When Offline Selectors Cannot Beat the Best Single Model: A Diagnostic Study on edX Dropout Prediction

Tyler Crosse [1] **Alan Nadelsticher Ruvalcaba** [1] **Dustin Khang LeDuc** [1] **Thomas Trask** [1] **Nicholas Lytle** [1] **David Joyner** [1]

Abstract

Different predictors often excel on different inputs, so picking the best one per instance promises higher accuracy than committing to a single model. In practice, selectors trained from logged data routinely fail to beat the strongest single predictor. Three causes typically go unseparated before more tuning is applied: a mismatched learner, a state that does not predict which model wins, or buffer-to-deployment label shift.

A three-stage diagnostic rules them out on a shared buffer. Stage 1 estimates a local ceiling on oracle recovery from k-NN label consistency. Stage 2 asks whether paired BC and offline-RL learners (BC, DQN, and CQL across penalty weights) reach that ceiling. Stage 3 ablates the selector state to test whether richer features would raise it. The combined verdict points to the most promising next step: tuning the learner, redesigning the state, or collecting new data.

We apply it to selecting among five dropout-prediction models on edX clickstream data. Across 16 windows, the oracle beats the strongest single base model by 9.7 accuracy points on average, yet BC, DQN, and CQL land in the same test-accuracy band below it (robust to a tenfold buffer sweep and N=2,000 held-out examples). The bottleneck is local representational ambiguity: CQL closes the imitation gap without a deployment gain (not conservatism), regret clusters tightly across learners (not tie-breaking), and the three learners converge on test accuracy (not shift). The next iteration should change the state or collect new data, not tune the offline learner further.

Accepted to the ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation (DEMO).
[1]Georgia Institute of Technology. Correspondence to: Tyler Crosse <tylerscottcrosse@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Decision-making from offline datasets is central to a growing class of applications where online interaction is expensive, slow, or ethically constrained, including scientific discovery, engineering design, healthcare, and education (Levine et al., 2020). The data take many forms (logged demonstrations, past trajectories, recorded interactions). A recurring special case is meta-learning over a pool of base models, where a policy chooses the best base model for each instance instead of committing to a single predictor (Rice, 1976; Cruz et al., 2018; 2015). The setup is attractive because individual base models can exhibit context-dependent competence, and prior educational prediction studies commonly compare several plausible model classes on clickstream and virtual-learning-environment data (Liu et al., 2023; Taylor et al., 2014; Casado Hidalgo et al., 2022). In practice, offline meta-learning routinely underperforms strong static baselines, and the reasons are opaque. The weakness could be *algorithmic* (offline RL's conservatism (Kumar et al., 2020) or reward misspecification), *representational* (the state does not contain the information needed to predict which model wins), or *distributional* (the oracle-label distribution available offline differs from the one induced at evaluation). Existing evaluations rarely separate these causes. As a result, practitioners tune algorithms that cannot fix representation failures, engineer features that add zero marginal signal, or attribute failure to shift without quantifying it.

We rule out the three hypotheses in turn with three diagnostics on a shared buffer (Figure 1). **Stage 1** measures local label consistency. The k-nearest-neighbor consistency quantifies how often nearby states in the buffer share the same oracle action, and a held-out 10-NN selector calibrates how that local ambiguity translates into test-time imitation. **Stage 2** separates algorithmic failure from representational failure. We train a supervised behavioral-cloning policy with hard-label cross-entropy and an offline Deep Q-Network (Mnih et al., 2013) on the same buffer. If both fail by similar margins, the bottleneck is more plausibly shared, pointing to representation or distribution rather than algorithm choice. **Stage 3** isolates the marginal value of

features. State ablations test whether the full behavioral state improves over the base-model probability vector, and whether disagreement-derived transforms of that vector add anything further.

The three stages constrain each other. Stage 1's local-consistency ceiling is what makes Stage 2's learner-agreement gap interpretable: learners converging close to a low ceiling implicates the representation, while learners separating below a high ceiling implicates the algorithm. Stage 3 then tests whether reachable features could raise the ceiling. Run together on a shared buffer, the three checks identify which intervention—more tuning, richer features, or upstream data collection—matters next. Each individual diagnostic (k-NN consistency, paired BC/RL ablation, feature ablation, and total-variation buffer-to-test shift) has antecedents in prior work (Cruz et al., 2018; Ko et al., 2008; Kumar et al., 2020); the contribution is the joint reading they enable.

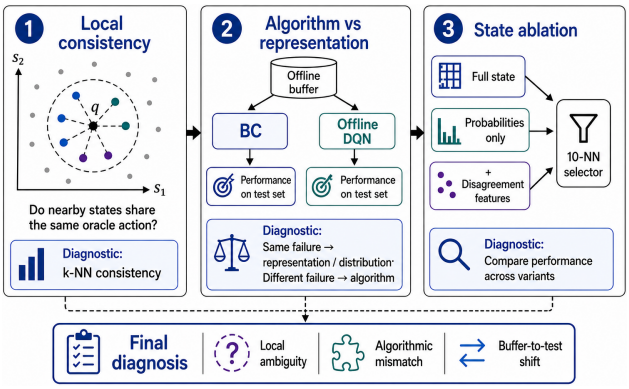

Figure 1. Three-stage diagnostic protocol for offline model selection. The same offline buffer is first used to measure local oracle consistency, then to compare algorithm-specific and shared failure modes across BC, offline DQN, and CQL at three penalty weights, and finally to ablate the selector state to test whether additional feature groups provide marginal value beyond base-model probabilities.

We apply the protocol to a concrete offline decision-making task: selecting among five static dropout-prediction models for MOOC and in-person computer science students on edX clickstream data. The task combines abundant observational data (84.5M events across 223,505 student-course pairs) with ethically constrained online experimentation, because any selected model eventually triggers a human intervention. Across 16 observation/prediction-window configurations, the per-instance oracle beats the strongest single base model by 9.7 accuracy points on average (range 4.5–15.5), yet no learned selector recovers that headroom on held-out accuracy. On the main $(14\text{d}, 14\text{d})$ configuration, every learner sits within ± 0.01 of 0.748 test accuracy, below the 0.762 static reference, while the local-consistency diagnostic is only 0.388 ± 0.010 (Table 1). State ablations show that probabilities-only BC nearly matches the full state, and that disagreement-derived transforms do not materially improve it. Buffer-to-test shift is small in aggregate (marginal $d_{\mathrm{TV}} = 0.063 \pm 0.011$) but locally substantial ($\mathbb{E}_s[d_{\mathrm{TV}}] \approx 0.29$ on the main configuration). The procedure converts an opaque negative result into a verdict on whether the next iteration should target the learner, the features, or the data-collection pipeline. Our contributions are:

- **C1.** A combined diagnostic procedure assembled from established checks (k-NN consistency, paired BC/RL ablation, state ablation, and marginal/conditional d_{TV}) for deciding whether offline data is sufficient before further offline tuning or online adaptation. We apply it as a single-task case study; whether it generalizes to other settings is left to future work.
- **C2.** An empirical case study on selecting among five pre-trained dropout-prediction models on edX clickstream data. Across BC, offline DQN, and CQL at three penalty weights, no learned selector beats the strongest single base model on held-out accuracy despite a 9.7-point per-instance oracle gap. Mechanism checks weigh against three candidate causes: algorithmic conservatism (CQL closes the imitation gap without a deployment gain), hard-label tie-breaking (regret clusters in $[0.089, 0.101]$ across BC/DQN/CQL while oracle agreement spreads over 0.36 to 0.52), and buffer-to-test marginal shift. The remaining candidate, consistent with the diagnostic readout, is local label ambiguity. Richer offline encodings of the same probability vector (disagreement-derived transforms and the full 38-d state) do not measurably improve deployment accuracy over the 5-d probability subspace alone, and the result is insensitive to training-buffer size.

Shared-failure patterns plausibly arise in other offline meta-learning settings that combine pre-trained predictors, including drug-response prediction, content recommendation, and offline hyperparameter selection. Whether the diagnostic procedure transfers to those settings is an open question we hope to test.

2. Related Work

This paper sits at the intersection of offline decision-making, dynamic model selection, and educational outcome prediction. Offline RL (Levine et al., 2020) must cope with support mismatch between the data distribution and the policy desired at deployment, while contextual-bandit and offline-policy-evaluation work in education provide closely related decision framings (Lan & Baraniuk, 2016; Mandel et al., 2014). Conservative or behavior-constrained methods such as CQL, IQL, BCQ, BRAC, and BEAR (Kumar

et al., 2020; Kostrikov et al., 2021; Fujimoto et al., 2019; Wu et al., 2019; Kumar et al., 2019) aim to reduce offline-RL failures caused by distribution shift, extrapolation error, or out-of-distribution action evaluation, but they do not by themselves resolve ambiguity in which action is optimal from the available state. Our contribution is therefore diagnostic. We quantify when the offline state and label construction are too ambiguous for the selectors we evaluate—a paired BC/offline-DQN/CQL family on a shared buffer—to recover oracle headroom, and we identify this as a failure mode that conservatism alone does not resolve.

The per-instance selection problem traces back to algorithm selection (Rice, 1976) and dynamic classifier selection (Cruz et al., 2018; Ko et al., 2008; Cruz et al., 2015), where different models dominate in different regions of feature space. In education, related work spans clickstream-based MOOC dropout prediction (Dalipi et al., 2018; Dass et al., 2021; Taylor et al., 2014; Xing et al., 2016), e-learning dropout prediction with model combinations (Lykourentzou et al., 2009), higher-education dropout prediction with administrative and LMS data (Goren et al., 2024), and meta-learning for student-performance prediction (Casado Hidalgo et al., 2022). We build on that observation but focus on diagnosing why offline adaptive selection fails even when oracle headroom exists. Each diagnostic in our procedure has antecedents elsewhere. The k-NN label consistency is closely related to neighborhood-purity measures used in dynamic classifier selection (Cruz et al., 2018; Ko et al., 2008), paired BC-versus-offline-RL ablations are a routine offline-RL diagnostic pattern, and feature ablations are standard ML practice. The contribution is the combined empirical application on this task, including the conservatism check against CQL (Kumar et al., 2020) and the conditional buffer-to-test shift estimate. The setup also resembles algorithm selection from ex post oracle-labeled buffers more than a full logged-bandit benchmark with historical action propensities (Dudík et al., 2011; Swaminathan & Joachims, 2015).[1]

The paper also contributes to evaluation protocol. Offline-learning benchmarks often report final policy quality without first quantifying whether the dataset contains actionable oracle headroom, whether simple local predictors can recover the relevant action labels, or whether distribution shift is large enough to dominate the result. This protocol turns those hidden assumptions into explicit measurements.

3. Problem Formulation

We formalize dynamic model selection as a contextual bandit. The oracle-label distribution is induced explicitly by the buffer construction. This formulation is sufficient for the one-step decision problem studied here and lets us separate algorithmic failure from distributional failure later in the protocol.

Decision task. We are given a pool of pre-trained base classifiers $\mathcal{M} = \{m_1, \ldots, m_K\}$. Here K=5 (logistic regression, random forest, gradient boosting, calibrated random forest, and a stacking ensemble). For each student-course pair, a 14-day observation window is summarized as a state $s \in \mathcal{S} \subseteq \mathbb{R}^d$, and the task is to select a single model $a \in \mathcal{A} = \{1, \ldots, K\}$ whose prediction for that sample is used at deployment.

State and reward. The state vector concatenates 28 engineered behavioral features (volume, consistency, temporal patterns, trends; see Section 5) with a binary modality flag, the K-dimensional vector of base-model dropout probabilities on that sample, their mean, and a 3-way one-hot bin derived from that mean, yielding the 38-dimensional state used in the main experiments. The deployment metric is the selected model's zero-one correctness on the 14-day prediction label, $R_{\text{eval}}(s, a) = \mathbb{I}[m_a(s) \text{ predicts the true label}]$. The offline DQN in the body is trained with the canonical oracle-match reward $R(s, a) = R_{\text{eval}}(s, a)$ on the buffer. The log-probability shaping variant used in earlier drafts is preserved as a sensitivity row in Appendix C. Throughout, *oracle agreement* is a label-imitation diagnostic against a single argmax action, *test accuracy* is the deployment-facing metric, and *regret* (Section 6.4) is the fraction of test samples on which the policy selects an incorrect model when at least one correct model exists.

Contextual-bandit reduction. Because each sample's observation window is fixed and does not evolve under the agent's choice, within-sample state transitions are degenerate. We set the discount factor γ=0, reducing the MDP to a contextual bandit (Lan & Baraniuk, 2016). The Deep Q-Network used in Section 4 retains its full architecture but learns only the immediate action-value $Q(s, a)$, with no bootstrapping term.

Offline buffer and buffer oracle distribution. The buffer $\mathcal{B} = \{(s_i, a_i^\star, r_i)\}_{i=1}^N$ is built by 4-fold cross-validation on the training split. For each fold, base models are fit on the other three folds and scored on the held-out fold. The *oracle action* is the hard label

$$a_i^\star \;=\; \arg\max_{a \in \mathcal{A}} \left[y_i \, p_a(s_i) + (1 - y_i)\,(1 - p_a(s_i)) \right], \quad (1)$$

that is, the base model assigning the highest probability to the true class. The induced distribution $\pi_\beta(a \mid s)$ is the categorical distribution over these oracle labels. It is reconstructed from cross-validation, so it has no observed action propensities and should not be read as a historical

[1] We therefore treat offline RL here as one comparator family among several, not as the sole framing.

logging policy. Every offline method in this paper (held-out 10-NN selection, behavioral cloning, offline DQN, and the state-ablation variants) sees $\mathcal{B}$ and nothing else.

Evaluation oracle and buffer-to-test shift. At test time we compute the *evaluation oracle* $\pi^\star(a \mid s)$ by the same argmax-over-base-models construction on the held-out 200-sample test set. The quantity the offline agent must close is the total-variation distance between the marginal action distributions:

$$d_{\mathrm{TV}}(\pi_\beta, \pi^\star) = \tfrac{1}{2} \sum_{a \in \mathcal{A}} \left| \mathbb{E}_{s \sim \mathcal{B}}\, \pi_\beta(a \mid s) - \mathbb{E}_{s \sim \mathcal{D}_{\text{test}}}\, \pi^\star(a \mid s) \right|. \tag{2}$$

Section 6 reports this quantity per window configuration and shows it is substantial. The CV-induced buffer concentrates probability mass on actions that are not dominant at test time, so an offline agent whose action distribution tracks π_β will, by construction, place mass on actions that $\pi^\star$ does not, with the gap lower-bounded by $d_{\mathrm{TV}}(\pi_\beta, \pi^\star)$. We treat $d_{\mathrm{TV}}(\pi_\beta, \pi^\star)$ as a buffer-to-test oracle-shift diagnostic, since it is not a full logged-policy support estimate.

Off-policy evaluation framing. The setup doubles as an off-policy evaluation problem, with π_β the behavior policy and $\pi^\star$ the target. Our test-accuracy and regret reports (Section 6.4) provide a direct deployment-value estimate for the selectors we evaluate, in lieu of importance-weighted OPE. Importance weighting (Dudík et al., 2011; Swaminathan & Joachims, 2015) would require action propensities, which the CV reconstruction does not provide. Section 7 returns to why conservative offline-RL methods do not close the failure mode our diagnosis uncovers.

4. Three-Stage Diagnostic Protocol

The protocol answers three questions in sequence. (1) Are the oracle's selections locally consistent in the current state representation? (2) If so, do a supervised and a reinforcement-learning approach succeed or fail together (pointing to representation or distribution) or differently (pointing to algorithm)? (3) Which state components carry signal beyond what the base-model probabilities already encode? Each stage produces a compact diagnostic. A reader who runs Stage 1 alone can already tell whether the state is locally ambiguous before investing in heavier learners.

4.1. Stage 1: Local label consistency via k-nearest neighbors

If neighboring states in the buffer disagree on the oracle action, no local selector can rely on a stable neighborhood rule. We measure this ambiguity directly. For each buffer sample $(s_i, a_i^\star) \in \mathcal{B}$, we locate its k nearest neighbors in state space under the Euclidean metric on the standardized d-dimensional state and compute the *consistency*

$$c_i \;=\; \tfrac{1}{k} \sum_{j \in \mathcal{N}_k(s_i)} \mathbb{I}\big[a_j^\star = a_i^\star\big]\,, \tag{3}$$

then average $\bar{c} = \mathbb{E}_i[c_i]$ over the buffer. We use k=10 throughout. Smaller k introduces sampling noise; larger k washes out local structure. The quantity $\bar{c}$ is a diagnostic, since it is not a formal upper bound on downstream BC or DQN performance. Low values indicate that nearby states often map to different oracle actions, so any local state-conditioned selector must resolve substantial ambiguity. To calibrate this diagnostic against an actual predictor, we also evaluate a held-out 10-NN selector that predicts the oracle action of each test sample from its nearest training-buffer neighbors in the same standardized state space. We report its oracle agreement and test accuracy alongside $\bar{c}$.

4.2. Stage 2: Algorithm versus representation via paired ablation

A low local-consistency diagnostic does not by itself distinguish algorithmic failure from representational failure. To separate them, we train two policies on the same buffer using architectures that would fail for different reasons if the bottleneck were algorithmic. The first is a supervised behavioral cloner that does not require reward engineering. The second is an offline Q-learner that does. If both fail by similar margins, the bottleneck is more plausibly shared, pointing to representation or distribution.

The behavioral-cloning policy $\pi^{\mathrm{BC}}(a \mid s)$ is a two-layer MLP ($d \to 64 \to K$) trained with cross-entropy loss against the hard oracle labels from the 4-fold cross-validation buffer, for 30 epochs with dropout 0.2 and weight decay 10^{-4}. This is a vanilla supervised multi-class classifier, so any failure here cannot be attributed to reward sparsity, bootstrapping instability, or target-network dynamics.

The Deep Q-Network (Mnih et al., 2013) is a larger MLP ($d \to 128 \to 64 \to K$) trained with the one-step bandit objective (γ=0), Adam, soft target updates with rate τ=0.05, and a replay buffer of the full offline $\mathcal{B}$. We use the canonical $\{0, 1\}$ oracle-match reward $R(s, a) = \mathbb{I}[m_a(s)$ predicts the true label$]$, which matches the deployment metric and avoids the policy collapse that arises under log-probability shaping (see Appendix C for the reward-sensitivity analysis). The network is trained for 50 epochs on the same buffer used by BC.

To control for offline-RL conservatism, we also train a Conservative Q-Learning (CQL) (Kumar et al., 2020) variant on the same buffer, network, and reward. CQL adds the standard penalty $\alpha\, \mathbb{E}_s[\log \sum_{a'} \exp Q(s, a') - Q(s, a^\star)]$ to the Bellman loss, suppressing Q-values for actions outside the buffer's support. We sweep $\alpha \in \{0.1, 1.0, 5.0\}$. We

anchor the second term on the hard buffer-oracle action $a^\star$ instead of $\mathbb{E}_{a\sim\hat{\pi}_\beta(a|s)}[Q(s,a)]$; under the deterministic CV-induced label distribution this collapses to the standard form and avoids estimating a stochastic behavior policy. Because $\gamma{=}0$, the conservatism term is the only thing distinguishing CQL from DQN here, which isolates the contribution of pessimism from any bootstrapping effect.

For each policy we report *oracle agreement* (the fraction of test samples on which the policy's selected action matches the hard evaluation-oracle label) and *test accuracy* (the accuracy of the selected base model's prediction on those samples). The first measures how well the policy imitates the oracle; the second measures whether imitation translates into useful predictions. A low oracle agreement combined with near-reference accuracy is consistent with a policy defaulting to a strong single model regardless of s.

4.3. Stage 3: Feature marginal value via state ablations

If the 28 engineered behavioral features carry unique signal for model selection (information not already encoded in the 5 base-model probabilities) *that the meta-learner can exploit*, we would expect removing them to reduce held-out accuracy. We construct a second pair of policies (BC and a probabilities-only MLP baseline) whose input is the 5-dimensional probability vector alone, training them on the same buffer with matched optimizer settings. The difference in test accuracy between full-state ($d{=}38$) and probabilities-only ($d{=}5$) variants isolates the marginal contribution of the engineered features to the meta-learning task, separately from their well-established contribution to the *base-model* prediction task. (The base models still use the full 28 behavioral features during their own training.) A null difference is consistent with the 28 features being redundant for model selection under our learners and sample size, despite their established value for the base-model prediction task; it does not, on its own, rule out an information-bearing signal that a different architecture or larger buffer could exploit.

We then augment the 38-dimensional state with 13 disagreement-derived transforms of the same base-model probability vector. These include the probability standard deviation, all pairwise absolute differences, the predictive entropy of the mean probability, and the top-1/top-2 margin. They serve as a representation-bias test. They are also deterministic transforms of quantities already in the state and so do not add independent information; a failure to improve held-out accuracy is evidence against this particular tweak, but not against any future augmentation.

5. Experimental Setup

This section specifies the dataset, sampling, features, base models, and buffer construction underlying every experiment in Section 6. Each choice is reproducible from the released configuration. The only source of variance across runs is the random seed controlling sample draws, model initialization, and the stochastic cross-validation partition.

5.1. Dataset and inclusion criteria

We use edX clickstream data from 52 offerings of the same introductory computer science course, covering 223,505 student-course pairs and 84.5 million events. We retain learners with at least 10 events, at least 3 active days, and at least 28 days of activity span, yielding 34,303 qualifying pairs from 21,990 students. For each experimental configuration we draw a stratified sample of 1,000 student-course pairs, so that all 16 window configurations and 5-seed sweeps are evaluated under the same per-configuration budget. The resulting diagnostics are estimates under this sampled regime, not full-corpus limits. The corpus mixes 38 MOOC offerings and 9 in-person sections, so the state includes a binary modality flag. Additional cohort and feature details are in Appendix A and Table 3.

5.2. Windows and labels

For each sampled pair we construct an observation window followed by a prediction window. The main configuration is $(14\text{d}, 14\text{d})$. Dropout is defined as no events in the prediction window, giving a 33.7% positive rate. We also evaluate the full 4×4 grid with $T_{\text{obs}}, T_{\text{pred}} \in \{7, 14, 21, 28\}$ days. We center on $(14\text{d}, 14\text{d})$ because shorter horizons are noisier and longer horizons are less actionable for intervention.

5.3. Engineered features

From each observation window we extract 28 aggregate behavioral features covering activity volume, content ratios, consistency, temporal placement, trends, and concentration of activity. All features are standardized for the linear components of the pipeline. The feature taxonomy and examples are in Table 3.

5.4. Base models

We train five base classifiers on the full training set using all 28 engineered features plus the binary modality flag. **LR** is L2-regularized logistic regression. **RF** is a random forest (400 trees, depth 16, min-split 2). **GB** is gradient boosting (300 trees, lr 0.05, depth 3). **Calibrated RF** is an isotonic-CV calibrated random forest. **Stacking** (Wolpert, 1992) is a logistic-regression meta-learner over LR, RF, and GB. Each model outputs a scalar dropout probability $p_i(s) \in [0, 1]$. The meta-learning action set $\mathcal{A} = \{1, \ldots, 5\}$ corresponds to these five models.

5.5. Buffer construction and evaluation protocol

We split each 1,000-pair sample into stratified train (800) and test (200) partitions. The offline buffer is built from 4-fold cross-validation on the training split, yielding out-of-fold base-model probabilities and hard oracle actions without leakage. The main state has 38 dimensions: 28 behavioral features, 1 modality flag, 5 base-model probabilities, 1 mean probability, and a 3-way one-hot risk bin. The disagreement augmentation adds 13 deterministic transforms of the probability vector, yielding a 51-dimensional augmented state. When we report the *best static* or *strongest single-model* reference, it is the single base model with the highest accuracy on that held-out test split. We report it as a hindsight reference, not as a deployable model-selection rule.

All reported numbers are mean ± standard deviation across 5 random seeds, controlling the sample draw, train/test split, and CV partition. We report test accuracy and oracle agreement throughout; Appendix H provides training details and the bootstrap protocol used for paired intervals. The main tables combine three uncertainty sources at once: resampling of examples, resplitting of train and test, and stochastic model training.

A sample-size sensitivity analysis on this main configuration, holding the test partition fixed and subsampling only the training buffer, is reported in Section 6.3. Code, experiment configurations, and aggregated CSV outputs underlying every table and figure are released at https://anonymous.4open.science/r/gtedm-icml26.[2]

6. Results

We first establish the amount of oracle headroom available, then summarize the three-stage diagnostic on the main configuration, and finally quantify how much buffer-to-test oracle shift remains after that diagnosis. Figure 3 gives a compact visual summary of the main case; Appendix B reports the raw sample counts and action marginals behind it.

6.1. Oracle headroom exists across window configurations

Across the 16 window configurations, the oracle improves on the strongest single base model by 4.5 to 15.5 points, with a mean gap of 9.7 points. On the main $(14\text{d}, 14\text{d})$ configuration, the strongest single base model reaches 0.762 ± 0.024 and the oracle reaches 0.825, leaving a 6.3-point gap for an adaptive selector to recover.

Table 6 shows that some other windows offer larger oracle gaps, but they correspond either to very early, noisier prediction settings or to later windows where the downstream intervention is less actionable. The main configuration sits near the middle of the headroom range while preserving a plausible educational intervention horizon.

[2]Anonymized for double-blind review. The link will be replaced with the public repository at camera-ready.

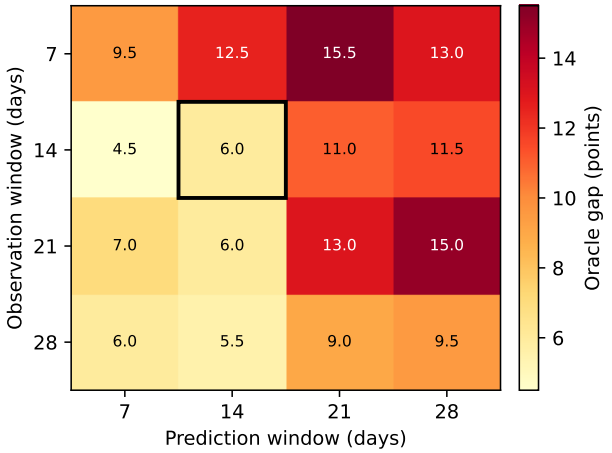

Figure 2. Oracle gap across the 16 observation/prediction window configurations. Values are absolute accuracy gains of the per-instance oracle over the strongest single-model reference on the same test split. The highlighted $(14\text{d}, 14\text{d})$ configuration lies near the middle of the observed oracle-gap range while preserving a plausible intervention horizon. Full window-level selector results are reported in Table 6.

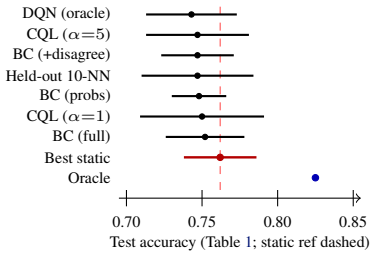

Figure 3. Every learned selector clusters within $[0.74, 0.77]$ near the static reference. The oracle headroom of 0.063 is not recovered. Bars: ± 1 std over five seeds.

6.2. Oracle imitation improves, but not deployment accuracy

The local-consistency diagnostic at $(14\text{d}, 14\text{d})$ is 0.388 ± 0.010. A held-out 10-NN selector in the same standardized state space reaches 0.486 ± 0.041 oracle agreement and 0.747 ± 0.037 test accuracy, which calibrates local recoverability from the current representation. BC, DQN, and CQL cluster between 0.36 and 0.51 oracle agreement, above random $(1/K = 0.20)$ but below the 10-NN ceiling. CQL matches BC's imitation (0.51 at α=1.0) while DQN reaches only 0.36. The gap between them isolates a missing pessimism term, since the supervisory signal is the same.

Table 1. Main $(14\text{d}, 14\text{d})$ results. "Best static ref." is the strongest single base model on that held-out test split, reported as a hindsight reference. The first block uses the 800-train/200-test regime that supports the BC/DQN/CQL mechanism comparison. The bottom block verifies the headline negative claim at N=2,000 test examples with an 8,000-buffer training set (full sweep in Table 2). DQN here uses the canonical $\{0, 1\}$ oracle-match reward; the original log-probability shaping reward yields 0.158 ± 0.025 oracle agreement and 0.749 ± 0.027 test accuracy on the same seeds (Appendix C). CQL adds the standard conservative-Q penalty to the same one-step bandit objective, swept across three penalty weights α. BC-full minus best-static-reference accuracy is -0.010 (95% CI $[-0.019, -0.001]$) in the small-N regime and 0.000 ± 0.001 at N=2,000. Probs-only BC minus BC-full is -0.004 (CI $[-0.017, 0.010]$). Disagreement-augmented BC minus BC-full is -0.005 (CI $[-0.014, 0.004]$).

Method	Oracle agreement	Test acc.
800-train / 200-test, 5 seeds		
Best static ref.	–	0.762 ± 0.024
Oracle	1.000	0.825
Held-out 10-NN	0.486 ± 0.041	0.747 ± 0.037
BC (full state)	0.508 ± 0.041	0.752 ± 0.026
DQN (oracle-match)	0.356 ± 0.038	0.743 ± 0.030
CQL (α=0.1)	0.497 ± 0.054	0.753 ± 0.038
CQL (α=1.0)	0.511 ± 0.033	0.750 ± 0.041
CQL (α=5.0)	0.505 ± 0.044	0.747 ± 0.034
BC (probs only)	0.386 ± 0.039	0.748 ± 0.018
BC (+disagreement)	0.516 ± 0.039	0.747 ± 0.024
8,000-train / 2,000-test verification, 5 seeds		
Best static ref. (hindsight)	–	0.753 ± 0.000
BC (full state)	–	0.753 ± 0.001
BC (probs only)	–	0.747 ± 0.001
DQN (oracle-match)	–	0.738 ± 0.003

The log-probability shaping reward originally used for DQN collapses oracle agreement below random because it is computed relative to LR; Appendix C shows test accuracy is roughly constant across reward variants (0.743 to 0.755).

Yet none of the four learners exceeds the static reference of 0.762 on held-out accuracy. Every method sits within ± 0.01 of 0.748 (Figure 3). The state ablations show the same pattern: probabilities-only BC nearly matches the full state, and disagreement-derived transforms improve oracle agreement slightly but leave test accuracy unchanged. Of the paired intervals in Table 1, only BC-full versus best static excludes zero, and in the wrong direction; both representation comparisons span zero. Richer offline encodings of the same base-model outputs do not measurably improve deployment accuracy here. The 200-example test split is not the cause: the bottom block of Table 1 shows that at N=2,000 held-out examples and an 8,000-example training buffer, BC's test accuracy (0.753 ± 0.001) is statistically indistinguishable from the hindsight static reference (0.753 ± 0.000) and DQN remains below both. Full 16-window results appear in Table 6.

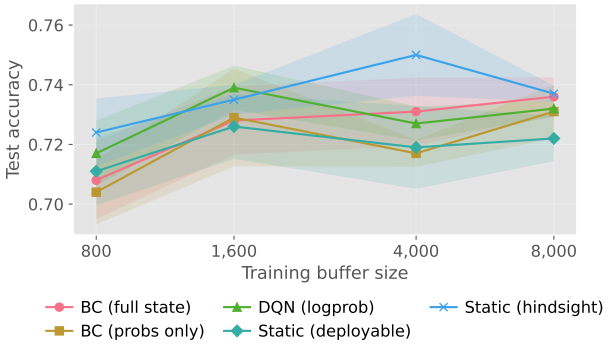

Figure 4. Sample-size sensitivity at $(14\text{d}, 14\text{d})$. Training buffer varied over $\{800, 1{,}600, 4{,}000, 8{,}000\}$ with the 20% test split held fixed across rows. Bands: mean $\pm$ std over five seeds.

6.3. The negative result is stable across windows and buffer sizes

Table 6 shows that the $(14\text{d}, 14\text{d})$ main case is not a one-off failure. Across all 16 observation/prediction windows, BC without disagreement augmentation trails the strongest single base model by 0.7 to 3.8 accuracy points, with a mean deficit of 1.6 points. The disagreement-augmented selector also fails to close the gap: across the same windows it trails that reference by 0.2 to 3.2 points, with a mean deficit of 1.9 points.

The window sweep also clarifies what disagreement features change. Augmentation improves oracle agreement in every window, by 0.4 to 11.9 points with a mean gain of 3.0 points; the corresponding held-out accuracy change averages -0.003 and never exceeds $+0.009$. Deterministic transforms of the probability vector increase oracle agreement without producing a meaningful deployment benefit, which supports the same representational diagnosis.

SAMPLE-SIZE SWEEP

A natural reading of the main-text gap of -0.010 between BC and the strongest single base model is that 200 test examples is too few to distinguish the two. To test that explanation, we hold the test partition fixed at 2,000 examples and vary only the training-buffer size. This addresses both the sample-size concern and the test-stability concern flagged by the diagnostic protocol's own design. Table 2 reports the result.

The negative deployment claim survives at every buffer size. BC test accuracy rises from 0.740 ± 0.004 at 800 training examples to 0.753 ± 0.001 at 8,000, but so does the hindsight static reference, and the gap between them collapses to a mean difference of roughly zero ($+0.000 \pm 0.001$ at 8,000) without crossing into a positive gain. Against the deployable static reference (the base model picked from training-set accuracy), BC actually *leads* by 0.6 to 0.9 points across all

Table 2. Sample-size sensitivity numbers underlying Figure 4. "static-hind." is the strongest single base model on the held-out test split, picked with hindsight. "static-deploy." is the base model selected on training accuracy and evaluated on the same test split, a deployable selector. The held-out test partition contains 2,000 examples and is fixed across rows.

buffer	k-NN full	k-NN probs	BC full acc	BC probs acc	DQN acc	d_{TV}	static-hind.	static-deploy.
800	0.381 ± 0.019	0.500 ± 0.023	0.740 ± 0.004	0.737 ± 0.007	0.731 ± 0.006	0.046 ± 0.018	0.739 ± 0.004	0.734 ± 0.003
1600	0.370 ± 0.025	0.511 ± 0.022	0.748 ± 0.006	0.744 ± 0.005	0.737 ± 0.003	0.045 ± 0.025	0.748 ± 0.003	0.739 ± 0.001
4000	0.349 ± 0.006	0.519 ± 0.007	0.749 ± 0.003	0.745 ± 0.004	0.737 ± 0.003	0.027 ± 0.010	0.752 ± 0.003	0.738 ± 0.006
8000	0.356 ± 0.004	0.540 ± 0.002	0.753 ± 0.001	0.747 ± 0.001	0.738 ± 0.003	0.028 ± 0.010	0.753 ± 0.000	0.744 ± 0.002

sizes. We treat that lead as suggestive because it is small and depends on a deployment story whose details (training-set picking rule, intervention budget, evaluation metric) are specific to this task.

The sweep also clarifies the diagnostic story underneath that result. Local k-NN consistency is roughly flat in the buffer size ($0.381 \rightarrow 0.356$ as the buffer grows by an order of magnitude), consistent with the local-ambiguity diagnosis not being a small-sample artifact. Buffer-to-test oracle shift, by contrast, decreases monotonically ($d_{\mathrm{TV}} = 0.046 \rightarrow 0.028$), consistent with a buffer that better approximates the test-time marginal as it grows. Neither change moves deployment accuracy materially, which mirrors the single-cache analysis in Section 6.5.

6.4. Hard-label tie-breaking does not drive the negative result

The hard-argmax oracle treats samples on which multiple base models are simultaneously correct as if the choice between them mattered. The tie rate (the buffer fraction with ≥ 2 correct base models) is 0.778 ± 0.008. We therefore also report mean per-sample regret: the fraction of test samples on which the policy selects an incorrect model when at least one correct model exists (the oracle attains 0).

Method	Oracle agree.	Mean regret
Best static ref.	–	0.059
BC (full state)	0.508 ± 0.041	0.089 ± 0.016
DQN (oracle-match)	0.356 ± 0.038	0.101 ± 0.009
CQL (α=1.0)	0.511 ± 0.033	0.094 ± 0.012

Two things follow. First, oracle agreement spans 0.36 to 0.52 while regret clusters tightly in $[0.089, 0.101]$, so hard-label agreement overstates how much BC, DQN, and CQL differ on a deployment-relevant metric. Second, the strongest base model still leads in regret (0.059), so the negative result is not a tie-breaking artifact. The high tie rate does qualify Stage 1: with 0.778 of buffer samples having ≥ 2 correct base models, the k-NN consistency of 0.388 partly reflects arg-max noise among near-tied probabilities, and is best read as an upper bound on the signal recoverable by any hard-label local rule.

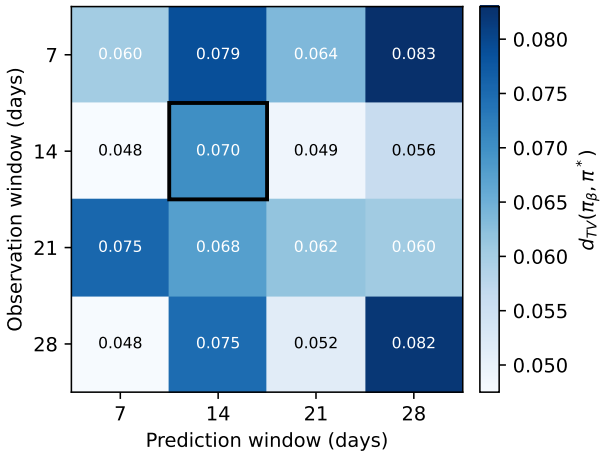

Figure 5. Buffer-to-test marginal $d_{\mathrm{TV}}(\pi_\beta, \pi^\star)$ across the 16 windows. All values are positive but modest. Main configuration: 0.070 ± 0.015. Full table and controlled-classifier analysis in Table 7.

6.5. Buffer-to-test oracle shift is locally non-trivial

Figure 5 shows that the oracle-label distribution induced by the offline buffer is not identical to the test-time oracle. Across all 16 window configurations, $d_{\mathrm{TV}}(\pi_\beta, \pi^\star)$ ranges from 0.048 to 0.083, with a mean of 0.063 ± 0.011. On the main configuration it is 0.070 ± 0.015.

The controlled classifier analysis summarized in Table 7 reaches the same conclusion. On $(14\mathrm{d}, 14\mathrm{d})$, a classifier trained to imitate the buffer oracle is closer to the buffer than to the test oracle by $+0.016 \pm 0.040$, confirming that shift is in the right direction to hurt deployment. A separate robustness check using external stratifiers is reported in Appendix G.

The marginal d_{TV} in Equation (2) averages over the state, so a small marginal can hide substantial pointwise shift. We therefore also estimate the conditional shift $\mathbb{E}_s[d_{\mathrm{TV}}(\pi_\beta(\cdot \mid s), \pi^\star(\cdot \mid s))]$ by approximating each per-state distribution from its k=10 nearest neighbors (buffer-side for π_β, test-side for $\pi^\star$). On the main configuration the conditional d_{TV} averages 0.288 ± 0.024, roughly four times the marginal, with 95th percentile 0.521 ± 0.040 and worst-state 0.860 ± 0.049. Pointwise shift is therefore much larger than the

marginal suggests, with a small tail of states approaching full disagreement.

If local shift were the dominant failure mode, CQL's conservatism term (which explicitly tightens toward π_β) should hurt deployment more than DQN's unregularized objective. It does not. BC, DQN, and CQL differ sharply in oracle imitation (0.36 to 0.51) but converge on test accuracy (0.743 to 0.753, within 0.01 across all three CQL penalty weights). Local support mismatch is substantial and works against deployment, but it is not the main driver of the result. The offline state simply does not identify a deployment-winning action reliably enough, and that limitation is shared across learners regardless of how aggressively they regularize toward the buffer.

7. Discussion: An Offline Diagnostic Procedure

The combined diagnostic procedure points to a locally ambiguous state representation as the primary failure mode. Held-out accuracy is nearly unmoved by disagreement-derived feature transforms, by adding the canonical offline-RL conservatism term, or by enlarging the buffer tenfold. Buffer-to-test oracle shift moves it only marginally, even when the shift is measured per-state.

7.1. Why offline-RL constraints may not close this gap

Conservative offline-RL methods (CQL (Kumar et al., 2020), IQL (Kostrikov et al., 2021), BRAC (Wu et al., 2019), and BEAR (Kumar et al., 2019)) target Q-value overestimation at unseen (s, a) under bootstrap. This task exhibits a different failure mode. Setting $\gamma{=}0$ eliminates bootstrap by construction, and the ambiguity is in the (s, a^*) labelling itself. Similar states disagree on the oracle action 61.2% of the time. Conservatism reweights the decision rule under the existing labels; the labels themselves are unchanged. It therefore cannot recover signal that the $(s, a^\star)$ pairs do not already contain.

Table 1 bears this out. Adding the CQL penalty to the same one-step objective recovers oracle agreement of 0.51 at $\alpha{=}1.0$, comparable to BC and well above DQN's 0.36. Across $\alpha \in \{0.1, 1.0, 5.0\}$, test accuracy moves by less than a point and never reaches the static reference of 0.762. Conservatism closes the imitation gap without moving deployment, even though the conditional shift $\mathbb{E}_s[d_{\mathrm{TV}}] \approx 0.29$ (Section 6.5) is substantial: the gap to the static reference appears to be limited by the representation rather than by distance to π_β.

7.2. Using the diagnostics and benchmark implications

As a pre-deployment checklist, the procedure first asks whether oracle headroom is large enough to justify any adaptive selector. If headroom exists, it asks whether a simple local predictor can recover oracle actions from the current state. When both k-NN consistency and held-out 10-NN accuracy are low, additional offline-RL tuning is unlikely to pay off; the next iteration should focus on state design, relabeling, or data collection. Heavier offline learners or offline-to-online adaptation (Mandel et al., 2014) are worth the cost only when local recoverability is already meaningful. The same four quantities (oracle gap, k-NN consistency, a held-out local selector, and a buffer-to-test oracle-shift measure) make a negative offline-selector result interpretable in benchmark settings.

7.3. Fairness and intervention-cost considerations

The inclusion criterion (≥ 10 events, ≥ 3 active days, ≥ 28-day activity span) excludes roughly 85% of enrolled learners, mostly casual browsers, so our claims are conditioned on the moderately engaged subpopulation. A learned selector also triggers downstream human intervention, so selection errors that systematically favor one base model in one subpopulation may translate into inequitable intervention coverage. An offline meta-learner whose failure modes are unmapped across demographic strata is not yet suitable for intervention-triggering use.

8. Limitations and Conclusion

All experiments draw from one introductory computer science curriculum at one institution; the diagnostics should be recomputed on any new dataset. The inclusion criteria bias the sample toward moderately engaged learners, and the contextual-bandit reduction ($\gamma{=}0$) cannot model within-course state transitions. The offline setting also precludes the online exploration our diagnosis motivates, so we can diagnose offline limits but cannot measure what online adaptation would achieve here.

On this task, the diagnostic procedure points to local label ambiguity as the primary offline bottleneck. Adding conservatism does not close the gap in our experiments, and accounting for buffer-to-test shift does not either. The combined procedure is a lightweight pre-deployment checklist for deciding whether the offline state is informative enough to support model selection: when it flags low consistency and shared BC/DQN/CQL failure, the next iteration is more likely to pay off by changing the data than by changing the learner.

References

Casado Hidalgo, Á., Moreno-Ger, P., and de la Fuente-Valentín, L. Using meta-learning to predict student performance in virtual learning environments. *Applied Intelligence*, 52(3):3352–3365, 2022. doi: 10.1007/s10489-021-02613-x. URL https://doi.org/10.1007/s10489-021-02613-x.

Cruz, R. M. O., Sabourin, R., Cavalcanti, G. D. C., and Ren, T. I. META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern Recognition*, 48 (5):1925–1935, 2015. doi: 10.1016/j.patcog.2014.12.003. URL https://doi.org/10.1016/j.patcog.2014.12.003.

Cruz, R. M. O., Sabourin, R., and Cavalcanti, G. D. C. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018. doi: 10.1016/j.inffus.2017.09.010. URL https://doi.org/10.1016/j.inffus.2017.09.010.

Dalipi, F., Imran, A. S., and Kastrati, Z. MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1007–1014, 2018. doi: 10.1109/EDUCON.2018.8363340. URL https://ieeexplore.ieee.org/document/8363340.

Dass, S., Gary, K., and Cunningham, J. Predicting student dropout in self-paced MOOC course using random forest model. *Information*, 12(11):476, 2021. doi: 10.3390/info12110476. URL https://doi.org/10.3390/info12110476.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104. Omnipress, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Goren, O., Cohen, L., and Rubinstein, A. Early prediction of student dropout in higher education using machine learning models. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, pp. 349–359, 2024. URL https://educationaldatamining.org/edm2024/proceedings/2024.EDM-short-papers.32/index.html.

Ko, A. H.-R., Sabourin, R., and Britto, A. d. S. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008. doi: 10.1016/j.patcog.2007.10.015. URL https://doi.org/10.1016/j.patcog.2007.10.015.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit Q-learning. *arXiv preprint arXiv:2110.06169*, 2021. URL https://arxiv.org/abs/2110.06169.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 11761–11771, 2019. URL https://papers.neurips.cc/paper/2019/hash/c2073ffa77b5357a498057413bb09d3a-Abstract.html.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1179–1191, 2020. URL https://papers.nips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Lan, A. S. and Baraniuk, R. G. A contextual bandits framework for personalized learning action selection. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pp. 424–429, 2016. URL https://educationaldatamining.org/EDM2016/proceedings/paper_63.pdf.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. doi: 10.48550/arXiv.2005.01643. URL https://arxiv.org/abs/2005.01643.

Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., and Wei, Y. Predicting student performance using clickstream data and machine learning. *Education Sciences*, 13(1):17, 2023. doi: 10.3390/educsci13010017. URL https://doi.org/10.3390/educsci13010017.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950–965, 2009. doi: 10.1016/j.compedu.2009.05.010. URL https://doi.org/10.1016/j.compedu.2009.05.010.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popović, Z. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014. URL https://www.ifaamas.org/Proceedings/aamas2014/aamas/p1077.pdf.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. doi: 10.48550/arXiv.1312.5602. URL https://arxiv.org/abs/1312.5602.

Rice, J. R. The algorithm selection problem. *Advances in Computers*, 15:65–118, 1976. doi: 10.1016/S0065-2458(08)60520-3. URL https://doi.org/10.1016/S0065-2458(08)60520-3.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823. PMLR, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Taylor, C., Veeramachaneni, K., and O'Reilly, U.-M. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*, 2014. doi: 10.48550/arXiv.1408.3382. URL https://arxiv.org/abs/1408.3382.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1. URL https://doi.org/10.1016/S0893-6080(05)80023-1.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019. URL https://arxiv.org/abs/1911.11361.

Xing, W., Chen, X., Stein, J., and Marcinkowski, M. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, 58:119–129, 2016. doi: 10.1016/j.chb.2015.12.007. URL https://doi.org/10.1016/j.chb.2015.12.007.

A. Feature and Setup Details

The filtered edX corpus contains 34,303 qualifying student-course pairs from 21,990 students after requiring at least 10 events, at least 3 active days, and at least 28 days of activity span. These filters bias the analysis toward moderately engaged learners, but they ensure that every sample supports the observation/prediction windows evaluated in the main paper.

From each observation window we derive 28 engineered behavioral features in six categories. These complement the modality flag and base-model probability vector used by the selector. Table 3 summarizes the taxonomy used throughout the paper.

Table 3. Feature categories with representative examples. All 28 features enter the selector state as standardized scalars.

Category	Examples	Count
Volume	total_events, active_days	4
Content ratios	video_ratio, problem_ratio	9
Consistency	event_variance, max_gap_days	4
Temporal	first/last-week ratio, days_since_last	4
Trends	engagement_slope, peak_to_avg	2
Advanced	streak_length, decay_rate, Gini	5
Total		**28**

B. Main-Case Diagnostics and Raw Counts

The body's Figure 3 compresses the main $(14\text{d}, 14\text{d})$ test-accuracy comparison into one view. BC, DQN, CQL, and held-out 10-NN all remain near the static reference of 0.762 while the per-instance oracle reaches 0.825. Each seed evaluates exactly 200 held-out examples, so the 0.010 accuracy gap between BC-full and the strongest single base model corresponds to roughly two test examples per seed. That scale is easy to lose in the mean-based summaries reported in the body. The corresponding buffer/test counts and main-case action marginals appear in Table 4.

Table 4. Main $(14\text{d}, 14\text{d})$ action marginals averaged over 5 seeds. Each seed uses an 800-example CV buffer and a 200-example held-out test split. The same regime yields k-NN consistency 0.388 ± 0.010, BC buffer-oracle agreement 0.541 ± 0.012, BC test-oracle agreement 0.508 ± 0.041, $d_{\mathrm{TV}}(\pi_\beta, \pi^\star) = 0.070 \pm 0.015$, and a classifier-to-buffer minus classifier-to-test distance of $+0.016 \pm 0.040$.

Action source	LR	RF	GB	CalRF	Stack
Buffer oracle π_β	0.271	0.182	0.331	0.128	0.088
Test oracle $\pi^\star$	0.265	0.217	0.321	0.111	0.086
Controlled clf π_{clf}	0.343	0.106	0.487	0.052	0.012

C. DQN Action-Marginal Diagnosis

The body Table 1 reports DQN under the canonical $\{0, 1\}$ oracle-match reward, which attains 0.356 ± 0.038 oracle agreement on the $(14\text{d}, 14\text{d})$ configuration. The log-probability shaping reward used in earlier drafts attains only 0.158 ± 0.025 oracle agreement on the same seeds, below the $1/K = 0.20$ uniform-random baseline on $K{=}5$ actions. This appendix explains why, by holding the buffer, base models, CV partition, and DQN architecture fixed and sweeping only the reward shaping function defined in Section 4. Five reward variants are evaluated on the same $(14\text{d}, 14\text{d})$ buffer with five seeds: $\mathrm{logprob}$ (the original variant), $\mathrm{clipped}$ correctness, $\mathrm{oracle\_match}$ (binary correctness on the chosen base model, used in the body), margin, and a hybrid correctness-plus-margin combination.

Table 5. DQN action-marginal sweep on $(14\text{d}, 14\text{d})$, five seeds. "oracle agree" is the fraction of test samples on which the DQN's chosen action equals the test oracle's. "top share" is the largest DQN selected-action marginal. The five right-most columns are the DQN's selected-action marginal on the test set, one column per base model. The "oracle (ref)" row reports the test-set oracle's marginal as a reference target.

reward	oracle agree	test acc	top share	LR	RF	GB	CalRF	Stack
oracle (ref)	1.000	—	—	0.270	0.200	0.321	0.122	0.087
logprob	0.158 ± 0.026	0.749 ± 0.029	0.422 ± 0.103	0.000	0.153	0.239	0.394	0.214
clipped	0.165 ± 0.046	0.749 ± 0.023	0.321 ± 0.052	0.077	0.305	0.147	0.207	0.264
oracle_match	0.356 ± 0.037	0.743 ± 0.033	0.364 ± 0.054	0.232	0.301	0.279	0.107	0.081
margin	0.248 ± 0.040	0.750 ± 0.025	0.561 ± 0.091	0.561	0.093	0.051	0.054	0.241
hybrid	0.349 ± 0.052	0.755 ± 0.031	0.361 ± 0.052	0.329	0.228	0.243	0.081	0.119

Two patterns explain the headline number. First, the $\mathrm{logprob}$ reward used in earlier drafts is computed relative to the LR baseline, $R_{\log}(s, a) = \log p_a^{(y)}(s) - \log p_{\mathrm{LR}}^{(y)}(s)$. On samples where LR is itself the oracle action, the chosen and baseline log-probabilities cancel and the buffer reward for that sample is exactly zero. On samples where another model dominates LR, the reward is positive. The DQN therefore observes that selecting LR is never net-positive in the buffer, and across all $1{,}000$ test predictions in the five-seed sweep it *never selects LR*, despite the test oracle picking LR 270 times. The LR-marginal cell in Table 5 for the $\mathrm{logprob}$ row is exactly 0.000. The 0.158 logprob oracle agreement is dominated by these unrecoverable LR samples.

Second, the failure is reward-specific. The $\mathrm{oracle\_match}$ reward (binary correctness, no LR-relative shaping) recovers an LR marginal of 0.232, close to the oracle's 0.270, and lifts oracle agreement from 0.158 ± 0.026 to 0.356 ± 0.037. The hybrid variant reaches 0.349 ± 0.052. Importantly, deployment-relevant test accuracy is roughly constant across reward variants. The five rows span 0.743 to 0.755, all within their mutual confidence intervals and all below the strongest single base model reference of 0.762 ± 0.024. The choice of reward shaping rearranges which samples the DQN gets right at the action level without materially shifting the deployment metric.

This decomposition tightens the Stage-2 reading in Section 6.2. The original logic ("BC and DQN fail by similar margins, indicating a shared bottleneck above both algorithms") was strained when oracle agreements differed by 0.508 (BC) vs. 0.158 (logprob DQN). With the reward sweep, the shared-failure claim is supported on the test-accuracy axis, where every reward and every comparator lands near 0.75, below the 0.762 static reference. It is qualified on the oracle-agreement axis, where reward shaping accounts for most of the original BC–DQN gap, and the body's $\mathrm{oracle\_match}$ row at 0.356 ± 0.038 already lifts the comparison above random. The representational diagnosis still holds, and the algorithmic-versus-representational separation is now sharper because the reward-shaping confound is identified.

D. Full 16-Window Aggregated Results

Table 6 gives the full 16-window summary underlying the body claims about local consistency, behavioral cloning, and disagreement augmentation.

Table 6. Full 16-window summary of local k-NN consistency, BC oracle agreement, BC held-out accuracy, and the best single-model reference, with and without disagreement augmentation.

obs	pred	augment	k-NN cons.	BC agree	BC acc	best single
7	7	disagreement	0.395 ± 0.036	0.525 ± 0.040	0.727 ± 0.017	0.739 ± 0.015
7	7	none	0.394 ± 0.035	0.406 ± 0.044	0.730 ± 0.020	0.739 ± 0.015
7	14	disagreement	0.375 ± 0.019	0.439 ± 0.043	0.708 ± 0.020	0.731 ± 0.021
7	14	none	0.374 ± 0.020	0.413 ± 0.043	0.722 ± 0.019	0.731 ± 0.021
7	21	disagreement	0.397 ± 0.021	0.478 ± 0.044	0.744 ± 0.015	0.765 ± 0.012
7	21	none	0.395 ± 0.021	0.472 ± 0.039	0.758 ± 0.008	0.765 ± 0.012
7	28	disagreement	0.442 ± 0.023	0.549 ± 0.029	0.819 ± 0.011	0.836 ± 0.014
7	28	none	0.440 ± 0.022	0.539 ± 0.032	0.822 ± 0.017	0.836 ± 0.014
14	7	disagreement	0.393 ± 0.014	0.475 ± 0.023	0.740 ± 0.020	0.758 ± 0.016
14	7	none	0.392 ± 0.013	0.419 ± 0.048	0.744 ± 0.007	0.758 ± 0.016
14	14	disagreement	0.389 ± 0.010	0.516 ± 0.039	0.747 ± 0.024	0.762 ± 0.024
14	14	none	0.388 ± 0.010	0.508 ± 0.041	0.752 ± 0.026	0.762 ± 0.024
14	21	disagreement	0.377 ± 0.015	0.522 ± 0.033	0.745 ± 0.031	0.747 ± 0.022
14	21	none	0.377 ± 0.015	0.487 ± 0.043	0.736 ± 0.026	0.747 ± 0.022
14	28	disagreement	0.409 ± 0.013	0.532 ± 0.036	0.754 ± 0.022	0.786 ± 0.013
14	28	none	0.409 ± 0.013	0.512 ± 0.031	0.753 ± 0.014	0.786 ± 0.013
21	7	disagreement	0.371 ± 0.013	0.481 ± 0.026	0.720 ± 0.021	0.730 ± 0.018
21	7	none	0.371 ± 0.012	0.430 ± 0.016	0.718 ± 0.018	0.730 ± 0.018
21	14	disagreement	0.365 ± 0.011	0.468 ± 0.039	0.740 ± 0.009	0.755 ± 0.012
21	14	none	0.364 ± 0.010	0.464 ± 0.030	0.747 ± 0.009	0.755 ± 0.012
21	21	disagreement	0.377 ± 0.017	0.499 ± 0.040	0.721 ± 0.022	0.751 ± 0.018
21	21	none	0.375 ± 0.018	0.482 ± 0.044	0.728 ± 0.028	0.751 ± 0.018
21	28	disagreement	0.367 ± 0.012	0.443 ± 0.016	0.713 ± 0.034	0.742 ± 0.004
21	28	none	0.365 ± 0.013	0.412 ± 0.010	0.704 ± 0.020	0.742 ± 0.004
28	7	disagreement	0.390 ± 0.015	0.472 ± 0.029	0.766 ± 0.026	0.788 ± 0.022
28	7	none	0.389 ± 0.015	0.451 ± 0.022	0.777 ± 0.025	0.788 ± 0.022
28	14	disagreement	0.393 ± 0.014	0.484 ± 0.078	0.755 ± 0.023	0.777 ± 0.022
28	14	none	0.392 ± 0.015	0.465 ± 0.056	0.758 ± 0.021	0.777 ± 0.022
28	21	disagreement	0.350 ± 0.013	0.463 ± 0.029	0.727 ± 0.020	0.740 ± 0.008
28	21	none	0.349 ± 0.014	0.448 ± 0.047	0.729 ± 0.023	0.740 ± 0.008
28	28	disagreement	0.383 ± 0.023	0.506 ± 0.049	0.742 ± 0.024	0.757 ± 0.019
28	28	none	0.381 ± 0.022	0.469 ± 0.045	0.736 ± 0.035	0.757 ± 0.019

E. k-NN Representation Robustness

Section 4.1 reports k-NN consistency under standardized Euclidean distance on the full d=38 state. With $n_{\text{buffer}} = 800$ and k=10, that geometry sits near the regime where neighborhood relationships can become metric-fragile. To check whether the local-ambiguity diagnosis depends on the metric or dimensionality, we recompute k-NN consistency $\bar{c}$ and the held-out 10-NN selector under three alternative representations of the same buffer state: the 5-dimensional base-model probability subvector alone (*probs-only*), a PCA projection to ten components fit on the buffer (*pca10*), and Mahalanobis distance using the buffer covariance with a small ridge (*mahalanobis*). All projections fit on the buffer only and are applied to held-out points without leakage.

Variant	$\bar{c}$	10-NN oracle agree.	10-NN test acc.
Full (d=51, Euclidean)	0.387 ± 0.010	0.491 ± 0.043	0.749 ± 0.030
Probs-only (d=5)	0.431 ± 0.020	0.477 ± 0.041	0.762 ± 0.031
PCA-10	0.382 ± 0.010	0.462 ± 0.030	0.750 ± 0.025
Mahalanobis (d=51)	0.430 ± 0.014	0.503 ± 0.030	0.738 ± 0.027

Across all four representations, $\bar{c}$ stays in $[0.38, 0.43]$ and the held-out 10-NN test accuracy stays in $[0.74, 0.76]$. The probability-subspace and Mahalanobis variants raise consistency by roughly five points relative to the full Euclidean baseline, but no representation breaks above 0.5 on either consistency or oracle agreement, and none beats the static reference of 0.762 on test accuracy. The Stage-1 diagnosis (that nearby states frequently disagree on the oracle action) is therefore not a metric artifact. It is robust to dimensionality reduction, subspace selection, and metric choice. These numbers are computed on the augmented d=51 state used in Section 4.3. On the unaugmented d=38 state, the full variant matches the body's headline 0.388 ± 0.010.

F. Distribution-Shift Tables

Table 7 reports the full buffer-to-test oracle-shift summary used for Figure 5, including the controlled classifier comparison to the buffer oracle.

Table 7. Full 16-window buffer-to-test oracle-shift summary. The "clf→buffer" column is the controlled classifier's difference in distance to the buffer oracle versus the test oracle. Positive values mean the classifier remains closer to the buffer than to deployment.

obs	pred	$d_{\mathrm{TV}}(\pi_\beta, \pi^*)$	clf agree	clf→buffer
7	7	0.060 ± 0.018	0.406 ± 0.044	0.028 ± 0.011
7	14	0.079 ± 0.032	0.413 ± 0.043	0.000 ± 0.057
7	21	0.064 ± 0.014	0.472 ± 0.039	0.024 ± 0.048
7	28	0.083 ± 0.013	0.539 ± 0.032	0.031 ± 0.033
14	7	0.048 ± 0.022	0.419 ± 0.048	0.015 ± 0.042
14	14	0.070 ± 0.015	0.508 ± 0.041	0.016 ± 0.040
14	21	0.049 ± 0.021	0.487 ± 0.043	-0.004 ± 0.038
14	28	0.056 ± 0.033	0.512 ± 0.031	0.036 ± 0.042
21	7	0.075 ± 0.028	0.430 ± 0.016	0.038 ± 0.021
21	14	0.068 ± 0.028	0.464 ± 0.030	0.021 ± 0.035
21	21	0.062 ± 0.010	0.482 ± 0.044	0.020 ± 0.030
21	28	0.060 ± 0.011	0.412 ± 0.010	0.004 ± 0.039
28	7	0.048 ± 0.011	0.451 ± 0.022	0.022 ± 0.019
28	14	0.075 ± 0.022	0.465 ± 0.056	0.043 ± 0.030
28	21	0.052 ± 0.026	0.448 ± 0.047	0.021 ± 0.032
28	28	0.082 ± 0.025	0.469 ± 0.045	0.013 ± 0.048

G. External-Stratifier Robustness

The slices below recompute the main specialization analysis using stratifiers that do not depend on the base-model probabilities themselves. The compact summary in Table 8 distills the full exported pairwise analysis into the comparisons most relevant to the workshop submission.

Table 8. Compact summary of the external-stratifier robustness analysis. External stratifiers correlate moderately with each other, but their agreement with the original probability-derived risk-bin construction is much weaker on the main configuration.

Stratifier pair	**Mean over 16 windows**	$(14\mathbf{d}, 14\mathbf{d})$
course_baserate × kmeans_k3	0.693	0.687 ± 0.128
course_baserate × kmeans_k5	0.657	0.654 ± 0.124
kmeans_k3 × kmeans_k5	0.673	0.641 ± 0.309
course_baserate × risk_bin	0.557	0.315 ± 0.090
kmeans_k3 × risk_bin	0.466	0.343 ± 0.373
kmeans_k5 × risk_bin	0.482	0.298 ± 0.143

At $(14\mathrm{d}, 14\mathrm{d})$, the strongest agreement is between the three external stratifiers, not between any external stratifier and the original risk-bin partition. We therefore treat the specialization story as fragile rather than as a main-paper result.

G.1. Full 16-window correlation matrix

Table 9 reports the per-window correlation matrix underlying the aggregate robustness summary.

Table 9. Full 16-window correlation matrix for the external-stratifier robustness analysis. 'cb' denotes course baserate, 'rb' denotes the probability-derived risk bin, and 'k3/k5' denote the two clustering-based stratifiers. The main $(14\text{d}, 14\text{d})$ setting is unusual in that all three correlations involving the original risk-bin partition are low.

obs	pred	cb×k3	cb×k5	cb×rb	k3×k5	k3×rb	k5×rb
7	7	0.607	0.656	0.708	0.606	0.451	0.492
7	14	0.605	0.538	0.519	0.666	0.344	0.534
7	21	0.823	0.767	0.524	0.884	0.517	0.473
7	28	0.543	0.437	0.458	0.739	0.586	0.571
14	7	0.758	0.702	0.711	0.602	0.467	0.666
14	14	0.687	0.654	0.315	0.641	0.343	0.298
14	21	0.691	0.643	0.217	0.746	0.381	0.115
14	28	0.756	0.688	0.283	0.693	0.100	0.187
21	7	0.655	0.689	0.840	0.648	0.499	0.670
21	14	0.711	0.675	0.579	0.642	0.660	0.551
21	21	0.708	0.796	0.475	0.801	0.283	0.442
21	28	0.778	0.743	0.499	0.611	0.466	0.309
28	7	0.604	0.710	0.835	0.628	0.291	0.635
28	14	0.708	0.730	0.755	0.688	0.708	0.708
28	21	0.810	0.545	0.787	0.453	0.721	0.409
28	28	0.640	0.535	0.411	0.715	0.638	0.652

G.2. Main-configuration slice winners

To make that fragility more concrete, Table 10 reports which base model is favored by the oracle within each external slice of the $(14\text{d}, 14\text{d})$ configuration, and which single base model actually attains the best held-out accuracy there.

Table 10. Main-configuration slice winners for the external-stratifier analysis. Several slices change which model looks best once the strata are defined independently of the original probability-derived risk bins, which is why the specialization claim should be treated cautiously.

Stratifier	**Stratum**	**Mean n**	**Top oracle model**	**Best held-out model / acc.**
risk bin	high	28.6	RF	All / 0.698
risk bin	low	103.2	GB	Stack / 0.900
risk bin	mid	68.2	LR	LR / 0.580
course baserate	high	55.4	LR	LR / 0.717
course baserate	low	39.0	LR	Stack / 0.901
course baserate	mid	105.6	GB	RF / 0.747
kmeans k3	c0	68.2	GB	CalRF/Stack / 0.762
kmeans k3	c1	69.4	LR	LR / 0.729
kmeans k3	c2	62.4	GB	Stack / 0.867
kmeans k5	c0	42.0	LR	LR / 0.728
kmeans k5	c1	37.8	GB	Stack / 0.758
kmeans k5	c2	46.4	GB	CalRF / 0.901
kmeans k5	c3	50.6	LR	Stack / 0.666
kmeans k5	c4	23.2	GB	GB / 0.876

H. Training and Uncertainty Details

All main-paper numbers are means and standard deviations over 5 random seeds. Each seed resamples $1{,}000$ student-course pairs, regenerates the stratified $800/200$ train-test split, and rebuilds the 4-fold CV buffer.

The behavioral-cloning model is a two-layer MLP ($d \to 64 \to 5$) trained with hard-label cross-entropy for 30 epochs, batch size 64, dropout 0.2, and weight decay 10^{-4}. The main-paper DQN is a one-step contextual-bandit reduction with $\gamma = 0$, architecture ($d \to 128 \to 64 \to 5$), soft target updates with $\tau = 0.05$, and 50 training epochs using the canonical $\{0, 1\}$ oracle-match reward defined in Section 4. The log-probability shaping variant used in earlier drafts is reported as a reward-sensitivity row in Appendix C. The probabilities-only and disagreement-augmented variants use the same BC objective with the corresponding input state.

Paired uncertainty intervals in the body are bootstrap intervals computed over the 5-seed per-configuration summaries. We use those intervals only for the comparisons that materially affect the paper's claims: BC-full versus best static accuracy,

probabilities-only BC versus BC-full accuracy, and disagreement-augmented BC versus BC-full accuracy.