When Is a Veto Worth Its Cost? Proposer Precision, Gate Selectivity, and Requalification under Distribution Shift

Anonymous authors
Paper under double-blind review

Abstract

Downstream gates can reject, defer, or override proposals, but every veto can also suppress a useful action. We study requalification under distribution shift: a proposer recommends reactivating a previously rejected alternative and a downstream gate either authorizes the switch or keeps the incumbent. We operationalize the classical cost-sensitive/net-benefit break-even form using independently measured blocking rates and blocked-outcome severities, and compare learned gating with a switch-rate-matched random blocker. Across three controlled domains, calibration-only operating points predict holdout sign and magnitude better than a naive fixed-0.5 rule in the informative domains, while fresh audits show within-domain drift. A targeted severity learner fails: its low-precision benefit is mostly refusal rate, while an equal-rate oracle reveals unrealized selectivity. Degrading that oracle yields a representation target of Spearman $\rho \approx 0.68$, versus $\rho \approx 0.10$–0.14 for the learned signal. Fresh feature-tier and small-budget audits further show that better ranking and probability calibration do not necessarily improve frozen-threshold regret. A frozen external stress test on real handwritten-digit observations under constructed corruption shows the audit's own refusal condition: the break-even is weakly identified there, yet matched-rate selectivity remains detectable in three corruption levels. The contribution is an empirical operating audit that separates ranking, calibration, threshold portability, and selective action value.

1 Introduction

A prediction or control system is often followed by a second mechanism that can refuse, defer, or override its proposal. Selective classifiers abstain, learning-to-defer systems hand a case to another decision maker, and safety shields block unsafe actions. These mechanisms are valuable only when the errors they prevent exceed the useful proposals they suppress. Under distribution shift, that balance can change abruptly because the proposer itself can move from unreliable to highly reliable.

The threshold algebra itself is classical. Cost-sensitive learning chooses actions by conditional risk under asymmetric error costs (Elkan, 2001), while decision curve analysis evaluates whether a prediction rule has positive net benefit across threshold probabilities that encode relative harms (Vickers & Elkin, 2006). We do not claim normalized cost-ratio logic as new. Classical analysis nevertheless answers the deployment question once, offline, for a fixed cost matrix or elicited preference and typically a calibrated score. Three things can break in a shifting learned system: the break-even can move across regimes; score calibration can fail to transfer even when ranking does; and a learned gate can change which failures it blocks, not merely how many. Our experiments isolate these three departures and test where learning actually adds value.

We study this problem through requalification. At an earlier state, a system selected an incumbent and rejected alternatives. After the state changes, fresh evidence can make a rejected alternative look better. A proposer ranks the incumbent together with retained alternatives and recommends a switch; a downstream gate decides whether that reversal of a prior exclusion is authorized. The problem is narrower than generic abstention but structurally useful: blocking returns to a known incumbent, so the cost of a veto can be measured directly against the proposal it prevents.

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.

An early version of this project was motivated by a Qualified Residual Ledger (QRL): retain rejected alternatives and their historical context so that later systems can reconsider them rather than rediscover them. Adversarial controls changed the scientific conclusion. Fresh evidence, not rich history, is the dominant current signal; an exact-stationarity case is solved by a simple context threshold; and a learned history-augmented veto is not uniformly better than a simpler context+evidence gate. We therefore do not present QRL as the empirical contribution of this paper. Instead, QRL supplies the requalification setting, while the paper asks the broader ML question: when is a downstream veto worth its cost? A complementary functional analogy comes from biochemical proofreading and dynamical hysteresis: a system can preserve or reopen access to a pathway without implying that the pathway should be taken, and an additional selective step can trade discrimination against intervention cost. We use this analogy only to locate the architecture among established mechanisms of selective transition; no molecular mechanism or thermodynamic equivalence is claimed.

This operating-audit framing is distinct from robot-pathology and HRI failure research. Those literatures concern maladaptive phenotypes, attribution, trust, or intervention; here requalification is a proposed control action, and the downstream veto is evaluated only by paired task regret. Evidence for one problem is not treated as evidence for the other.

The answer has a simple decision-theoretic form but nontrivial empirical consequences. Let π denote the precision of the proposer's switch recommendations. A non-discriminating blocker that removes correct and wrong proposals equally is useful only below a cost-dependent threshold, equal to 0.5 under symmetric gains and losses. A discriminating gate can move that threshold upward by blocking wrong proposals more often, or by preferentially blocking especially harmful wrong proposals. This yields a direct baseline that is often missing in gated systems: a random blocker matched to the learned gate's intervention rate.

Our contributions are: (1) an outcome-measured operating decomposition (q_c, q_w, H_c, H_w) for a learned proposal-and-veto system; (2) a same-intervention-rate random control separating selective value from generic refusal; (3) proposer-precision intervention under a frozen gate with calibration-only holdout prediction; (4) fresh drift and threshold-portability audits; (5) a pre-specified severity-target intervention with matched-random, rank, and oracle diagnostics; (6) fresh feature-tier, risk-coverage, and small-budget target-calibration audits that distinguish probability repair from decision repair; and (7) a reusable six-step operating-audit protocol, stress-tested under a frozen pre-hashed real-data corruption shift where the break-even is weakly identified but matched-rate selectivity remains detectable at three corruption levels.

2 Proposal-and-Veto Theory

Consider episodes in which a proposer either keeps an incumbent a0 or proposes a switch a1. Conditional on proposing a switch, let π=P(C) be the probability that the proposal is correct, where ' correct' means that a1 improves true post-shift utility over a0 by more than a fixed margin δ. Let a gate block a correct proposal with probability qc and a wrong proposal with probability qw. If a correct proposal is blocked, let Hc$\geq$0 denote the mean benefit lost by reverting to the incumbent; if a wrong proposal is blocked, let Hw$\geq$0 denote the mean harm avoided. Let ρ be the proposal rate.

$$\Delta = \mathbb{E}[\text{regret}_{\text{gated}} - \text{regret}_{\text{proposer}}] = \rho\left[\pi q_c H_c - (1-\pi) q_w H_w\right].$$

Proposition 1 (operating break-even; classical cost-ratio form). The gate reduces expected regret iff

$$\pi < \pi^* = \frac{q_w H_w}{q_w H_w + q_c H_c}.$$

Proof. Unblocked proposals have identical loss under the gated and ungated systems. Blocking a correct proposal adds Hc, while blocking a wrong proposal removes Hw. Weighting the two events by their conditional probabilities and proposal rate gives $\Delta = \rho[\pi q_c H_c - (1-\pi) q_w H_w]$, and rearranging $\Delta < 0$ yields $\pi < \pi^*$.

The normalized ratio $\pi^* = q_w H_w/(q_w H_w + q_c H_c)$ is not new decision theory: it mirrors the optimal cost-sensitive threshold structure in Elkan (2001) and the net-benefit logic of decision curve

analysis (Vickers & Elkin, 2006). Our experiments instead treat q_c, q_w, H_c, H_w as measured operating quantities of a frozen learned gate rather than stipulated costs.

Corollary 1. If blocking is independent of correctness ($q_w = q_c$) and blocked correct/wrong proposals have symmetric magnitude ($H_w = H_c$), then π^*=0.5. Thus, within a fixed operating regime, an exact sign change at 0.5 is evidence for an approximately non-discriminating, symmetric-cost blocker—not a universal law. Conversely, $q_w > q_c$ or larger harm among the wrong proposals selected for blocking can raise π^* above 0.5. This is precisely what a useful learned gate should do.

This formulation motivates a switch-rate-matched random blocker: it removes exactly as many proposed switches as the learned gate, but uniformly and without correctness information. Failure to beat this control indicates conservatism rather than learned selectivity.

3 Requalification Benchmark

3.1 Episode structure

At t0, K candidates are evaluated under an old state; noisy observations select an incumbent. A bounded archive retains M=12 rejected candidates. At t1, the state changes and one inexpensive evidence observation is available. EvidenceRank proposes the candidate with the largest fresh score among the incumbent and archive. The primary learned gate, ContextEvidenceML, receives only the evidence gap, its standard-error proxy, evidence count, and signed/absolute context change. A history-augmented FullValidityVeto additionally receives the old score gap, old rank, descriptor distance, and a near-old-winner indicator. Both are histogram gradient-boosting classifiers trained on disjoint training episodes. Operational thresholds are selected on a validation split only and then frozen.

A proposed switch is correct if the chosen archived candidate improves true post-shift utility over the incumbent by more than δ=0.025; otherwise it is wrong. Regret is the difference between the utility of the globally best post-shift candidate and the executed candidate. The gate either accepts the proposal or retains the incumbent. This makes the gate's intervention cost observable episode by episode.

3.2 Three controlled domains

Domain	Candidates / shift	Purpose
Droplet	17 positions; precursor-exchange context	Paper-anchored surrogate of context-dependent abiological self-preservation.
Bio	20 mechanisms from four response families	Mechanism ambiguity under perturbation; new generator and new outcomes.
Generic	20 contextual policies with heterogeneous centers/widths/slopes	Broad stress test with tunable evidence noise and context-shift magnitude.

Table 1. Controlled benchmark domains, candidate sets, shifts, and experimental purpose.

The droplet domain is inspired by Egbert (2021)'s public motile-oil-droplet study, whose code varies precursor exchange over χP$\in$[0.1,12.5]. Our current benchmark does not execute that simulator; it uses a headless paper-anchored surrogate that reproduces only the qualitative change in preferred resource region. We therefore treat it as development scaffolding, not external biological validation.

The bio domain shares only a high-level mechanism-ambiguity motif with a prior conference benchmark (Anonymous, 2026). To preserve double-blind review, identifying bibliographic details are withheld. The present generator, episode identifiers, candidate utilities, learning objective, regret endpoint, gating experiments, figures, and tables are new; no numerical result, table, or figure from that prior work is reused. Full provenance can be disclosed to the chairs if requested.

3.3 Splits, threshold selection, and random-blocker control

Training, validation, ordinary shifted test, micro-shift test, exact-stationarity test, calibration, and controlled-precision holdout identifiers are disjoint. The primary gate threshold (0.30 for ContextEvidenceML; 0.40 for the history-augmented variant) was selected on validation under a risk-matched non-inferiority criterion before the analyses in this version. For each evaluated cell, the random blocker is matched after the learned gate is applied only in the number of proposed switches blocked; its blocked indices are sampled uniformly. The controlled-precision evaluation uses three domain-specific calibration pools and, within each domain, five non-overlapping 800-episode holdout mixtures.

3.4 Six-step operating-audit protocol

The experiments instantiate the following audit. The classical cost-ratio form is an input to the procedure rather than the claimed novelty.

Step	Operating audit
1	Freeze the proposer, gate, action rule, and outcome loss before target evaluation.
2	On a disjoint calibration set, measure q_c, q_w, H_c, H_w from realized outcomes; declare weak identification rather than imputing an absent blocked class.
3	Compute the gate-specific break-even π^* and predict holdout sign/magnitude before inspecting target outcomes.
4	Compare with a switch-rate-matched random blocker to separate selective value from generic refusal.
5	Under shift, re-audit score calibration and threshold portability separately from ranking quality.
6	Report ranking, calibration, threshold portability, and action value separately; retain failed or non-identifiable cases.

4 Three-Domain Controlled Proposer-Precision Test

The confirmatory design was motivated by an exploratory pass over 12 cells (Appendix C); all primary claims below rest on the disjoint 15-cell protocol.

4.1 Calibration-only break-even prediction

For each domain we fix one generator regime near balanced proposer precision. A calibration pool, disjoint from all evaluation episodes, estimates q_c, q_w, H_c, H_w for the frozen primary gate and thereby computes $\hat{\pi}^*$ before holdout mixtures are evaluated. Fresh proposer-switch episodes are partitioned into five non-overlapping 800-episode mixtures with target precisions $\{0.20, 0.35, 0.50, 0.65, 0.80\}$. Within a domain, the generator, gate, threshold, and class-conditional episode distributions are fixed; only the mixture proportion of correct and wrong proposer switches changes. This yields 4,000 evaluation episodes per domain and 12,000 total. Calibration predicts sharply different break-even points: $\hat{\pi}^*$=0.877 for droplet, 0.440 for bio, and 0.706 for generic. The same calibration quantities also predict an effect magnitude, $\widehat{\Delta} = (q_c H_c + q_w H_w)(\pi - \hat{\pi}^*)$, for these proposal-conditioned mixtures.

Across the 15 primary holdout cells, the calibration-derived sign agrees with 13/15 point estimates and all 8 decisive bootstrap cells. Pooled magnitude prediction is strong (Table 2), with clear within-domain improvement over the fixed-0.5 comparator in generic and bio; droplet is resolution-limited. A holdout-fitted break-even is reported only as an optimistic upper bound. Calibration recovers bio as the lowest-break-even domain but does not recover the exact cross-domain ordering, showing that the operating point is estimable but not a portable domain constant.

The fixed-0.5 comparator uses the same calibration-derived slope, while the holdout-fitted comparator is intentionally optimistic because it sees evaluation outcomes. The primary claim is limited to calibration-only improvement where the effect is resolvable.

Table 2: Calibration-only break-even model versus the fixed-0.5 comparator and a holdout-fitted upper bound across the 15 controlled-precision cells.

Break-even model	Direct R^2	MAE	$\lvert\Delta\rvert$ R^2 + domain FE	AIC
Calibration-only q/H	0.943	0.00186	0.717	-111.6
Fixed π^*=0.5 (same slope)	0.427	0.00556	0.501	-103.1
Holdout-fitted π^* upper (same slopes)	0.984	0.00092	—	—

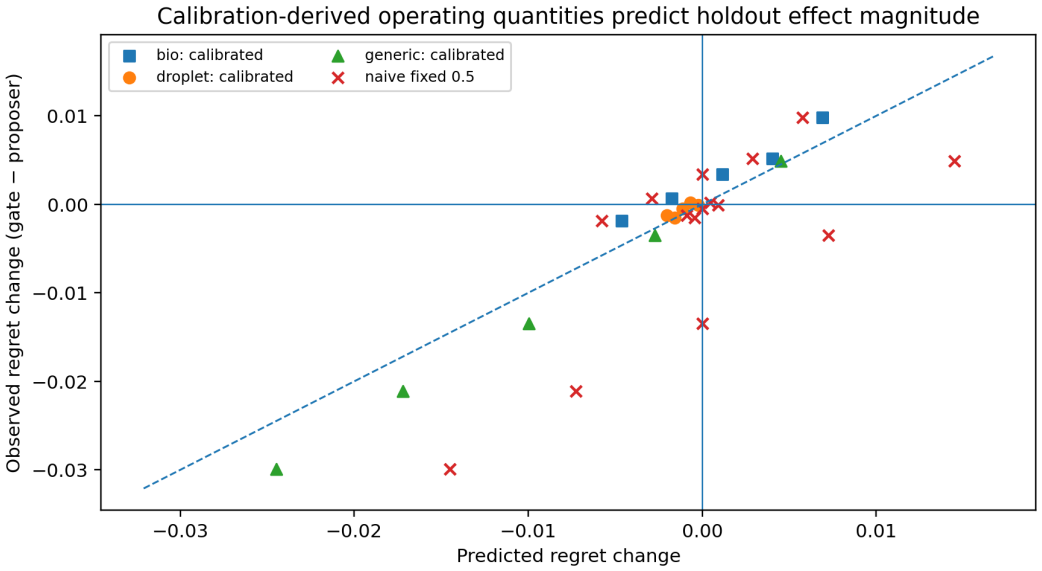

Figure 1: Observed regret change versus calibration-only prediction for the original 15 controlled-precision cells. The pooled fit is strong but scale-heterogeneous; domain-specific R^2/MAE and an in-sample fitted-break-even upper bound are reported in § 4.1 and Table 2. Crosses show the fixed-0.5 comparator with the same calibration-derived slope.

4.2 Targeted discrimination against a fixed 0.5 rule

The original grid was sparse where the calibrated and fixed-0.5 rules disagree. We therefore fixed disagreement intervals and primary target precisions before evaluating new independent seed ranges; one later cell per domain was rerun as replication.

Generic π=0.60 favors the calibrated rule (Δ=-0.00339, 95% CI [-0.00650,-0.00013]) and independently replicates (-0.00448 [-0.00764,-0.00129]); bio is directionally consistent but mixed on replication, and droplet remains unresolved because the predicted effect is at the bootstrap resolution limit. Full point estimates, confidence intervals, and replication details are retained in the supplement.

4.3 Learned selectivity is real but domain-dependent

Class selectivity alone is insufficient. In bio, the gate blocks wrong proposals more often than correct ones, yet the blocked wrong proposals are relatively mild while blocked correct proposals are valuable; the resulting severity-weighted terms reverse the apparent advantage. Droplet and generic have the opposite severity balance. Proposition 1 therefore requires both conditional blocking rates and consequence magnitudes, not proposer precision or q_w/q_c alone.

5 Controls That Bound the Claim

5.1 Fresh evidence dominates; history is not the current positive result

Permutation importance on candidate-level stale-exclusion prediction assigns by far the largest AP decrease to the fresh evidence gap (0.582), followed by absolute context change (0.0457), descriptor distance (0.0183), and old score gap (0.0149). A fresh 800-episode-per-domain feature-tier test sharpens the result. Descriptor/history features improve proposal AUROC from 0.770 to 0.837 in droplet and from 0.635 to 0.678 in bio, and improve the common-coverage regret curve in both. At validation-frozen thresholds, however, their regret differences from ContextEvidenceML are -0.00019, $+0.00047$, and -0.00034 in droplet, bio, and generic, with all paired 95% bootstrap intervals crossing zero. Current history helps ranking in some domains but does not establish a stable action-level advantage. Typed rejection reasons, state-version predicates, age, and provenance remain untested variables rather than relabeled current features.

5.2 Retention and requalification are separate bottlenecks

A gate cannot revive an alternative that the archive failed to retain. At M=5, the QRL-inspired archive-oracle regret is 0.00034 in droplet but 0.127 in bio and 0.131 in generic. At M=12 those latter ceilings fall to 0.0283 and 0.0140. The retention heuristic is not uniformly best against random or k-center alternatives, so we treat retention policy and gate quality as distinct questions rather than attributing both to one mechanism.

5.3 Negative controls: exact stationarity and delay matching

Exact stationarity is not a victory for learning: a hand rule that requires nonzero context change and an evidence gap makes zero unnecessary switches in all three stationary tests. Conversely, an earlier sequential LearnedDeferred policy appeared beneficial only because it waited for more observations. A DelayMatchedRank baseline collects the same number of evidence rounds and then ranks without the learned deferral logic; it reduces regret relative to LearnedDeferred by 0.0164 (droplet), 0.1242 (bio), and 0.1878 (generic), with all paired bootstrap intervals excluding zero. We retain both failures because they identify the mechanism more sharply.

6 What Learning Adds Beyond a Fixed Cost Table

Classical cost-sensitive and net-benefit analysis supplies the correct threshold algebra. The empirical question is whether the operating quantities of a learned veto remain valid after shift and whether the gate targets consequential errors rather than merely refusing more often. Sections 6.1-6.4 test those two requirements.

6.1 Regime drift makes a one-time operating point expirable

Fresh disjoint tiny/small/moderate-shift phases show that a one-time operating point can expire within a domain. Regime-specific re-estimation improves pooled prediction overall, but not uniformly: sparse gate actions and finite-sample variance can make local q/H estimates noisy. The result is a bias-variance warning, not a claim that continual recalibration is always superior; detailed ranges and fits are in Appendix D.

6.2 Ranking is not threshold portability

Leave-one-domain-out tests show that ranking can remain useful while deployment thresholds fail to transfer. The fresh small-budget audit sharpens this distinction. With 128 labeled target proposals, target Platt calibration reduces Brier error from 0.280 to 0.226 in droplet, 0.310 to 0.224 in bio, and 0.153 to 0.033 in generic. Yet the resulting outcome-tuned threshold improves mean test regret only in generic (0.0295 to 0.0246); droplet and bio remain slightly worse than their source-only thresholds. Ranking, probability calibration, and action-threshold portability are therefore separate deployment properties (Appendix F).

6.3 SEVERITY-AWARE LEARNING: A PRE-SPECIFIED REPAIR ATTEMPT THAT DOES NOT FULLY WORK

The severity asymmetry identified above motivates a pre-specified bio intervention: the proposer and decision-time features remain frozen, but the gate target changes from correctness to realized switch gain.

The strong prediction fails. Gain magnitude is not predictable (validation R^2=-0.041), although weak ordinal information survives. The learned severity gate moves the operating point toward symmetry but does not cross the pre-specified target; Table 3 reports the full operating decomposition.

Table 3: Bio objective intervention on fresh disjoint data. The severity objective changes the operating point but does not achieve the pre-specified >0.5 repair; matched-random significance for the severity gate uses 10,000 same-block-count draws.

Method	Validation signal	$q_c H_c$	$q_w H_w$	$\hat{\pi}^*$	Cells $\Delta < 0$	Sig. < random
Probability gate	P(correct), AUROC .602	0.00871	0.00559	0.391	2/5	1/5
Severity gate	gain R^2 -.041; ρ .135; AUC+ .579	0.08978	0.07676	0.461	3/5	2/5
Cost-weighted clf.	|gain|-weighted block label	—	—	~0.464	2/5	0/5

A 10,000-draw same-rate random audit shows that the large low-precision benefit is mostly a refusal-rate effect. At higher precision, modest selectivity can coexist with net harm relative to no veto. Selection quality and deployment value are therefore distinct.

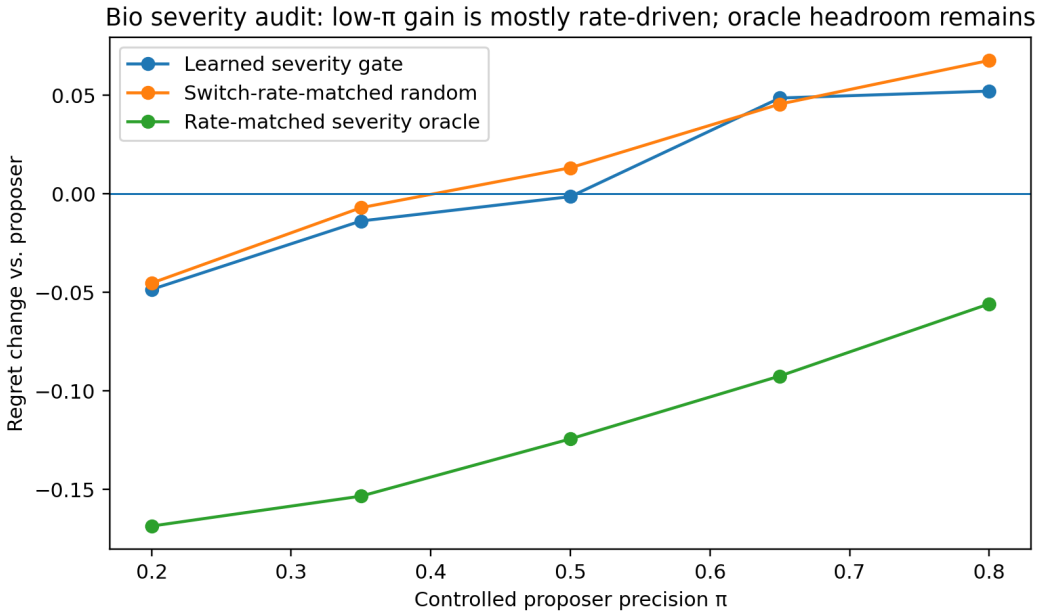

Figure 2: Bio severity audit. The learned severity gate is compared with a switch-rate-matched random blocker and a diagnostic same-rate severity oracle. The large benefit at π=0.20 is statistically indistinguishable from random blocking at the same rate, whereas the oracle remains beneficial across all five cells. At high proposer precision the learned gate can retain selectivity relative to random yet still be harmful relative to no veto.

A same-rate oracle is beneficial in all five bio cells, while a cost-weighted classifier does not rescue the learned result. The failure is therefore not lack of headroom; it is inadequate severity information in the current representation.

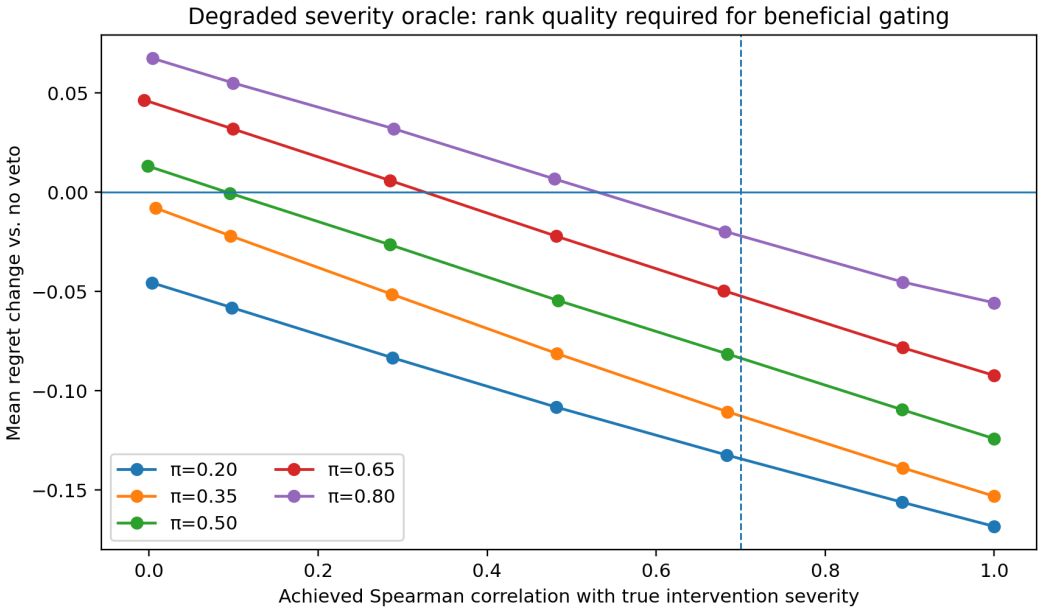

Figure 3: Degraded severity oracle. With the learned severity gate's block count fixed, increasingly noisy access to true intervention severity is used only as a diagnostic sensitivity analysis. A common target latent rank correlation of 0.5 achieves mean Spearman $\rho\approx0.48$ but remains slightly harmful at π=0.80 (mean $\Delta\approx+0.0067$). At target 0.7, achieved mean $\rho\approx0.68$ and mean regret change is beneficial in all five controlled-precision cells. The current learned severity representation reaches only $\rho\approx0.10$–0.14, quantifying the representation gap rather than merely asserting oracle headroom.

6.4 Oracle break-even and a degraded-oracle representation specification

The rate-matched oracle is consistent with Proposition 1 because the break-even is gate-specific: an oracle that almost never blocks beneficial switches has a break-even near one (Table 4). We then degrade oracle severity ranking while fixing the learned gate's block count to estimate how much ordering information is needed for useful selective veto.

Controlled π	**Oracle block rate**	**Oracle π^***	**First tested Spearman with mean benefit**
0.20	0.250	1.000	≈0.00
0.35	0.293	1.000	≈0.01
0.50	0.217	1.000	≈0.10
0.65	0.230	1.000	≈0.48
0.80	0.220	0.997	≈0.68

Table 4. Gate-specific oracle break-even and degraded-oracle sensitivity. The first three oracle break-evens are exactly one because no correct proposals are blocked; π=0.80 remains 0.997. The last column is a diagnostic specification, not a learned-model result: true intervention severity ranks are degraded with Gaussian-copula noise while the block count is held equal to the learned severity gate.

The degraded-oracle audit therefore turns headroom into a representation specification: across all five bio cells, the first tested shared level with beneficial mean effect is about Spearman $\rho\approx0.68$, far above the current learned signal.

7 Claim-Evidence Audit

The full claim-by-claim ledger is retained in Appendix B. The audit records both positive and negative findings rather than promoting only successful hypotheses: the fixed 0.5 break-even is contradicted

by domain-specific operating points; class selectivity without severity is insufficient in bio; the severity-target repair fails despite oracle headroom; and descriptor/history or target calibration can improve ranking or Brier score without a stable action-level gain. These failures delimit the positive result—realized operating quantities and matched-rate controls can measure when selective veto adds value and when the audit should refuse a portable threshold.

8 Relation to Prior Work

The closest formal lineage is cost-sensitive decision theory and net-benefit analysis (Elkan, 2001; Vickers & Elkin, 2006), selective classification/defer/shield mechanisms (Chow, 1970; Geifman & El-Yaniv, 2017; Madras et al., 2018; Mozannar & Sontag, 2020; Alshiekh et al., 2018), and calibration under distribution shift (Ovadia et al., 2019; Liang et al., 2024). Those traditions provide threshold, abstention, and calibration machinery; our contribution is the outcome-measured operating audit, including the matched-rate control and controlled proposer-precision intervention.

Proofreading, hysteresis, hybrid dynamics, and machine-behavior/HRI literatures supply useful structural or evaluative analogies, but no molecular, dynamical, pathology, or human-trust equivalence is claimed. The extended comparison and citations are retained in Appendix E.

9 Limitations and Scope

Controlled mixtures permit causal intervention on proposer precision while holding the frozen gate and class-conditional outcome distributions fixed; natural observational data do not provide that sweep. External generalization nevertheless remains open. Appendix G therefore stress-tests the audit on real handwritten-digit observations under constructed corruption shift; this is portability evidence, not causal or deployment validation. Break-even is an operating-regime quantity rather than a portable domain constant, and sparse blocks can make it weakly identified. The feature tiers still lack typed rejection predicates, event age, and cryptographic state-version binding. We do not establish a QRL-specific effect, biological equivalence, robot-pathology taxonomy, or molecular/dynamical mechanism.

10 Conclusion

A veto is not free. Classical cost-sensitive logic gives the threshold form; deployment value depends on realized selectivity and consequences. Calibration-only measurements beat a fixed-0.5 rule where effects are resolvable, while fresh audits show operating-point drift. Feature-tier, target-budget, and severity audits show that better ranking, Brier score, or apparent gain need not improve selective action value. The six-step operating audit therefore separates ranking, calibration, threshold portability, and consequence-weighted action value. The positive result is that these quantities are separately measurable: calibration-only operating estimates predict holdout decision value where effects are resolvable, while the frozen external Digits audit correctly refuses a weakly identified break-even yet still detects matched-rate selectivity at three corruption levels.

Reproducibility Statement

All results use fixed scripts with disjoint training, validation, calibration, and evaluation identifiers. The supplement contains generators, frozen thresholds, matched-rate controls, primary and fresh audits, Appendix-G external-audit protocols/results, figures, and SHA-256 manifests. The final strengthening script reproduced byte-identical CSV outputs; the droplet benchmark is a surrogate; oracle sensitivity uses fixed block counts and 300 draws per non-perfect correlation level.

AI Use Statement

Generative AI tools were used to propose and refine hypotheses, critique experimental design, implement and debug experiment scripts, derive and check mathematical formulations, interpret numerical outputs, and draft/edit manuscript text. AI-assisted code and claims were not accepted at

face value: numerical claims in the manuscript were recomputed from saved artifacts; critical scripts were rerun; result-file collisions and reporting mismatches were audited; matched baselines and independent-seed holdouts were added after adversarial review; and the authors remain responsible for all final claims, code, citations, and disclosures. Primary controlled benchmarks are simulated or synthetic; Appendix G additionally uses a public, non-sensitive handwritten-digit dataset under constructed corruption. No human-subject or confidential data were generated or analyzed by an AI tool.

REFERENCES

E. Alpaydin and C. Kaynak. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. doi:10.24432/C50P49.

M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe Reinforcement Learning via Shielding. In *AAAI*, pp. 2669–2678, 2018.

Anonymous. Prior virtual-cell conference benchmark. Identifying bibliographic details withheld for double-blind review, 2026.

J.-P. Aubin. A survey of viability theory. *SIAM Journal on Control and Optimization*, 28(4):749–788, 1990. doi:10.1137/0328044.

M. S. Branicky. Multiple Lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):475–482, 1998. doi:10.1109/9.664150.

H. R. Cameron, S. Castle-Green, M. Chughtai, et al. A Taxonomy of Domestic Robot Failure Outcomes: Understanding the Impact of Failure on Trustworthiness of Domestic Robots. In *Second International Symposium on Trustworthy Autonomous Systems*, pp. 1–14, 2024. doi:10.1145/3686038.3686050.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi:10.1109/TIT.1970.1054406.

M. Egbert. Self-preserving mechanisms in motile oil droplets: a computational model of abiological self-preservation. *Royal Society Open Science*, 8(12):210534, 2021. doi:10.1098/rsos.210534.

C. Elkan. The Foundations of Cost-Sensitive Learning. In *IJCAI*, pp. 973–978, 2001.

Y. Geifman and R. El-Yaniv. Selective Classification for Deep Neural Networks. In *NeurIPS 30*, 2017.

T. A. Henzinger. The theory of hybrid automata. In *11th Annual IEEE Symposium on Logic in Computer Science*, pp. 278–292, 1996. doi:10.1109/LICS.1996.561342.

S. Honig and T. Oron-Gilad. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology*, 9:861, 2018. doi:10.3389/fpsyg.2018.00861.

J. J. Hopfield. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *PNAS*, 71(10):4135–4139, 1974. doi:10.1073/pnas.71.10.4135.

IUPAC. Catalyst. In *Compendium of Chemical Terminology (the Gold Book)*, 5th ed., 2025. doi:10.1351/goldbook.C00876.

H. Liang, L. Peng, and J. Sun. Selective Classification Under Distribution Shifts. *Transactions on Machine Learning Research*, 2024.

D. Madras, T. Pitassi, and R. Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *NeurIPS 31*, 2018.

H. Mozannar and D. Sontag. Consistent Estimators for Learning to Defer to an Expert. In *ICML*, pp. 7076–7087, 2020.

A. Murugan, D. A. Huse, and S. Leibler. Speed, dissipation, and error in kinetic proofreading. *PNAS*, 109(30):12034–12039, 2012. doi:10.1073/pnas.1119911109.

J. Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57(5):587–595, 1975. doi:10.1016/S0300-9084(75)80139-8.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *NeurIPS 32*, 2019.

I. Rahwan, M. Cebrian, N. Obradovich, et al. Machine behaviour. *Nature*, 568:477–486, 2019. doi:10.1038/s41586-019-1138-y.

E. Roesler, M. Vollmann, D. Manzey, and L. Onnasch. The dynamics of human–robot trust attitude and behavior: Exploring the effects of anthropomorphism and type of failure. *Computers in Human Behavior*, 150:108008, 2024. doi:10.1016/j.chb.2023.108008.

P. Sartori and S. Pigolotti. Kinetic versus energetic discrimination in biological copying. *Physical Review Letters*, 110(18):188101, 2013. doi:10.1103/PhysRevLett.110.188101.

W. Sha, J. Moore, K. Chen, A. D. Lassaletta, C.-S. Yi, J. J. Tyson, and J. C. Sible. Hysteresis drives cell-cycle transitions in Xenopus laevis egg extracts. *PNAS*, 100(3):975–980, 2003. doi:10.1073/pnas.0235349100.

A. J. Vickers and E. B. Elkin. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, 26(6):565–574, 2006. doi:10.1177/0272989X06295361.

N. Watson and A. Hessami. Psychopathia Machinalis: A Nosological Framework for Understanding Pathologies in Advanced Artificial Intelligence. *Electronics*, 14(16):3162, 2025. doi:10.3390/electronics14163162.

W. Xiong and J. E. Ferrell Jr. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature*, 426:460–465, 2003. doi:10.1038/nature02089.

X. Zhang, S. K. Lee, W. Kim, and S. Hahn. "Sorry, it was my fault": Repairing trust in human-robot interactions. *International Journal of Human-Computer Studies*, 175:103031, 2023. doi:10.1016/j.ijhcs.2023.103031.

A Derivation of Proposition 1

Condition on episodes where the proposer recommends a switch. With probability π the proposed switch is correct. The gate blocks such a proposal with probability qc, in which case returning to the incumbent increases regret by Hc on average. With probability $1-\pi$ the proposal is wrong. The gate blocks it with probability qw, decreasing regret by Hw on average. Multiplying by proposal rate ρ yields $\Delta=\rho[\pi\text{qcHc}-(1-\pi)\text{qwHw}]$. Because $\rho\geq 0$, $\Delta<0$ iff $\pi\text{qcHc}<(1-\pi)\text{qwHw}$, which rearranges to Proposition 1. For a matched random blocker, qc=qw in expectation. A learned gate can exceed the random blocker by class discrimination (qw>qc), severity discrimination (larger Hw and/or smaller Hc among blocked proposals), or both.

B Claim-Evidence Audit v0.17

Claim	Status	Evidence
A universal π=0.5 threshold governs veto value	Contradicted	Calibration-only $\hat{\pi}^*$ values vary 0.440–0.877; fixed-0.5 quantitative prediction R^2=0.427 vs. 0.943 for calibrated operating quantities.
Calibration-derived operating quantities predict holdout effect magnitude	Supported in generic/bio; droplet resolution-limited	Pooled R^2=0.943/MAE=0.00186. Within-domain calibrated vs fixed-0.5 R^2: generic 0.925 vs -0.074; bio 0.645 vs 0.233. Droplet is numerically better (0.163 vs -0.152) but predicted effects are at the bootstrap resolution limit, so it is not counted as a third informative improvement.
$\hat{\pi}^*$ adds information beyond a fixed 0.5 break-even	Supported in generic; directional bio; droplet near-null	Generic π=0.60 decisively favors calibrated $\hat{\pi}^*$=0.706 and independently replicates. Bio π=0.45 is decisively harmful once with same-sign non-decisive replication. Droplet π=0.80 has predicted $\lvert\Delta\rvert\approx$0.00023, comparable to CI half-width$\approx$0.00027, and remains unresolved.
Learned gate is only a conservatism dial	Contradicted in 9/15 original cells	Gate beats switch-rate-matched random blocking significantly in 5/5 generic and 4/5 droplet original cells; not supported in bio.
Class selectivity q_w>q_c is sufficient	Contradicted	Bio has q_w/q_c=1.85 but q_wH_w=0.00848<q_cH_c=0.01077 and $\hat{\pi}^*$=0.440; severity determines the operating threshold.
Fresh evidence dominates prediction	Supported	Permutation AP importance 0.582 versus 0.0457 for context_abs and smaller history features.
Rich QRL history is necessary	Not supported	Context+evidence gate is at least as strong; typed QRL fields are not yet tested.
Cross-domain ranking implies threshold portability	Contradicted	LODO AUROC/AP can remain high while operational thresholds diverge.
Retention and requalification are one bottleneck	Contradicted	Archive capacity imposes a separate oracle ceiling in bio/generic.
Learned deferral explains waiting gains	Contradicted	Delay-matched ranking outperforms LearnedDeferred in all three domains.
Egbert simulator reproduced	Not supported	Current droplet domain is a paper-anchored surrogate only.
Calibration exactly recovers the holdout-optimal break-even	Not supported	Not supported. Holdout-fitted break-evens are 0.709/0.322/0.761 versus calibration 0.877/0.440/0.706. Errors are +0.168/+0.118/-0.055 and the droplet/generic ordering reverses.
Proposition 1 is novel decision theory	Contradicted / not claimed	The normalized threshold form follows the classical cost-sensitive/net-benefit lineage (Elkan, 2001; Vickers & Elkin, 2006). Novelty is empirical operationalization for a learned gate, matched-rate random blocking, and controlled proposer-precision intervention.

Claim	Status	Evidence
Continual regime-specific recalibration is always better than one-time calibration	Not supported	Fresh exploratory drift audit: overall R^2 0.821 vs 0.506 and MAE 0.00322 vs 0.00431 favor regime-specific estimates, but pooled calibration is better in droplet/generic and regime-specific re-estimation is better in bio. Operating drift exists, but estimation variance prevents a universal claim.
Break-even is constant within a domain	Contradicted / conditional	Fresh regime estimates move within droplet (0.500-0.758) and generic (0.655-0.834); bio moderate is weakly identified. Regime-specific re-estimation improves pooled nine-cell prediction overall but not every domain.
Severity-aware training repairs bio anti-selection	Contradicted	Gain regression validation R^2=-0.041; validation/test Spearman ρ=0.135/0.104 and AUROC(gain>0)=0.579/0.543. $\hat{\pi}^*$ moves 0.391→0.461 but does not exceed 0.5. At π=0.20, the large benefit does not beat 10,000-draw same-rate random blocking.
Exploitable severity-selective headroom exists	Supported as oracle upper bound	A rate-matched oracle blocks the same number of proposals using realized switch value and lowers regret in 5/5 bio precision cells; at π=0.80 it changes Δ from +0.0521 (learned) to -0.0558 (oracle).
Large low-π severity gain demonstrates learned selectivity	Contradicted	At π=0.20, learned Δ=-0.0485 versus same-rate random mean -0.0453; p=0.335. Most of the apparent gain is refusal rate.
Severity gate has no ordinal information	Contradicted weakly	Magnitude fit fails, but validation/test Spearman ρ=0.135/0.104 and AUROC(gain>0)=0.579/0.543 show weak ordering signal.
Current paper validates robot pathology	Not claimed / unsupported	No pathology diagnosis, mental-state inference, human trust, abnormality, or observer-intervention endpoint is measured; robot pathology is treated as an upstream complementary program.
Descriptor/history ranking gain implies frozen-threshold regret gain	Contradicted	Fresh test: AUROC and common-coverage curves improve in droplet/bio, but paired regret intervals include zero in all domains.
Small target calibration is sufficient for portable action thresholds	Contradicted / conditional	At 128 labels Brier falls 19%–79%, but regret improves only in generic; droplet/bio remain slightly worse than source-only.

C Exploratory 12-Cell Analysis

This appendix preserves the exploratory 12-cell analysis and matched-rate control that motivated the confirmatory protocol; no primary claim depends on these cells.

We first re-analyzed 12 previously generated cells: three ordinary shifted tests plus tiny/small/moderate shift bins in each domain. These cells were exploratory; the precision pattern was discovered after the cells existed and is not presented as confirmatory evidence.

Across the 12 cells, proposer precision and relative regret change have Spearman ρ=0.916 (p=2.8e-05); the benefit sign matches the empirical π=0.5 split in 12/12 cells. The mechanism becomes clearer after conditioning on what the gate blocks. For the history-augmented veto, the median conditional block rate is 0.688 for wrong proposals versus 0.315 for correct proposals. Therefore the gate is not, in aggregate, blocking the two classes equally.

The apparent pooled split around π=0.5 should not be interpreted as a common break-even: these exploratory cells mix domains and regimes whose later calibration-only estimates span $\hat{\pi}^*$=0.440–0.877, while the learned gate itself is strongly class-selective.

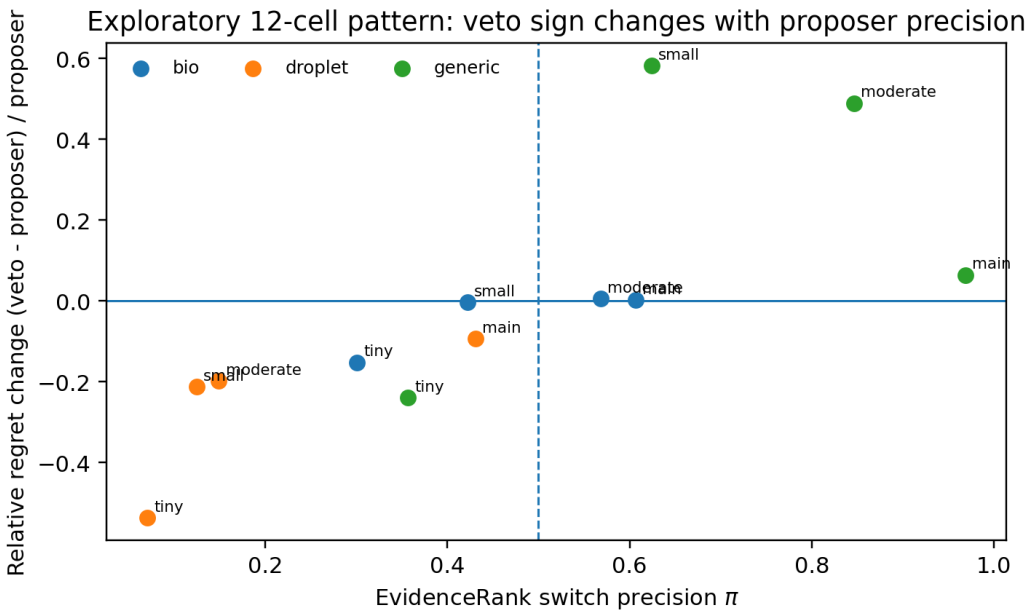

Figure 4: Exploratory 12-cell relationship. Negative relative regret change favors the history-augmented learned veto. The sign matches π<0.5 in all 12 cells, but Proposition 1 shows that 0.5 is only a symmetric special case.

C.1 Matched random blocking distinguishes learning from conservatism

The random-blocker baseline changes the interpretation substantially. Matching only the number of vetoes, the history-augmented learned gate has lower regret than the mean matched-random blocker in 11/12 exploratory cells and is significantly better by one-sided randomization in 8/12. Repeating the same audit with the simpler ContextEvidenceML gate yields the same 11/12 and 8/12 counts. This establishes a positive result that the stationary threshold control could not: the learned gates contain information about which proposals to block, not merely how many to block.

D Calibration, Transfer, and Drift Details

Domain	$\hat{\pi}^*$	qcHc	qwHw	Orig. sign	Sig. < random	n
Droplet	0.877	0.00037	0.00263	4/5	4/5	4,000
Bio	0.440	0.01077	0.00848	4/5	0/5	4,000
Generic	0.706	0.01420	0.03414	5/5	5/5	4,000
Total	—	—	—	13/15	9/15	12,000

Table D1. Calibration-only operating quantities and matched-random summary for the 15 controlled-precision cells.

Held-out	AUROC	AP	Brier	Source th.	Diagnostic th.
droplet	0.950	0.715	0.060	0.41	0.40
bio	0.785	0.737	0.228	0.44	0.30
generic	0.937	0.865	0.190	0.35	0.05

Table D2. Leave-one-domain-out ranking, calibration, and threshold-transfer diagnostics.

Leave-one-domain-out ranking remains useful, but operational calibration does not transfer cleanly. The strongest example is generic: held-out AUROC is 0.937 and AP 0.865, yet a source-selected

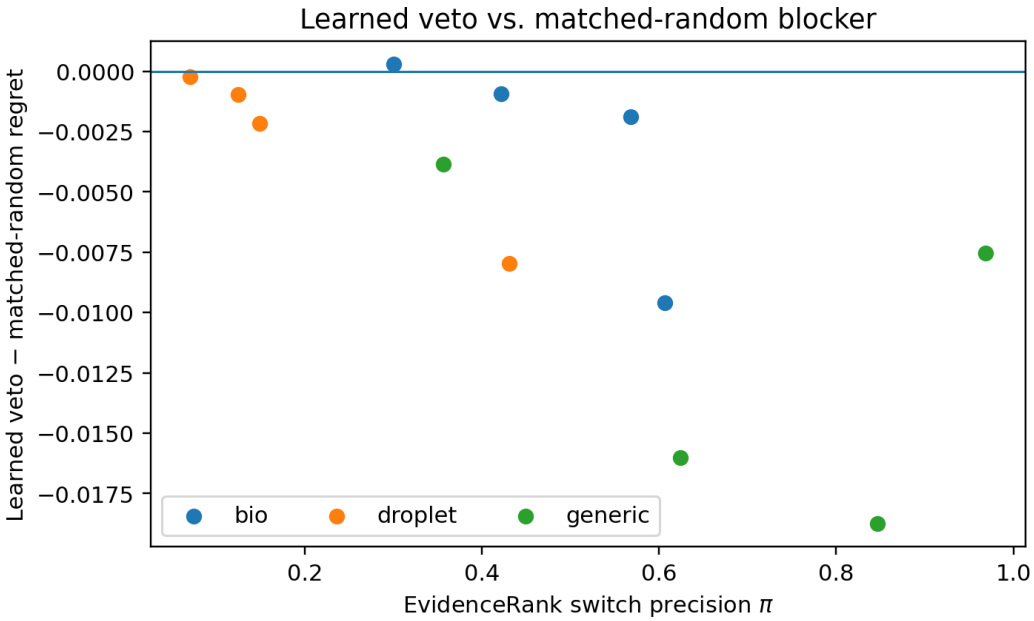

Figure 5: Exploratory comparison to a switch-rate-matched random blocker. Values below zero indicate that the learned history-augmented veto blocks a more useful subset of proposals than random blocking at the same intervention rate.

threshold of 0.35 is far from a diagnostic threshold of 0.05 computed from held-out labels. This matters for the classical cost-sensitive prescription: ranking quality can survive while the score calibration needed for a portable threshold does not. Our operating audit therefore treats gate scores as decision inputs, not calibrated posterior probabilities.

Exploratory regime-drift audit. Using fresh disjoint identifiers, we re-estimated operating quantities in tiny, small, and moderate shift phases. These nine cells are a separate evaluation set from the 15 controlled-precision holdout cells in § 4.1, so the two pooled R^2 values are not directly comparable. Break-even estimates move within domain: droplet 0.500-0.758, generic 0.655-0.834, while bio becomes weakly identified in the moderate phase because the frozen gate almost never blocks and estimated avoided harm approaches or crosses zero. Across the nine fresh cells, regime-specific estimates improve pooled prediction overall (R^2=0.821 vs. 0.506; MAE 0.00322 vs. 0.00431), but not in every domain: one-time pooled calibration is better for magnitude prediction in droplet and generic, while regime-specific re-estimation is substantially better in bio. Importantly, droplet's regime-specific range (0.500-0.758) brackets the holdout-fitted optimum 0.709 whereas the one-time primary estimate 0.877 does not. Re-estimation can therefore improve the location of the break-even even when sampling variance worsens pointwise magnitude prediction. We treat this as a bias-variance result, not as evidence that continual recalibration is universally superior.

Domain	Primary $\hat{\pi}^*$	Fresh regime $\hat{\pi}^*$	Fresh: regime R^2	Fresh: pooled R^2
Droplet	0.877	0.500-0.758	-2.284	-0.328
Bio	0.440	0.466-0.468*	0.788	0.085
Generic	0.706	0.655-0.834	0.780	0.888

Table D3. Fresh within-domain drift audit. *Bio moderate is omitted from the displayed range because the gate almost never blocks and the estimated avoided-harm term becomes weakly identified. The nine fresh cells are distinct from the 15 primary controlled-precision cells. Negative R^2 in droplet reflects very small effects and high estimation variance, not absence of movement in the estimated break-even.

E Prior-Work Comparison Table

Lineage	Primary object	Typical endpoint	Relation here
Machine behaviour	Artificial agents as behavioral objects	Behavioral regularities in context	Broad methodological umbrella
Machine psychopathology	Persistent maladaptive AI patterns	Nosology / diagnostic criteria	Upstream phenotype taxonomy; not tested
HRI failure / trust	Robot failure in interaction	Attribution, trust, repair	Human-centered consequence; no human endpoint here
Present operating audit	Proposer + downstream veto	Paired regret, selectivity, π^*	Downstream intervention-policy audit

Table E1. Comparison with adjacent machine-behavior, machine-pathology, and HRI-failure lineages.

E.1 Extended conceptual relation to prior work

The closest formal lineage is cost-sensitive decision theory and net-benefit analysis (Elkan, 2001; Vickers & Elkin, 2006): those traditions specify decisions once costs or utilities are given. Here the costs are empirical properties of a particular learned gate-task operating pair, estimated from realized conditional blocks and their consequences.

Selective classification, learning to defer, and safety shielding likewise allow abstention or intervention (Chow, 1970; Geifman & El-Yaniv, 2017; Madras et al., 2018; Mozannar & Sontag, 2020; Alshiekh et al., 2018). Our distinction is evaluative: the same-rate random control asks whether a learned veto targets better cases than generic refusal, while proposer precision is manipulated independently.

Calibration degradation under distribution shift is well established (Ovadia et al., 2019; Liang et al., 2024). Our LODO and fresh-regime audits use that fact as a boundary condition: preserved ranking does not imply a portable action threshold.

The droplet domain is a paper-anchored surrogate inspired by Egbert (2021), not a reproduction of the original simulator. The bio domain likewise uses a new generator and disjoint evaluation; neither domain supports a claim of external biological validation.

E.2 Catalysis, kinetic proofreading, and selective gating

Biochemical analogies are functional, not mechanistic. Catalysis changes reaction kinetics without changing thermodynamic ordering (IUPAC, 2025); similarly, retention can lower the barrier to reconsideration without making a transition warranted. Kinetic proofreading adds selective stages that trade error reduction against time and dissipation (Hopfield, 1974; Ninio, 1975; Murugan et al., 2012; Sartori & Pigolotti, 2013). Our analogue is only evaluative: the matched-rate blocker asks whether a gate suppresses consequentially wrong transitions better than generic refusal. Thus the comparison motivates selectivity under intervention cost, not molecular equivalence or ' more filtering' as a virtue.

E.3 Cellular memory, hysteresis, and hybrid dynamical switching

Cellular bistability and hysteresis provide a structural precedent for history-dependent transitions (Xiong & Ferrell, 2003; Sha et al., 2003). In such systems, the same present input can support different state transitions because prior state changes the effective transition landscape. Requalification has the same abstract temporal form but makes the history explicit: a previously rejected candidate is retained together with exclusion context and can re-enter evaluation after a context change. This is stronger than ordinary replay but weaker than a biological-mechanism claim; our experiments do not identify molecular memory, and fresh evidence—not rich history—is the dominant positive signal in the present benchmark.

Hybrid and switched systems formalize mode changes, guards, stability, and admissible trajectories, while viability theory asks whether trajectories can remain within acceptable state sets (Henzinger, 1996; Branicky, 1998; Aubin, 1990). Those frameworks principally characterize when a transition is dynamically admissible. Our question is orthogonal: once a transition has already been proposed, is an additional learned authorization layer worth its realized false blocks and avoided harms? Thus

dynamic admissibility, requalification, and authorization are distinct stages rather than competing formalisms.

Taken together, the lineages are complementary rather than substitutable: cost-sensitive theory supplies asymmetric consequence algebra; proofreading supplies selective checking; hysteresis supplies path dependence; and hybrid dynamics supplies mode-transition structure. What is specific here is a history-bearing requalification pipeline—preserve → requalify → propose → veto → act—whose downstream veto is audited from realized outcomes under shift. The novelty claim is therefore not ' digital catalysis' or a new threshold formula, but an empirical method for deciding whether a learned rejector adds selective deployment value beyond its refusal rate.

E.4 Relation to robot pathology and HRI failure research

Machine-behavior, machine-pathology, and HRI failure research study behavioral phenotypes, diagnosis, attribution, trust, and repair (Rahwan et al., 2019; Watson & Hessami, 2025; Honig & Oron-Gilad, 2018; Zhang et al., 2023; Roesler et al., 2024; Cameron et al., 2024). Our unit of analysis is downstream: given an already proposed action change, should a learned veto allow it? The endpoint is paired task regret and the qc/qw/Hc/Hw decomposition, not pathology labels or human judgments.

F Fresh Feature-Tier and Calibration-Budget Audits

This appendix reports fresh seed ranges not used by the earlier feature-importance, controlled-precision, severity, or drift analyses. The protocol fixes M=12, δ=0.025, one fresh observation, identical histogram gradient-boosting hyperparameters, validation-only risk-matched threshold selection, and 800 test episodes per domain. Four nested tiers are compared: EvidenceOnly; ContextEvidence; MinimalHistory, which adds old gap, old rank, and the near-old-winner flag; and DescriptorHistory, which additionally adds descriptor distance. The last tier is the current nine-feature model, not typed QRL provenance.

F.1 Feature tiers and risk coverage

Domain	Proposal AUROC	Frozen-threshold regret	Δ regret vs CE 95% CI	Common-coverage AUR
Droplet	0.770→0.837	0.01459→0.01439	-0.00019 [-0.00066, +0.00034]	0.01766→0.01667
Bio	0.635→0.678	0.34716→0.34763	+0.00047 [-0.00238, +0.00334]	0.38998→0.37504
Generic	0.972→0.965	0.02742→0.02708	-0.00034 [-0.00083, +0.00011]	0.03257→0.03301

Table F1. Fresh four-tier audit. Negative Δ regret favors DescriptorHistory over ContextEvidence. Proposal ranking and common-coverage curves improve in droplet and bio, but the validation-frozen action comparison is unresolved in all three domains.

F.2 Small target budgets repair calibration before decisions

Held-out domain	Brier source→128	Brier reduction	Regret source→128	128-label repeat range
Droplet	0.280→0.226	19.3%	0.01532→0.01546	[0.01483, 0.01689]
Bio	0.310→0.224	27.7%	0.33869→0.34379	[0.33806, 0.35457]
Generic	0.153→0.033	78.5%	0.02952→0.02463	[0.02436, 0.02592]

Table F2. Leave-one-domain-out adaptation with 30 repeats. The target Platt fit uses 128 labeled target switch proposals; the action threshold is selected on the same target budget and evaluated on 900 disjoint target episodes. The displayed range is the 2.5th–97.5th percentile across repeats, not a confidence interval for a population parameter.

These results delimit the remaining QRL claim. Existing descriptor/history fields can improve ordering without producing a robust thresholded action gain, and a small labeled budget can repair

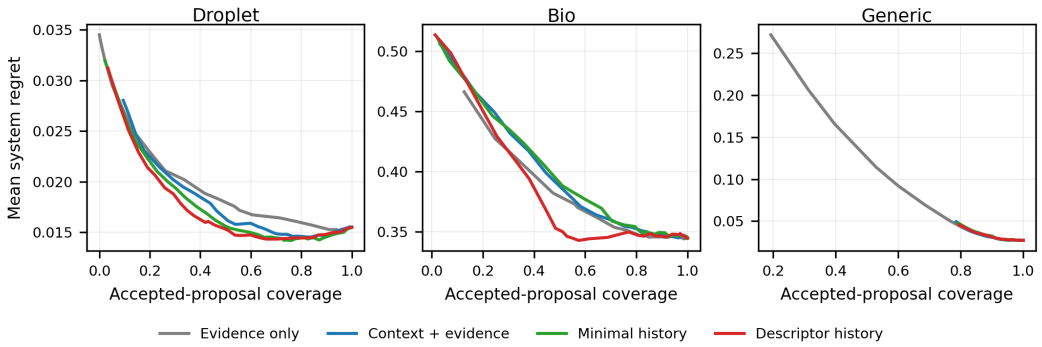

Figure 6: Risk-coverage/requalification curves on the fresh holdout. Coverage is the fraction of proposer switches accepted by the gate; regret is computed over all episodes. DescriptorHistory improves the curve over a meaningful range in droplet and bio, while no tier dominates uniformly.

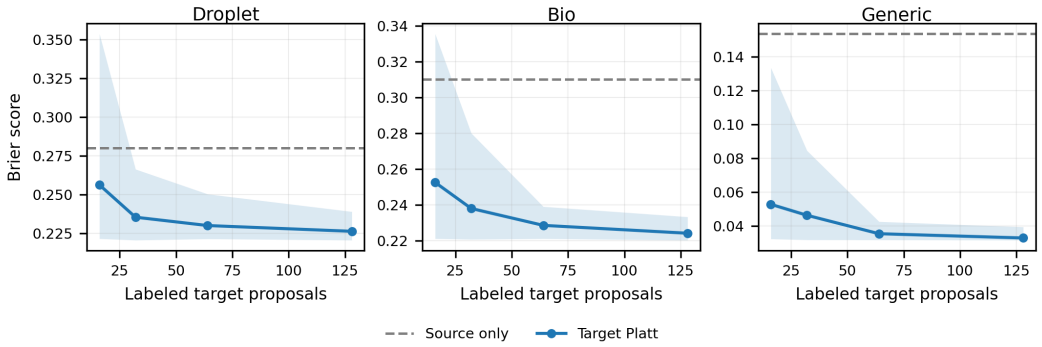

Figure 7: Target-domain Platt calibration with 16–128 labeled switch proposals. Lines show mean Brier score across 30 draws and bands show the 2.5th–97.5th percentile. Dashed lines are source-only calibration. Probability calibration improves in all three domains, but Table F2 shows that action regret improves only in generic.

probability estimates before it repairs outcome-sensitive intervention. A genuine typed-history test must therefore add rejection reason and predicate validity, event age, and verifiable state-version binding, then compare them against the four tiers above on a locked external model.

G Operating-Audit Protocol and External Real-Data Stress Test

This appendix separates two roles that should not be conflated. The controlled domains in the main paper are intervention instruments: they allow proposer precision to be swept while the gate and class-conditional outcome distributions remain fixed. The external audit below instead asks whether the same audit mechanics behave sensibly off those generators. Because its distribution shift is constructed and proposer precision is observed rather than manipulated, it is a stress test of portability, not causal confirmation or deployment validation.

G.1 Protocol instantiation and refusal rule

Section 3.4 states the reusable six-step protocol. This appendix instantiates it without changing the proposer, gate, split, feature set, threshold grid, or endpoint after outcomes are observed. The pre-hashed protocol also treats an absent blocked class as weak identification rather than imputing a break-even.

G.2 Pre-specified real-image corruption stress test

We use the 1,797-image, 64-feature 8×8 test-set copy of the UCI Optical Recognition of Handwritten Digits data distributed with scikit-learn (Alpaydin & Kaynak, 1998). A fixed stratified partition assigns 35%/25%/15%/10%/5%/10% of images to incumbent training, proposer training, gate training, threshold validation, operating calibration, and final evaluation. The incumbent is a standardized multinomial logistic regression trained on clean images; the proposer uses the same architecture trained on a fixed Gaussian-corruption realization (σ=4). A binary logistic gate is trained only on model-disagreement events pooled over $\sigma\in\{2,4,6\}$, using the corrupted pixels plus the two models' maximum probabilities, entropies, and confidence gap. Multiclass Brier loss defines whether a proposal is beneficial. The validation-only gate threshold is chosen on a fixed 0.05–0.95 grid; the calibration cell is σ=4; the final evaluation uses σ=1,...,8 with independent deterministic corruption streams. Every evaluated cell uses a 10,000-draw exact-block-count random blocker.

The complete protocol was hashed before any Digits model fitting or corruption-outcome evaluation (SHA-256 0d5ebbc2e90832767b3cfb93cb612a734fe067933819a47a8dbf55d1bcfa092d; data-array SHA-256 f6d9e39f37dc45d327f6db33428ee58970ccceabb2535a5c179de35886b70443). The firewall forbids split, model, feature, corruption-level, threshold-grid, or endpoint changes after outcomes are seen. A separate chronological Victoria-electricity pilot, also locked in advance, terminated before target evaluation because only 15 gate-training proposal events were available; its split/model was not retuned.

G.3 Results: selectivity transfers more readily than the break-even estimate

The Digits gate had 376 proposal events for training (85.9% beneficial) and 251 for threshold validation; validation selected the lowest grid threshold, 0.05. The disjoint σ=4 operating-calibration cell contained 39 proposals (π=0.923), but the frozen gate blocked none. Consequently q_c=q_w=0 in calibration and H_c/H_w are undefined: π^* is weakly identified, so the protocol correctly refuses to manufacture a calibration-only break-even prediction for this external stress test. This is a failure of operating identifiability, not a failed sign prediction.

σ	Proposal n	π	Block rate	Gate Δ	Random Δ	Holm p
1	17	0.588	0.0%	+0.000000	+0.000000	1.0000
2	46	0.804	0.0%	+0.000000	+0.000000	1.0000
3	68	0.868	0.0%	+0.000000	+0.000000	1.0000
4	94	0.862	1.1%	-0.000430	+0.001246	0.4220
5	128	0.852	2.3%	-0.001975	+0.002729	0.0128
6	144	0.882	3.5%	-0.000663	+0.003722	0.0444
7	147	0.837	0.0%	+0.000000	+0.000000	1.0000
8	159	0.836	3.8%	-0.001561	+0.003588	0.0189

Table G1. Frozen external Digits stress test. Δ is mean gated Brier minus proposer Brier on proposal events; negative values favor the gate. Random Δ is the mean from 10,000 switch-rate-matched random blockers. Holm p corrects the eight one-sided randomization tests. The gate rarely intervenes, but at σ=5, 6, and 8 its small blocked subset beats equal-rate random blocking after Holm correction (adjusted p=0.0128, 0.0444, and 0.0189).

The external result is therefore deliberately narrower than a claim that π^* transfers. On real images the calibration cell itself reveals that the consequence-weighted break-even is not estimable for this frozen gate, while the same-rate control still detects selective action value in three high-corruption cells. This is exactly the distinction the operating audit is intended to enforce: ranking or occasional selective blocking can remain measurable when a portable decision threshold is not.

G.4 What the external audit does and does not add

The external stress test removes one ambiguity but not all external-validity concerns. The observations are real handwritten digits, but Gaussian corruption and the incumbent/proposer construction are experimental choices; no deployed guardrail, human deferral workflow, or physical shield is claimed. Its contribution is instead adversarial: the protocol carries its own refusal condition. Rather than tuning until a break-even appears, it records weak identification and still permits a matched-rate selectivity test. The controlled domains remain necessary for the causal π sweep in the main paper, while Appendix G shows how the same audit can return a bounded positive result—or no break-even estimate—on data outside those generators.