger.

misaligned metric would require evaluating it on a metric aligned with its training objective, which the current pipeline does not include (section 7).

5.5 Cross-dataset reproduction on Junyi: the PF-diff vs. Imm ranking is calibration-dependent

The narrower claim that survived subsection 5.3—DT(PF-DIFF, $\lambda = 2.0$) achieves higher FQE than DT(IMM, $\lambda = 2.0$) at the same MOPO penalty on the same data—is a relative-ranking statement on a single estimator and a single dataset. We test its generalization by running the same 20-seed bootstrap on Junyi Academy [6], a primary-school mathematics interaction log of 16.2M (student, exercise, correctness) interactions over 1,326 unique exercises organized into a four-level topic hierarchy. The DKT ensemble uses the same 1L-LSTM-128 architecture as on XES3G5M; on Junyi it reaches validation AUC 0.762, an ~ 0.09 gap below the XES3G5M result and close to the Junyi DKT ceiling reported in the literature. Trajectory generation, DT training, single-seed OPE, and the 20-seed bootstrap follow the protocol of subsection 5.2 unchanged. The cumulative-return correlation between PF-DIFF and IMM on Junyi is -0.9931, a near-mirror of the -0.9959 on XES3G5M, confirming the structural anti-correlation is a property of the reward shape and not a dataset accident.

The relative ranking inverts. DT(IMM, $\lambda = 2.0$) achieves mean FQE 10.042 ± 0.133; DT(PF-DIFF, $\lambda = 2.0$) achieves 8.041 ± 0.129. The -2.001 mean difference yields a Welch $t = -48.30$, opposite direction from the XES3G5M result and approximately five times its magnitude. Against the empirical teacher's 8.515 ± 0.072, DT(PF-DIFF) falls below by -0.474 ($t = -14.33$), while DT(IMM) clears the teacher by $+1.527$ ($t = +45.21$). The single-seed OPE that preceded the bootstrap also captured WIS and DualDICE: WIS calls DT(IMM) and DT(PF-DIFF) tied (43.77 vs. 43.44), and DualDICE at $\gamma = 0.99$ rank-orders DT(PF-DIFF) above DT(IMM). Three estimators, three orderings between the same two policies on the same evaluation protocol. OPE rankings depending on world-model calibration are documented in the methodology literature; the cross-dataset check surfaces the dependence on this specific comparison, which a single-dataset evaluation would miss.

The cross-dataset reproduction corroborates the action-overlap reading from subsection 5.3 via a different mechanism. A DT trained to maximize cumulative PF-DIFF should, in the limit, learn the inverse of the immediate-reward DT — this is what the cumulative-return correlation precondition predicts. The XES3G5M result that DT(PF-DIFF) scored higher than the immediate baseline on FQE-immediate despite the structural anti-correlation depended on the XES3G5M world model's calibration (test AUC 0.851) absorbing the mismatch; Junyi's noisier model (val AUC 0.762) breaks the alignment, and the FQE comparison goes where the reward's training objective predicts. The precondition diagnostic, the FQE-direction reversal between datasets, and the three-estimator disagreement on Junyi all point to the same reading: the FQE comparison between DT(PF-DIFF) and DT(IMM) at $\lambda = 2.0$ measures the world model's calibration regime, not teaching quality.

The data-infrastructure gap underneath is that there is no behavioral anchor outside the world model itself. Real students appear only as the data the world model was trained on; the policy is trained on synthetic rollouts in the world model; the policy is evaluated on the same world model's hidden-state replay of test students. The world model's calibration is the load-bearing variable in every step of that chain, and a comparison that depends on the calibration regime as much as on the reward shape carries no signal about the reward shape. A logged real-teacher trajectory, or a post-intervention transfer assessment scored by a third party, would each give the offline evaluation a fulcrum outside the world model. Neither exists in either dataset. The cross-dataset check enforces honesty: a single-dataset reading would have credited the XES3G5M PF-diff vs. Imm FQE result as evidence about the policies; the inversion on Junyi exposes it as evidence about the world model.

5.6 Pairwise inter-DT diagnostic identifies policy-collapse on Junyi

We test whether different pedagogically-grounded reward shapes produce different policies on Junyi at $\lambda = 2.0$. Beyond PF-DIFF and PFL-GAIN, we trained three additional learning-sciences-grounded reward shapes — sliding-window learning progress [8, 16], moving-target difficulty [24], and ensemble information gain. All three clear the cumulative-return precondition (per-(traj, step) correlations with IMM: $+0.429$, -0.942, $+0.622$); two of them clear DT(IMM, $\lambda = 2.0$) on the 20-seed FQE bootstrap at $t \geq +12.9$. The pairwise inter-DT action-overlap diagnostic, computed on the same 500 Junyi test states as subsection 5.3, settles the question. Across all ten pairs of the five Junyi DTs (IMM, PF-DIFF, PF-DECAY, LP-WINDOW, VARIANCE-REDUCE), argmax agreement is 99.6% or higher on real test states; the deployable policy is invariant across reward shapes. The bootstrap t-statistics measure FQE Q-net evaluation noise across reward-shape RtG conditioning on what is essentially a single underlying policy.

The reading for this collapse is structural rather than algorithmic. Junyi's action space is a flat 1,326-way list of exercises without a hierarchical structure separating within-topic from between-topic decisions, and the world model's validation AUC of 0.762 leaves the per-step ensemble disagreement u_t large across most of the state space. One candidate mechanism: the MOPO penalty $\lambda \cdot u_t$ at $\lambda = 2.0$ swamping the reward term across the five shapes. The isolation experiment below tests this directly. A separate candidate — the world-model calibration regime, addressable by retraining Junyi's DKT until validation AUC clears the XES3G5M band — remains open.

Isolation experiment. The MOPO-dominance reading predicts that lowering λ should partially recover reward-shape sensitivity. We tested this directly. We re-trained DT(IMM) on the same Junyi trajectory blob at $\lambda = 0.5$ (four times smaller than the $\lambda = 2.0$ baseline) and computed the pairwise argmax agreement against the existing DT(IMM, $\lambda = 2.0$) on 500 real Junyi test states. The argmax agreement is 99.2%, with top-5 overlap $3.21/5$; both DTs land at the same 8.6% argmax agreement with the empirical teacher. Lowering MOPO by a factor of four barely moves the policy. The MOPO-dominance mechanism is therefore not supported on its own; the structural cause is action-space flatness combined with the low-AUC world-model regime, with the pessimism penalty incidental. This converges with the trajectory-filter isolation on XES3G5M (subsection 5.3): across the two datasets, action-space flatness is the load-bearing structural variable.

6 Data-Infrastructure Gaps

Each diagnostic firing in section 5 traces to a specific gap in the data infrastructure available to offline-RL for instructional sequencing. The catalogue specifies what a faithful next attempt would have to satisfy.

No directed prerequisite graph. PFL's transfer-of-learning construct demands a directed edge: action a teaches about a downstream topic b that depends on a. XES3G5M's four-level KC hierarchy is undirected and curated by similarity, which the precondition exposed as the $+0.9956$ collapse of SIB-MEAN to IMM. PFL-GAIN's sibling-belief signal is the same proxy (section 6). A directed prerequisite graph, with explicit edges $a \to b$, would let both rewards be operationalized faithfully; neither XES3G5M nor Junyi exposes one.

No transfer-aligned action structure. A pedagogically-grounded reward over a graph implies a structured action space (stay-on-node vs. advance), but the action spaces here are flat: 865 KCs on

XES3G5M, 1,326 exercises on Junyi. The action-overlap diagnostic and the two isolation experiments converge on this flatness as the load-bearing cause of OPE pathology: filtering training data to the top-10% slice on XES3G5M and dropping the MOPO penalty by $4\times$ on Junyi both leave the policy unchanged at 99.2% argmax agreement with the original. A hierarchical action space — within-node intervention plus advance to a small neighborhood — would tighten the support FQE fits on and give the reward term a margin to drive policy difference.

No post-intervention transfer assessment. PFL's empirical metric is a transfer assessment, scored on held-out related material after the intervention. The offline pipeline lacks one. PFL-GAIN substitutes sibling-belief change in the world model, so the PFL-GAIN bootstrap is evidence about a sibling-aggregation reward under FQE-immediate, not evidence about a faithful PFL reward under PFL's own construct.

No behavioral anchor outside synthetic trajectories. The pipeline routes everything through the world model: it simulates students for trajectory generation, computes rewards, and supports OPE on real test states recovered via DKT replay. The cross-dataset rank inversion (subsection 5.5) traces to the model's calibration regime (XES3G5M test AUC 0.851 vs. Junyi val AUC 0.762), not to the policies. A real-teacher trajectory log would give FQE an anchor decoupled from the model's miscalibrations. The BC-filtered-top-10% baseline we trained via `d3rlpy` (WIS 45.0, above the empirical teacher) imitates the synthetic teacher, not a real one — F01 confirms that filtered synthetic is not equivalent to a behavioral anchor.

7 Future Work

Immediate next experiments. Four experiments would directly address what this paper does not establish. (i) A faithful PFL reward over a knowledge graph with prerequisite edges would test PFL on a reward consistent with its transfer-of-learning construct (section 6). (ii) A calibration-regime sweep on Junyi (DKT retrained to higher AUC, or λ swept down) would test whether reward-shape sensitivity recovers (subsection 5.6). (iii) World-model rollout under each reward's own metric would let PFL-GAIN be evaluated outside FQE-immediate. (iv) A distribution-matching boundary-of-capability reward — KL divergence from a moment-matched target distribution of predicted correctnesses across the trajectory — would clear the precondition by construction and could be evaluated under the remaining three diagnostics on its merits.

The broader research program. Three pillars sit between this paper and a deployable system: a pedagogically-faithful PFL reward over a prerequisite graph (closing section 6 and section 6); automatic knowledge-graph construction from class assets — slide decks, worksheets, video recordings; and a hierarchical natural-language action space, with a foundation-model interaction layer executing teaching turns (closing section 6). Each pillar carries through the screening protocol before any claim of teaching quality.

A radical direction. One direction worth parking sits outside the offline-RL setting: an online RL system updated against a pedagogical LLM-as-judge. The construction borrows from RLHF [18] and Constitutional AI [2], with the constitution being a pedagogical perspective (Productive Failure, Preparation for Future Learning, Vygotsky's zone of proximal development) rather than helpfulness-harmlessness. The judge replaces the off-policy estimators with a reward signal grounded in a teaching corpus that exemplifies the chosen pedagogy. The risks — judge calibration, distillation failure, corpus-pedagogy alignment — are open.

8 Conclusion

The four diagnostics — a cumulative-return precondition, an action-overlap check, a pairwise inter-DT diagnostic, and a cross-dataset check, all backed by a 20-seed FQE bootstrap matched to policy-class variance — caught every comparison between a pedagogically-grounded policy and the immediate baseline. The precondition caught SIB-MEAN at $+0.9956$ (similarity-curated tree edges) and PF-DIFF at -0.9959 (foreseeable algebra). The action-overlap diagnostic caught DT(PF-DIFF, λ=2.0)'s teacher-band FQE as Q-bootstrap inflation on actions the flat 865-way space leaves uncovered. The

cross-dataset check caught the PF-DIFF-vs-IMM ranking as calibration-dependent. The pairwise diagnostic caught five differently-rewarded DTs on Junyi at $\lambda = 2.0$ collapsing to a single policy at 99.6%+ argmax agreement. Two isolation experiments (filtering trajectories to the top-10% slice on XES3G5M; dropping the MOPO penalty by $4\times$ on Junyi) converge on action-space flatness as the load-bearing structural cause.

The paper offers no positive claim about either reward shape. Its contribution is the screening protocol — diagnostics applied before training, isolation experiments applied after — paired with a specification of the data infrastructure a faithful evaluation would require: a directed prerequisite graph, a transfer-aligned action structure, a post-intervention transfer assessment, and a behavioral anchor outside synthetic trajectories. None of these are currently available; the paper names them as the prerequisites for the next attempt.

Contributions

Ozan Bayiz. Conceived the project; designed the four-diagnostic screening protocol and the reward-shape comparisons; implemented the codebase (DKT ensembles, world-model environment, Decision Transformer training, off-policy evaluation pipeline, screening diagnostics, and two isolation experiments); ran all experiments on Modal; performed all analysis; made all figures; wrote the paper.
Narges Norouzi. Faculty advisor.

References

[1] Richard C. Atkinson. Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, 96(1):124–129, 1972.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback, 2022.

[3] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C. Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In *CHI Conference on Human Factors in Computing Systems*, 2020.

[4] Elizabeth L. Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Morton Ann Gernsbacher, Richard W. Pew, Leaetta M. Hough, and James R. Pomerantz, editors, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth Publishers, 2011.

[5] John D. Bransford and Daniel L. Schwartz. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24(1):61–100, 1999.

[6] Hsiang-Sheng Chang, Hsuan-Ju Hsu, and Kuan-Ta Chen. Modeling exercise relationships in e-learning: A unified approach. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*, pages 532–535, 2015. Junyi Academy dataset. The version used here is the Kaggle public release: https://www.kaggle.com/datasets/junyiacademy/learning-activity-public-dataset-by-junyi-academy.

[7] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Benoît Clément, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, 7(2):20–48, 2015.

[9] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. Where's the reward? a review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29(4): 568–620, 2019.

[10] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, 2019.

[11] Manu Kapur. Productive failure. *Cognition and Instruction*, 26(3):379–424, 2008. Cited for contrast in §1.3, §2.5: Productive Failure is a related but distinct learning-sciences framework that intervenes via initial-failure tasks.

[12] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[13] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, 2019. The FQE algorithm appears here as a building block for batch constrained policy learning.

[15] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Yang. XES3G5M: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[16] Manuel Lopes and Pierre-Yves Oudeyer. Strategic student approach for life-long exploration and learning. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–8, 2012.

[17] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[19] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[20] Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, 2000.

[21] Daniel L. Schwartz and John D. Bransford. A time for telling. *Cognition and Instruction*, 16(4):475–522, 1998.

[22] Daniel L. Schwartz and Taylor Martin. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2):129–184, 2004.

[23] Daniel L. Schwartz, Catherine C. Chase, Marily A. Oppezzo, and Doris B. Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4):759–775, 2011.

[24] David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976.

[25] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.