Auditing Actionability in AI Forecasting Interfaces

Anonymous Authors[1]

Abstract

AI forecasts are increasingly used as decision interfaces: scores trigger coupons, notifications, clinical prompts, and resource allocation. Standard forecast evaluation asks whether an interface predicts a target Y. Deployment asks a different question: does the same interface preserve the information needed for an action query q? We formalize this distinction as decision insufficiency and propose a validity-controlled audit for actionability claims. The audit states the decision unit, loss, exposed forecast interface, and evidence source; checks identification and prospective provenance; rules out forecast-only repairs; and refines the interface only when a certificate gives positive held-out regret reduction. The output is a typed verdict: certify, refine, request response evidence, lower the claim, or refuse. Across controlled, semi-real, and real treatment/control settings, the audit separates deployable refinements from unsupported actionability claims. X5 RetailHero yields score-cell certificate value CI $[0.077, 0.111]$, Drink Less yields user-level value CI $[0.027, 0.045]$, while myBPmyLife fails the deployable-value gate despite a response-source upper bound. The result is a decision-theoretic evaluation protocol for forecast interfaces, not another forecasting architecture.

1. Introduction

Forecasting systems are no longer used only as passive predictors. A purchase-risk score may decide whether a customer receives a coupon. An app-open forecast may trigger a notification. A clinical-risk score may route outreach. In such settings, the forecast is an interface between history and action. Accuracy on the forecast target is useful, but it is not a certificate that the interface is safe to act on.

[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.

The failure mode is simple. Two customers may receive the same purchase-risk score. One would buy without intervention; the other buys only if contacted. The score can be calibrated for purchases and still hide the distinction that determines the correct action. This is not merely a calibration error. It is a failure of decision sufficiency: the exposed interface collapses information relevant to the downstream action query.

We study this problem as an evaluation question for AI forecasting. Let $I(X)$ be the interface exposed to a decision maker and $q(X)$ be the decision query, such as response, treatment effect, or net intervention value. The paper asks when $I(X)$ supports an actionability claim and when extra prospective evidence must be appended as a certificate. The audit is deliberately conservative. It refuses claims that are not identified, lowers claims to the level supported by the data, and distinguishes genuine response-bearing evidence from ordinary repairs of the forecast interface.

Contributions.

- **A decision-sufficiency formulation.** We formalize the gap between predictive validity and actionability, and show that a perfect forecast of Y can be insufficient for a distinct decision query q.
- **A validity-controlled audit.** We define certificate value as held-out regret reduction after identification, provenance, and no-gap-control checks.
- **A mixed-verdict empirical study.** The same protocol yields positive refinements on X5 RetailHero and Drink Less, while refusing or lowering unsupported claims on myBPmyLife, HeartSteps, and iFlex.

2. Decision sufficiency

Let X denote the available history, Y the forecasting target, and $I(X)$ the forecast interface exposed to a decision maker. Let $q(X)$ denote the action query and $L(a, q)$ the loss for action $a \in \mathcal{A}$. A policy observing representation $J(X)$ has risk

$$\mathcal{R}(\pi; J, q) = \mathbb{E}[L(\pi(J(X)), q(X))], \quad \mathcal{R}^*(J; q) = \inf_{\pi} \mathcal{R}(\pi; J, q). \tag{1}$$

Definition 1 (Decision insufficiency)**.** The interface I is

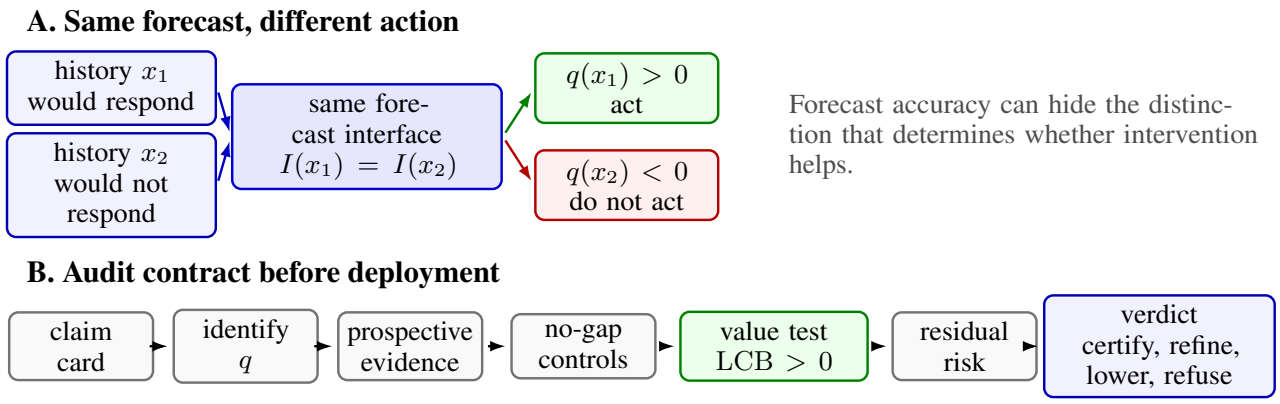

Figure 1. **Forecast actionability audit.** Panel A shows the information-bottleneck failure: two histories share the same forecast interface but require different actions. Panel B shows the audit. Actionability is accepted only after the decision query is identified, evidence is prospective, forecast-only repairs fail to close the gap, and held-out certificate value has a positive lower confidence bound.

decision-sufficient for q when $\mathcal{R}^*(I;q) = \mathcal{R}^*(X;q)$. Its decision-insufficiency gap is

$$\mathrm{DIG}(I,q) = \mathcal{R}^*(I;q) - \mathcal{R}^*(X;q) \geq 0. \tag{2}$$

A deployable certificate $C(X)$ is prospective evidence appended to I before action. Its value is

$$\mathrm{CV}(C;I,q) = \mathcal{R}^*(I;q) - \mathcal{R}^*((I,C);q). \tag{3}$$

Positive value means that C recovers action-relevant distinctions collapsed by the exposed forecast interface.

Theorem 1 (Accuracy is not actionability)**.** *There is a distribution over (X,Y,q) and an interface $I(X)$ such that I is a perfect forecast of Y, yet I is not decision-sufficient for q. With binary actions and zero-one loss,* $\mathrm{DIG}(I,q) = 1/2$.

The proof is in Appendix A. The construction separates the forecast target from the response variable: I reveals Y exactly but hides q. This mirrors deployments where a forecast score tracks event risk while intervention response depends on a different latent distinction.

Proposition 1 (Estimator-interface separation)**.** *A response estimator may consistently recover $m(X) = \mathbb{E}[q \mid X]$ while the deployed interface $I(X)$ remains insufficient. If $m(X)$ is not measurable with respect to $I(X)$, histories with the same interface value can require different optimal actions under a threshold loss.*

The proposition motivates no-gap controls. A certificate should receive credit only when it adds response-bearing information that cannot be recovered by recalibration, bin refinement, smoothing, or another function of the forecast interface alone.

3. Audit procedure

The audit evaluates a declared actionability claim. The analyst specifies the decision unit, forecast interface, action query, loss, and evidence source. The protocol then applies five gates: identification of q at the claimed level; prospective availability of the certificate; forecast competence relative to a degenerate baseline; no-gap controls using only I; and positive held-out regret reduction.

Audit gates. The operational report is short. It first states the unit, interface, query, loss, and evidence source. It then applies the following gates in order: identify q at the claimed unit; verify that certificate features are available before action; compare I with a degenerate forecasting baseline; run no-gap controls that use only transformations of I; cross-fit each prospective certificate; compute held-out $D_i = \ell_i(I) - \ell_i(I,C)$; and refine only if a multiplicity-adjusted lower confidence bound for $\mathbb{E}[D]$ is positive. If identification fails, the claim is lowered or response evidence is requested. If residual regret remains too high, the strong actionability claim is refused or lowered.

The test statistic is paired and decision-level specific. For held-out units,

$$D_i = \ell_i(I) - \ell_i(I,C), \qquad H_0 : \mathbb{E}[D] \leq 0. \tag{4}$$

The audit refines only when a lower confidence endpoint for $\mathbb{E}[D]$ is positive. We use paired bootstrap intervals in experiments; Appendix A gives a bounded-difference alternative.

4. Empirical audit

The experiments are organized by the strongest claim the evidence supports. Controlled and semi-real settings test whether the audit detects response information when ground truth is available. Real treatment/control settings test whether the same gates remain conservative when individual counterfactuals are unobserved. Figure 2 summarizes regret reductions; Figure 3 isolates the real-data confidence intervals.

Table 1. Audit summary. Real rows are reported at the decision level supported by the evidence, not promoted to individual counterfactual actionability.

Setting	Level	Interface I	Certificate C	Verdict and evidence
CVSim	individual	pressure forecast	response trajectory	REFINE: regret $0.246 \to 0.064$
ETTh1	semi-real window	forecast representation	response context	REFINE: regret $0.417 \to 0.019$
Tumor sim.	individual	tumor forecast	action head	REFINE: regret $0.222 \to 0.083$
X5 RetailHero	score cell	purchase-risk score	purchase history	REFINE: $0.324 \to 0.231$, CI $[0.077, 0.111]$
Drink Less	randomized user	app-open risk	user/context response	partial REFINE: $0.134 \to 0.098$, CI $[0.027, 0.045]$
myBPmyLife	participant-time	forecast/context	no deployable gain	EVIDENCE NEEDED / REFUSE
HeartSteps	participant slot	activity context	profile response	weak value; refuse strong claim
iFlex	aggregate event	load forecast	event response	LOWER CLAIM: aggregate only

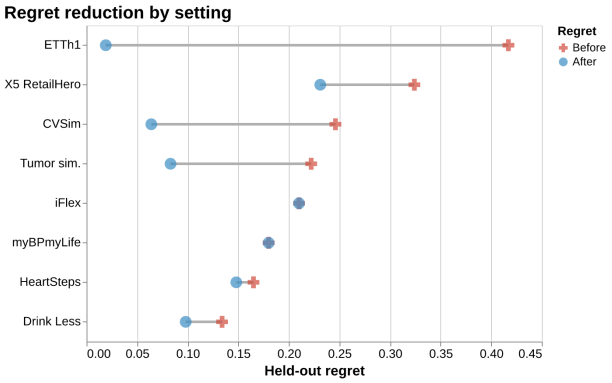

Figure 2. Held-out regret before and after certificate refinement.

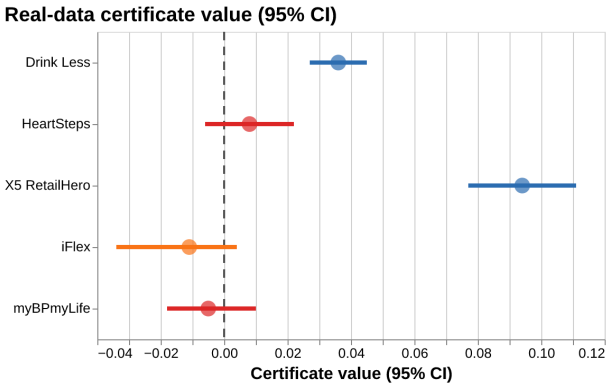

Figure 3. Real-data certificate value with 95% confidence intervals. Intervals to the right of zero support refinement at the declared level.

Real-data refinements. X5 RetailHero is the strongest real-data case. At a 3 percentage-point communication cost, a prospective purchase-history certificate reduces score-cell regret from 0.324 to 0.231, with paired value interval $[0.077, 0.111]$ and positive value on 29 of 30 splits. The supported claim is score-cell refinement under the treatment/control design. Drink Less gives a smaller randomized-user result: user/context certificates reduce regret from 0.134 to 0.098, with value interval $[0.027, 0.045]$ and positive value on 44 of 50 splits.

Refusals and lowered claims. myBPmyLife illustrates why refusal is a scientific output. Ordinary forecast/context features yield no deployable value, while an outcome-derived response row would close the gap. The audit marks that row as a non-deployable upper bound and returns response evidence needed. HeartSteps has weak positive value but high residual regret, so the strong actionability claim is refused. iFlex supports aggregate event-level reasoning, not individual household actionability.

No-gap controls and reporting. For every real row, we compare response-bearing certificates with interface-only controls: monotone calibration, bin refinement, local smoothing, and context expansions using only features already present in I. If such controls close the gap, the verdict is interface repair rather than certificate evidence. The reproducibility release should include split definitions, held-out D_i values, confidence intervals, no-gap control results, residual regret, and certificate provenance.

5. Related work and conclusion

The audit complements decision-focused learning, uplift modeling, treatment-rule estimation, and calibration. Those literatures train decision-aware predictors, estimate response from full covariates, or validate forecast probabilities. Our question is narrower and deployment-facing: given the interface already exposed to the decision maker, does it contain enough information for the declared action query? The answer is a statistical claim about an information interface, not a property inherited from predictive accuracy.

Forecasting is increasingly evaluated in agentic, economic, clinical, and operational settings where predictions trigger actions. This paper gives a decision-theoretic audit for that setting. A forecast interface is actionable only when

the query is identified, evidence is prospective, forecast-only repairs fail, and a certificate reduces held-out regret. Otherwise the correct output is to lower the claim, request response evidence, or refuse.

References

Amos, B. and Kolter, J. Z. OptNet: Differentiable optimization as a layer in neural networks. ICML, 2017.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. PNAS, 2016.

Blackwell, D. Equivalent comparisons of experiments. Annals of Mathematical Statistics, 1953.

Elmachtoub, A. N. and Grigas, P. Smart predict, then optimize. Management Science, 2022.

Ferns, N., Panangaden, P., and Precup, D. Metrics for finite Markov decision processes. UAI, 2004.

Klasnja, P. et al. Efficacy of contextually tailored suggestions for physical activity. Annals of Behavioral Medicine, 2019.

Kuenzel, S. R. et al. Metalearners for estimating heterogeneous treatment effects. PNAS, 2019.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for MDPs. ISAIM, 2006.

Nie, Y. et al. A time series is worth 64 words: Long-term forecasting with transformers. ICLR, 2023.

Qian, M. and Murphy, S. A. Performance guarantees for individualized treatment rules. Annals of Statistics, 2011.

Rossellini, R. et al. Can a calibration metric be both testable and actionable? arXiv, 2025.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. JASA, 2018.

Wilder, B., Dilkina, B., and Tambe, M. Melding the data-decisions pipeline. AAAI, 2019.

Zhou, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. AAAI, 2021.

A. Proof details

Proof of Theorem 1. Let $X = (Z, U)$ where Z and U are independent Bernoulli variables with parameter $1/2$. Set $Y = Z$, $q = U$, and $I(X) = Z$. Then I is a perfect forecast of Y. For binary actions with $L(a, q) = \mathbf{1}\{a \neq q\}$, a policy observing X chooses $a = U$ and has zero risk. A policy observing only $I = Z$ has no information about U and makes error with probability $1/2$. Thus $\mathcal{R}^*(I; q) = 1/2$, $\mathcal{R}^*(X; q) = 0$, and $\mathrm{DIG}(I, q) = 1/2$.

Held-out lower bound. Conditional on the fitted certificate and held-out fold, suppose $D_i \in [-B, B]$ are independent. Hoeffding's inequality gives

$$\Pr\left(\widehat{\Delta} - \mathbb{E}[D] \geq t\right) \leq \exp\left(-\frac{nt^2}{2B^2}\right). \tag{5}$$

Setting $t = B\sqrt{2\log(1/\alpha)/n}$ yields a valid lower confidence endpoint for $\mathbb{E}[D]$. When several certificate families are evaluated, we use Bonferroni adjustment or a separate selection fold.

B. Action-cost sensitivity

For X5 RetailHero, certificate value remains positive across communication costs from 2 to 6 percentage points.

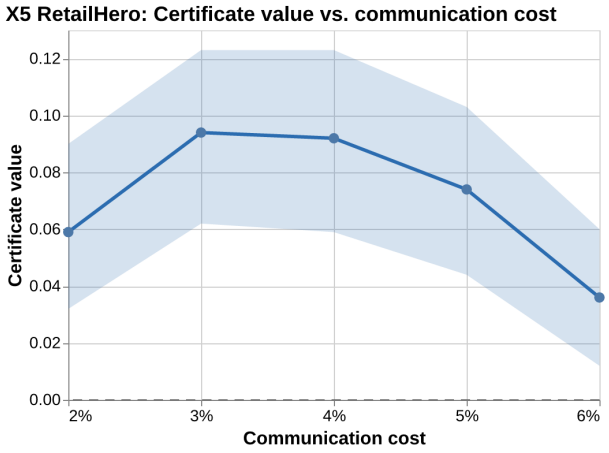

Figure 4. X5 score-cell certificate value across communication costs.

C. Reproducibility checklist

For each row, the release should include preprocessing scripts, split generators, interface construction code, candidate certificate construction code, no-gap control definitions, held-out D_i values, confidence intervals, residual regret, and the verdict-generation script. The central anti-leakage check is that every deployable certificate feature is available before the intervention occurs. Outcome-derived response information is reported only as an upper bound.