Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
DOI: https://doi.org/10.51903/jtie.v5i1.503

Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small)

Tong Ye*[1], Jinyi Mu[2], James Hunter[3]
Email: akiraye1999@gmail.com
[1]*Computer Science, Northeastern University, CA, USA*
[2]*Computer Science and Engineering, UCSD, CA, USA*
[3]*Computer Science, University of Colorado Boulder, CO, USA*
*Corresponding Author

Abstract
Dynamic bidding and ranking systems must improve revenue or engagement while avoiding harmful regressions during deployment. This paper presents an end-to-end offline OPE and conservative policy-selection workflow for slot-level contextual bandit approximations of ranking decisions. Using the small Open Bandit Dataset (OBD-small) from ZOZOTOWN (ZOZO, Inc.), each logged row is treated as a context-dependent choice among discrete actions (items), with binary click rewards and logged propensity. This formulation is suitable at the slot level but does not capture full listwise ranking or multi-step offline reinforcement learning. Dynamic bidding and ranking systems must improve revenue or engagement while avoiding harmful regressions during deployment. This paper presents an end-to-end offline OPE and conservative policy-selection workflow for slot-level contextual bandit approximations of ranking decisions. Using the small Open Bandit Dataset (OBD-small) from ZOZOTOWN (ZOZO, Inc.), each logged row is treated as a context-dependent choice among discrete actions (items), with binary click rewards and logged propensity. This formulation is suitable at the slot level but does not capture full listwise ranking or multi-step offline reinforcement learning. Empirically, highly deterministic evaluation policies exhibit extreme variance under sparse clicks, while the logistic reward model remains weak (ROC-AUC ≈ 0.5), limiting DM/DR interpretability. Clipped-DR mixing yields only limited certified improvements: in the women's campaign, gains appear only at moderate confidence (δ=0.10) and for caps up to M=5, whereas stricter or looser settings revert to baseline; in the men's campaign, certification is largely absent. These findings demonstrate that OPE diagnostics and conservative mixing enable reproducible offline selection under uncertainty, but do not indicate deployment-ready improvements.

Keywords: *Off-Policy Evaluation, Contextual Bandits, Safe Policy Improvement, Dynamic Ranking, Dynamic Bidding.*

I. INTRODUCTION

Modern dynamic bidding and dynamic ranking pipelines in advertising and recommendation can be viewed as decision systems that repeatedly map user and item context to an action, such as selecting which item to show, choosing a ranker, or selecting a bid multiplier (Kim Sa Ram et al., 2025; Oktavia et al., 2026; Simon et al., 2026; Siswanto et al., 2024). Many of these decisions are effectively one-step choices during serving time: the system observes a context (user, time, page position, and candidate-set signals) and chooses an action whose reward is observed only for that action (e.g., a click or a conversion). This structure is naturally modeled as a contextual bandit (Li et al., 2010; Dudik et al., 2011).

In practice, organizations cannot freely explore online because exploration can reduce revenue, harm user experience, or violate contractual constraints. As a result, most innovation happens offline using historical logs. Offline learning and evaluation are difficult because logs are biased

Received on February, 2026; Revised on March, 2026; Accepted on April, 2026; Published on April, 2026

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

toward the behavior policy that generated them, and the reward is missing for unchosen actions (Bottou et al., 2013). Off-policy evaluation (OPE) provides tools to estimate how a new policy would perform without deploying it. Importance sampling (IPS) corrects for the logging policy but can be highly variable when the evaluation policy differs from the logging policy or when propensities are small (Swaminathan & Joachims, 2015). The direct method (DM) can be low-variance but biased if the reward model is misspecified, whereas doubly robust (DR) estimators combine both approaches and are consistent when either the propensity model or the reward model is accurate (Dudik et al., 2011; Dudik et al., 2014).

Even with strong OPE estimators, deployment requires risk control. Point estimates can indicate improvement while confidence intervals include large negative values, especially when rewards are sparse or evaluation policies are deterministic. Conservative Policy Iteration (CPI) addresses this by updating policies conservatively. Instead of switching directly to a greedy policy, CPI mixes the current policy with a candidate policy using a step size that can be chosen to guarantee improvement under appropriate bounds (Kakade & Langford, 2002). In offline settings, safe policy improvement methods, including high-confidence policy improvement and baseline bootstrapping approaches (Thomas et al., 2015; Laroche et al., 2019; Petrik et al., 2016), formalize this intuition and provide mechanisms to fall back to a baseline policy when uncertainty is high.

This paper studies a concrete "OPE + CPI" chain for dynamic ranking and bidding in a real logged bandit dataset: the Open Bandit Dataset (OBD) from ZOZOTOWN (Saito et al., 2020, 2021). OBD is distinctive because it contains multiple logs collected on the same platform, using different logging policies (uniform random and Bernoulli-Thompson sampling) and across different campaigns. We use the small-size release (OBD-small) and focus on the men and women campaigns. Each row corresponds to an item displayed at one of three positions and includes the logged propensity for that item, enabling IPS-style OPE. We therefore interpret each row as a contextual bandit round, where the action is the displayed item_id conditioned on the context and position, which directly maps to a ranking decision for a particular slot. This slot-level reduction is appropriate for OBD-small, but the resulting conclusions should be understood as applying to a per-slot contextual bandit approximation rather than to full listwise-ranking optimization or multi-step offline RL. Because each log contains only 10,000 rounds with very low click rates, we use OBD-small primarily as a small-data stress test for uncertainty-aware offline evaluation and conservative selection.

Our goal is to build and evaluate candidate policies under explicit risk constraints. We implement reward modeling, policy construction, and four OPE estimators with bootstrap confidence intervals, then apply CPI-style conservative mixing with a one-sided lower confidence bound on

improvement. In our implementation, the target policy is fixed in advance and the conservative mechanism searches only over the mixing coefficient alpha, so the contribution is best understood as risk-controlled policy selection under uncertainty rather than full iterative policy improvement. We also incorporate a practical variance-control mechanism by capping importance weights within DR, thereby operationalizing conservative behavior in the presence of heavy-tailed weights. The paper reports detailed experiments (nine tables and six figures) with all results computed directly from OBD-small logs. Our main takeaway is that chaining OPE and conservative mixing yields a transparent, reproducible workflow that clarifies when small-data offline gains are supportable and when the data are insufficient to certify improvement.

Operationalizing "dynamic bidding/ordering" as a contextual bandit. Although bidding and ranking are often framed as sequential or listwise decision problems, many production decisions in sponsored search and recommendation can be approximated as per-request choices: given a request context (user features, time, page type) and a set of candidates, the system chooses either (i) which candidate to place in a slot (dynamic ranking) or (ii) a discrete bid multiplier or strategy class (dynamic bidding). Under this view, the action is discrete and the reward is observed only for the selected action. OBD-small matches this structure at the slot level: item_id is the discrete action, position is an observable covariate, and click is the reward. This approximation does not model cross-slot interactions, listwise objectives, or downstream state transitions, which we leave to future work.

Why chaining OPE and CPI matters. Offline policy learning alone does not mitigate deployment risk because a learned policy can exploit modeling errors and concentrate probability mass on regions with weak support. OPE alone does not solve decision-making, because evaluation without a mechanism to choose among candidate policies leaves deployment decisions ad hoc. Chaining OPE and CPI creates a disciplined loop: propose a candidate policy, evaluate it with uncertainty, then choose the largest conservative mixture with a known baseline that satisfies a risk criterion. In this paper, that loop is implemented for a fixed-target policy and a grid over alpha, so it should be read as a conservative selection procedure rather than a full multi-iteration CPI algorithm. This framing is particularly relevant for ranking systems where even small CTR regressions can be costly.

Reproducibility considerations. We fixed the dataset version (OBD-small), the train/evaluation split protocol (time-ordered 70/30), and all hyperparameters and random seeds. All reported numbers in this manuscript were generated from the downloaded CSV logs and deterministic scripts, and the tables and figures are direct outputs of those scripts.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

II. LITERATURE REVIEW

Offline evaluation and conservative policy improvement for ranking and bidding, under a slot-level contextual bandit approximation, sit at the intersection of contextual bandits, counterfactual evaluation, and offline reinforcement learning.

A. Contextual bandits and ranking

Contextual bandits model decision-making in which a learner repeatedly chooses an action based on observed context and receives reward only for the chosen action (Auer et al., 2002). They have been used for personalization and recommendation at scale, including news recommendation with linear contextual bandits (Li et al., 2010). In ranking and recommendation, feedback is typically biased by exposure and presentation effects. Learning-to-rank from implicit or biased feedback, therefore, relies on explicit randomization or propensity correction to achieve unbiased evaluation and learning (Joachims et al., 2017; Wang et al., 2018).

B. Off-policy evaluation

OPE estimates the expected value of an evaluation policy using data generated by a different behavior policy. IPS reweights observed rewards by the ratio of evaluation to behavior probabilities, providing an unbiased estimator when propensities are correct but often exhibiting high variance (Swaminathan & Joachims, 2015). SNIPS (self-normalized IPS) reduces variance and improves empirical stability at the cost of small bias and loss of exact unbiasedness (Swaminathan & Joachims, 2015). DM replaces reweighting with a reward model and can be low variance but biased if the model is misspecified. DR estimators combine IPS and DM, yielding consistency when either the propensity model or the reward model is accurate, and often reducing mean squared error in practice (Dudik et al., 2011; Dudik et al., 2014). Extensions to sequential decision making include doubly robust estimators for reinforcement learning (Jiang & Li, 2016).

C. Confidence intervals and high-confidence evaluation

Deployment decisions require uncertainty quantification rather than point estimates alone. High-confidence off-policy evaluation provides lower confidence bounds on policy value, which can support safe deployment decisions (Thomas et al., 2015; Thomas & Brunskill, 2016). Bootstrap methods are widely used to construct approximate confidence intervals when analytical variance is complex, and they can be applied to OPE estimators to estimate uncertainty (Efron & Tibshirani, 1993; Davison & Hinkley, 1997).

D. Conservative policy improvement and safe offline RL

CPI (Kakade & Langford, 2002) improves policies by taking conservative steps toward a candidate policy, often expressed as a mix of the baseline and candidate policies. Safe policy

improvement research in batch RL includes robust baseline-regret formulations (Petrik et al., 2016) and baseline-bootstrapping approaches, such as SPIBB, which restrict changes to regions supported by data and fall back to a baseline policy elsewhere (Laroche et al., 2019). Parallel work in offline RL emphasizes pessimism and distributional shift, for example through conservative value estimation (Kumar et al., 2020); broader context is summarized by Levine et al. (2020). In the present study, we borrow these ideas only at a reduced contextual-bandit level, not as a full sequential offline RL treatment.

E. Gaps motivating this study

While the literature offers strong theory and large-scale benchmarks, practitioners often need a concrete, end-to-end offline workflow that connects (i) the choice of OPE estimators, (ii) uncertainty quantification, and (iii) conservative policy updates under explicit risk constraints, especially when data are limited. The Open Bandit Dataset and Open Bandit Pipeline were introduced to enable realistic and reproducible OPE research with multiple logging policies on the same platform (Saito et al., 2020, 2021). The small-size release provides a convenient testbed for studying the practical tradeoffs of OPE and safe improvement under sparse rewards. This paper fills that gap by implementing an OPE-to-conservative-mixing chain on OBD-small, providing detailed empirical comparisons across estimators, policies, and risk settings in a deliberately small-data benchmark. The aim is not to validate deployment-ready improvement, but to study how uncertainty-aware offline selection behaves when data are limited.

F. Offline RL connections

In full offline RL, actions influence future states, and conservative methods address distributional shift between the dataset and the learned policy (Kumar et al., 2020; Levine et al., 2020). In contextual bandits, the "state" does not transition, but an analogous support mismatch remains. If a learned policy assigns high probability to actions rarely taken by the logger, IPS-style reweighting becomes unstable. In this sense, offline RL ideas appear here only in reduced form as variance control (weight clipping), pessimistic selection (lower bounds), and constrained updates (mixing with a baseline).

G. Variance control via clipping and normalization

Empirically, production counterfactual evaluation systems frequently apply clipping or truncation to importance weights to bound variance, and self-normalization to stabilize estimators (Swaminathan & Joachims, 2015). While these interventions introduce bias, they are often acceptable when the alternative is unbounded variance and unusable confidence intervals. This

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

tradeoff motivates our clipped-DR CPI design: we use weight capping as an explicit risk constraint, rather than as an undocumented engineering trick.

H. Benchmarking in OBD

The Open Bandit Dataset and Pipeline were explicitly designed to support comparisons of OPE methods under real logging policies and realistic feature distributions (Saito et al., 2020, 2021). In the full pipeline, users can replay the Bernoulli-Thompson sampling policy and compute the evaluation policy distributions. In this paper, we focus on small-size logs and study the behavior of OPE and CPI in a small-data regime, highlighting uncertainty and risk constraints.

I. Safe improvement beyond CPI

Several lines of work propose explicit constraints to ensure that an offline policy does not deviate too much from a baseline. SPIBB uses counts or confidence sets to identify state-action regions where the data are reliable and restricts policy changes elsewhere (Laroche et al., 2019). Robust MDP approaches treat transition or reward models as uncertain and optimize against worst-case realizations, yielding conservative updates by construction (Petrik et al., 2016). In the contextual bandit setting of OBD-small, our clipped-DR CPI can be viewed as a lightweight instance of the same philosophy: it restricts the effective influence of low-support events via weight capping and the size of the policy update via alpha selection.

J. Practical uncertainty estimation

While high-confidence off-policy evaluation provides theoretical guarantees, it often requires concentration inequalities and carefully derived variance bounds that can be conservative in practice (Thomas et al., 2015). Bootstrap intervals are easier to apply across multiple estimators and can be paired with practical variance controls such as self-normalization and clipping (Efron & Tibshirani, 1993; Davison & Hinkley, 1997). The pipeline in this paper follows a practical approach, using bootstrap quantiles of improvement as a risk criterion for conservative updates.

III. RESEARCH METHOD

A. Problem formulation

We consider a contextual bandit with context x, discrete actions $a \in \{0,1,\dots,K-1\}$, and bounded reward $r \in \{0,1\}$. A stochastic policy $\pi(a \mid x)$maps each context to a distribution over actions. The policy value is $V(\pi) = \mathbb{E}_{x\sim\mathcal{D},\, a\sim\pi(\cdot|x)}[r]$. Logged data $\mathcal{D} = \{(x_i, a_i, r_i, b_i)\}_{i=1}^{n}$are collected under a behavior policy with propensity $b_i = b(a_i \mid x_i)$, i.e., the probability that the behavior policy selected the logged action $a_i$under context x_i. In OBD-small, this formulation is applied at the slot level by treating each displayed item-position row as an independent

contextual-bandit decision. Our objective is therefore to evaluate candidate slot-level policies and select conservative mixtures using only $\mathcal{D}$and off-policy confidence intervals, rather than to solve full listwise ranking or multi-step offline RL.

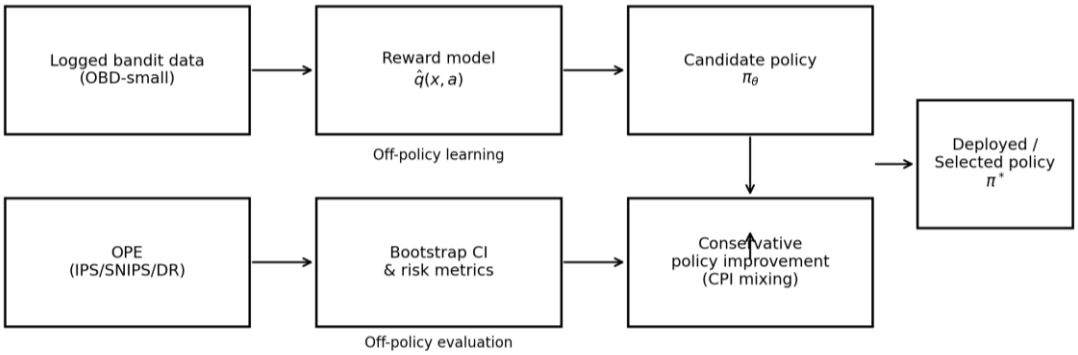

Figure 1. End-to-end Offline Workflow Chaining OPE and Conservative Policy Selection via CPI-Style Mixing

B. Dataset and preprocessing (OBD-small)

We used the small-size Open Bandit Dataset (OBD-small) distributed with the Open Bandit Pipeline repository (Saito et al., 2020). We downloaded four logs: random/men, bts/men, random/women, and bts/women. Each log contains 10,000 rows (Table 1). Each row includes a timestamp, item_id (action), position (1–3), a binary click reward, and the logged propensity score for the displayed item. Context features include hashed categorical user_feature_* fields and numeric user–item_affinity_* features. The men campaign has $K = 34$actions and 34 affinity features; the women campaign has $K = 46$actions and 46 affinity features. We sorted each log by timestamp and applied a time-ordered split: the first 70% of rows for training and the last 30% for evaluation (Table 2). This split mimics a realistic setting where future traffic is evaluated using models trained on past logs. Given the small sample size and low click rates, we use OBD-small primarily as a small-data stress test for OPE uncertainty and conservative selection.

C. Reward model

We trained a logistic regression reward model $\hat{q}(x,a) = \mathbb{P}(r = 1 \mid x,a)$on the training split. The feature vector concatenated (i) one-hot encoded user_feature_* and position, (ii) numeric user–item_affinity_* features, and (iii) a one-hot action indicator for item_id. We used L2 regularization with $C = 1.0$, solver = lbfgs, and max_iter = 200. For each evaluation context, we predicted $\hat{q}(x,a)$for all actions to form a matrix of predicted rewards. This matrix is used for both DM and DR estimators and for defining greedy and softmax policies. Because DM and DR depend directly on $\hat{q}$, we later report held-out reward-model performance and interpret downstream OPE/CPI results in light of model quality.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

1. Baseline: For CPI experiments we used the uniform baseline $\pi_b(a \mid x) = \frac{1}{K}$corresponding to the random logger, because this baseline distribution is known exactly and ensures full support.
2. DM Greedy: We defined a deterministic greedy policy that selects the action with the highest predicted reward, i.e., $a^*(x) = \arg\max_a \hat{q}(x, a)$.
3. Epsilon-greedy: We formed a stochastic policy that chooses the greedy action with probability $1 - \epsilon$and otherwise selects a random action; we set $\epsilon = 0.1$, i.e., $\pi(a \mid x) = (1 - \epsilon)\mathbf{1}[a = a^*(x)] + \epsilon\frac{1}{K}$.
4. Softmax: We defined a stochastic policy $\pi_\tau(a \mid x) = \frac{\exp(\hat{q}(x,a)/\tau)}{\sum_{a'}\exp(\hat{q}(x,a')/\tau)}$with $\tau = 0.05$.
5. IPW learner: We implemented a context-only softmax policy, learned via weighted multinomial logistic regression with SGD. The model predicts actions from context and position and uses sample weights of $\frac{r_i}{b_i} + 0.1$, which corresponds to an inverse-propensity-weighted learning objective used in batch learning from bandit feedback (Swaminathan & Joachims, 2015; Dudík et al., 2011). We fixed all hyperparameters and the random seed (Table 2) to ensure reproducibility.

D. Off-policy evaluation (OPE)

We implemented four OPE estimators (Table 4): IPS, SNIPS, DM, and DR (Dudík et al., 2011). IPS estimates $V(\pi)$via $\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\pi(a_i|x_i)}{b_i}$. SNIPS normalizes by the sum of weights to stabilize variance, i.e., $\hat{V}_{\text{SNIPS}}(\pi) = \frac{\sum_{i=1}^{n} r_i \frac{\pi(a_i|x_i)}{b_i}}{\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{b_i}}$. DM uses the reward model $\hat{q}$to compute the expected reward under π, i.e., $\hat{V}_{\text{DM}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \sum_a \pi(a \mid x_i)\hat{q}(x_i, a)$. DR combines DM with an importance-weighted correction term, i.e.,

$$\hat{V}_{\text{DR}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_a \pi(a \mid x_i)\hat{q}(x_i, a) + \frac{\pi(a_i \mid x_i)}{b_i}(r_i - \hat{q}(x_i, a_i))\right].$$

For each policy and campaign, we computed these estimates on the evaluation split. To quantify uncertainty, we used nonparametric bootstrap: we resampled the evaluation indices with replacement $B = 200$times, recomputed the estimator on each resample, and reported the 2.5th and 97.5th percentiles as a 95% confidence interval (Efron & Tibshirani, 1993; Davison & Hinkley, 1997). We also report one-sided lower bounds on improvements for CPI selection.

Conservative policy improvement (CPI) and risk constraints. CPI updates a baseline policy conservatively by mixing it with a target policy (Kakade & Langford, 2002):

$$\pi_\alpha(a \mid x) = (1-\alpha)\pi_b(a \mid x) + \alpha\pi_{\text{target}}(a \mid x),$$

where $\alpha \in [0,1]$. We used $\pi_b$as the uniform baseline and π_{target}as a fixed epsilon-greedy policy unless otherwise specified. For each αon a grid $\{0.00,0.05,\ldots,1.00\}$, we evaluated the improvement $\Delta V(\alpha) = \hat{V}(\pi_\alpha) - \hat{V}(\pi_b)$using bootstrap resamples. We defined the conservative constraint via a one-sided lower confidence bound: $\text{LCB}_\delta(\alpha)$is the δ-quantile of the bootstrap distribution of $\Delta V(\alpha)$. Given a confidence level δ, CPI selects α^*as the largest αsuch that $\text{LCB}_\delta(\alpha) \geq 0$; if no αsatisfies the constraint, CPI returns $\alpha^* = 0$(no change). Because $\pi_{\text{target}}$is fixed and only $\alpha$is searched, this experiment should be interpreted as conservative selection of a safe mixture under uncertainty rather than as a full iterative policy-improvement algorithm.

E. Clipped doubly robust evaluation inside CPI

In small, logged datasets with sparse rewards, DR can exhibit heavy tails because the correction term can multiply potentially large importance weights $w_i = \frac{\pi(a_i|x_i)}{b_i}$. To control variance and operationalize conservative behavior, we used a clipped-DR estimator in CPI by replacing the importance weight $w_i$with $\min(w_i, M)$, where $M$is a user-specified cap. Weight clipping is a common variance control in counterfactual learning and provides a practical pessimism mechanism in offline evaluation (Swaminathan & Joachims, 2015; Levine et al., 2020). We evaluated CPI under $M \in \{1,2,5,10\}$and confidence levels $\delta \in \{0.10,0.05,0.01\}$.

F. Evaluation metrics

We report three classes of metrics. First, estimated policy value under IPS, SNIPS, DM, and DR. Second, off-policy confidence intervals and one-sided improvement bounds from bootstrap resampling. Third, a diagnostic regret proxy: because the optimal policy value is not identifiable from bandit feedback, we measure this proxy as the gap between the best estimated candidate in our evaluated policy set (under a fixed estimator and risk setting) and each policy's value. This quantity is not a true regret estimate and is used only for comparative diagnostics. Finally, we report the sensitivity of the confidence interval width to the number of evaluation samples and the computational costs of the full pipeline. Algorithmic summary. The implemented pipeline follows six deterministic steps:

1. Load an OBD-small campaign-policy log and sort by timestamp.
2. Split into train (first 70%) and evaluation (last 30%).

3. Fit a logistic reward model on (context, action) pairs from the train split.
4. Construct candidate policies from the reward model (greedy, epsilon-greedy, and softmax) and from weighted action prediction (IPW learner).
5. Compute OPE estimates on the evaluation split using IPS, SNIPS, DM, and DR, and compute 95% bootstrap confidence intervals by resampling the evaluation rows.
6. For conservative selection, define a baseline uniform policy and a fixed target policy, then search over mixing coefficients alpha and select the largest alpha whose one-sided bootstrap improvement bound is nonnegative under clipped-DR.

G. Regret proxy computation

For each campaign and estimator setting, we define a diagnostic regret proxy as the gap between the best estimated policy value among the candidate set and the estimated value of a given policy. Because this proxy uses OPE estimates rather than ground-truth counterfactual rewards, it should be interpreted only as a within-study comparative diagnostic and not as an absolute measure of regret or optimality.

H. Bias-variance tradeoff in clipped-DR

Weight clipping introduces bias because it truncates the contribution of high-weight samples. However, in small logged datasets, this bias can be dominated by variance, and clipping can improve decision quality by tightening confidence bounds. We treat the cap M as an explicit risk-control knob: small M produces tighter but more biased estimates, while large M approaches standard DR and can reintroduce heavy-tailed behavior. Our CPI experiments make this tradeoff explicit by reporting both the selected alpha and the lower confidence bound of the improvement across different caps.

I. Implementation details for reproducibility

All experiments were executed with fixed random seeds for model fitting and bootstrap resampling, and any stochastic policy construction used deterministic seeds. We used deterministic time sorting by the provided timestamp field. For each campaign, action indices correspond exactly to item_id values in the log (0..K-1), enabling direct indexing of qhat matrices and policy distributions. Bootstrap resampling was performed over rows of the evaluation split with replacement. To keep the study reproducible and lightweight, we used a fixed number of bootstrap resamples for each table and figure, as reported in Table 2.

IV. RESULT

A. Overview of experimental settings

Table 1 summarizes the four OBD-small logs used in this study. Random logs have constant propensities by construction, while BTS logs have heavy-tailed propensity distributions (Figure 6). Because CPI requires a baseline distribution defined for all actions, our conservative-mixing experiments use the random logs and treat the uniform policy as baseline. Given that each log contains only 10,000 rounds and the evaluation split is the last 30% by time, we interpret OBD-small here primarily as a small-data stress test for OPE uncertainty rather than as evidence of deployment readiness. We report OPE results on the evaluation split for both men and women campaigns, then report risk-controlled alpha selection outcomes.

Table 1. Dataset Statistics for OBD-Small Logs Used in This Study

Log	n	K	Click rate	pscore mean	pscore min	pscore max
random-men	10000	34	0.0046	0.029412	0.029412	0.029412
bts-men	10000	34	0.0069	0.163594	0.000165	0.725290
random-women	10000	46	0.0046	0.021739	0.021739	0.021739
bts-women	10000	46	0.0046	0.135504	0.000001	0.962800

Table 2. Experimental Setup, Preprocessing, and Hyperparameters

Component	Setting
Train/evaluation split	Time-ordered 70% / 30% within each campaign-policy log
Reward model	Logistic regression (L2), solver=lbfgs, C=1.0, max_iter=200
Context features	One-hot: user_feature_*, position; Numeric: user-item_affinity_*
Candidate policies	DM Greedy, Epsilon-Greedy (ε=0.1), Softmax (τ=0.05), IPW Learner (SGD)
IPW Learner	SGDClassifier log-loss, alpha=1e-4, max_iter=50; sample_weight = click/pscore + 0.1
OPE estimators	IPS, SNIPS, DM, DR (Dudik et al., 2011)
Confidence intervals	Nonparametric bootstrap, 200 resamples for Tables 5-6
Conservative improvement	Clipped-DR CPI mixing with baseline uniform and fixed target policy; weight caps M ∈ {1,2,5,10}; α grid step 0.055
Risk levels	One-sided bootstrap lower confidence bound (LCB) on improvement at δ ∈ {0.10,0.05,0.01}
Random seed	12345 (all randomized procedures)

B. OPE comparison on OBD-small (men)

Table 5 reports estimated policy values on the men campaign. The baseline uniform policy achieved a DR estimate of 0.0061 with a 95% bootstrap CI of [0.0034, 0.0088]. DM Greedy produced a higher model-based DM estimate (0.0085) but its IPS and SNIPS estimates were near zero because it assigns probability one to a single action; overlap with the logged action is infrequent under 34 actions, and the low click rate makes the resulting estimates unstable. DR for DM Greedy had a negative point estimate (-0.0008) with a CI that spanned negative values, highlighting the variance amplification caused by rare matches and large correction terms. In contrast, the softmax policy (tau=0.05) retained full support and produced DR estimates close to baseline with narrower CIs. The IPW learner produced intermediate values, improving over DM Greedy under IPS/SNIPS but not surpassing baseline under DR in this campaign.

C. OPE comparison on OBD-small (women)

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 6 reports corresponding results for the women campaign. Here, DM Greedy and epsilon-greedy produced higher IPS and SNIPS point estimates than the baseline (IPS 0.0153 and 0.0144 versus 0.0060; SNIPS 0.0143 and 0.0135 versus 0.0060). However, their bootstrap confidence intervals were wide: for DM Greedy, the DR CI included negative values and extended to a high upper tail, reflecting sensitivity to heavy-tailed importance-weighted corrections (Figure 2). The softmax policy and the IPW learner were more stable but closer to baseline. These findings show that in sparse-reward, small-data settings, estimator choice and stochasticity of the evaluation policy strongly affect uncertainty; apparent gains should therefore be interpreted cautiously, especially because the reward model used by DM and DR is weak (Table 10).

Table 3. Evaluation Policies Constructed from OBD-Small Logs

Policy	Definition (text)	Purpose
Behavior (Uniform)	π_b(a|x)=1/K	Baseline policy from random logger
DM Greedy	argmax_a q̂(x,a)	Deterministic greedy w.r.t. reward model
Epsilon-Greedy (ε=0.1)	(1-ε)·1[a=a*] + ε/K	Adds exploration support for OPE stability
Softmax (τ=0.05)	softmax(q̂(x,·)/τ)	Stochastic ranking/bidding policy
IPW Learner (SGD)	softmax(f_θ(x))	Offline policy learned by weighted multinomial logistic regression
CPI (mixing)	π_α = (1-α)π_b + α π_target	Risk-controlled mixture selection with fixed target policy

Table 4. Off-Policy Evaluation (OPE) Estimators Implemented in This Study

Estimator	Formula (ASCII)	Notes
IPS	$\hat{V}_{\text{IPS}} = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\pi(a_i \mid x_i)}{b_i}$	Unbiased under correct logging; high variance
SNIPS	$\hat{V}_{\text{SNIPS}} = \frac{\sum_{i=1}^{n} r_i\, w_i}{\sum_{i=1}^{n} w_i}$, where $w_i = \frac{\pi(a_i \mid x_i)}{b_i}$	Lower variance than IPS; biased but consistent
DM	$\hat{V}_{\text{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a} \pi(a \mid x_i)\, \hat{q}(x_i, a)$	Model-based; bias if q̂ is misspecified
DR	$\hat{V}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a} \pi(a \mid x_i)\, \hat{q}(x_i, a) + w_i(r_i - \hat{q}(x_i, a_i))\right]$	Doubly robust; reduces bias/variance tradeoff
Clipped-DR	$\hat{V}_{\text{DR-clip}} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a} \pi(a \mid x_i)\, \hat{q}(x_i, a) + \min(w_i, M)(r_i - \hat{q}(x_i, a_i))\right]$	Conservative variant for risk control

D. Conservative selection via clipped-DR CPI

Table 7 reports results under different risk constraints. For each weight cap M and confidence delta, the procedure selected a mixing coefficient alpha* from the grid and reported the resulting clipped-DR value and the improvement LCB. The women campaign showed limited certified improvements only at moderate risk: at delta=0.10, alpha*=1.00 was selected for caps M=1,2,5, yielding clipped-DR values of 0.0081, 0.0082, and 0.0086, respectively. In these settings, the one-sided improvement bounds LCB_delta were positive but small (0.0002-0.0006), so the certified gains are modest and sensitive to the clipping level. When the cap increased to M=10, or when

stricter confidence levels were used, the procedure reverted to alpha*=0.00, indicating that the data were insufficient to certify improvement beyond the baseline.

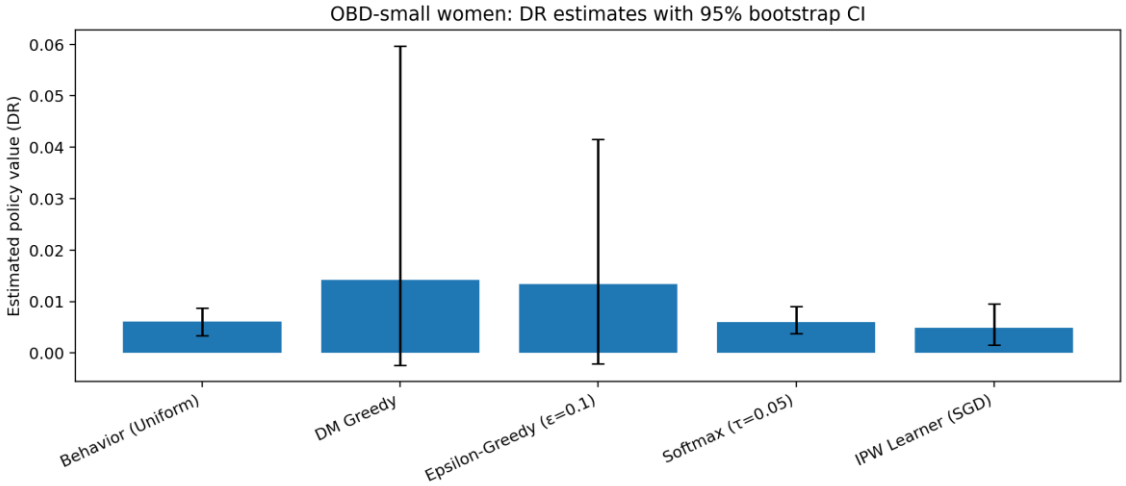

Figure 2. OBD-Small Women Campaign: DR Policy Value Estimates with 95% Bootstrap Confidence Intervals

For the men campaign, certified improvements were even more limited. At delta=0.10 with the strongest clipping (M=1), the procedure selected alpha*=1.00 and achieved a clipped-DR value of 0.0080 with a nonnegative improvement bound. For larger caps or stricter confidence levels, it reverted to alpha*=0.00. Thus, for men, the workflow largely preferred the baseline except under the most aggressive variance control. This asymmetry between campaigns is consistent with differences in action space size, click distribution, and how well the reward model aligns with logged outcomes.

Table 5. OBD-Small Men Campaign: OPE Estimates on the Evaluation Split (Last 30% by Time)

Policy	IPS	SNIPS	DM	DR	DR_CI_low	DR_CI_high	SNIPS_CI_low	SNIPS_CI_high
Behavior (Uniform)	0.0060	0.0060	0.0046	0.0061	0.0034	0.0088	0.0040	0.0093
DM Greedy	0.0000	0.0000	0.0085	-0.0008	-0.0036	0.0014	0.0000	0.0000
Epsilon-Greedy (ε=0.1)	0.0006	0.0006	0.0081	-0.0001	-0.0028	0.0024	0.0003	0.0009
Softmax (τ=0.05)	0.0060	0.0060	0.0048	0.0061	0.0034	0.0087	0.0036	0.0087
IPW Learner (SGD)	0.0027	0.0028	0.0053	0.0033	0.0007	0.0058	0.0011	0.0050

E. Risk-return patterns

Figure 3 summarizes the risk-return tradeoff for women at delta=0.10. As M increases from 1 to 5, the selected clipped-DR value increases modestly because the estimator allows larger corrections. However, increasing M beyond 5 breaks the conservative constraint and forces alpha*=0, producing a sharp transition back to baseline. Figure 4 visualizes alpha* across (M,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

delta) and shows that stricter confidence requirements sharply reduce allowable updates in OBD-small. Overall, the pattern is one of fragile, setting-dependent certification rather than uniformly robust improvement.

Table 6. OBD-Small Women Campaign: OPE Estimates on the Evaluation Split (Last 30% by Time)

Policy	IPS	SNIPS	DM	DR	DR_CI_low	DR_CI_high	SNIPS_CI_low	SNIPS_CI_high
Behavior (Uniform)	0.0060	0.0060	0.0041	0.0061	0.0034	0.0087	0.0037	0.0090
DM Greedy	0.0153	0.0143	0.0082	0.0142	-0.0024	0.0597	0.0000	0.0514
Epsilon-Greedy (ε=0.1)	0.0144	0.0135	0.0078	0.0134	-0.0021	0.0416	0.0004	0.0401
Softmax (τ=0.05)	0.0059	0.0060	0.0041	0.0060	0.0037	0.0090	0.0033	0.0092
IPW Learner (SGD)	0.0046	0.0049	0.0049	0.0049	0.0015	0.0095	0.0014	0.0093

F. Sensitivity to evaluation sample size

Table 8 and Figure 5 report the width of the DR bootstrap CI as a function of the number of evaluation samples. For the baseline policy, CI width decreased with more data. For epsilon-greedy, CI widths remained substantially larger, especially in the women campaign, illustrating that sparse clicks and large correction terms dominate uncertainty. This sensitivity analysis reinforces the interpretation of OBD-small as a small-data stress test: when CI widths remain large at available sample sizes, apparent policy improvements should be treated as exploratory rather than as evidence of deployment readiness.

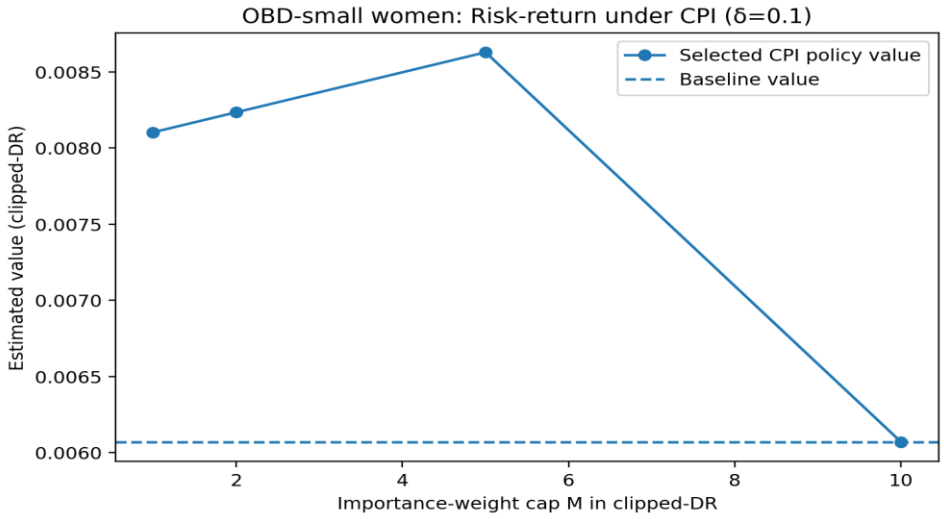

Figure 3. OBD-Small Women Campaign: Risk-Return Pattern for Clipped-DR CPI at δ=0.10

G. Diagnostic regret proxy

We computed a regret proxy within each campaign and risk setting as the gap between the best certified value among our candidate set and the value of each candidate. Under women clipped-DR CPI with M=5 and delta=0.10, the selected CPI policy achieved the highest certified value in our candidate set, yielding a proxy of zero by definition. Under stricter constraints, alpha*=0 increased this within-set gap. Because this quantity is entirely determined by the evaluated candidate set and the chosen estimator, it should be read only as a comparative diagnostic and not as a real regret estimate or a statement about missed deployment value.

Table 7a. Men Campaign: Clipped-DR CPI Results Across Risk Constraints (M and δ)

Campaign	M	δ	α*	Vhat (clipped-DR)	95% CI	LCBδ(ΔV)	P(ΔV<0)
men	1	0.10	1.00	0.0080	[0.0073, 0.0089]	0.0000	0.100
men	1	0.05	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	1	0.01	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	2	0.10	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	2	0.05	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	2	0.01	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	5	0.10	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	5	0.05	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	5	0.01	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	10	0.10	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	10	0.05	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000
men	10	0.01	0.00	0.0061	[0.0036, 0.0093]	0.0000	0.000

Table 7b. Women Campaign: Clipped-DR CPI Results Across Risk Constraints (M and δ)

Campaign	M	δ	α*	Vhat (clipped-DR)	95% CI	LCBδ(ΔV)	P(ΔV<0)
women	1	0.10	1.00	0.0081	[0.0076, 0.0090]	0.0005	0.067
women	1	0.05	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	1	0.01	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	2	0.10	1.00	0.0082	[0.0074, 0.0101]	0.0006	0.057
women	2	0.05	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	2	0.01	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	5	0.10	1.00	0.0086	[0.0067, 0.0134]	0.0002	0.083
women	5	0.05	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	5	0.01	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	10	0.10	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	10	0.05	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000
women	10	0.01	0.00	0.0061	[0.0035, 0.0091]	0.0000	0.000

H. Computational cost

Table 9 reports runtime breakdowns. Reward model fitting took approximately 0.9-1.4 seconds per campaign, full-action prediction took 0.7-0.9 seconds, and bootstrap evaluation (200 resamples) was under 0.02 seconds. These timings demonstrate that the OPE + CPI chain is computationally lightweight on OBD-small and can be integrated into iterative offline development and model selection loops.

I. Reward model quality and its implications

Table 10 reports predictive performance of the logistic reward model on the evaluation split using the probability assigned to the logged action. The ROC-AUC values are 0.496 (men) and 0.511

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

(women). Values close to 0.5 indicate that this simple linear reward model provides little discrimination between clicked and non-clicked impressions in OBD-small.

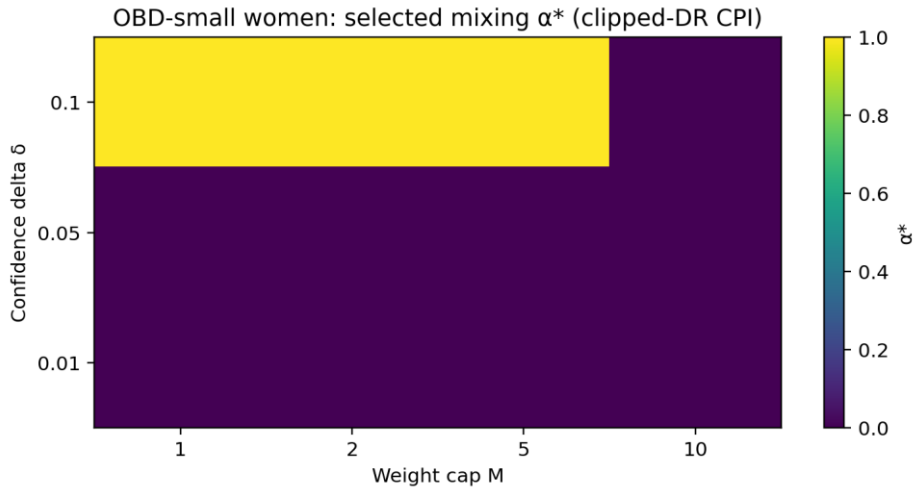

Figure 4. OBD-Small Women Campaign: Selected CPI Mixing Coefficient α* Across Risk Constraint

Table 8. Sensitivity Analysis: DR 95% CI Width Versus Evaluation Sample Size (Time-Ordered Prefix)

n_eval	Women baseline DR CI width	Women ε=0.1 DR CI width	Men baseline DR CI width	Men ε=0.1 DR CI width
500	0.0131	0.0115	0.0102	0.0072
1000	0.0092	0.1308	0.0105	0.0070
2000	0.0061	0.0657	0.0061	0.0057
3000	0.0052	0.0438	0.0052	0.0053

This weakness materially limits the interpretation of downstream DM, DR, and clipped-DR results. DM relies entirely on qhat and therefore may be biased when qhat is misspecified. DR adds an importance-weighted correction term, which can reduce bias when either qhat or the propensity model is accurate (Dudik et al., 2011), but with sparse rewards and variable importance weights it can also amplify noise rather than stabilize estimation. Consequently, the CPI results in Table 7 should be read as the behavior of a conservative selection workflow under a weak reward model, not as evidence that a strong reward predictor has been learned. Weight capping (M) reduces the magnitude of the correction term and is the mechanism that makes conservative selection feasible in this small-data setting.

Table 9. Runtime Breakdown for Major Pipeline Components on OBD-Small random Logs

Campaign	Reward fit (s)	qhat scoring (s)	Bootstrap (s)	Total (s)
men	0.910	0.701	0.009	1.633
women	1.376	0.908	0.009	2.305

J. Diagnostic comparison under a fixed risk setting

Table 11 reports clipped-DR values (M=5) for the candidate policies together with the diagnostic regret proxy. In the women campaign, DM Greedy has the highest clipped-DR point estimate

(0.0089) and its one-sided improvement bound at delta=0.10 is positive, so under this specific estimator-setting pair it is both the highest-valued and certified candidate. In the men campaign, DM Greedy has the highest point estimate (0.0072) but its improvement bound is negative at delta=0.10, meaning that the conservative criterion would still reject deployment. This contrast illustrates the difference between optimistic point estimates and claims that remain supportable after uncertainty quantification.

Table 10. Reward Model Predictive Performance on the Evaluation Split (Logged-Action Prediction)

Campaign	Log loss	ROC-AUC
men	0.0417	0.496
women	0.0384	0.511

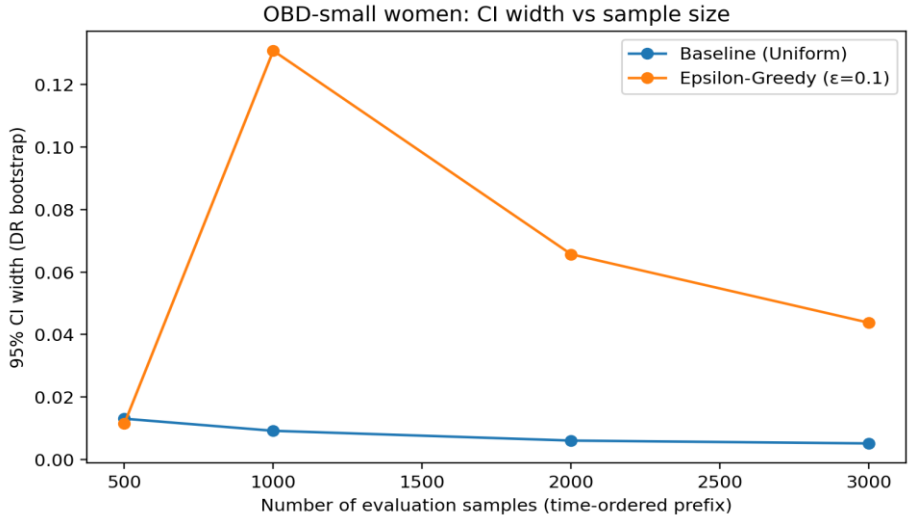

Figure 5. OBD-Small Women Campaign: DR CI Width Versus Evaluation Sample Size

K. Interpreting differences between campaigns

The women campaign has a larger action space (K=46) than the men campaign (K=34), and the BTS logger shows a broader range of propensities (Figure 6). In our experiments, the women evaluation split also exhibited a slightly higher baseline click rate, which increases the effective sample size for estimating improvements. These factors, combined with stochastic policies that maintain support, make it easier to identify settings where conservative bounds permit non-trivial updates.

L. What the certified improvements mean

In this paper, a certified improvement is defined relative to a specific estimator (clipped-DR), a fixed target policy, and a specific uncertainty model (bootstrap quantiles). Therefore, certification should be interpreted as "supported by the chosen estimator and bootstrap risk criterion on OBD-small" rather than as a universal guarantee across estimators, datasets, or deployment settings. In the present benchmark, certified gains are small and fragile, and their absence often leads the

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 178-199
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

procedure to revert to the baseline. This is precisely the behavior one should expect in a small-data stress test with sparse rewards and a weak reward model.

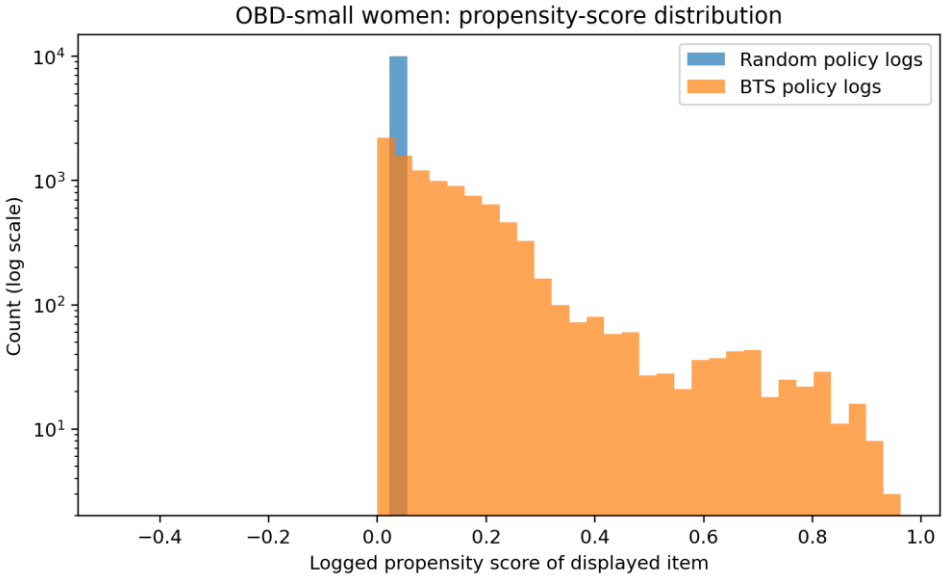

Figure 6. OBD-Small Women Campaign: Propensity Score Distributions for Random and BTS Loggers

Table 11. Diagnostic Regret Proxy Under Clipped-DR (M=5) and δ=0.10. The Proxy is the Gap to the Best Estimated Candidate within Each Campaign and Is Reported Only as a Comparative Diagnostic, Not as a True Regret Estimate

Campaign	Policy	Vhat	CI95	LCB_delta	RegretProxy_estbest
men	DM Greedy	0.0072	[0.0063, 0.0082]	-0.0008	0.0000
men	Epsilon-Greedy (ε=0.1)	0.0069	[0.0061, 0.0079]	-0.0008	0.0002
men	Behavior (Uniform)	0.0061	[0.0037, 0.0087]	0.0000	0.0011
men	Softmax (τ=0.05)	0.0061	[0.0037, 0.0086]	-0.0002	0.0011
men	IPW Learner (SGD)	0.0053	[0.0033, 0.0073]	-0.0023	0.0019
women	DM Greedy	0.0089	[0.0070, 0.0123]	0.0003	0.0000
women	Epsilon-Greedy (ε=0.1)	0.0086	[0.0067, 0.0121]	0.0002	0.0002
women	IPW Learner (SGD)	0.0074	[0.0042, 0.0120]	-0.0008	0.0015
women	Behavior (Uniform)	0.0061	[0.0034, 0.0091]	0.0000	0.0028
women	Softmax (τ=0.05)	0.0060	[0.0033, 0.0090]	-0.0001	0.0028

V. CONCLUSION AND RECOMMENDATION

This study implemented and evaluated an offline OPE + conservative selection workflow for dynamic ranking and bidding decisions under a slot-level contextual-bandit approximation. Using the OBD-small men and women campaigns, we trained reward models, constructed multiple

evaluation policies, compared IPS, SNIPS, DM, and DR estimators with bootstrap confidence intervals, and applied CPI-style conservative mixing by selecting alpha with one-sided improvement bounds. Because the target policy was fixed and only alpha was optimized, the empirical contribution is best interpreted as conservative policy selection under uncertainty rather than as full iterative policy improvement.

The experiments produced three practical findings. First, in small logged datasets with sparse rewards, deterministic evaluation policies can yield highly unstable OPE, with wide confidence intervals and heavy-tailed behavior under DR. Second, the simple logistic reward model was weak (ROC-AUC 0.496 on men and 0.511 on women), which materially limits the meaning of DM-, DR-, and clipped-DR-based downstream conclusions. Third, conservative mixing produced only limited certified gains: on the women campaign, certification occurred only at moderate confidence (delta=0.10) and only for caps up to M=5, while stricter confidence levels or looser caps reverted to the baseline; on the men campaign, the method almost always reverted to the baseline except under the strongest clipping. Accordingly, the study demonstrates a reproducible and informative offline workflow under small-data uncertainty, rather than robust policy improvement in a broad deployment sense.

Recommendations

For practitioners developing ranking or bidding policies offline, we recommend three steps. (1) Report OPE point estimates together with uncertainty, including bootstrap confidence intervals and one-sided lower bounds. (2) When offline evidence is limited, prefer conservative mixtures with a known baseline and select the mixing coefficient using an explicit risk criterion, rather than switching directly to a greedy policy. (3) In small-data regimes, treat variance control (e.g., weight clipping or other pessimistic estimators) and reward-model validation as first-class design choices, and expand data collection or exploration when conservative constraints systematically block improvement. These recommendations translate classical OPE and safe-improvement ideas into a concrete workflow for slot-level contextual bandit problems derived from real-world recommendation logs.

Limitations

This study uses the OBD-small release, which intentionally restricts data size. Each log contains only 10,000 rounds, click rates are very low, and the evaluation split uses only the last 30% of each log, all of which increase uncertainty. The simple logistic reward model achieved near-random discrimination (Table 10), limiting the reliability of DM-, DR-, and clipped-DR-based conclusions. In addition, we modeled each row independently, while real bidding and ranking decisions are often slate-based, interactive, and sequential. The present results therefore apply to

a slot-level contextual-bandit approximation on OBD-small and should be interpreted as a small-data stress test, not as evidence of deployment-ready full-ranking optimization or full offline RL.

Future work

Two directions are immediately actionable: (i) incorporate richer reward models and doubly robust variants designed for variance reduction (e.g., switch-DR or more careful propensity clipping), and (ii) extend the approach from per-slot contextual bandits to slate/ranking OPE and to genuinely sequential offline RL settings where dynamic bidding decisions influence future user state. A further extension is to study conservative policy improvement with iterative target-policy updates rather than selection over a fixed candidate/baseline mixture.

REFERENCES

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, *47*(2), 235–256. https://doi.org/10.1023/a:1013689704352

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013). Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, *14*(101), 3207–3260. http://jmlr.org/papers/v14/bottou13a.html

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press. https://doi.org/10.1017/cbo9780511802843

Dudík, M., Langford, J., & Li, L. (2011). Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 1097–1104. https://arxiv.org/abs/1103.4601

Dudík, M., Langford, J., & Li, L. (2014). Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, *29*(4), 485–511. https://doi.org/10.1214/14-sts485

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593

Jiang, N., & Li, L. (2016). Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 652–661. https://proceedings.mlr.press/v48/jiang16.html

Joachims, T., Swaminathan, A., & de Rijke, M. (2017). Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*, 781–789. https://doi.org/10.1145/3018661.3018704

Kakade, S., & Langford, J. (2002). Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, 267–274. https://doi.org/10.5555/645531.656011

Ram, K. S., Hoon, P. J., & Yeon, H. J. (2025). A Hybrid Noise Reduction And Normalization Framework For Improving Multimodal Sensor Data Quality In Real-Time Systems. *Journal of Technology Informatics and Engineering*, *4*(3), 350-368. https://doi.org/10.51903/jtie.v4i3.440

Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1179–1191. https://proceedings.neurips.cc/paper/2020/hash/0d2b1ed03d5448366938a9d18dfc63a5-abstract.html

Laroche, R., Trichelair, P., & Tachet des Combes, R. (2019). Safe Policy Improvement with Baseline Bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3652–3661. https://proceedings.mlr.press/v97/laroche19a.html

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643*, 1–31. https://arxiv.org/abs/2005.01643

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 661–670. https://doi.org/10.1145/1772690.1772758

Oktavia, D. Z., Hidayat, D. A., Natalia, D., Prabantara, S. K., & Arfriandi, A. (2026). Machine Learning Performance Comparison for Web Application Security Threat Detection: A Systematic Review. *Jurnal Ilmiah Sistem Informasi*, *5*(1), 326-339. https://doi.org/10.51903/dhayjg79

Petrik, M., Chow, Y., & Ghavamzadeh, M. (2016). Safe Policy Improvement by Minimizing Robust Baseline Regret. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2298–2306. https://proceedings.neurips.cc/paper/2016/hash/30018695029e2832a829141f23788a87-abstract.html

Saito, Y., Aihara, S., Matsutani, M., & Narita, Y. (2020). Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *arXiv preprint arXiv:2008.07146*, 1–45. https://arxiv.org/abs/2008.07146

Saito, Y., Aihara, S., Matsutani, M., & Narita, Y. (2021). Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*, 1–14. https://openreview.net/forum?id=99o-9YpWXv

Simon, M., Din, S. M., & Chib, R. J. (2026). A Comparative Study on Self-Organization in Wireless Sensor Networks. *Jurnal of Technology Informatics and Engineering*, *5*(1), 39-53. https://doi.org/10.51903/jtie.v5i1.483

Siswanto, E., Wahyuning, S., Qosidah, N., Huda, H. I., & Asti, P. (2024). Enhancing Employee Engagement through Gamified Digital Platforms: A Case Study Approach in the Technology Sector. *Journal of Management and Informatics*, *3*(3), 531-548. https://doi.org/10.51903/jmi.v3i3.59

Swaminathan, A., & Joachims, T. (2015). Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 814–823. https://proceedings.mlr.press/v37/swaminathan15.html

Swaminathan, A., & Joachims, T. (2015). The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3231–3239. https://proceedings.neurips.cc/paper/2015/hash/39027dfad5102b9d14ce1447a734966e-abstract.html

Thomas, P. S., Theocharous, G., & Ghavamzadeh, M. (2015). High-Confidence Off-Policy Evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 3000–3006. https://doi.org/10.1609/aaai.v29i1.9602

Thomas, P. S., & Brunskill, E. (2016). Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2046–2054. https://proceedings.mlr.press/v48/thomas16.html

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., & Najork, M. (2018). Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, 610–618. https://doi.org/10.1145/3159652.3159733