From Offline Evidence to Online Action: A Decision Framework for Imperfect Offline Evaluation

Yuhao Wang [1] **Lorenzo Masoero** [1]

Abstract

Decision-making from offline datasets often requires acting before reliable online evidence is available. Logged data may be selective, outcomes may be delayed, deployment populations may differ from historical ones, and offline metrics may only imperfectly reflect the true objective. Methods such as off-policy evaluation, sensitivity analysis, robustness analysis, and uncertainty quantification improve parts of this pipeline, but they do not by themselves answer the central offline-to-online question: what level of online action does the current offline evidence justify?

This position paper frames that question as a distinct decision problem. We propose a compact protocol organized around three dimensions of evidence—relevance, transportability, and reliability—and use it to distinguish when offline evidence supports deployment, targeted validation, constrained rollout, or deferral. The contribution is not a new estimator, but a decision-oriented framework for translating imperfect offline evidence into appropriate online action.

1. Introduction

Machine learning systems are often evaluated offline before they are exposed online. This is typically cheaper, safer, and faster than deployment: teams can test candidate policies on historical logs, simulations, benchmark suites, or retrospective datasets. Yet the quantity that ultimately matters is almost always online: revenue, user satisfaction, clinical outcomes, or long-term impact. Offline evidence is therefore useful only to the extent that it supports conclusions about this online behavior (Gilotte et al., 2018). This extrapolation is rarely trivial: logs are selective, proxies are imperfect, deployment populations shift, and long-term outcomes may not yet be observed.

Existing work substantially improves how offline evidence is estimated and stress-tested. Off-policy evaluation (OPE) (Precup et al., 2000) estimates policy value from logged data; High-confidence off-policy evaluation (Thomas et al., 2015; Kallus & Uehara, 2020) provides confidence guarantees; distribution-shift and robustness methods (Tibshirani et al., 2019; Barber et al., 2023; Quiñonero-Candela et al., 2009; Tibshirani et al., 2019) characterize validity under changing environments; and causal transportability (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2013) formalizes when conclusions learned in one population can be carried to another. These approaches answer a descriptive question: *what does the evidence suggest?*

The offline-to-online decision requires a further normative judgment: *what scope of online action does this evidence justify?* This is not merely a question of estimation accuracy. An estimate may be statistically precise yet irrelevant to the decision objective, valid on the logged population but not in deployment, or stable under one evaluation protocol but fragile under reasonable alternatives. The missing layer is therefore an evidence-to-action protocol: a way to translate imperfect offline evidence into admissible online action.

In practice, offline evaluation outputs are often treated as implicit decision rules. A benchmark gain may authorize an A/B test; an OPE estimate with a narrow confidence interval may justify deployment; a short-term lift may trigger long-term rollout. Yet such mappings are typically ad hoc, relying on heuristics, precedent, or local risk tolerance. The same evidence may justify deployment in one setting, targeted validation in another, and deferral in a third. What is missing is not better estimates alone, but a principled way to determine what actions the evidence is sufficient to support.

We illustrate the framework with a running example. Consider replacing the language model in a customer-facing assistant with a new model. Offline benchmarks show improvements in accuracy, factual grounding, and preference metrics, with stable gains across seeds and prompt variants (Liang et al., 2022). Under current practice, such results often justify an A/B test or phased rollout. Under the proposed protocol, however, the evidence may still be insufficient for

[1]Amazon. Correspondence to: Yuhao Wang <wanyuhao@amazon.co.jp>.

Accepted at ICML 2026 Workshop on *Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning*.

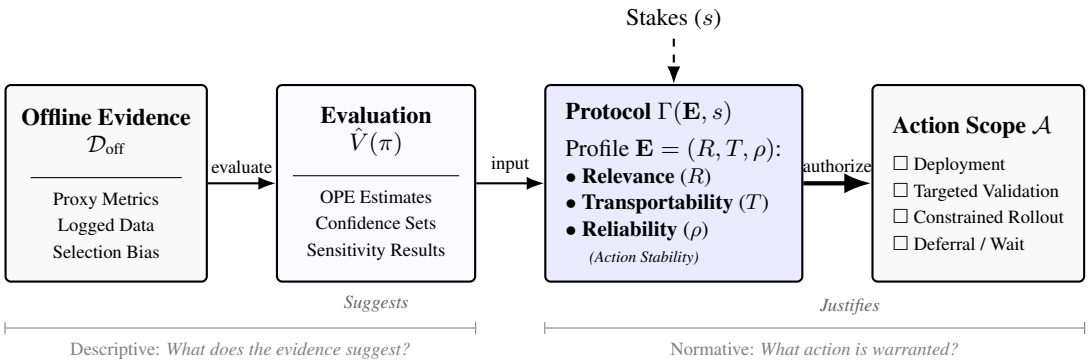

Figure 1. The Offline-to-Online Decision Architecture. The protocol introduces an intermediate admissibility layer between offline evaluation outputs and online action.

deployment-level action. The benchmarks may not reflect the true objective (e.g., user trust, task completion, latency), raising relevance concerns. The evaluation distribution may differ from production traffic, raising transportability concerns. The results may vary under reasonable prompt or evaluation changes, raising reliability concerns.

The protocol therefore does not reject the benchmark evidence; it assigns it a narrower decision role. The evidence is inadmissible for deployment-level action, but sufficient for targeted validation: evaluation on production-representative prompts, objective-aligned metrics, and stability checks before user-facing rollout.

These issues are not specific to language models. In recommendation systems, logged data may limit transportability due to exposure bias (Schnabel et al., 2016). In clinical decision support, retrospective validation may not transfer across patient populations (Futoma et al., 2020). In causal evaluation, conclusions may be sensitive to unobserved confounding (Rosenbaum, 2002). In delayed-outcome settings, short-term gains may not reflect long-term impact (Athey et al., 2019).

In this position paper, we propose a structured decision protocol that replaces ad hoc sufficiency judgments with an explicit assessment of evidence quality. Our contributions are threefold. First, we frame the offline-to-online transition as an *admissibility problem*, separating evaluation (what the evidence suggests) from authorization (what actions the evidence justifies). Second, we introduce a compact protocol organized around three axes: relevance, transportability, and reliability. The protocol maps an evidence profile, together with decision stakes, to an action scope such as deployment, targeted validation, constrained rollout, or deferral. Third, we show how existing tools, including OPE, sensitivity analysis, causal transportability, robustness analysis, and uncertainty quantification, can serve as inputs to this protocol without by themselves specifying the justified action scope.

2. Why evaluation outputs are not decision rules

Existing offline evaluation outputs provide statistically meaningful summaries of evidence — estimates, confidence sets, policies, or sensitivity bounds — but they do not resolve the decision problem of what action is warranted. The limitation is structural: these outputs characterize performance under assumptions, whereas deployment decisions must determine admissible exposure under uncertainty, stakes, and potential failure modes. This gap cannot be eliminated by improving estimation alone.

This distinction persists even for highly precise outputs. A statistically significant lift in a proxy metric may be inadmissible for deployment if the proxy is misaligned with the true objective. An OPE estimate with a narrow interval may be inadmissible for rollout if it relies on support or overlap assumptions that fail in the deployment population. A benchmark result stable across random seeds may be inadmissible for launch if the model ranking reverses under reasonable changes to prompt format, evaluator pool, or outcome horizon. These failures correspond to the three axes used below: relevance, transportability, and reliability. In each case, the statistical artifact is legitimate, but the action it licenses remains underspecified.

Existing tools can therefore be understood as modular inputs to a decision protocol rather than substitutes for one. OPE and high-confidence OPE quantify value and uncertainty from logged data (Precup et al., 2000; Thomas et al., 2015; Kallus & Uehara, 2020); safe policy improvement and pessimistic RL encode uncertainty or support limitations into conservative policies (Laroche et al., 2019; Kumar et al., 2020b); causal transportability and sensitivity analysis diagnose setting mismatch, confounding, and action stability (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2013; Rosenbaum, 2010); and distribution-shift and conformal methods characterize validity under changing environments (Tibshirani et al., 2019; Barber et al., 2023). These

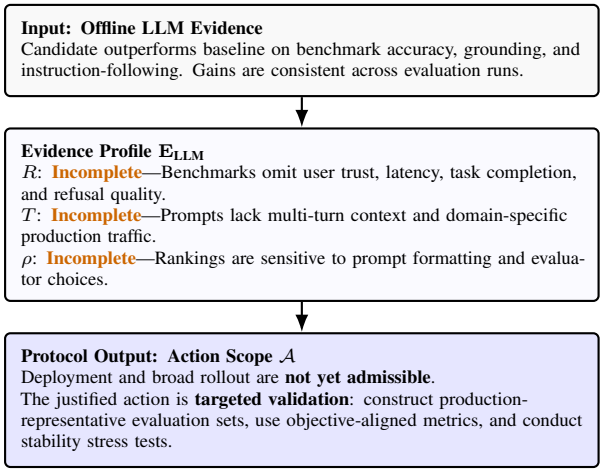

Figure 2. Instantiation of the protocol for the LLM running example. The protocol assigns a positive benchmark signal to a restricted action scope: targeted validation before user-facing exposure.

tools strengthen the evidentiary basis for action, but they return estimates, intervals, policies, sets, or bounds—not a decision about whether the warranted action is deployment, targeted validation, constrained rollout, or deferral.

Consider two candidate systems with the same estimated offline gain and the same confidence interval. In the first case, the estimate targets the true deployment objective but is based on limited overlap with the deployment population. In the second case, the estimate is statistically precise on the logged distribution but measures a proxy weakly related to the true objective. A scalar evaluation output may treat these cases similarly. An admissibility protocol does not: the first case calls for constrained rollout to collect higher-overlap evidence, while the second calls for targeted validation or redesign of the evaluation objective. The issue is not only the amount of uncertainty, but the type of evidentiary failure. Reducing admissibility to uncertainty alone therefore collapses qualitatively distinct evidentiary failures into a single scalar notion of confidence, even though these failures warrant different weaker actions.

3. A decision framework for imperfect evidence

We formalize the proposed layer as a mapping from evidence quality to admissible action scope. For a candidate policy or intervention π, let

$$\mathbf{E}(\pi; \mathcal{D}_{\text{off}}) = (R, T, \rho)$$

denote an evidence profile, where R summarizes relevance, T transportability, and ρ reliability. These coordinates may be quantitative scores, qualitative labels, or diagnostic summaries derived from the tools discussed above. The goal is not to impose a universal metric, but to make explicit which aspects of evidence quality are being used to authorize action.

The protocol is a map

$$\Gamma(\mathbf{E}, s) \subseteq \mathcal{A}, \quad \mathcal{A} = \left\{ \begin{array}{l} \text{deployment, targeted validation,} \\ \text{constrained rollout, deferral} \end{array} \right\},$$

where s denotes decision stakes such as downside risk, reversibility, exposure size, and cost of delay. Unlike an estimator, Γ does not output a value estimate or a learned policy. It restricts the set of actions whose evidentiary requirements are met.

The three coordinates correspond to distinct admissibility requirements. *Relevance* asks whether the offline evidence measures the decision objective rather than only a proxy. *Transportability* asks whether the conclusion applies in the deployment setting rather than only in the logged, simulated, or benchmark distribution. *Reliability* asks whether the warranted action remains stable under residual uncertainty and reasonable perturbations. Thus reliability is not generic statistical precision; it is action stability.

The three axes are useful because admissibility failures are typed rather than merely scalar. A relevance failure suggests that the evaluation objective or metric should be redesigned. A transportability failure suggests that deployment-representative evidence or restricted exposure is needed. A reliability failure suggests stress testing before stronger action. The protocol therefore does not simply lower confidence in a conclusion; it maps different failure types to different weaker action scopes.

One simple instantiation is threshold-based. Let $\tau(s, a)$ denote the minimum evidence requirements for action a under stakes s. Then

$$\Gamma(\mathbf{E}, s) = \{a \in \mathcal{A} : \mathbf{E} \succeq \tau(s, a)\},$$

where $\succeq$ denotes coordinatewise satisfaction of the required relevance, transportability, and reliability levels. Higher-stakes or less reversible actions require stronger evidence.

A natural consistency property is monotonicity: if the current evidence authorizes a stronger action, such as deployment, it should also authorize weaker actions such as constrained rollout or targeted validation. Equivalently, stronger actions should require weakly stronger evidence. This highlights the distinction between evaluation and authorization: evaluation ranks candidate systems, whereas admissibility determines the strongest online action the evidence can support.

Other implementations, such as audit-style checks, risk-budget rules, or paradigm-specific diagnostics, can be viewed as alternative ways of specifying the requirements $\tau(s, a)$; we leave these examples to Appendix **??**.

The resulting action scopes have different commitments. *Deployment* authorizes broad use; *constrained rollout* authorizes limited exposure while collecting higher-fidelity evidence; *targeted validation* authorizes tests aimed at a specific weak axis; and *deferral* indicates that the current evidence is not yet sufficient for online action.

Figure 2 gives a worked instantiation for the running LLM example. The input is a positive benchmark signal for a candidate model; the evidence profile marks relevance, transportability, and reliability as incomplete; and the resulting action scope is targeted validation rather than deployment. Thus the benchmark signal is not discarded, but assigned a narrower role before user-facing rollout.

4. Relation to existing safe and robust approaches

This framework is compatible with, but distinct from, safe or conservative approaches in offline learning (Kumar et al., 2020b; Jin et al., 2021; Xie et al., 2021; Laroche et al., 2019). Such methods typically encode uncertainty, support mismatch, or confounding into the objective, policy update, or lower-bound estimate. Their goal is often to find a policy that remains safe under a specified model of uncertainty.

The distinction is that conservative optimization typically operates within an already admissible action class: it asks which policy should be chosen or how much the update should be penalized. Admissibility asks a prior question: whether that action class should be available at all under the current evidence. Thus a pessimistic method may recommend a conservative deployment, while the proposed protocol may instead recommend targeted validation or deferral when the evidence is precise but misaligned, non-transportable, or action-unstable.

The protocol is therefore not a replacement for robust methods, but a decision layer that clarifies when and how their outputs should be acted upon.

5. Conclusion

Offline evidence is indispensable in machine learning, but evaluation outputs are not themselves decision rules. The central question is therefore not only what the data suggest, but what level of online action the evidence can support. We framed this as an admissibility problem and proposed a compact protocol based on relevance, transportability, and reliability. The key point is that evidentiary failures are typed rather than merely scalar: proxy mismatch, lack of transportability, and action instability call for different weaker actions. A useful offline evaluation pipeline should therefore report not only estimates or uncertainty, but the strongest action scope warranted by the available evidence.

References

AI, N. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pp. 100–1, 2023.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working Paper 26463, National Bureau of Economic Research, 2019.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Bareinboim, E. and Pearl, J. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 2013.

Ben-Tal, A. and Nemirovski, A. Robust optimization–methodology and applications. *Mathematical programming*, 92(3):453–480, 2002.

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

Berger, J. O. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

D'Amour, A., Heller, K., Moldovan, D., , et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 22(280):1–61, 2021.

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

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

Gilotte, A., Calauzènes, C., Thomas, T., et al. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.

Imbens, G. W. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132, 2003. doi: 10.1257/000282803321946921.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *JMLR*, 2020.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33: 1179–1191, 2020a.

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

Laroche, R., Trichelair, P., and Tachet des Combes, R. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C. W., Parvez, M. R., et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13785–13816, 2024.

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.

Liang, P., Bommasani, R., Lee, T., , et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Little, R. J. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

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 *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2016.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 759–766, 2000.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Radi, H., Hanna, J. P., Stone, P., and Taylor, M. E. Safe evaluation for offline learning: Are we ready to deploy? *arXiv preprint arXiv:2212.08302*, 2022.

Rosenbaum, P. R. *Observational Studies*. Springer, 2nd edition, 2002.

Rosenbaum, P. R. *Design of Observational Studies*. Springer, 2010.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning (ICML)*, 2016.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *AAAI*, 2015.

Tibshirani, R. J., Foygel Barber, R., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Trimmer, P. C., Houston, A. I., Marshall, J. A., Mendl, M. T., Paul, E. S., and McNamara, J. M. Decision-making under uncertainty: biases and bayesians. *Animal cognition*, 14 (4):465–476, 2011.

Wilm, T. and Normann, P. Identifying offline metrics that predict online impact: A pragmatic strategy for real-world recommender systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 967–970, 2025.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zheng, L., Chiang, W.-L., Sheng, Y., , et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A. A weakly formal view of admissibility

The main text introduces an admissibility layer between offline evaluation and online action. Let π denote a candidate policy or intervention, and let $\mathcal{D}_{\text{off}}$ denote the offline data available at decision time. Existing pipelines often aim to produce an object such as

$$\widehat{V}(\pi; \mathcal{D}_{\text{off}})$$

or an associated interval, ranking, or sensitivity statement. Such outputs do not by themselves determine the strongest admissible online action.

We summarize the admissibility logic as

$$\mathcal{D}_{\text{off}} \longrightarrow \mathbf{E}(\pi; \mathcal{D}_{\text{off}}) \longrightarrow \Gamma(\mathbf{E}(\pi; \mathcal{D}_{\text{off}}), s) \subseteq \mathcal{A},$$

where $\mathbf{E}(\pi; \mathcal{D}_{\text{off}}) = (R, T, \rho)$ is the evidence profile, s denotes the stakes or decision context, $\mathcal{A}$ is the set of possible online actions, and Γ returns the warranted action scope under the current evidence profile and stakes.

The key point is that Γ restricts action scope rather than selecting a single optimal action. This distinction matters because similar offline estimates may warrant different actions when their failures occur along different axes.

B. Operational diagnostics and weaker actions

The diagnostics are not intended to produce universal scores. Their purpose is to identify which part of the offline evidence blocks stronger action, and therefore what weaker action is appropriate.

Relevance. Relevance asks whether the offline result measures the quantity that the deployment decision actually cares about. Typical checks include whether the proxy metric agrees with the target objective, whether the model ranking is stable under alternative objectives, and whether the offline metric is predictive of downstream outcomes. These checks are connected to work on surrogate outcomes, proxy metrics, offline-to-online evaluation and metric alignment (Athey et al., 2019; Gilotte et al., 2018; Liang et al., 2022; Gilotte et al., 2018; Wilm & Normann, 2025). If relevance is weak, the appropriate weaker action is usually targeted validation or metric redesign rather than broader exposure. Additional deployment may otherwise collect more evidence about the wrong objective.

Transportability. Transportability asks whether the offline conclusion applies to the population or environment where the action will be taken. Typical checks include support overlap, covariate or policy shift, external validity across environments, and horizon mismatch for delayed outcomes. These checks are connected to causal transportability, dataset shift, off-policy evaluation, exposure-bias correction, and invariance across environments (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2013; Quiñonero-Candela et al., 2009; Precup et al., 2000; Schnabel et al., 2016; Peters et al., 2016). If transportability is weak while the objective is otherwise aligned, the natural weaker action is often constrained rollout or deployment-representative data collection. The issue is not objective mismatch, but uncertainty about transfer to the deployment population or environment.

Reliability. Reliability asks whether the recommended action class is stable under reasonable changes to the evaluation pipeline. Typical checks include sensitivity to modeling assumptions, support restrictions, outcome horizon, missing outcomes, and evaluator or prompt choices. These checks are connected to sensitivity analysis, missing-data analysis, algorithmic stability, underspecification, and evaluation bias in language-model assessment (Rosenbaum, 2002; Imbens, 2003; Little & Rubin, 2019; Bousquet & Elisseeff, 2002; D'Amour et al., 2021; Zheng et al., 2023). If reliability is weak, the appropriate weaker action is usually stress testing before stronger exposure. The issue is not only estimator variance, but whether reasonable perturbations would change the recommended action class.

C. Cross-paradigm instantiations

The same admissibility structure appears across several offline evaluation settings. The goal of Table 1 is not to provide a complete taxonomy, but to illustrate how different evidentiary failures can lead to different weaker actions even when the offline signal itself appears positive.

Paradigm	R: relevance	T: transportability	ρ: reliability	Action implication
Bayesian decision making	Utility or loss mismatch	Prior or environment shift	Posterior sensitivity	Targeted validation
RL / OPE	Reward proxy mismatch	Support or overlap limitation	CI / estimator sensitivity	Constrained rollout
Causal evaluation	Outcome proxy mismatch	Failed transportability	Unobserved-confounding sensitivity	Targeted validation
LLM evaluation	Benchmark-objective gap	Prompt / user-distribution shift	Prompt or evaluator sensitivity	Deferral or targeted validation

Table 1. Examples of typed admissibility failures across evaluation settings. The final column indicates the weaker action typically suggested by the dominant failure type.

Bayesian decision making. Bayesian decision analysis makes explicit that an action is justified relative to a utility or loss function and a posterior distribution (Berger, 1985; Trimmer et al., 2011). In offline-to-online settings, this creates three natural failure modes: the utility used for evaluation may not match the deployment objective; the prior or data-generating environment may differ from the deployment setting; and the posterior recommendation may be sensitive to prior, likelihood, or loss specification. Robust Bayesian analysis studies such sensitivity directly (Berger, 1994). Under the admissibility view, posterior evidence may therefore justify targeted validation even when it is not yet strong enough to authorize deployment.

RL / off-policy evaluation. Offline reinforcement learning and off-policy evaluation use data collected under a behavior policy to evaluate or learn a target policy before deployment (Precup et al., 2000; Levine et al., 2020; Radi et al., 2022). A central evidentiary bottleneck is support overlap: if the behavior policy rarely visits the state-action regions used by the target policy, the resulting estimate can be high-variance, biased, or poorly justified for deployment. High-confidence and doubly robust OPE methods provide uncertainty quantification and variance reduction (Thomas et al., 2015; Kallus & Uehara, 2020), but a positive estimate may still be inadmissible for full deployment when overlap is weak or estimator sensitivity is high. In such cases, the weaker action is often constrained rollout or additional data collection under safer exploration.

Causal evaluation. Causal evaluation asks whether an estimated effect supports an intervention in the target setting (Pearl, 2009). Relevance failures arise when the measured outcome is only a proxy for the decision objective; transportability failures arise when an effect learned in one population or environment does not carry to another (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2013); and reliability failures arise when conclusions depend strongly on untestable assumptions such as no unobserved confounding (Rosenbaum, 2002). Under the admissibility view, causal evidence can be sufficient for targeted validation or policy refinement without being sufficient for broad deployment.

LLM evaluation. Language-model evaluation often relies on benchmark suites, preference tests, or model-based evaluators to summarize offline performance (Liang et al., 2022; Chang et al., 2024; Laskar et al., 2024). These evaluations can fail to authorize deployment when benchmark tasks do not match the intended user objective, when prompt or user distributions differ from production traffic, or when rankings are sensitive to prompt wording and evaluator choice. Recent work on LLM-as-a-judge documents evaluator biases and instability that are directly relevant to this reliability axis (Zheng et al., 2023). Under the admissibility view, a positive benchmark result may therefore support targeted validation or deferral rather than user-facing rollout.

D. Relation to constraints and pessimistic objectives

A natural question is whether admissibility is merely another form of constraint or pessimism in offline optimization. This question is natural because robust optimization and pessimistic offline learning also restrict actions under uncertainty (Ben-Tal & Nemirovski, 2002; Kumar et al., 2020a; Jin et al., 2021).

The distinction is one of decision level. Constraints and pessimistic objectives typically operate after an action class has already been accepted as available. By contrast, admissibility asks whether that action class is warranted at all under the current evidence.

This distinction matters because admissibility may recommend a different kind of weaker action rather than a more

conservative deployment. For example, proxy mismatch may call for targeted validation, whereas weak transportability may call for constrained rollout. Even highly accurate evaluation does not remove this step: a precise estimate of the wrong objective, the wrong population, or an unstable action recommendation may still be inadmissible for deployment.

E. Implementation choices

The main text uses a threshold-based admissibility rule because it makes the connection between evidentiary failures and weaker actions explicit. Other implementations are possible, such as checklist-based review or continuous risk scores, as in broader work on model cards, datasheets, and risk management frameworks (Mitchell et al., 2019; Gebru et al., 2021; AI, 2023). These are implementation choices rather than part of the paper's core conceptual claim.