Calibrated Pessimism in Offline Reinforcement Learning: Auditing Support-Conditioned Reliability and Control Transfer

Anonymous authors
Paper under double-blind review

Abstract

Offline reinforcement learning often relies on pessimism to reduce the influence of value estimates in weakly supported regions, yet the uncertainty signals used to construct pessimistic penalties are rarely tested as calibrated error surrogates. We study this reliability question directly in a focused five-seed HalfCheetah Medium case study. Using leakage-controlled train/calibration/validation splits, we calibrate global, uncertainty-ratio, and local-support-aware bounds for an observable held-out Bellman-residual proxy, and separately evaluate marginal coverage, support-conditioned reliability, and sharpness. Calibrated constructions attain near-nominal marginal coverage, but this aggregate view hides substantial redistribution of reliability across support regimes: at 90% nominal coverage, the global bound remains relatively narrow while under-covering low-support transitions, whereas adaptive scaling raises low-support coverage at the cost of substantially wider bounds and lower high-support coverage. We then reuse the frozen calibration objects in an evaluation-time candidate-reranking study with matched proposals, no retraining, and no return-based tuning. The reranking interface itself substantially underperforms the deterministic actor, and within this restricted interface we find no evidence that calibrated penalties improve return over unpenalized Q reranking. Our results therefore show, within the scope of this case study, that uncertainty magnitude, marginal statistical calibration, and downstream control utility can diverge substantially. The standard split-conformal guarantee applied here concerns only marginal coverage of the observable residual proxy under exchangeability assumptions; it is not a guarantee for true Bellman error, policy value, or return.

1 Introduction

Offline reinforcement learning (RL) must improve a policy using only a fixed dataset. Because a learned policy may place weight on state–action pairs that are weakly represented in that dataset, a broad family of methods introduces pessimism: uncertain values are reduced, constrained, or otherwise treated more conservatively. Ensemble-based pessimism is one prominent realization of this idea Ghasemipour et al. (2022). The principle is compelling, but it leaves a basic measurement question unresolved: when an offline-RL method assigns a larger pessimistic penalty to a state–action pair, what justifies the numerical magnitude of that penalty?

In practice, uncertainty surrogates such as ensemble disagreement are often used as if a larger score directly implied a proportionally larger estimation error. That interpretation is stronger than what the raw signal guarantees. A disagreement score may rank examples by difficulty without being calibrated to any observable error quantity, and a global multiplier can simultaneously over-penalize some regions and under-penalize others. The issue is especially important in offline RL because the same support imbalance that motivates pessimism also creates heterogeneous reliability: a method can appear well calibrated on average while remaining systematically unreliable where data support is weakest.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study this reliability problem before asking whether calibration improves control. Using a frozen IQL backbone Kostrikov et al. (2022), we treat held-out one-step Bellman residuals as an observable proxy rather than as true Bellman error. We then apply split-conformal-style calibration Angelopoulos & Bates (2023) to map frozen uncertainty or support features to residual-proxy bounds. The resulting objects are evaluated along three separate axes: marginal coverage, support-conditioned reliability, and sharpness. This decomposition is deliberate. Marginal validity can be statistically correct while operationally uninformative about where errors occur; support-conditioned diagnostics can expose that hidden allocation; and neither form of statistical reliability by itself implies that subtracting the resulting bound from a critic will improve a policy.

Recent work has already brought both calibrated uncertainty and residual information into offline RL. Conformal approaches such as UNIQ Upadhyay (2026) use calibrated uncertainty during policy learning to adapt conservatism, while residual-based methods use empirical residuals as algorithmic objects in Bellman-style optimization. Ensemble-based approaches such as MSG Ghasemipour et al. (2022) likewise couple uncertainty to pessimistic policy optimization. We therefore do not claim novelty for conformal calibration, ensemble pessimism, or residual use in offline RL. Our focus is narrower: we freeze the learned backbone, treat held-out residuals as observable diagnostic targets, audit how uncertainty-derived bounds allocate reliability across support regimes, and then test those same frozen objects in a separate evaluation-time control interface.

Across five HalfCheetah Medium seeds, global, uncertainty-ratio, and local-support-aware calibration all achieve near-nominal marginal residual-proxy coverage. Yet the same marginal coverage corresponds to markedly different support-conditioned behavior. At 90% nominal coverage, the global quantile is the sharpest calibrated construction but undercovers low-support transitions, whereas adaptive scales shift coverage toward those transitions at a substantial cost in width and high-support coverage. We then reuse the same frozen calibration objects in evaluation-time candidate reranking with matched proposals. The downstream study does not support a return-improvement claim: none of the calibrated penalties improves on unpenalized Q reranking, and all reranking variants underperform the deterministic actor mean. Within this focused case study, the results show that marginal calibration, support-conditioned reliability, and downstream control behavior should not be treated as interchangeable evidence.

Our contributions are:

- A leakage-controlled reliability audit for uncertainty-based pessimism. We calibrate frozen uncertainty and local-support features against held-out Bellman-residual proxies using disjoint episode splits, while explicitly limiting the statistical interpretation to the observable proxy.
- A support-conditioned diagnosis of marginal calibration. Across five HalfCheetah Medium seeds, near-nominal marginal coverage conceals substantial differences between high- and low-support regions; adaptive scaling changes this allocation while trading off sharpness and high-support coverage.
- A restricted control-transfer test. Reusing the frozen calibration objects in a matched evaluation-time candidate-reranking study, we find no evidence that calibrated residual-proxy penalties outperform unpenalized Q reranking under an interface that itself substantially underperforms the deterministic actor.

Figure 1 summarizes the study design: we first calibrate frozen uncertainty and support features against a held-out residual proxy, and then evaluate the resulting quantities separately as statistical objects and as inputs to a downstream control interface.

2 Related Work

Pessimism and uncertainty in offline RL. Offline-RL algorithms use constraints, conservative objectives, or uncertainty penalties to reduce exploitation of poorly supported actions. Point-wise uncertainty penalties are provably efficient under partial data coverage Jin et al.

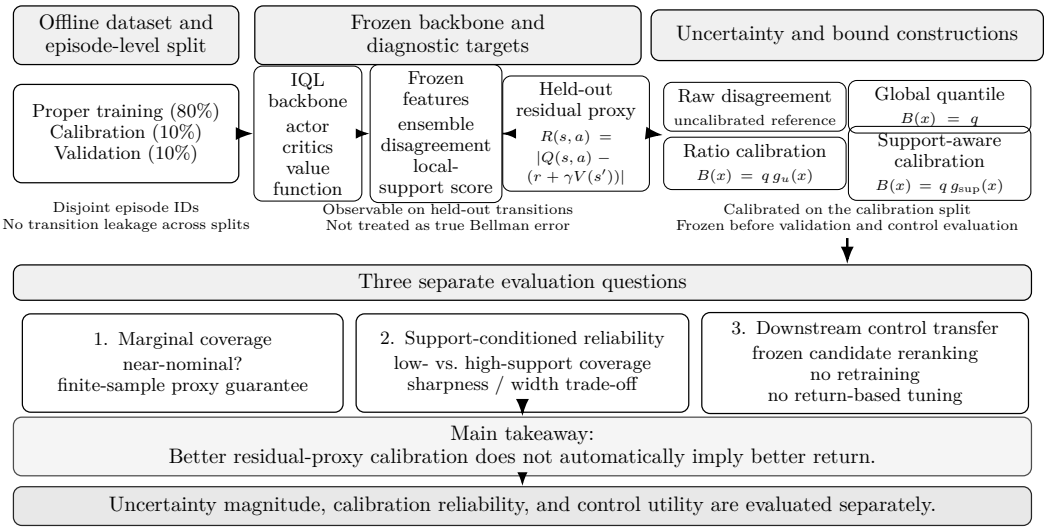


Figure 1: Reliability audit of uncertainty-based pessimism in offline RL. Disjoint episode partitions separate proper training, calibration, and validation. Proper training fixes the IQL backbone, ensemble disagreement, and local-support score; calibration maps the frozen features to residual-proxy bounds; validation evaluates marginal coverage and support-conditioned reliability. The same frozen objects are separately tested in an evaluation-time control interface. The study distinguishes uncertainty magnitude, marginal statistical calibration, and downstream control behavior without treating them as interchangeable guarantees.

(2024), while policy-constraint and conservative methods penalize out-of-distribution value estimates directly Kumar et al. (2020). Recent theoretical analyses formalize the geometric and statistical structure of pessimism in offline RL Weltevrede et al. (2026), and related empirical strategies explore adaptive penalty scaling Liu et al. (2024) or scaled uncertainty penalties Srinivasan & Knottenbelt (2024). Other frameworks integrate Lyapunov stability principles to regulate uncertainty growth during policy updates Jiang et al. (2026). MSG is especially relevant because it studies independent Q ensembles as the primary source of pessimism and optimizes a policy with respect to a lower-confidence-style action-value objective Ghasemipour et al. (2022). Our question is orthogonal: given an uncertainty signal that may be used for pessimism, is its numerical scale meaningful as a held-out error surrogate, and where does that interpretation fail? IQL Kostrikov et al. (2022) provides our backbone; its role here is to isolate the calibration study on a stable offline-RL learner rather than to contribute a new policy-optimization method.

Conformal calibration in offline RL. Conformal prediction provides finite-sample marginal coverage under exchangeability Angelopoulos & Bates (2023). Recent conformal work has studied learned uncertainty features, robustness under distribution shift, transformations of conformity scores for improved conditional behavior, and explicit diagnostics of conditional coverage Kiyani et al. (2024); Ai & Ren (2024); Plassier et al. (2025); Braun et al. (2025). Recent work has also adapted this perspective to offline-RL conservatism. UNIQ Upadhyay (2026), for example, uses an IQL backbone, a multi-expectile value ensemble, split-conformal uncertainty, and a state-dependent mapping from uncertainty to conservatism. We therefore do not claim that conformal calibration in offline RL is new. Our focus is the reliability audit itself: we calibrate an observable held-out residual proxy, inspect how marginal validity is allocated across fixed support groups, and separately test whether the calibrated quantities remain useful under a frozen downstream decision rule.

Residual-based offline RL. Residuals-based Offline RL Zhu & Yu (2026) incorporates empirical residuals directly into a Bellman-style operator and develops theoretical and algorithmic consequences for offline policy optimization. We use residual information differently.

The residual in our study is a held-out observable diagnostic target for auditing uncertainty features; it is not treated as true Bellman error, does not define our policy-optimization operator, and does not certify a lower bound on policy value.

Marginal versus support-conditioned reliability. Standard split conformal validity is marginal. In a heterogeneous offline dataset, such validity need not reveal where miscoverage occurs. We therefore report support-conditioned coverage only as a diagnostic and make no claim of conditional conformal validity. Likewise, our support-aware scale is a frozen adaptive construction used to study reliability allocation, not a proof that conditional coverage has been solved.

3 Problem Setup

Consider a discounted Markov decision process with states s, actions a, reward r, transition kernel P, and discount $\gamma \in (0, 1)$. The learner receives only an offline dataset $\mathcal{D}$ collected by an unknown behavior policy. Complete trajectories, rather than individual transitions, are divided into proper-training, calibration, and validation partitions. The proper-training partition determines all learned value functions, actor parameters, uncertainty features, normalization statistics, and the reference set used by the support score. The calibration partition determines conformal quantiles and support-group thresholds. The validation partition is reserved for the frozen reliability evaluation.

Let $\bar{Q}(s, a)$ denote a frozen critic estimate and let $\bar{V}(s)$ be its associated state-value estimate. On a held-out transition $z_i = (s_i, a_i, r_i, s_i')$, define the observable one-step residual proxy

$$R_i = \left|\bar{Q}(s_i, a_i) - \left(r_i + \gamma \bar{V}(s_i')\right)\right|. \tag{1}$$

This quantity contains transition and reward noise and need not equal the true Bellman error. Our primary reliability metrics are empirical coverage and absolute deviation from a nominal coverage target. Sharpness is summarized by mean, median, and 90th-percentile bound width, both marginally and within fixed support groups.

4 Calibrated Residual-Proxy Bounds

Let $x = (s, a)$ and let $u(x)$ be ensemble disagreement. We evaluate four frozen constructions on the same critic features and validation transitions.

Raw disagreement. The uncalibrated baseline uses $B_{\text{raw}}(x) = u(x)$ and is included to test whether ensemble spread is already meaningful on the scale of the held-out residual proxy.

Global residual quantile. A global method ignores x and uses a finite-sample quantile of calibration residuals R_i. It is coverage calibrated but cannot adapt its width to local features.

Ratio calibration. We use $\epsilon_u = 10^{-8}$. Define $g_u(x) = u(x) + \epsilon_u$ and calibration scores $Z_i = R_i / g_u(X_i)$. For target miscoverage α, let

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad q = Z_{(k)}, \tag{2}$$

with $q = \infty$ when $k > n$. The ratio-calibrated bound is $B_u(x) = q g_u(x)$.

Support-aware ratio calibration. We use the following support-aware construction as a diagnostic probe rather than as a proposed optimal calibration rule. We set $\epsilon_d = 10^{-8}$. We standardize state and action coordinates using proper-training statistics. From proper training we deterministically sample 100,000 reference transitions using NumPy random seed zero. Let $d(x)$ be the mean Euclidean distance from standardized x to its 50 nearest reference neighbors; larger $d(x)$ denotes lower local support but is not interpreted as an exact density. With m_d the median of d on calibration data,

$$f_{\text{sup}}(x) = \sqrt{\frac{d(x) + \epsilon_d}{m_d + \epsilon_d}}, \qquad g_{\text{sup}}(x) = (u(x) + \epsilon_u)\, f_{\text{sup}}(x). \tag{3}$$

We then apply the same rank calibration in Eq. 2 to $R_i/g_{\text{sup}}(X_i)$. There is no additional fitted β parameter in the evaluated method.

Support groups. For diagnostics only, the 20th and 80th percentiles of d on the calibration split define high- and low-support thresholds. Validation points at or below the former are high support; those at or above the latter are low support. Residuals and validation labels never define these groups, and we do not fit separate group-conditional calibrators.

Finite-sample proxy coverage. Condition on the frozen training output and residual-free preprocessing. If the n calibration pairs and one test pair are exchangeable, the order-statistic choice in Eq. 2 yields

$$\Pr\{R_{\text{test}} \le B(X_{\text{test}})\} \ge 1 - \alpha. \tag{4}$$

This is the standard split-conformal marginal statement under exchangeability. In our offline-RL setting, however, transitions within trajectories are temporally dependent, so exact transition-level exchangeability is an idealization. The disjoint episode split prevents trajectory leakage across proper-training, calibration, and validation partitions, but it does not make transitions within trajectories independent or exchangeable. We therefore use Eq. 4 as a nominal marginal benchmark for the observable residual proxy rather than claiming an exact finite-sample guarantee for Markov-dependent transitions. It is not a guarantee for unobserved true Bellman error, Q-function error, support-conditional coverage, learned-policy occupancy, policy value, or return. In particular, subtracting B from a critic does not become a certified lower confidence bound on policy value merely because Eq. 4 holds.

5 Experimental Protocol

Backbone and dataset. We first train and validate an Implicit Q-Learning backbone Kostrikov et al. (2022) without the proposed calibration objects. Every run records an immutable configuration, dependency snapshot, learning curve, and final checkpoint. Evaluation uses the final policy rather than selecting a checkpoint by test return. The benchmark is the versioned Minari Younis et al. (2024) mujoco/halfcheetah/medium-v0 dataset with its recovered Gymnasium environment; we do not transfer normalization constants or published scores from legacy D4RL tasks to this version.

Leakage-controlled splits. For each training seed, complete episodes are split 80/10/10 into proper training, calibration, and validation, yielding 800/100/100 episodes and 800,000/100,000/100,000 transitions. Episode-ID sets are pairwise disjoint. The calibration study evaluates raw disagreement, the global residual quantile, ratio calibration, and support-aware ratio calibration at nominal coverage levels 0.70, 0.80, 0.90, and 0.95. No method is selected by validation return.

Frozen downstream control test. The control-transfer study reuses each seed's frozen actor, critics, normalization, and calibration objects. For every environment state, we form $K = 8$ action candidates: candidate zero is the deterministic actor mean and the remaining seven are stochastic actor proposals generated by a method-independent RNG. The common Q score is $\min(Q_1, Q_2)$. At nominal coverage 0.90 and fixed penalty weight $\lambda = 1$, we compare the deterministic actor mean, unpenalized Q reranking, raw-uncertainty penalty, ratio-calibrated penalty, and support-aware penalty. A global constant quantile is omitted from rollouts because it is action-independent and therefore induces exactly the same ranking as Q reranking. No method is retrained and no return-based tuning is performed. Within each training seed, all methods share the same 20 environment seeds and candidate randomness. The training seed ($n = 5$), not the episode, is the primary statistical unit.

Table 1: HalfCheetah Medium residual-proxy calibration over five training seeds. Entries are empirical marginal coverage and mean width (mean ± sample standard deviation).

Method	Nominal	Coverage	Mean width
Raw disagreement	—	0.1985 ± 0.0307	0.1555 ± 0.0080
Global quantile	0.70	0.6991 ± 0.0076	0.4720 ± 0.1189
Global quantile	0.80	0.8004 ± 0.0059	0.6169 ± 0.1444
Global quantile	0.90	0.9004 ± 0.0049	0.8875 ± 0.1438
Global quantile	0.95	0.9500 ± 0.0038	1.2665 ± 0.0920
Ratio calibration	0.70	0.6973 ± 0.0070	0.9962 ± 0.2439
Ratio calibration	0.80	0.7978 ± 0.0043	1.5713 ± 0.3749
Ratio calibration	0.90	0.8991 ± 0.0015	3.2515 ± 0.7517
Ratio calibration	0.95	0.9489 ± 0.0011	6.5139 ± 1.5292
Support-aware	0.70	0.6977 ± 0.0053	1.2610 ± 0.3058
Support-aware	0.80	0.7988 ± 0.0034	2.0083 ± 0.4645
Support-aware	0.90	0.8983 ± 0.0017	4.1392 ± 0.9728
Support-aware	0.95	0.9488 ± 0.0008	8.3538 ± 1.9467

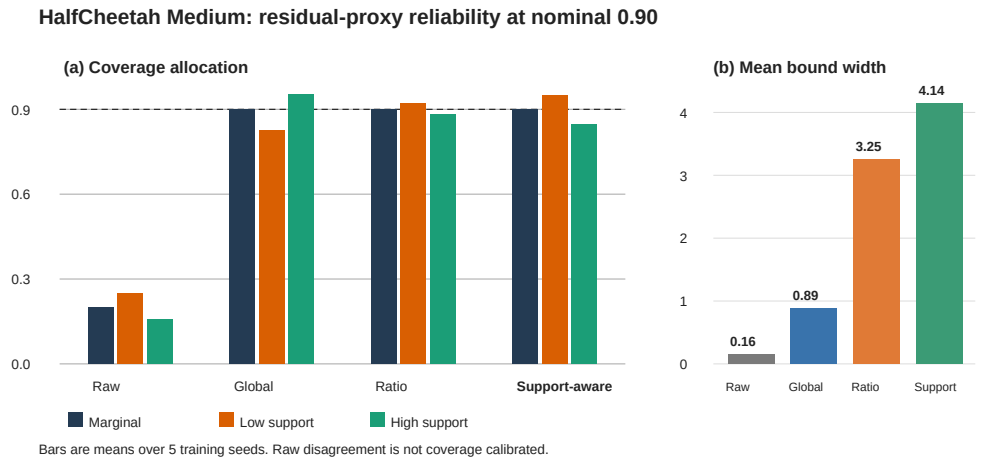


Figure 2: HalfCheetah Medium calibration reliability and sharpness at nominal coverage 0.90 over five training seeds. Raw disagreement is shown as an uncalibrated reference. Similar marginal coverage among calibrated methods conceals markedly different low-/high-support allocation and width. Error bars summarize variation across training seeds.

6 Results

6.1 Marginal calibration is necessary but not sufficient for diagnosis

Across the five training seeds, we verified finite outputs, disjoint episode partitions, and reproducibility of the frozen evaluation pipeline.

All calibrated constructions closely track their marginal targets, whereas raw disagreement is not calibrated to the residual-proxy scale. At nominal 0.90, the global, ratio, and support-aware variants reach mean marginal coverages of 0.900, 0.899, and 0.898, respectively. Marginal coverage alone therefore provides little basis for distinguishing their reliability behavior.

6.2 Marginal validity hides support-conditioned reliability allocation

Figure 2 shows that the marginal summary hides substantial heterogeneity. The global quantile is the narrowest calibrated construction, with mean width 0.8875 ± 0.1438, but its low-support coverage is only 0.8268 ± 0.0133 while high-support coverage reaches $0.9542 \pm$

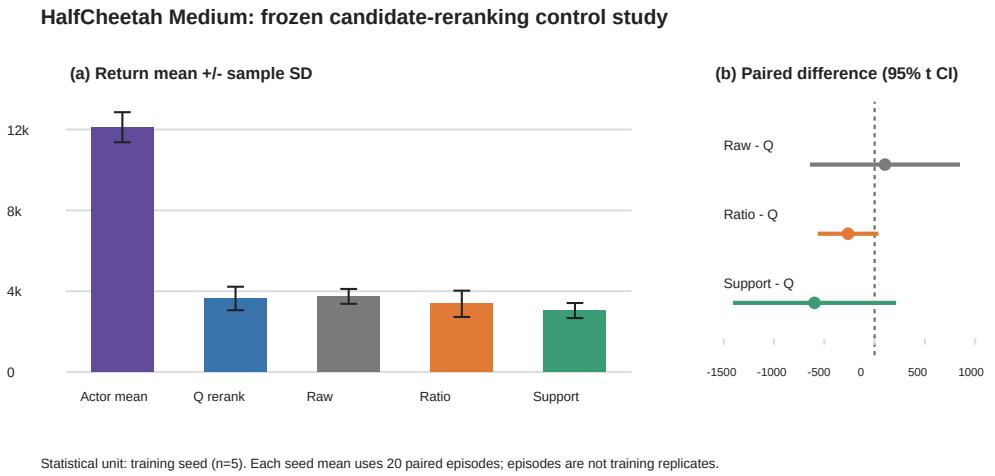


Figure 3: Frozen HalfCheetah Medium control-transfer study. Left: return (mean ± sample standard deviation) across five training seeds. Right: seed-level paired 95% intervals for each penalty method relative to unpenalized Q reranking. Each training seed uses 20 matched evaluation episodes per method; the training seed, not the episode, is the statistical unit. No retraining or return-based tuning is used.

0.0243. Ratio calibration changes this allocation to 0.9197 ± 0.0242 low-support and 0.8811 ± 0.0171 high-support coverage, while mean width rises to 3.2515 ± 0.7517. The support-aware construction pushes the redistribution further: low-support coverage reaches 0.9495 ± 0.0162, high-support coverage falls to 0.8480 ± 0.0184, and mean width grows to 4.1392 ± 0.9728.

The important observation is therefore not that an adaptive construction is uniformly superior. Rather, near-identical marginal validity can coexist with very different allocations of reliability and sharpness across support regimes. The support-aware scale improves low-support allocation in this benchmark only by paying a substantial width and high-support-coverage cost. This is the failure mode that a marginal coverage number alone cannot reveal.

6.3 Statistical calibration does not imply control utility

The control experiment evaluates a deliberately restricted question. The unpenalized Q-reranking interface itself substantially underperforms the deterministic actor mean, indicating that this decision rule is poorly aligned with the frozen actor–critic pair in our setting. We therefore do not treat the experiment as a general test of whether calibrated pessimism improves offline RL. Instead, it asks whether the frozen residual-proxy penalties improve action ranking relative to unpenalized Q reranking within this same evaluation-time interface. The deterministic actor mean achieves $12{,}114.76 \pm 744.07$ across the five training seeds. Unpenalized Q reranking achieves $3{,}640.26 \pm 581.09$, while raw, ratio-calibrated, and support-aware penalties achieve $3{,}743.22 \pm 367.77$, $3{,}376.20 \pm 650.94$, and $3{,}042.62 \pm 373.01$, respectively. The seed-level paired 95% intervals for penalty minus Q reranking all include zero: raw $[-642.96, 848.88]$, ratio $[-565.84, 37.71]$, and support-aware $[-1{,}408.72, 213.43]$.

Under this restricted test, we find no evidence that calibrated residual-proxy penalties outperform unpenalized Q reranking. This result does not establish that calibrated uncertainty is unhelpful for control more generally, nor does it identify the mechanism responsible for the reranking degradation. It establishes only the narrower point supported by the experiment: statistical reliability of a residual-proxy bound is insufficient, by itself, to justify a downstream control claim.

7 Discussion

The experiments separate three questions that are easy to conflate. First, an uncertainty score should not be interpreted as an error bound without validating its numerical scale against an observable target. On this benchmark, raw ensemble disagreement is not calibrated to the held-out residual proxy. Second, marginal calibration is not a complete description of reliability in heterogeneous offline data: constructions with almost identical marginal coverage can allocate miscoverage very differently across support regimes. Third, statistical calibration and decision utility require separate validation. In our restricted reranking study, a statistically interpretable residual-proxy bound does not yield a return advantage over the corresponding unpenalized ranking rule. These observations are best read as a focused case study of reliability allocation and control transfer, rather than as a benchmark-wide principle or a claim that calibrated uncertainty is generally ineffective for control.

8 Limitations

This paper is a focused case study on HalfCheetah Medium rather than a benchmark-wide evaluation of offline-RL pessimism. Five training seeds strengthen the reliability of the reported within-task effects, but a single continuous-control task cannot establish that the same support-conditioned behavior will persist across environments, dataset qualities, or learning backbones. The empirical conclusions should therefore be interpreted as evidence that these failure modes can occur, not as estimates of how frequently they occur across offline RL.

The calibrated target is an observable one-step Bellman-residual proxy. It contains transition and reward noise and need not equal true Bellman error, Q-function error, policy-value error, or return loss. The finite-sample argument therefore applies only to marginal coverage of this proxy under exchangeability. Complete episodes are kept disjoint across proper-training, calibration, and validation splits, but temporal structure still makes transition-level exchangeability an approximation.

The support score is a diagnostic construction rather than a density estimator: it is the mean standardized distance to 50 nearest neighbors from a fixed proper-training reference set. The p20/p80 groups are frozen from calibration data and used only for analysis. The support-aware scale improves low-support allocation in this benchmark but widens bounds and lowers high-support coverage; it is not established as a uniformly preferable method.

Finally, the downstream experiment evaluates only one control interface: fixed-actor proposal generation followed by evaluation-time reranking. Its negative result does not rule out alternative uses of calibrated uncertainty during training, policy improvement, or constrained optimization. It instead shows that a statistical calibration result should not be promoted directly into a control claim without a separate evaluation of how the calibrated quantity interacts with the decision rule.

9 Conclusion

Uncertainty-based pessimism is often motivated by a simple intuition: uncertain state–action pairs should receive larger penalties. Our results show that several different questions are hidden inside that intuition. An uncertainty signal may correlate with error without being calibrated; a calibrated construction may achieve correct marginal coverage while distributing reliability unevenly across support regimes; and a statistically reliable residual-proxy bound may still fail to improve action selection.

On five HalfCheetah Medium seeds, split calibration converts raw uncertainty into near-nominal marginal residual-proxy coverage, but marginal validity alone obscures large support-conditioned differences. Adaptive scaling reallocates coverage toward low-support transitions only by paying a substantial sharpness and high-support-coverage cost. When

the same frozen calibration objects are inserted into a matched candidate-reranking study, that statistical improvement does not translate into better return.

Within the scope of this study, the results support a methodological distinction: uncertainty magnitude, marginal calibration reliability, and downstream control behavior should be measured separately. Treating them as interchangeable can turn a statistically meaningful uncertainty signal into an unsupported control claim. Broader evidence will require testing the same questions across additional datasets, backbones, and control interfaces.

AI use statement

Generative AI tools were used to aid writing and polishing, to support retrieval and discovery of related work, and for research ideation and execution. AI-assisted research suggestions were treated as hypotheses rather than evidence: relevant claims and citations were checked against primary sources, generated code was inspected and exercised with deterministic checks, and all reported numerical results are taken from immutable experiment outputs rather than generated text. The author reviewed the AI-assisted work and takes responsibility for the final text, claims, code, and artifacts.

Reproducibility statement

The method definition, data-splitting protocol, finite-sample assumptions, support diagnostic, control-transfer interface, and evaluation metrics are specified above. All reported tables and figures are generated from immutable seed-level or transition-level artifacts with recorded hashes. The project records complete configurations, dependency snapshots, seed-wise results, and the commands used to reproduce each figure. Any supplementary archive included with the final submission will contain only materials actually uploaded with that submission; the main paper does not rely on unpublished numerical claims.

References

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 641–665, 2024.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction, 2025.

Kamyar Ghasemipour, Shixiang Gu, and Ofir Nachum. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In Advances in Neural Information Processing Systems, volume 35, pp. 18267–18281, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html.

Ke Jiang, Yao Li, and Xiaoyang Tan. Toward reliable offline reinforcement learning via Lyapunov uncertainty control. IEEE Transactions on Neural Networks and Learning Systems, 37(3):1347–1361, 2026. doi: 10.1109/TNNLS.2025.3616159.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline reinforcement learning? Mathematics of Operations Research, 2024. doi: 10.1287/moor.2022.0216.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 24749–24769, 2024.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems, volume 33, pp. 1179–1191, 2020.

Jie Liu, Yinmin Zhang, Chuming Li, Yaodong Yang, Yu Liu, and Wanli Ouyang. Adaptive pessimism via target Q-value for offline reinforcement learning. Neural Networks, 180: 106588, 2024. doi: 10.1016/j.neunet.2024.106588.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 49459–49492, 2025.

Padmanaba Srinivasan and William J. Knottenbelt. SpOiLer: Offline reinforcement learning using scaled penalties. In Proceedings of the 6th Annual Learning for Dynamics and Control Conference, volume 242 of Proceedings of Machine Learning Research, pp. 825–838, 2024. URL https://proceedings.mlr.press/v242/srinivasan24a.html.

Aditya Upadhyay. UNIQ: Conformal calibration for adaptive conservatism in offline reinforcement learning, 2026.

Max Weltevrede, Matthijs T. J. Spaan, and Wendelin Böhmer. Generalization in offline RL: The structure is more important than the amount of pessimism, 2026.

Omar G. Younis, Rodrigo Perez-Vicente, John U. Balis, Will Dudley, Alex Davey, and Jordan K. Terry. Minari: A standard format for offline reinforcement learning datasets, 2024. URL https://doi.org/10.5281/zenodo.13767625.

Qing Zhu and Xian Yu. Residuals-based offline reinforcement learning, 2026.