Position: Offline-Dataset Evaluation for Online Decision-Making Needs an Identification Standard

Anonymous Authors[1]

Abstract

Offline reinforcement learning, offline black-box optimization, contextual bandits, and Bayesian optimization all share a common evaluation challenge: a method's reported performance on a logged offline dataset is treated as evidence about how the method will perform when deployed online. We argue this transfer is observationally equivalent under three distinct mechanisms that current evaluation cannot separate: distribution-shift artifact, dataset-coverage artifact, and genuine policy improvement. The same offline-evaluation score can be rationalized by any combination of these three mechanisms, so a high offline score alone cannot identify which one will generalize to online deployment. We formalize this as a non-identification result, conduct a cross-domain audit of the principal offline-to-online literature documenting where the identification gap is largest in each subfield, propose a four-item identification standard that offline-to-online papers should disclose, and argue that the cross-domain perspective sharpens the standard by showing how the same identification problem manifests differently across offline RL, offline BBO, off-policy evaluation, and Bayesian optimization. The standard is cheap to apply, draws on quantities offline papers already collect, and gives reviewers a common vocabulary for evaluating online-deployment claims.

1. Introduction

Offline decision-making research has become a central program at the intersection of machine learning, optimization, and applied science. Offline reinforcement learning trains policies from logged trajectories without environment interaction (Levine et al., 2020; Kumar et al., 2020; Kostrikov et al., 2022). Offline model-based optimization fits surrogate models to a fixed dataset and proposes designs by optimizing the surrogate (Trabucco et al., 2022). Off-policy evaluation in contextual bandits estimates the expected return of new policies from logged decisions made under a behavior policy (Dudík et al., 2014; Swaminathan & Joachims, 2015). Bayesian optimization on offline initializations decides where to sample next from a prior dataset (Shahriari et al., 2016; Snoek et al., 2012; Frazier, 2018). The shared task across these subfields is to use offline data to inform online decisions.

Each subfield evaluates progress by reporting performance on standard offline-evaluation protocols and concluding that the proposed method is preferable for downstream online deployment. What this evaluation pattern does not establish is that the reported gain identifies a policy improvement that will transfer online rather than something else. A 5 percent improvement in offline-policy value estimate could reflect a genuinely better policy. It could equally reflect a distribution-shift artifact whose direction would reverse under different state visitation. It could reflect a dataset-coverage artifact that would disappear with broader logged behavior. The three explanations are observationally equivalent under the headline statistic. The methodological literature has produced strong fidelity results in many cases. It has not systematically established identification of online improvement.

Position. Offline-to-online decision-making papers should disclose an identification strategy alongside their offline-evaluation results. A reported offline score is not, by itself, evidence of online improvement. It becomes such evidence only when the paper articulates which observationally equivalent alternatives are excluded by the experimental design.

This position is adapted from econometrics (Angrist et al., 1996; Imbens, 2020; Pearl, 2009), where the distinction between prediction and identification is foundational. Our contribution is to specify what an offline-to-online identification disclosure should contain across the diverse subfields the workshop convenes, and to use cross-domain comparison to show that the same identification problem manifests in subtly different ways across offline RL, offline BBO,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

off-policy evaluation, and Bayesian optimization.

Contributions. First, we formalize a three-mechanism observational equivalence result for offline-to-online transfer and show that score-based evidence alone does not separate the mechanisms (Section 2). Second, we conduct a cross-domain audit comparing how the identification gap manifests in offline RL, offline BBO, off-policy evaluation, and Bayesian optimization, and present a landscape view of which mechanism dominates in each (Section 3). Third, we propose a four-item identification standard suitable for inclusion in workshop and conference papers across these subfields (Section 4). Fourth, we discuss limitations and alternative views (Section 6).

2. The Observational Equivalence Result

Let $\hat{V}(\pi; \mathcal{D})$ denote the offline-evaluation score of a candidate policy or proposed design π given offline dataset $\mathcal{D}$ collected under behavior policy β. The standard offline-to-online paper reports the difference $\Delta = \hat{V}(\pi; \mathcal{D}) - \hat{V}(\pi'; \mathcal{D})$ between two methods π and π' on a chosen dataset $\mathcal{D}$. The substitution claim is that this difference identifies the relative online-deployment advantage of π over π'.

We decompose the offline-evaluation score into three additive components:

$$\hat{V}(\pi; \mathcal{D}) = f_{\text{shift}}(\pi, \beta) + g_{\text{cov}}(\pi, \mathcal{D}) + h_{\text{policy}}(\pi) \qquad (1)$$

where $f_{\text{shift}}(\pi, \beta)$ captures distribution-shift artifact arising from the gap between the candidate's induced state distribution and the behavior policy's logged distribution, $g_{\text{cov}}(\pi, \mathcal{D})$ captures dataset-coverage artifact arising from how well $\mathcal{D}$ supports counterfactual evaluation of π's decisions, and $h_{\text{policy}}(\pi)$ is the genuine online-deployment value of interest.

The observed offline score gap is then

$$\Delta = \underbrace{f_{\text{shift}}(\pi, \beta) - f_{\text{shift}}(\pi', \beta)}_{\text{distribution-shift}} + \underbrace{g_{\text{cov}}(\pi, \mathcal{D}) - g_{\text{cov}}(\pi', \mathcal{D})}_{\text{coverage}} + \underbrace{h_{\text{policy}}(\pi) - h_{\text{policy}}(\pi')}_{\text{online improvement}}. \qquad (2)$$

Observational Equivalence Proposition. *For any observed Δ on a single $(\mathcal{D}, \beta)$ pair, multiple combinations of $(f_{shift}, g_{cov}, h_{policy})$ rationalize the same Δ. The reported gap alone does not identify the share attributable to genuine online improvement.*

This is the standard non-identification result for additively separable systems without exclusion restrictions. The substitution claim that $\Delta > 0$ implies π will outperform π' online requires the additional assumption that distribution-shift and coverage artifacts contribute zero, on average, to Δ. This assumption is rarely stated and almost never tested. Conservative offline-RL methods such as CQL and IQL (Kumar et al., 2020; Kostrikov et al., 2022) address one slice of this concern by constraining policies to stay near the data, but conservative behavior is itself a methodological choice that introduces its own coverage assumption rather than dissolving the identification problem. Pessimism-based offline RL (Xie et al., 2021; Jin et al., 2021) similarly addresses one mechanism without identifying the others.

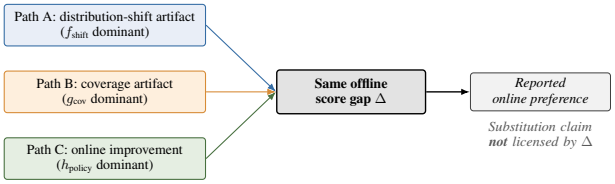

Figure 1. Observational equivalence in offline-to-online evaluation. Three distinct mechanisms can produce the same observed offline score gap Δ. A reported online-deployment preference is licensed only when the paper provides exclusion restrictions that rule out two of the three mechanisms.

3. Cross-Domain Audit: Where Identification Fails Most

Different subfields of offline-to-online research have different dominant identification gaps. Understanding the cross-domain landscape sharpens the proposed standard because it shows which exclusion restriction is most load-bearing in each setting.

3.1. Offline Reinforcement Learning

The dominant gap in offline RL is distribution-shift. Policies trained on a fixed trajectory dataset induce state-visitation distributions that differ from the behavior policy that collected the data, and the offline value estimate over the behavior distribution can diverge sharply from the online return. The literature has responded with conservative methods (CQL), implicit Q-learning (IQL), and pessimism-based regularizers that constrain candidate policies to stay near the data manifold. Cross-comparison studies of offline RL algorithms (Fu et al., 2021) have repeatedly surfaced that headline ordering on D4RL (Fu et al., 2020) is sensitive to behavior-policy regime in ways the headline tables do not communicate. The relative ordering of CQL, IQL, BCQ, and BC variants can shift across medium, medium-replay, and expert variants of the same task.

3.2. Offline Black-Box Optimization

The dominant gap in offline BBO is coverage. The surrogate model is fit on a fixed offline dataset of designs and their evaluations, and the offline-evaluation score of a proposed design is the surrogate's prediction at that design. If the proposed design is in a region of low data support, the sur-

rogate prediction can be high without the true value being high. Design-Bench (Trabucco et al., 2022) surfaces this through tasks where the offline dataset's coverage of the design space is the binding constraint, including superconductor and TF-Bind tasks where the optimum lies outside the training-data envelope. The literature has responded with conservative model-based optimization, normalized maximum likelihood approaches, and explicit constraint formulations that restrict proposed designs to stay near the data, but these responses themselves embed coverage assumptions rather than identifying online improvement.

3.3. Off-Policy Evaluation in Contextual Bandits

The dominant gap in OPE is the joint variance of estimators that share the offline-online reduction. Doubly robust estimation (Dudík et al., 2014) and counterfactual risk minimization (Swaminathan & Joachims, 2015) reduce variance under specific assumptions, but reported policy values can still differ across estimators on the same logged data, indicating that the online value is not identified by the offline-evaluation alone. Open Bandit Pipeline (Saito et al., 2021) provides standardized logged-bandit feedback datasets where comparing OPE estimators on the same data reveals the joint sensitivity to logging policy and estimator class.

3.4. Bayesian Optimization

The dominant gap in offline-warm-started Bayesian optimization is sequential decision interaction: the next-point selection depends on the prior dataset's coverage, the GP kernel choice, and the acquisition function, and the offline-validation score of a BO policy is in general not what the same BO policy will produce online. Reviews of BO practice (Shahriari et al., 2016; Frazier, 2018) document this through the well-known sensitivity of BO trajectories to kernel and acquisition choice in the early-iteration regime where offline initializations dominate the search.

3.5. Cross-Domain Comparison

Figure 2 positions the four subfields in a (coverage gap, distribution-shift gap) plane. Offline RL occupies the high-shift region; offline BBO occupies the high-coverage region; OPE and Bayesian optimization sit at intermediate positions. The implication for the four-item standard is that the relative weight of the four items differs across subfields. An offline RL paper is most usefully challenged on cross-behavior-policy stability. An offline BBO paper is most usefully challenged on cross-coverage stability. An OPE paper is most usefully challenged on within-estimator-class robustness across logging policies. A Bayesian optimization paper is most usefully challenged on sensitivity of acquisition behavior across initialization regimes.

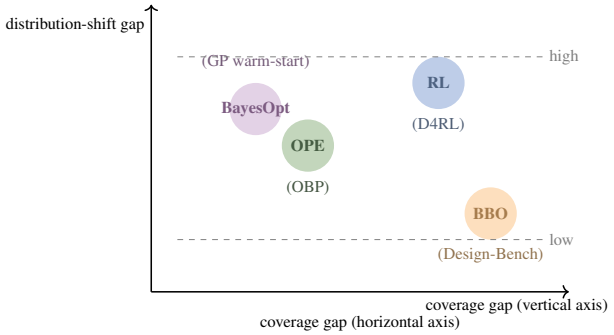

Figure 2. Cross-domain identification landscape. Each subfield occupies a different region of the (coverage gap, distribution-shift gap) plane. Offline RL faces the highest distribution-shift challenge. Offline BBO faces the highest coverage challenge. OPE and warm-started Bayesian optimization sit in intermediate regimes where both gaps matter. The four-item standard adapts to each subfield by emphasizing the exclusion restriction most relevant to its location.

4. Four-Item Identification Standard

We propose that offline-to-online papers report four items alongside their offline-evaluation results.

Item one. Cross-behavior-policy stability. The paper reports the headline contrast on at least two datasets collected under substantively distinct behavior policies (for example, medium-replay versus expert-replay variants of the same D4RL task; or for OPE, logged data from at least two distinct logging policies). The exclusion restriction being imposed is that distribution-shift artifact does not reverse direction across behavior policies. If the relative ordering of methods does reverse, the paper states this explicitly and discusses what the reversal implies about the underlying mechanism.

Item two. Cross-dataset-coverage stability. The paper reports the headline contrast on dataset variants that span at least three substantively distinct coverage regimes (for example, narrow expert data, broad medium data, and random data; or for offline BBO, datasets stratified by initial design diversity). The exclusion restriction is that coverage artifact does not reverse direction across regimes.

Item three. Online-validation evidence. Whenever feasible, the paper reports at least one online-rollout evaluation of the policies whose offline gap it reports, even at small scale. The exclusion restriction is that the offline gap is consistent in sign with the online-rollout gap. When online evaluation is infeasible, the paper explicitly states this and discusses how online behavior should be expected to track or diverge from the offline result.

Item four. Honest gap assessment. The paper reports the across-seed standard error of the offline score and explicitly states whether the headline gap exceeds it. If the gap is

Table 1. Four-item identification standard for offline-to-online decision-making. Each item provides an exclusion restriction that rules out one mechanism as the dominant explanation of the observed offline gap.

Item	Mechanism excluded	Required evidence
Cross-behavior-policy stability	Distribution-shift artifact	Same ordering across ≥ 2 behavior-policy regimes
Cross-coverage stability	Coverage artifact	Same ordering across ≥ 3 coverage regimes
Online-validation evidence	Offline-online divergence	At least one online-rollout consistency check
Honest gap assessment	Noise misattribution	Headline gap $>$ across-seed standard error

comparable to the standard error, the paper either reports more seeds to tighten the estimate or marks the contrast as not yet identified.

These four items are computable from quantities offline-to-online papers already collect. The marginal effort of disclosure is the explicit articulation of which exclusion restrictions are being assumed and which are being tested.

5. Why Offline-to-Online Needs This Now

The case for adopting an identification standard now is structural. Offline-to-online evaluation operates in a small-gap regime where reported method differences are routinely comparable to the noise floor of the evaluation. Reproducibility audits across offline RL benchmarks have repeatedly surfaced that headline results depend on seeds, hyperparameters, and dataset version in ways that headline tables do not communicate. The combination of small effect sizes, dataset-version sensitivity, and online-deployment relevance is exactly the configuration in which observational equivalence becomes a practical rather than a philosophical concern.

The cross-domain perspective also strengthens the case. When the same identification problem manifests across offline RL, offline BBO, OPE, and Bayesian optimization, a domain-specific patch in any one subfield risks missing the deeper issue. A standard that travels across subfields gives the four communities a shared vocabulary for discussing transferable methodological lessons.

Adoption of an identification standard would also strengthen the cumulative character of the field. Papers that report exclusion-restriction tests give downstream researchers and reviewers a basis on which to integrate findings across studies. Papers that report only headline contrasts force subsequent work to repeat or guess at the implicit exclusion structure.

6. Limitations

The proposal is a reporting standard rather than a new technical contribution. The four items prescribe disclosure rather than experimental design; authors who satisfy the items mechanically without engaging the underlying logic will produce uninformative reports. The prescribed thresholds are starting points rather than canonical values; community calibration is an open question. The proposal does not address training-data contamination of LLM-based offline policies, which is a separate identification failure with its own disclosure literature (Ishida et al., 2026). The proposal is also focused on aggregate metrics and does not directly address fairness or subgroup considerations, although the same identification logic applies subgroup by subgroup. Online-validation evidence may be infeasible in some application domains; the disclosure remains valuable in such cases because it forces explicit acknowledgment of the constraint. The cross-domain landscape sketched in Figure 2 is qualitative rather than quantitative; calibrating subfield positions empirically is a useful direction.

7. Alternative Views

Existing offline-evaluation protocols are sufficient. One view is that protocols such as D4RL and Design-Bench have established themselves as adequate proxies for online deployment, and that the field's iteration speed is itself evidence that this signal is informative. We accept that these protocols have driven progress; we argue that the small-gap regime in which the offline-to-online community now operates is precisely where observational equivalence becomes binding, and that past adequacy does not guarantee future adequacy.

Conservative offline methods already address this. A second view is that conservative offline RL and constraint-based offline BBO already address the distribution-shift concern through their methodological design. We agree that conservative behavior is an important methodological response; we argue it is a response within one of the three mechanisms (a coverage-aware modeling choice) rather than a way to identify the contributions of the other two. Conservative methods can themselves benefit from explicit identification disclosure.

Disclosure is unenforceable. A third view is that voluntary disclosure standards are routinely ignored, so the proposal will not change practice. We accept the risk and note that the proposal is structured to be enforceable at the workshop and venue level through reviewer checklists, in the manner of existing reproducibility checklists.

Cross-domain framing is too broad. A fourth view is that subfields are different enough that a unified standard will not fit any of them well. We accept the pluralism objection in part; we argue that the four-item standard is intentionally lightweight and adapts to subfield context (Section 3), and that the cross-domain framing surfaces the shared mechanism that subfield-specific patches risk obscuring.

8. Related Work

The four-item identification standard is structurally analogous to documentation standards elsewhere in machine learning evaluation. Datasheets for datasets (Gebru et al., 2021) prescribe metadata for the training artifacts. Model cards (Mitchell et al., 2019) prescribe disclosure for the resulting models. Audit cards (Staufer et al., 2025) extend this to evaluation pipelines. Benchmark design audits (Reuel et al., 2024) prescribe internal-quality criteria for individual benchmarks. Within the offline RL and offline BBO communities, surveys including Levine et al. (2020) and Trabucco et al. (2022) have emphasized the importance of careful evaluation. Within Bayesian optimization, methodological reviews (Shahriari et al., 2016; Frazier, 2018) have surfaced the sensitivity of reported gains to design choices that the headline numbers do not communicate. No existing instrument addresses the specific question of whether the offline score-gap evidence presented in support of a method actually identifies an online improvement rather than a behavior-policy or coverage interaction. Our proposal complements rather than replaces these existing instruments.

9. Conclusion

Offline-to-online decision-making research has produced methodologically important advances in offline RL, offline BBO, off-policy evaluation, and Bayesian optimization. The evaluation literature on which these advances are assessed has not yet absorbed the lesson that small reported offline gains in a noisy benchmark ecosystem do not, by themselves, identify online improvement. The cross-domain audit shows that the same identification problem manifests in different but recognizable ways across the four subfields. The four-item identification standard is a small intervention in this gap. It asks offline-to-online papers to make explicit what their experimental design assumes and what it tests. The marginal cost is modest. The downstream value is the ability of reviewers, downstream researchers, and practitioners to read a reported contrast as evidence about online deployment behavior rather than as a single observation drawn from an unknown joint distribution of mechanisms.

Disclosure of LLM Use

Large language models were used during manuscript preparation for grammatical revision and citation formatting. All argumentative content, claims, and the position advanced in this paper were authored by the human author, who takes full responsibility for the content.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. In *Statistical Science*, 2014.

Frazier, P. I. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., Paduraru, C., Levine, S., and Paine, T. Benchmarks for deep off-policy evaluation. *International Conference on Learning Representations (ICLR)*, 2021.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Imbens, G. W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179, 2020.

Ishida, T., Lodkaew, T., and Yamane, I. How can I publish my LLM benchmark without giving the true answers away? In *International Conference on Machine Learning (ICML)*, 2026.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning (ICML)*, 2021.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Saito, Y., Aihara, S., Matsutani, M., and Narita, Y. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *NeurIPS Datasets and Benchmarks Track*, 2021.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2016.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Staufer, L., Yang, Y., Reuel, A., and Casper, S. Audit cards: Toward documentation standards for AI evaluation audits. *arXiv preprint*, 2025.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning (ICML)*, 2015.

Trabucco, B., Geng, X., Kumar, A., and Levine, S. Design-Bench: Benchmarks for data-driven offline model-based optimization. In *International Conference on Machine Learning (ICML)*, 2022.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.